University of Alberta

Modelling phonetic reduction in a corpus of spoken English using Random Forests and Mixed-Effects Regression

by

Philip Dilts

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics

Fall 2013
Edmonton, Alberta

To my families, who made this possible:

The Diltses, who gave me purpose,

And the Lenios, who made sure I always had a home.

And especially to George.

Abstract

In this thesis, phonetic reduction in the Buckeye Corpus (Pitt *et al.* 2005) of conversational speech is modelled using advanced statistical techniques.

Two measures of phonetic reduction are modelled, reduction in the duration of words and deletion of segments from words. Statistical modelling techniques are used to predict how much of each type of reduction is observed in the corpus. Predictor variables are selected from a number of broad classes, including demographic, phonetic, predictability, syntactic, semantic, and pragmatic variables. The broad scope of these variables leads to a generalizable picture of the factors leading to reduction in spontaneous speech.

Two modelling techniques with complementary properties are applied to the modelling task: Random Forest (RF) models (Breiman 2001), and Linear Mixed-Effect Regression (LMER) Models. RF models can be used to model complex interactions and highly co-linear predictor variables much more easily than LMER models can. Conversely, LMER models allow each word form and speaker to differ in their response to reduction-predicting variables. LMER models can also easily incorporate predictor variables composed of a large number of unordered categories. Both of these properties of LMER models are effectively impossible to incorporate into current RF models on the scale required for the present study.

Results relating to the variables or combinations of variables that correlate with reduction or improve model prediction are described. Possible explanations for the results and implications for the nature of the processes underlying reduction during spontaneous speech are explored. Results relating to the modelling process are also discussed. In particular, random forest modelling indicated that several potential interactions between variables were overlooked in initial LMER modelling. When these interactions were included in a second round of LMER modelling, several were found to improve prediction significantly.

The results of the present study may lead to improvements in speech recognition and speech production technologies. The results also suggest that random forests can be used to improve regression models of language data.

Acknowledgements

I am overwhelmed with gratitude. It would be impossible to name everyone without whom this thesis could not exist. My supervisors Gary Libben, Ben Tucker, and Harald Baayen found opportunities, provided guidance and insights, fostered discussion and thought, and demonstrated infinite patience. In the Department of Linguistics at the University of Alberta, the support staff did more work for me than I had any right to expect. The faculty fostered an environment of curiosity and collaboration, let me find out that I love teaching, and made me feel like I could be good at something again. The students and PostDocs and friends in Linguistics and Philosophy (and Mary Hocaliuk) provided intellectually stimulating discussion, invaluable advice, hair cuts, and good times. Members of the Alberta Phonetics Laboratory and participants at the Alberta Conference on Linguistics and Nijmegen Workshop on Conversational Speech provided valuable comments on presentations of this work, as did Ryan Podlubny and Andrea Tam at the practice defence that Pud was kind enough to organize for me. The Social Sciences and Humanities Research Council, the Faculty of Graduate Studies and Research, and the Department of Linguistics all generously provided me with funding. And Anne-Michelle Tessier, of course, for existing, and for the everything, because obviously. Thank you all.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Phonetic Reduction

The present study investigates phonetic reduction in the Buckeye Corpus (Pitt et al. 2005), exploring the conditions under which words are reduced in spontaneous speech, and drawing inferences about the processes underlying reduction.

Reduction is defined here as the difference between word tokens produced in connected speech and word tokens produced in citation form (e.g., in isolation). The term 'reduction' is used because when such a difference is found, words produced in connected speech tend to be shorter or to contain less phonetic information than their associated citation forms. This reduction is widespread throughout connected speech, at least in English: Johnson (2004) found that more than 25% of the word productions in a corpus of conversational English had at least one fewer segment than their citation forms predicted. This reduction can also reach extreme levels: Bybee (2006:p.720) notes that *I'm going to* is often pronounced as [aimənə], for example, and Warner (2011b:p.1866) describes a production of *but I was like* that took the form [b̥ɹʌʒləɪ]. Remarkably, speakers not only correctly understand such productions, but also seem not to notice that any drastic reduction is taking place.

The widespread and extensive nature of reduction means that any automatic speech recognition (ASR) software attempting to recognize English speech must treat reduction as extremely important. ASR software must be able to convert reduced phonetic forms into their underlying lexical representations. It must do so even when a form is extremely reduced, and the large proportion of word tokens that are likely to be reduced means that it must do so very often. Indeed, to complete its task ASR software must be able to detect (or predict) whether reduction is or is not occurring for each token in the first place.

Automatic speech production (ASP) software must also incorporate reduction. To produce speech that is as natural-sounding as possible, ASP systems must know when and how to reduce word productions. ASP systems must also know to perform this reduction without hindering listener comprehension.

Current ASR and ASP systems have not yet been able to master these reduction-related capabilities. As the references above indicate, however, human adults have mastered the production and comprehension of reduced forms to the extent that they process reduction in real time *without being conscious that they are doing so* (Warner 2011b). This gap

in performance between humans and computers suggests a concomitant, fundamental gap in the current understanding of reduction processes. As Warner (2011b) notes, however, reduction has often been seen as merely a surface process, irrelevant to models of underlying linguistic processing. The pervasive and unconscious nature of reduction appears to indicate otherwise: Namely, that a model of linguistic processing that fails to explain or incorporate reduction must be considered far from complete. This suggests in turn that the study of the nature of reduction may reveal important properties of underlying linguistic and cognitive processes, informing theoretical models of speech production and perception. The present work is intended as a contribution to this study of reduction.

1.2 Overview

Two types of reduction are modelled in the present study: Reduction in the duration of words, and deletion of segments from words.

A set of predictor variables spanning several levels of linguistic processing are evaluated in terms of how they affect reduction. Word- and syllable-level properties like frequency and number of stressed syllables, utterance-level properties like conditional probability and speaking rate, discourse-level properties like topicality and dispersion, and speaker-level properties like age and gender are each used to model phonetic reduction. While the set of variables chosen does not cover all possible factors that might affect reduction in speech, it does cover a broad variety of ways in which reduction could be influenced. Predictors were selected across a wide spectrum of linguistic factors, allowing for a more generalizable picture of what leads to reduction.

A more generalizable study might be argued to come at the cost of decreased confidence in the results, however. The sets of words examined in several previous studies (see Section 1.4 below) are limited to small sets of carefully selected tokens or types in order to control for any systematic confounds. This leads to a great deal of confidence that the results are valid for the words under study, but also to less confidence that the picture of reduction they describe is true of words in general. By contrast, the present study begins by investigating every intelligible content word in a corpus of natural speech, excepting only those word tokens for which accurate values of the predictors can not be calculated. To control for variables that may affect reduction but are not included in the current study, linear mixed-effects regression (LMER) is used. LMER models allow for predictors to relate to reduction in different ways depending on the speaker or word form under study, controlling for the effects of word- and speaker-level predictors that may have been overlooked.

1.3 Methods Overview

This overview is divided into three parts. In Section 1.3.1, the ways in which reduction is operationalized are described. Section 1.3.2 describes the variables used to predict the reduction measures. Finally, Section 1.3.3 describes the statistical modelling techniques used to analyze the effect of the predictor variables on phonetic reduction.

1.3.1 Reduction Measures

Reduction is operationalized here in two ways: Reduction in word duration, and number of segments deleted from a word. Separate models are constructed for each of these measures; The results for each measure are described in separate chapters. The reduction measures are described in more detail below.

1.3.1.1 Number of Deleted Segments

Each word token in the Buckeye Corpus is tagged with both a citation phonological form and an actual output phonological form. The number of segments deleted from a word can be calculated by simply subtracting the number of segments in the actual output form from the number of segments in the citation form. There is some imprecision in this measure of reduction: Segment transformations are ignored, and the assimilation of two segments into one will appear to be a simple deletion. The simpler segment-count measure was chosen because it avoids the theoretical difficulties involved in detecting individual phonetic features in a corpus with hundreds of thousands of segments.

1.3.1.2 Reduction in Word Duration

A word token may also be considered reduced if it is pronounced more quickly than expected. Unfortunately, the question of how quickly a word is expected to be pronounced has no definite answer. That is, there is no principled way to choose what form a word can be said to be reduced *from*. In the present study, the citation form provided in the Buckeye corpus is taken as the expected form, primarily because no other type of expected form is available.

A word's most common pronunciation in the corpus could theoretically be used a baseline for duration reduction calculations. Using these common forms as reference points would allow the population under study or each individual speaker to determine their own set of expected phonological forms. The present study, however, aims to describe reductions from the citation forms that are thought to reflect a word's phonological representations in a speaker's mental lexicon. In this view, many of the most common forms in conversation are likely reduced from their complete lexical forms, and thus they represent the type of reduction under study. As a result, taking common conversational forms as unreduced baselines makes their study as reduced forms impossible. (The prevalence of such forms in the corpus is explored in the following chapter.)

In an ideal solution to this problem, the speakers who produced the conversational speech would also have produced a more formal type of speech (e.g. by reading word lists) that could be taken as an approximation of lexical form and used as a baseline. Indeed, some speech corpora currently in production (e.g., the Pacific Northwest English Project - see www.artsci.washington.edu/NWenglish/ (retrieved June 6, 2013)) are collecting multiple genres of production at varying levels of formality from each participant, in part to allow for such a measure of reduction to be calculated. As these corpora are not currently available, the present study is forced to rely on an estimation of lexical-form pronunciation.

Citation pronunciations provided with the corpus are taken as the expected underlying phonological forms of each lexeme. The average unreduced-form duration for each word type

is then calculated as follows: Word tokens produced with the same number of segments as their citation form are collected, and the average duration of these tokens is taken as the unreduced duration for that word form type. To account for variation between speakers, the average unreduced duration for each token is calculated separately for each speaker.

This measure treats all citation-length tokens as unreduced. As a result, tokens with segments that have changed but not completely elided are still considered unreduced. (Indeed, tokens that have undergone metathesis, or deletion and insertion in equal numbers, will also be considered unreduced.) This choice was made to increase quantity and quality of the average unreduced duration measures. Without this compromise, word forms that are never produced in *exact* citation form by a speaker would have to be excluded from analysis. Word forms produced only rarely in exact citation form would have few tokens over which averages could be drawn, and these averages are thus more susceptible to undue influence from unusual productions.

1.3.1.3 Comparison of Reduction Measures

Another objective of the present study is to examine the similarities and dissimilarities between the models of these two reduction measures. Each predictor variable may differ in the way it effects deletion and duration reduction. This comparison is performed in Chapter 5.

1.3.2 Predictors

A total of 18 fixed-effect predictors are included in the present study. Descriptions of the predictors are provided in this section. Previous studies examining the effects of these predictors on reduction are described in Section 1.4.2 below. Details regarding the calculation and distribution in the corpus of each predictor are provided in the following chapter.

1.3.2.1 Demographic Predictors

Speaker age and gender, along with interviewer gender, are each included as predictors in the models.

1.3.2.2 Phonetic Predictors

Four phonetic predictors of potential interest are examined for their effect on reduction: Two speaking rates (by speaker and by intonational unit), the number of stressed syllables expected in a word, and a word's expected length in segments. (A word's 'expected' lengths in segments syllables are based on the citation forms provided with the Buckeye Corpus)

1.3.2.3 Predictability

A word's frequency might be thought of as a rough estimate of its predictability. Two frequency measures are considered here, one based on the Contemporary Corpus of American English (COCA - Davies (2009)) and one based on a word's frequency in the Buckeye Corpus itself.

Two measures of predictability from lexical context are included as predictors as well. The measures are the conditional probabilities shown by Jurafsky *et al.* (2001) and Bell *et al.* (2003) to have an effect on reduction: Conditional probability given the previous word and conditional probability given the following word.

Conditional probability, also called transitional probability, is an estimate of how likely a target word is to appear given a certain neighbouring word. For example, a word's conditional probability given the previous word is calculated by dividing the 2-gram frequency of the two words - the target word and the preceding word - by the frequency of the previous word in the corpus. Roughly speaking, this measures what proportion of the word tokens following the preceding word are the word of interest.

1.3.2.4 Structural Constituency

A word's position in some larger linguistic structure might affect reduction. Words at phrase boundaries, for example, show distinctive behaviour as described in section 1.3.2.3 above.

In the present study, more detailed and gradient measures of a word token's constituency in its linguistic structures are adopted. Agnostic definitions of 'structure' and 'constituency' are used, and only a small set of structures that can be specified completely and calculated with relative ease are included. This of course precludes an analysis based on a more powerful or comprehensive syntactic theory, but it may still yield interesting results.

Five predictors related to structural constituency are included in the present study.

The most naive measure of constituency used here simply counts how far along in an intonational unit the speaker had proceeded before saying the target word. (The operational definition of *intonational unit* is provided in the following chapter.) This measure counts how many words have passed between the beginning of the intonational unit and the target word. Such a measure is not capable of capturing higher order dependencies, but it can give an approximation of how far the speaker has proceeded into a planning unit.

A word's part of speech (POS) could also be thought of as a highly simplified version of its syntactic constituency. Thus, a token's POS is included as a predictor.

Conditional probabilities based on part of speech can also be calculated, just as Jurafsky *et al.* (2001) and others have done for word forms. These probabilities can be used to model a kind of syntactic predictability effect, capturing the probability that a determiner will be followed by a noun, for example, and looking for any effect that this probability might have on reduction.

After a selection process described in the following chapter, three POS-based conditional probabilities were selected for inclusion in the models here: Conditional probability of a POS given the previous POS ('forwards POS predictability'), conditional probability of a POS given the following POS ('backwards POS predictability'), and conditional probability of a POS given the two neighbouring parts of speech ('surrounding POS predictability').

1.3.2.5 Topicality

A word's relevance to the current conversation might have an effect on how it is produced. Previous studies of reduction have found a related measure (*givenness*) useful in predicting

reduction, as described in Section 1.4.2.5 below. Computational linguistics research into search engine optimization and automatic summarization has produced several techniques for estimating the relevance of a term to a particular document. The present study uses an easy to calculate but consistently effective (Robertson & Spärck Jones 1994) measure of relevance, term frequency-inverse document frequency (tf-idf - (Luhn 1958; Robertson & Jones 1976)). Details of the calculation of this measure are described in the following chapter.

1.3.2.6 Time

The time at which the target word appears in an interview is also included as a predictor. A relationship between this measure and reduction rates would indicate a change in the way a speaker reduces their words or deletes phones as they proceed through a conversation.

1.3.3 Modelling Techniques

Two modelling techniques are applied to the prediction of each reduction measure. The first technique is linear mixed-effects regression (LMER, as implemented in Bates *et al.* (2011)). LMER modelling is becoming well-established in linguistic research (see e.g. Baayen (2008); Baayen *et al.* (2008); Jaeger (2008)). LMER models are relatively easy to construct and understand, though they are also easy to misuse (Gelman & Hill 2007; Barr *et al.* 2013).

In a simple linear model of reduction, the reduction is described as a linear combination of the properties of the words under study (the fixed effects), along with some amount of noise. In a mixed-effects regression model like LMER, each of these fixed effects is allowed to take on different values for each level of a given *grouping variable*. The grouping variables in the present study are word forms and speakers. That is, in the current study each speaker and word form is permitted to respond to each of the predictor variables in their own way. The variation among these by-speaker and by-word form contributions form the *random effects structure* of a mixed-model. The presence of this random effects structure is what differentiates LMER models from simple linear models. Indeed, the grouping variables, or the variation found between levels of a grouping variable, are often simply called the *random effects* in an LMER model.

LMER models have some limitations: LMER specification requires that the analyst make certain assumptions about the distribution of the predictors and the response variable. Moreover, both computational and modelling constraints limit the ways in which the predictors are allowed to interact with each other in determining the response.

The second technique is Random Forest (RF - Breiman (2001)) modelling. RF modelling is a non-parametric modelling technique that makes fewer assumptions about the properties of the individual predictors and the response variable. RF models also automatically allow the predictors to interact with each other in arbitrarily complex ways, as long as these interactions improve the predictive power of the model. Due to the relatively recent development and non-parametric nature of RF models, however, their behaviour under various conditions is still being studied, and discussions of how best to apply RF models form an active area of research (Archer & Kimes 2008; Strobl *et al.* 2008;

Nicodemus *et al.* 2010). Some proposed improvements to the technique (in particular, conditional permutation-based variable importance calculation (Strobl *et al.* 2008)) are not yet available in the most flexible software implementation of RF modelling, the R package **randomForest** (Liaw & Wiener 2002). Due to certain constraints on computational power, it is difficult to allow individual speaker and word types (the random-effects predictors in LMER models) to be included as sources of variation in RF models. Still, RF models have been shown to provide an improvement in predictive power over LMER and other modelling techniques in some linguistic research (Tagliamonte & Baayen 2010).

These modelling techniques have complementary properties, then: RF models can be used to model complex interactions and highly collinear predictor variables much more easily than LMER models can. LMER models, on the other hand, allow each word form and speaker to differ in their response to reduction-predicting variables, and can easily incorporate predictor variables composed of a large number of unordered categories. Both of these properties of LMER models are effectively impossible to incorporate into current RF models on the scale required for the present study.

In the following chapters, the results of RF and LMER modelling are compared, in terms of both their overall predictive power and the ways in which each predictor affects reduction.

1.4 Previous Studies

Many existing studies have examined phonetic reduction (see Warner (2011b) for a general overview, or Warner (2011a) for an overview of the methods applied to the study of reduction).

1.4.1 Large Scale Studies

The present study represents an attempt to model reduction in as large a set of words as is practical, using a large set of predictors that span a broad range of linguistic factors.

Some recent studies have constructed models of large collections of data. Johnson (2004), for example, demonstrated widespread and 'massive' reduction in spontaneous speech, showing that in some cases multiple syllables are deleted from a single word. Johnson also used the Buckeye corpus (Pitt *et al.* 2005), then only partially transcribed, demonstrating its value for corpus studies of reduction. Johnson's study considers the corpus as a whole, looking for reduction in every word in the corpus that was available at the time.. The study's broad focus came at the cost of more detailed analysis of the causes of reduction, however, and considered only two predictors: citation-form length in number of syllables, and a coarse division into function vs. content words.

Other studies have also constructed models of large collections of data, using a large number of predictor variables. Bell *et al.* (2009) examine predictability effects on reduction in nearly 7,000 tokens (all of the tokens in the corpus that met their inclusion criteria) in a subset of the Switchboard Corpus (Godfrey *et al.* 1992), using several predictor variables. Gahl *et al.* (2012) examine the effect of phonological neighbourhood density on reduction using over 9,000 tokens (all of the monomorphemic CVC content words in the Buckeye Corpus that met their inclusion criteria), using several predictor variables and LMER mod-

elling. Gahl (2008) modelled reduction in approximately 80,000 world tokens taken from the Switchboard corpus (Godfrey *et al.* 1992), finding that the more frequent member of a pair of homophonous words tended to be produced significantly shorter.

In most other cases, researchers have chosen a more limited set of words or speech units to study in detail. Scheibman & Bybee (1999), for example, observed greater reduction of the word *don't* in more predictable contexts. Gregory *et al.* (1999) and Jurafsky *et al.* (2001) looked for contexts in which word-final t/d deletion occurred, showing that higher frequency and conditional probability led to higher rates of deletion. Raymond *et al.* (2006) modelled the links between several predictors and word-medial t/d deletion in the Buckeye Corpus. In two studies of reduction in careful speech, Aylett and Turk looked at how syllables with the same citation form manifest in different words, showing that several types of predictability from linguistic context affected syllable length (Aylett & Turk 2004) and the spectral quality of vowels (Aylett & Turk 2006).

Aylett and Turk's syllables of interest shared phonological content but not necessarily meaning, a problem that Pluymaekers *et al.* (2005b) addressed by investigating reduction in a set of affixes attached to carrier stems that differed in frequency. Jurafsky *et al.* (2001) and Bell *et al.* (2003) looked for reduction in the 10 most frequent words in the Switchboard corpus (Godfrey *et al.* 1992), finding that words preceding disfluencies, words in less predictable positions, and words falling at the beginning or end of an utterance tend to be less reduced than their counterparts.

The works mentioned above looked for reduction in a tightly controlled set of examples. And for good reason: Words are different from each other in ways that are not yet completely understood. By limiting the set of word types under study, the set of systematic or idiosyncratic properties of words that could skew models in unpredictable ways can be minimized. A study of the frequency effect at the word level, for example, might consider 'oak' and 'elm' similar enough in phonology (one onset-less syllable) morphology (monomorphemic) and semantics (types of tree) that the main difference between them comes from their frequency (medium for 'oak', low for 'elm'). For a study of detailed phonetic properties, of course, 'oak' and 'elm' must be considered completely different, and reduction researchers have had to find words that are phonologically similar (or even identical, in some cases) but that also have different frequencies. Pluymaekers *et al.* (2005c), for example, looked at the duration of affixes, which share both phonology and semantics but differ in whole-word frequency in the context of the stems with which they are combined. Gahl (2008) uses homonyms, which can differ in frequency while maintaining the same phonological form.

Mixed-effects models (Pinheiro & Bates 2000) provide a statistical tool that addresses mathematically the problem that previous studies have had to work around. Instead of finding stimuli that are both different and the same, a mixed-effects model factors out the variation found among individual words. Baayen (2008) shows how mixed-effects models can be used to study linguistic phenomena, modelling items and speakers as random effects rather than fixed effects. In this way, items are treated as random samples of English words, and speakers are treated as random samples of English speakers in general, rather than repeatable treatments that can be manipulated. Mixed-effects modelling techniques are applied to reduction here, allowing for the study of a large set of words while compensating

for the variation between them.

1.4.2 Predictors

1.4.2.1 Demographic Predictors

The Buckeye Corpus is coded for three demographic variables: Age, gender, and interviewer gender. Previous studies of these predictors' relationships to reduction have found weak or mixed results.

1.4.2.1.1 Age Some studies show older speakers reducing less than younger speakers. Bell *et al.* (2009) found this effect in the Switchboard corpus. Two studies have found that older speakers are less likely to delete segments than younger speakers: Raymond *et al.* (2006) found an age effect in the Buckeye Corpus in certain contexts, and Strik *et al.* (2008) found an age effect in Dutch. Gahl *et al.* (2012) and Yao (2011), however, found no significant effect of age on reduction in the Buckeye corpus.

1.4.2.1.2 Gender Strong evidence for a general difference in reduction between men and women has not been reported. Bell *et al.* (2009) do find an interactive effect between speaker gender and speech rate in duration reduction, with men speaking more quickly on average. Other studies have sought a link between gender and segment deletion rates (Patterson *et al.* 2003; Pluymaekers *et al.* 2005a; Raymond *et al.* 2006; Strik *et al.* 2008; Zimmerer 2009). Gender is found to affect reduction in some of these studies: Raymond *et al.* (2006) found English-speaking men deleting fewer segments in certain segmental contexts, while Zimmerer (2009) found German-speaking women deleting fewer segments than men in certain other segmental contexts. No broadly-applicable difference in reduction across genders is attested to, then, in these studies.

1.4.2.1.3 Interviewer Gender None of the studies cited in the present work find any relationship between interviewer gender and reduction.

1.4.2.2 Phonetic Predictors

1.4.2.2.1 Local Speaking Rate A few studies find no relationship between local speaking rate and phonetic reduction. Tily *et al.* (2009) found no effect of speaking rate on the duration of productions of *to*, for example. Patterson *et al.* (2003) find no link between speaking rate and schwa deletion in some contexts, and Pluymaekers *et al.* (2005a) find no link between speech rate and segment deletion for seven particular word forms in Dutch.

Most existing studies, however, have found that higher speaking rates lead to more phonetic reduction. Word productions surrounded by fast speech were both shorter in duration (Gahl *et al.* 2012; Yao 2011; Gahl 2008; Pluymaekers *et al.* 2005a) and more likely to contain deletions (Fosler-Lussier & Morgan 1999; Raymond *et al.* 2006; Guy *et al.* 2008; Bürki *et al.* 2011) in the studies described here.

1.4.2.2.2 Global Speaking Rate Global speaking rate is calculated here by speaker. Each speaker has a (mean) average rate at which they speak, calculated across all of their

utterances in the corpus. The effect of this measure on reduction is relatively understudied. None of the studies cited here compare this measure to changes in word duration. One study (Raymond *et al.* 2006) considers a ratio between global and local speaking rates as a predictor. In this study, speakers were found to delete more segments when they while they were speaking faster than they do on average.

The paucity of existing studies considering this measure provides one of the motivations for including it in the present work.

1.4.2.2.3 Word Length The effects of word length on reduction depend heavily on both the type of reduction under study and the operationalization of the term 'length'.

Deletion studies have found that words whose citation-forms contain more segments or syllables show higher rates of deletion (Patterson *et al.* 2003; Raymond *et al.* 2006; Van Bael *et al.* 2007).

Studies modelling word duration have found more mixed results. Some studies taking orthographic length as a predictor (Bell *et al.* 2009; Gahl *et al.* 2012; Yao 2011) find no relationship between orthographic word length and word production duration.

Studies that take the average duration of a word form as a predictor, on the other hand, (Bell *et al.* 2009) or the summed average of that word form's constituent segments (Gahl *et al.* 2012), do find a relationship between expected length and duration. Both studies find, unsurprisingly, that word token duration is well predicted by the average word form duration as calculated over similar tokens. In short, words with higher expected durations were found to have higher observed durations. Average word duration is thus taken as a (completely necessary) control variable, accounting for the obvious effect it has on the durations of individual word productions and allowing other effects to be examined with greater confidence.

In the present study, controlling for expected duration is incorporated into the dependent variable itself. Rather than modelling word duration itself as a measure of reduction, the present study takes the difference between expected and observed word duration as its reduction measure. In this way, a different type of question about length can be asked through modelling: A duration-based reduction measure can be used to show that long words tend to be long. A reduction-based measure can be used to ask further whether these long words are more likely to be shortened in conversational speech than short words.

1.4.2.2.4 Number of Stressed Syllables Many studies have shown that segments in stressed syllables are less likely to be deleted than segments in unstressed syllables (Greenberg 1999; Pluymaekers *et al.* 2005a; Raymond *et al.* 2006; Van Bael *et al.* 2007; Zimmerer 2009), suggesting that stress is an important predictor of reduction. This result is only attested in deletion studies, however: None of the studies cited here that model word duration have found a link between stress patterns and word duration reduction. Indeed, no such study includes the number of stressed syllables as a predictor.

1.4.2.3 Predictability

1.4.2.3.1 Local (Buckeye) Frequency Some reduction studies have considered a word form frequency measure that was based on the corpus being modelled. Most of these studies have found that higher local frequency leads to greater reduction, in terms of both word duration (Bell *et al.* 2009; Gahl 2008) and segment deletion (Fosler-Lussier & Morgan 1999; Jurafsky *et al.* 2001; Guy *et al.* 2008; Zimmerer 2009; Meunier & Espesser 2011) (though *c.f.* Priva (2008)).

1.4.2.3.2 External (COCA) Frequency Studies incorporating a measure of reduction calculated over a separate corpus from the one under study have found similar results: High-frequency words are more likely to be reduced in duration than low-frequency words are (Gahl *et al.* 2012; Baker *et al.* 2011; Aylett & Turk 2004; Baker & Bradlow 2009; Yao 2011). Deletion results are less conclusive, with several studies failing to find a link between externally-calculated frequency and the number of deletions in word productions (Raymond *et al.* 2006; Guy *et al.* 2008; Schuppler *et al.* 2009). Some of these authors ascribe this apparent lack of frequency effect to genre differences between the corpora.

1.4.2.3.3 Word-Based Conditional Probability As mentioned above, two conditional probability measures were selected as predictors. Jurafsky *et al.* (2001) and Bell *et al.* (2003) found that conditional probability given the preceding word and conditional probability given the following word are the predictability measures that are the most effective in predicting reduction.

1.4.2.4 Structural Constituency

1.4.2.4.1 Position in Phrase A word token's linear position within its intonational unit has been found to relate to reduction rates. Fougeron & Keating (1997) summarize previous studies on prosodic domain strengthening, and Bell *et al.* (2003) replicates this result in a broader, more recent study of reduction.

Phrase-position has only been shown to affect reduction in a limited domain, however. The studies cited find that words at the beginning or end of an intonational unit are more likely to experience shortening or lengthening than words near the middle of a unit.

The present study applies a predictor that is different in two important ways. First, for technical reasons (described in Chapter 2), words at the boundaries of intonational units are not included in the present analysis. Second, the measure is more gradient. A word's phrase position is taken as the number of words preceding it in the current intonational unit. Thus, the position-in-phrase predictor here controls for those positions already known to relate to reduction, and looks in greater detail at those positions for which no effect has been shown.

1.4.2.4.2 Part of Speech Previous studies (Gahl *et al.* 2012; Yao 2011) found different parts of speech undergoing duration reduction at different rates, with nouns showing the shortening and verbs showing the most. No study of deletion rates cited here has used part of speech as a predictor, however.

1.4.2.4.3 Part-of-Speech-Based Conditional Probability

Part-of-speech n-grams have been found to be effective in automatic style and author attribution for written texts (Argamon *et al.* 1998; Koppel *et al.* 2003; Gamon 2004) , but remain relatively under-explored in studies of reduction, or spoken language generally. None of the studies cited here use part-of-speech n-grams as predictors in reduction studies.

1.4.2.5 Topicality

Topicality has not been operationalized using tf-idf (Luhn 1958; Robertson & Jones 1976) in the studies cited here.

Instead, studies of reduction have focused on operationalizations of a similar (and likely correlated) measure: Givenness. Several studies find that previous mentions of a word form associate with higher levels of duration reduction, whether that word form had been mentioned by the speaker under study (Aylett & Turk 2004; Bard *et al.* 2000; Baker & Bradlow 2009; Bell *et al.* 2009; Fowler & Housum 1987; Lam & Watson 2010), mentioned by another participant in the conversation, (Kahn & Arnold 2012), or even evoked by an image representing the word (Anderson & Howarth 2002).

The topicality measure applied here is likely to absorb any such effects of givenness, due to its high correlation with the number of previous mentions over a short stretch of time.

1.4.2.6 Time

Time-of-utterance is not included as a predictor of reduction in any of the studies cited here. The measure is used here to measure changes in reduction rates that take place as a speaker proceeds through a conversation.

1.5 Possible Applications

Any increase in understanding of reduced speech gained by this study may have interesting applications to both theoretical and applied linguistics. For example, reduction is likely to be an integral part of language change, and predicting reduction would thus be a necessary step towards predicting language change (See e.g. Ohala (1993)).

Reduction must also be accounted for in any complete psycholinguistic model of language production. Gahl (2008), for example, showed how differential reduction in homophone pairs might lead to a refinement of Levelt *et al.* (1999)'s influential "Speaking" model of language production. By looking at the way a large number of variables interact to lead to increased likelihood of reduction, then, information about the cognitive processes underlying reduction may be revealed. For example, if there were any context in which frequency was found to have no effect on reduction, it would be a remarkable result. Newmeyer (2006:p.401), for example, alleges that the relationship between frequency and reduction is indicative of a universal underlying cognitive process, stating that "It is a truism that the more often we do something, the faster we are able to do it." Linguistic situations failing to show an effect of frequency might provide a surprising counterexample, or at least a slight refinement, to Newmeyer's 'truism'.

There are also several potential practical applications for the results of this research project, including contributions to both commercial applications and research infrastructure. To analyze the way in which words and segments are reduced, we must first decide what we consider them reduced from. As a result, a kind of database tracking the durations of word types in the Buckeye Corpus, as well as a set of the most common forms of each word type in the spontaneous speech of Central Ohio, both of which could be provided to the corpus compilers for possible inclusion in further distributions of the corpus. Indeed, any mark-up, and any program created to mark up the corpus, will be made as freely available as the licensing terms of the corpus allows.

The results of the present study may also find some application in speech recognition and speech production research. Since reduction abounds in natural speech, automatic speech recognition systems might be aided by a better ability to predict and adapt to reduction. Speeding and eliding segments like a native speaker could also contribute to the naturalness of automatic speech production systems. A model of native-like reduction might also lead to applications in foreign-language pedagogy.

In addition to the findings related to reduction, the present study provides insights into the use of the modelling techniques themselves. Random Forests in particular are relatively understudied in terms of their application to language modelling. In the present study, Random Forest models are constructed, and their results are compared to and combined with those of the LMER models. Through this process, valuable insights may be gained into the effective use of Random Forest modelling on linguistic data.

1.6 Thesis Outline

The remainder of this thesis is divided into four chapters. Chapter 2 describes the methodology used in greater detail, describing how the predictors and response variables are calculated and further specifying the modelling procedures applied. Chapters 3 and 4 each describe the process, results, and implications of the modelling of a single response variable: Chapter 3 reports the results of word duration reduction modelling, and Chapter 4 reports the results of deletion modelling. Chapter 5 summarizes the findings of Chapters 3 and 4, and outlines the general conclusions arrived at during the present study.

Chapter 2

Methods

2.1 Corpus Overview

The Buckeye speech corpus (Pitt *et al.* 2005) is made up of 40 sociolinguistic-style interviews transcribed to the level of the individual speech sounds. Each interview was categorized according to the genders of the interviewer and interviewee and a stratification of age (under 30 or over 40). Each word token in the corpus is listed along with its orthography, its time of occurrence, its part of speech, and two lists of phones representing pronunciations. One list of phones describes the canonical, citation-form pronunciation of the word form. The other list of phones describes the way in which the word token was actually pronounced.

Part-of-speech tags are also provided for each word token. A modified version of the tags used by the Penn Tree-bank project (Marcus *et al.* 1993) are used. A modified C&C tagger (Curran *et al.* 2007) was used, with an expected accuracy on the order of 90% (Kiesling *et al.* 2006) A total of 43 parts of speech are listed in the corpus, including 11 compound tags that are used to label contractions like *gonna* and *shouldn't*. This level of detail is not required for the present study. Instead, the content words are divided into four broad part-of-speech categories as shown in Table 2.1.

Category	Tags	Parts of speech
Noun	nn, nnp, nns, nnps	singular, plural, mass, and proper nouns
Verb	vb, vbd, vbp, vbn, vbz, vbg	all non-modal verb forms
Adjective	jj, jjr, jjs	bare, comparative, and superlative adjective forms
Adverb	rb, rbr, rbs	bare, comparative, and superlative adverb forms

Table 2.1: Part-of-Speech Categories

These part-of-speech classes are far from equally represented in the corpus. There are more than twice as many noun tokens (47,234) as adjective tokens (14,948), for example, and more than *ten times* as many noun types (5,523) as adverb types (365). Type and token counts are illustrated in Figure 2.1

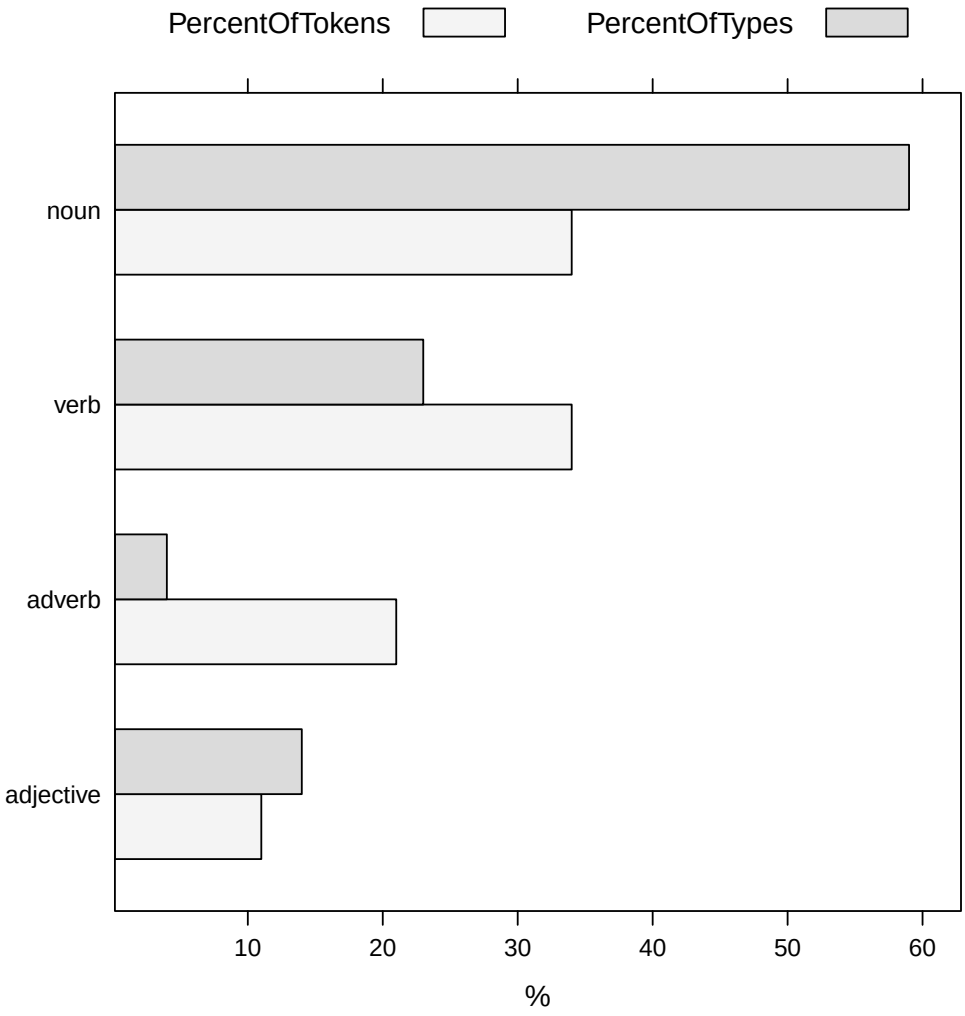

Figure 2.1: Number of word tokens and types for each part of speech

Previous studies (Tree & Clark 1997; Jurafsky *et al.* 1998; Bell *et al.* 1999; Bell *et al.* 2009) have found qualitative differences in the way that content words and function words reduce. For example, some function words tend to have fewer reductions preceding a filled pause or other planning-related disfluency. A study of reduction, then, should be careful to consider content and function words separately. The present study addresses this concern by limiting its scope to reduction in content words. (i.e., words in the classes that are listed in Table 2.1) There are a total of 137,319 such content words in the corpus.

2.2 Reduction Measures

The two reduction measures, reduction in word duration, and number of segments deleted from a word, are described in the previous chapter. Findings related to the properties of these measures in the Buckeye Corpus are described below.

2.2.1 Number of Deleted Segments

A total of 29,888 word forms, or 22% of the tokens in the corpus, contain at least one deletion. The most deletions encountered in a single token is 8, but such heavily deleted forms are very rare. For example, 80% of the word tokens with deletions contain only a single deletion. In general, as the number of deletions goes up, the number of word tokens with that many deletions goes down exponentially, as illustrated in Figure 2.2

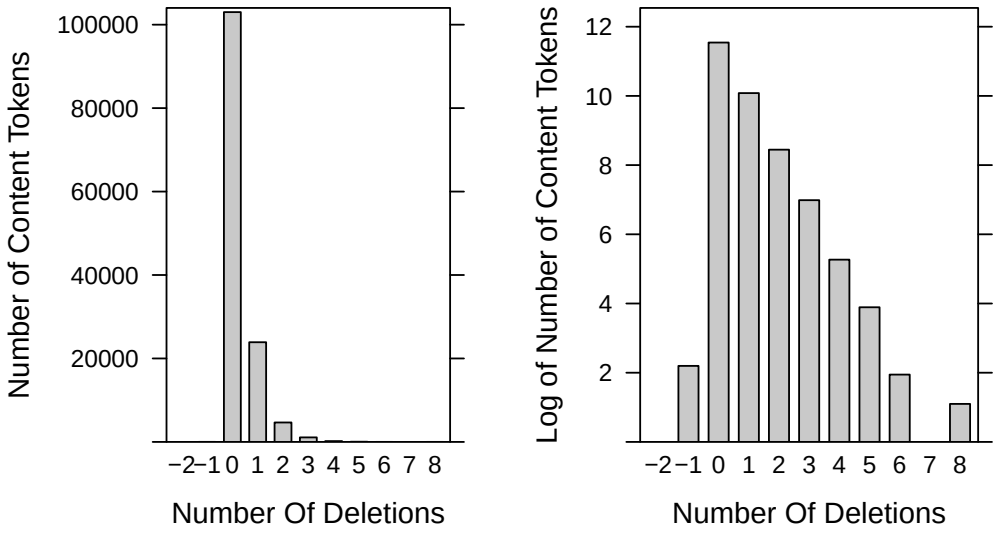

Figure 2.2: Number of deleted segments per word

2.2.2 Reduction in Word Duration

Duration reduction is operationalized here as reduction from the average duration of word forms produced with the number of segments found in the citation form, calculated

separately for each speaker, as described in the previous chapter. This operationalization is motivated partly by the high prevalence of non-citation forms, as described below.

There is a large amount of divergence between the most commonly produced forms and the citation forms provided with the corpus, as illustrated in Figure 2.3. Fully 38% of the word types in the corpus are more often pronounced in a non-citation form, and the average speaker produces 44% of their word form types in non-citation form most often. Moreover, the degree to which speakers tend to produce (Buckeye-provided) citation-form pronunciations of words varies widely: One speaker favoured citation forms for only 36% of word types, while another speaker favoured citation forms for 67% of word types.

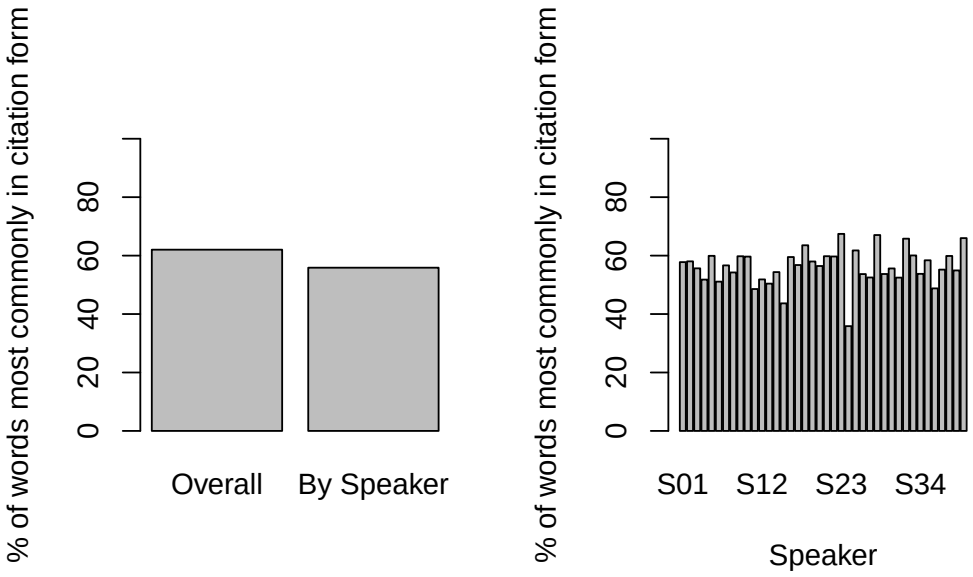

Figure 2.3: Prevalence of citation forms, across the entire corpus and by speaker

This reduction measure is unavailable for word types that a speaker only says once, and word types that a speaker never says in citation form. There is a fairly significant number of these word tokens: 26,625, or 19% of all content words in the corpus.

When compared to citation-length forms, words in the Buckeye Corpus tend to be reduced in duration on average, as illustrated in Figure 2.4. The average reduction is not very large, however, at about 14 milliseconds, or about 5% of the average word duration.

A small number of tokens (36) appeared to increase in length by more than a full second, but closer inspection revealed these to be coding errors in the original corpus. All 36 are excluded from the analysis.

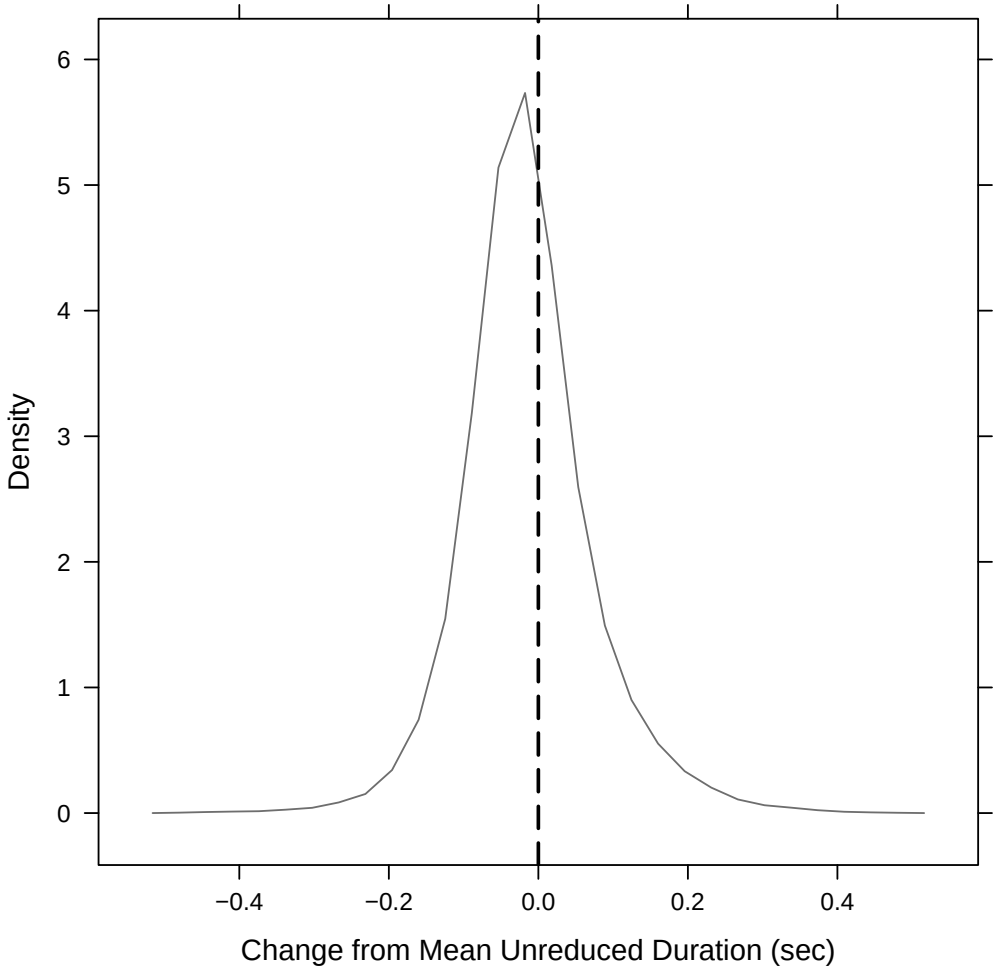

Figure 2.4: Distribution of Word Tokens by Reduction in Word Duration

2.3 Predictors

2.3.1 Demographic Predictors

The corpus includes only a small amount of personal information about its participants and interviewers, preferring instead to control for demographic factors where possible. For example, all interview participants were middle-class Caucasian people who had lived in Columbus Ohio since they were 10 years old or younger. As a result, only three sociolinguistic predictors, each with only two possible values, are included in the present study: Sex of the participant (Male or Female), sex of the interviewer (Male or Female), and age of the participant. Age is divided by the creators of the corpus into 'Old' (over 40 years old) and 'Young' (under 30 years old).

The Corpus was constructed with these three variables fully crossed, with 5 interviews for each of the 8 possible combinations of age, participant sex and interviewer sex.

2.3.2 Phonetic Predictors

Two speaking rates (by speaker and by intonational unit), the number of (expected) stressed syllables in a word, and a word's expected length in segments are included in this category. Expected lengths were unavailable for 4,409 word tokens, or 3.2% of the content words in the corpus. All of these tokens were excluded from the analysis.

Johnson (2004) showed how the transcription of the corpus can be used to easily determine, for a working definition of 'nucleus' and 'margin' at least, whether a segment is in a syllable margin or nucleus. Vowels do not appear in syllable margins in English, and the transcription for each nasal, lateral, and rhotic indicates whether it is in the syllable nucleus (i.e., whether it is syllabic) or not. Thus, syllable position can be read directly from the transcription.

This clear transcription of syllable nuclei aids in the calculation of local and global speaking rates. A measure of local speaking rate similar to that used in Jurafsky *et al.* (2001) was adopted here (namely, the number of syllables-per-second in the intonational unit surrounding the target word.) An operational definition of 'intonational unit' was used, marking intonational unit boundaries wherever a word or tag indicates a potential disfluency in speech. Boundaries include words indicating filled pauses like 'uh' or 'um' and tags indicating silence, laughter, hesitation, non-fluent lengthening of a word, partially produced words, and interviewer speech. Regrettably, proper names redacted from the corpus are also counted as intonational unit boundaries due to ambiguous tagging.

Once these units were extracted, local speaking rate was estimated by simply counting the number of syllable nuclei, as defined above, in each intonational unit, and dividing that number by the duration of the intonational unit in seconds.

The duration and number of syllables in the target word itself are excluded from this calculation, as in Gahl (2008). As a result, one-word utterances are excluded from the analysis. Removing one-word utterances, as well as other tokens for which no speaking rate could be calculated, led to a loss of 2,408 words, or 1.8% of the content words in the corpus.

Figure 2.5 shows how each speaker's speaking rate changes over time. There is no consistent pattern across speakers, with some decreasing in speed near the beginning (e.g. S09)

or end (e.g. S23) of their conversation, some increasing in speed during their conversation (e.g. S37, S07), and many appearing to remain relatively stable.

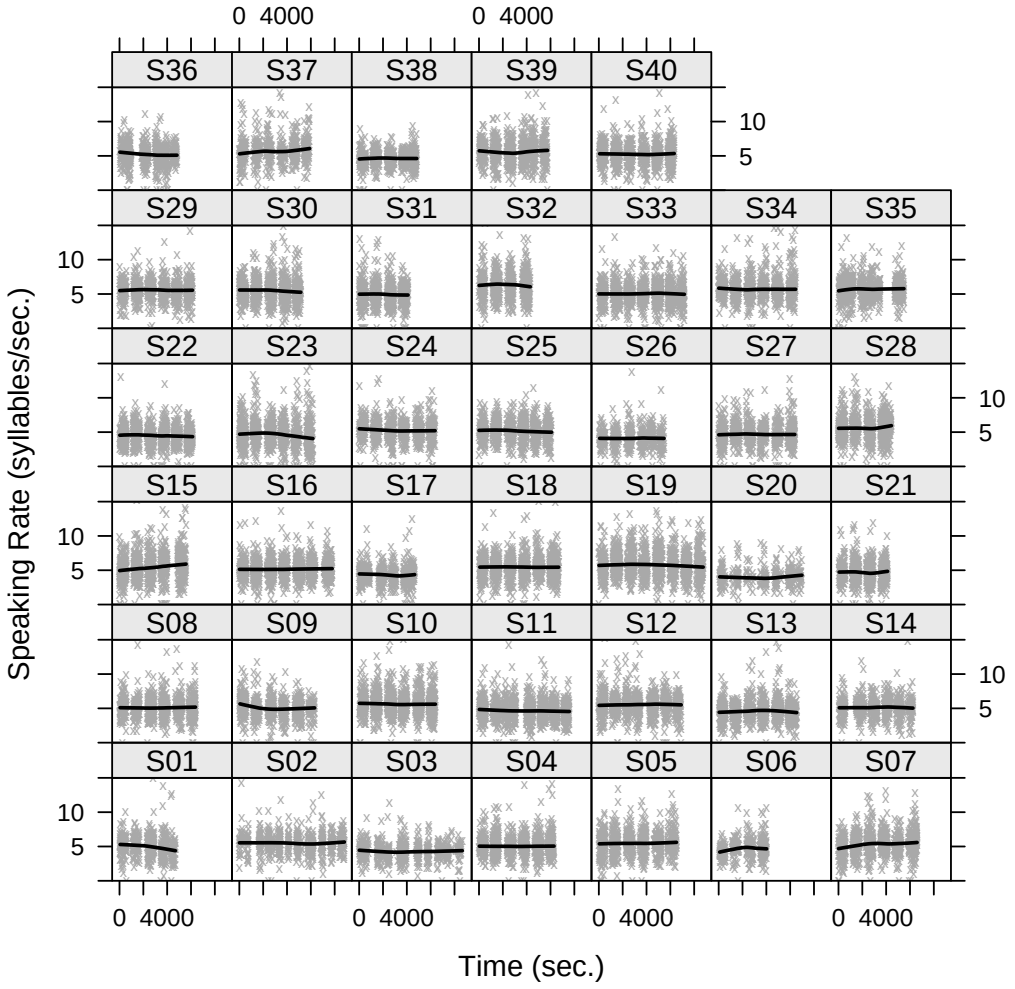

Figure 2.5: Speaking rate over time by speaker

The number of syllables uttered during the entirety of a speaker's interview was then divided by the amount of time taken during that speaker's intonational units, providing a global speaking rate for each participant. The average global speaking rate was about 5 syllables per second.

The number of syllables expected to have either primary or secondary stress was also included as a predictor. This is approximated in the current study by cross-referencing words in the corpus to stress-marked words in the CMU Pronouncing Dictionary where possible.

Stress patterns could not be found for 133 of the tokens in the corpus, and these tokens are not included in the analysis.

2.3.3 Predictability

Four measures of predictability are included here: COCA frequency, Buckeye ('local') frequency, conditional probability given the previous word ('forward word predictability'), and conditional probability given the following word ('backwards word predictability')

COCA frequencies were retrieved from the COCA website (corpus.byu.edu/coca/) on February 3rd, 2010, and include all genres in the corpus. Both frequencies are log transformed, since both are expected to have Zipfian (Zipf 1935) distributions. That is, with a linear decrease in frequency, an exponential decrease in the number of tokens with that frequency is expected. 51 word form types in the Buckeye corpus, totaling 3,623 word form tokens, were not listed (or not listed as single entries) in COCA. In most cases, this was because contractions and possessives are listed as two entries in COCA but one entry in the Buckeye Corpus. Word forms without frequency information are excluded from the analysis

Two predictability measures based on conditional probability are included as predictors: Conditional probability given the previous word and conditional probability given the following word. Both measures were calculated using n-grams taken from the Buckeye corpus itself. These conditional probabilities can not be calculated for words at the beginning or end of intonational units, unless the edge of an intonational unit is itself taken as a kind of word form token. In the present study, however, phrase-initial and phrase-final words are simply excluded from the analysis. Words have been shown to differ in phonetic reduction at the boundaries of phrases. For example, Bell *et al.* (2003) showed that words resist shortening in conversational speech if they come at the beginning or end of an intonational unit. Their work builds on previous research on prosodic domain strengthening (see Fougeron & Keating (1997) for a review), primarily involving laboratory speech, showing final lengthening, final weakening, and initial strengthening (Bell *et al.* 2003). A conditional probability measure that also tracks phrase boundaries would be conflated with, and might be overwhelmed by, phrase-boundary effects. Avoiding such words comes at a cost, though: fully 26,022 word tokens are at phrase boundaries, or 19% of the content tokens in the corpus.

Both conditional probability measures are heavily skewed towards zero, so log transformed versions of these variables are used as predictors during modelling. The effect of this transformation is illustrated in Figure 2.6.

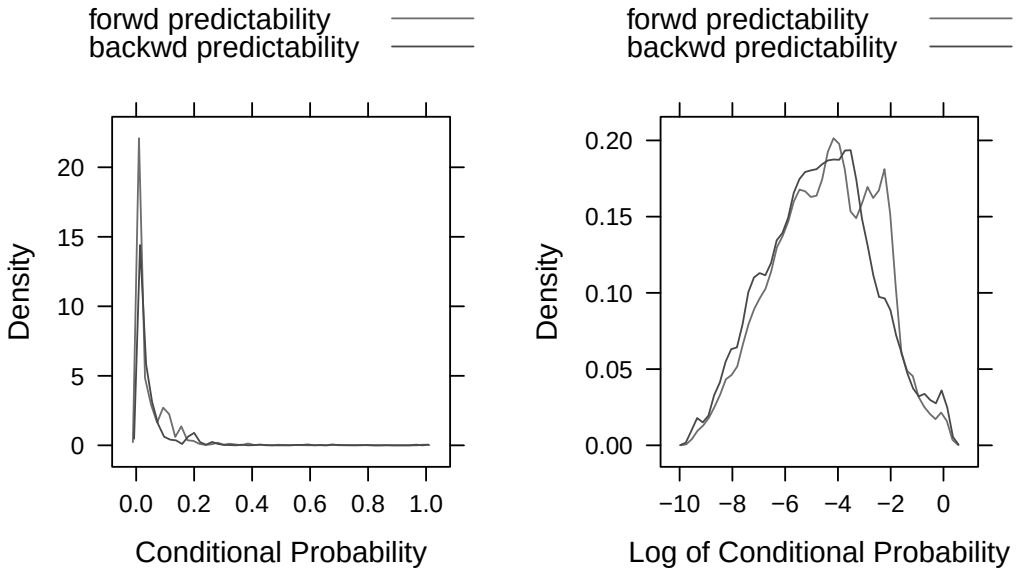

Figure 2.6: Skew in Log Transformed Conditional Probability Distribution

2.3.4 Structural Constituency

A word's position in its carrier phrase ('Phrase order') is included as a naive structural constituency predictor.

A word's part of speech ('POS') is also included as a predictor in the models. With four distinct, unordered values, POS is modelled as an unordered 4-level factor in the models presented here. Unordered factors require a reference value against which the other values in the factor are compared. In this case, some part of speech must be chosen as the POS to which other POSes are compared. To mitigate bias, the **lmer** function's default behaviour was applied to choose the reference POS. As a result, the POS that falls first alphabetically (adjectives, in this case) was taken as the reference level for this variable.

Conditional probabilities based on part of speech n-grams are also included. Three such probabilities were selected for inclusion in the models, for the reasons described below.

While the analysis of word predictability used here is based on 2-grams, the part-of-speech predictability measure was also calculated for larger n-grams. These larger n-grams can capture longer-distance dependencies or larger syntactic structures to use in predicting reduction. These larger structures come at a cost, however: predictability given a large n-gram window can only be calculated for words in long utterances. Predictabilities based on all possible 4-gram windows, for example, can only be calculated for targets in utterances of at least 7 words. Moreover, only the central word of each 7-word utterance could be considered, and in general only words at least three words from both edges of an utterance could be assigned a 4-gram-based predictability measure.

As a compromise, only sequences of three or fewer parts of speech containing a target word were examined in the initial evaluation presented here. Limiting the analysis to 2- and 3-gram parts of speech still comes at a heavy cost, however: 22,486 words must be excluded

from the data, or 16% of the content tokens in the original data set. It is therefore worth evaluating how much extra information the larger POS windows provide.

As with the word-based conditional probabilities, the POS-based conditional probabilities are skewed towards zero. The probabilities are thus log transformed before proceeding with the analysis, reducing their skewness as illustrated in Figure 2.7.

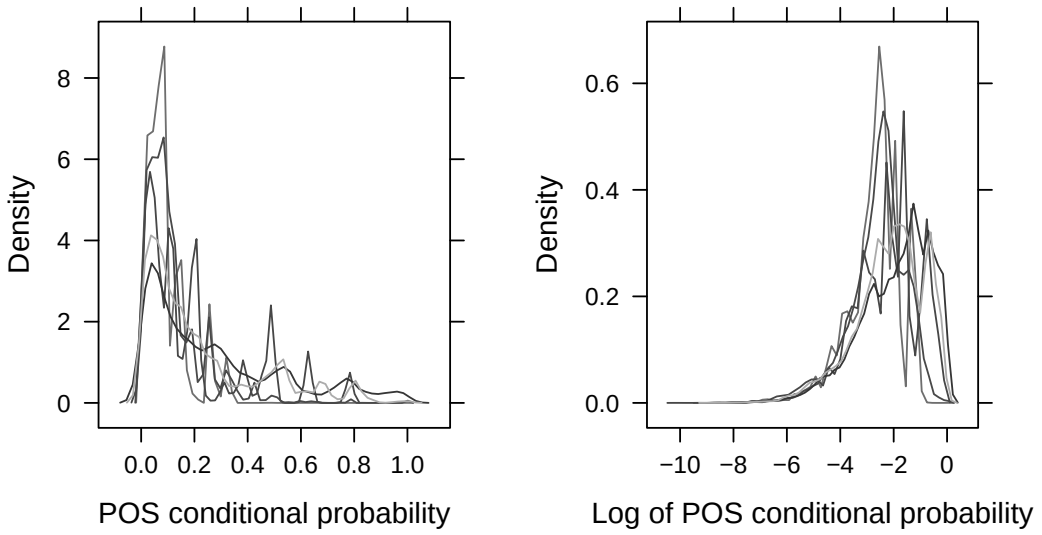

Figure 2.7: Skew in Log Transformed POS Conditional Probability Distribution. Each line represents one of the remaining POS probability predictors.

A hierarchical cluster analysis was then performed, using the **varclus** function (Sarle 1990) in the **Hmisc** package (Harrell *et al.* 2013) using the squared Pearson coefficient to show the correlational structure of the POS conditional probabilities. As shown in Figure 2.8, there are strong ($r^2 > 0.7$) correlations both between the measures of predictability given a word's preceding parts of speech, and between the measures of predictability given a word's following parts of speech. There is also a moderate ($r^2 > 0.5$) correlation between the measure of predictability given a word's surrounding parts of speech and the predictabilities given the preceding parts of speech. In both cases, the correlation is large enough to reduce confidence in random forest analysis (Archer & Kimes (2008) - see Sections 2.5 and 2.6.2 for details). The strong correlations also suggest that the small amount of information potentially gained by including larger POS windows may not be worth the heavy data loss that such inclusion incurs. Thus, the predictability measures given the two preceding or two following parts of speech are excluded from the analysis.

Including the predictability measure based on the surrounding parts of speech, however, may add information without requiring the exclusion of any data points. To address the correlation between this measure and the predictability given the preceding POS, a residualized version of the surrounding POS measure was calculated. For details on the residualization process, see Section 2.5 below.

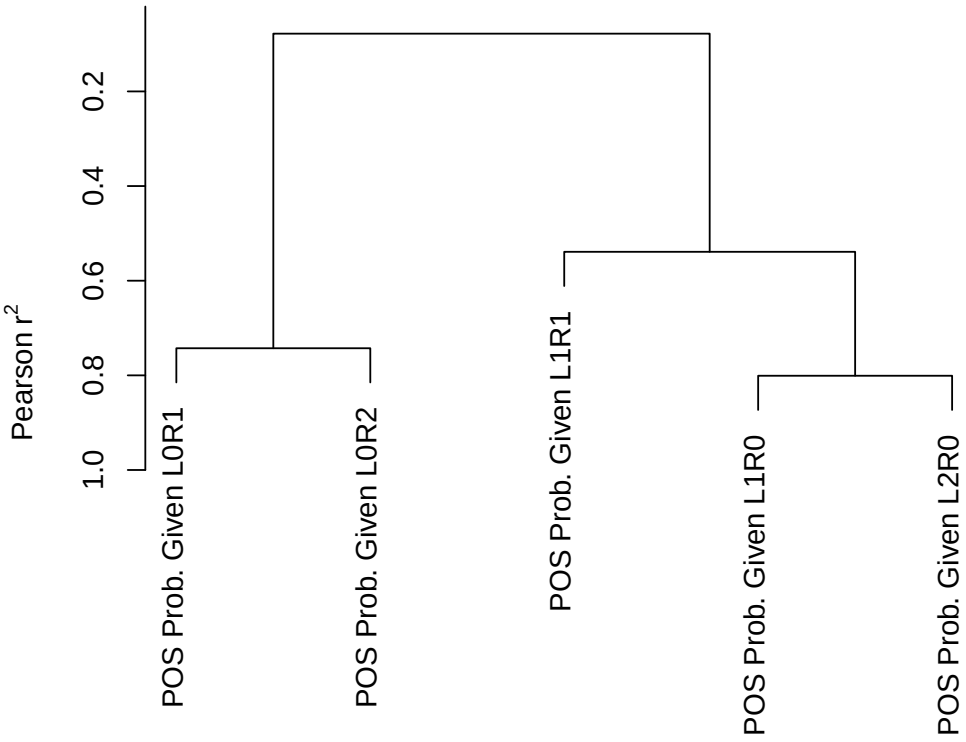

Figure 2.8: Cluster Analysis of Correlation Among POS Conditional Probabilities. 'LnRm' Indicates Conditional Probability given the n Preceding and m Following Parts of Speech

2.3.5 Topicality

Topicality is operationalized in the present study using a measure called term-frequency-inverse-document-frequency (tf-idf). Tf-idf compares the frequency of a word in a document (term frequency) to the number of documents in which that word occurs (document frequency).

More formally, the term frequency of a word form type W_i in a document d_j is calculated here as:

$$tf(W_i, d_j) = \frac{|\{w \in d_j : w \in W_i\}|}{|\{w \in d_j\}|} \tag{2.1}$$

where w represents a word form token in the corpus.

The document frequency of a word form type W_i in the Buckeye Corpus C is calculated here as:

$$df(W_i) = \log\left(\frac{|\{d \in C : W_i \in d\}|}{|\{d \in C\}|}\right) \tag{2.2}$$

The tf-idf of each token in each document can then be calculated by taking the ratio of these two values. That is, for each token w of word form type W_i in document d_j, the tf-idf topicality of w is given by

$$tfidf(W_i, d_j) = tf(W_i, d_j) \div df(W_i) \tag{2.3}$$

Each conversation in the Buckeye corpus comes already divided into chunks that may be considered 'documents' for the purpose of tf-idf calculation: The interviews are broken into a series of files, each composed of roughly 10 minutes of that interview. In calculating tf-idf here, each of these files is taken as a separate document. This allows the relevance of a word to vary within a conversation, and thus within speaker. The tf-idf values appear to be logarithmically distributed, as shown in Figure 2.9, so the raw tf-idf scores were log transformed before further analysis.

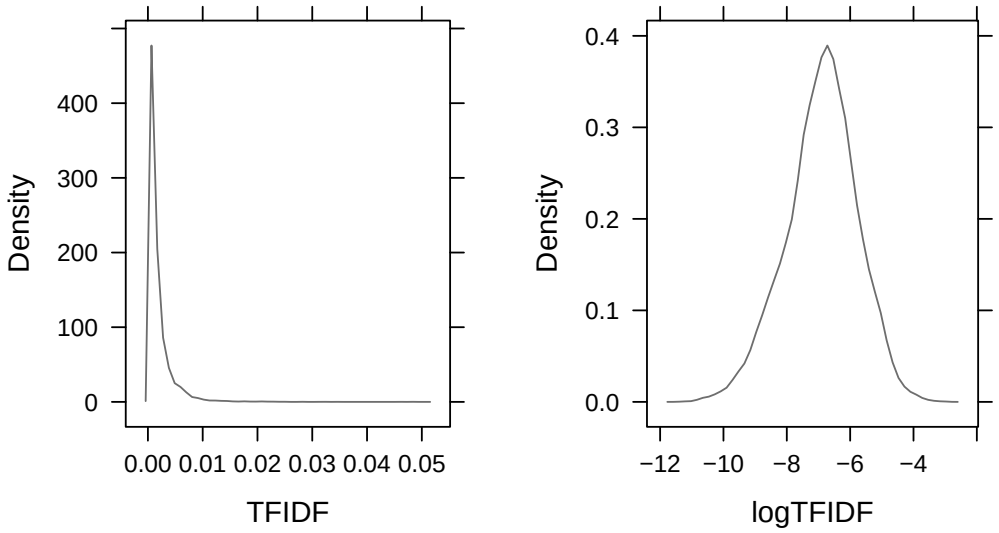

Figure 2.9: Distribution of tf-idf and log-tf-idf values

2.3.6 Time

The time at which the target word appears in an interview is also included as a predictor.

2.4 Remaining Corpus

A total of 74,096 word tokens remain after the trimming processes described above. The remainder of this work attempts to model phonetic reduction in these tokens.

2.5 Addressing Correlation Among Predictors

Correlation between linguistic predictors has long been a significant and confounding problem for psycholinguists (Köhler 1986; del Prado Martín 2003) In the present study, the problem of correlation is compounded by the inclusion of predictors strongly related to each other (e.g. local frequency and COCA frequency) or partially derived from each other (e.g. local frequency and tf-idf). As mentioned in Section 2.3.4 above, Random Forest modelling is resistant to problems normally associated with weak or moderate correlation. This resistance is difficult to quantify, however. In a simulation study, Archer & Kimes (2008) found that a true predictor p of a response was selected as most important in RF models over predictors correlated with p at $\rho \leq 0.5$. For stronger correlations the true predictor still reaches a high importance value, though some of its correlated predictors may achieve a higher importance. Random Forests are also applied to sets of predictors with correlations as high as 0.9 (Díaz-Uriarte & De Andres 2006). Archer & Kimes (2008) also found that RF models converge (and thus provide at least some information about predictor importance) much more often than linear regression models even when predictor correlation reaches $\rho = 0.95$. Based on the results of Archer & Kimes (2008), correlation among predictor variables in the

current study will be reduced below $r = 0.5$. This restriction may seem overly conservative, especially when compared to current practices in random forest modelling. However, it reduces correlation to a level that can be considered acceptable even for linear mixed-effects modelling, allowing both types of model under study to be constructed using the same set of predictor variables.

Correlation will be reduced by the following procedure: First, (squared Spearman rank) correlation among the numeric predictors is estimated using hierarchical cluster analysis, again using the **varclus** function (Sarle 1990; Harrell *et al.* 2013). Next, one variable is selected from each cluster of variables that exceed $r^2 = 0.25$. A set of linear models can then be constructed: For each non-reference variable in a cluster, a linear regression is performed in which the reference variable is used to predict the non-reference variable. The residual (error) values from each these models will then be taken as a new predictor for the larger analysis. Such predictors will have a much lower correlation with the reference variable than the original predictors, while maintaining some of the information encoded by the original variable. These residualized predictors describe how each word form token differs in value between the variable of interest and the variable against which they are residualized.

Results of the first cluster analysis are shown in Figure 2.10.

Figure 2.10 shows two problematic clusters of variables, a part-of-speech conditional probability cluster (Forward POS probability and surrounding POS probability), and a frequency/topicality cluster of three predictors (COCA frequency, local frequency, and topicality). In the first cluster, POS predictability given the preceding POS is chosen as the reference variable, and the POS predictability given the surrounding POSes is residualized against it. The resulting predictor can then be considered a measure of how much extra information is provided by adding the following POS to the context by which a word's part of speech can be predicted:

For the frequency/topicality cluster, COCA Frequency is chosen as the reference variable, as it is the only measure not derived from the Buckeye corpus. The residualized measure of local frequency captures how each word's Buckeye Corpus frequency differs from its COCA frequency.

The residualized measure of tf-idf shows how a word's topicality differs from its overall frequency.

In total, then, three predictors are residualized before further analysis: One part-of-speech predictability predictor, one frequency predictor (local frequency), and one topicality predictor.

A second hierarchical cluster analysis is shown in in Figure 2.11.

This analysis reveals one remaining high-correlation cluster: COCA frequency is correlated with citation word length at a rate of $r^2 = 0.3$. (i.e., $r \approx 0.55$, slightly above the $r = 0.5$ threshold)

For consistency, COCA frequency is again taken as the reference variable. The resulting variable reflects how a word's length differs from the length one would expect it to have given its frequency.

A final cluster analysis is shown in Figure 2.12.

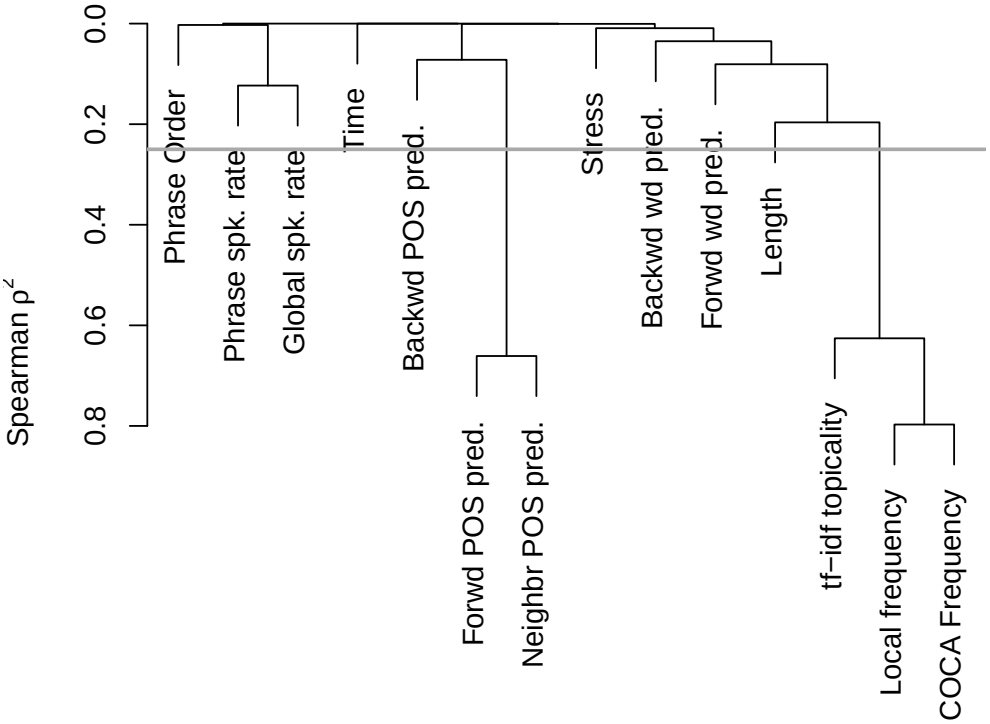

Figure 2.10: Cluster Analysis showing correlation among numeric predictors

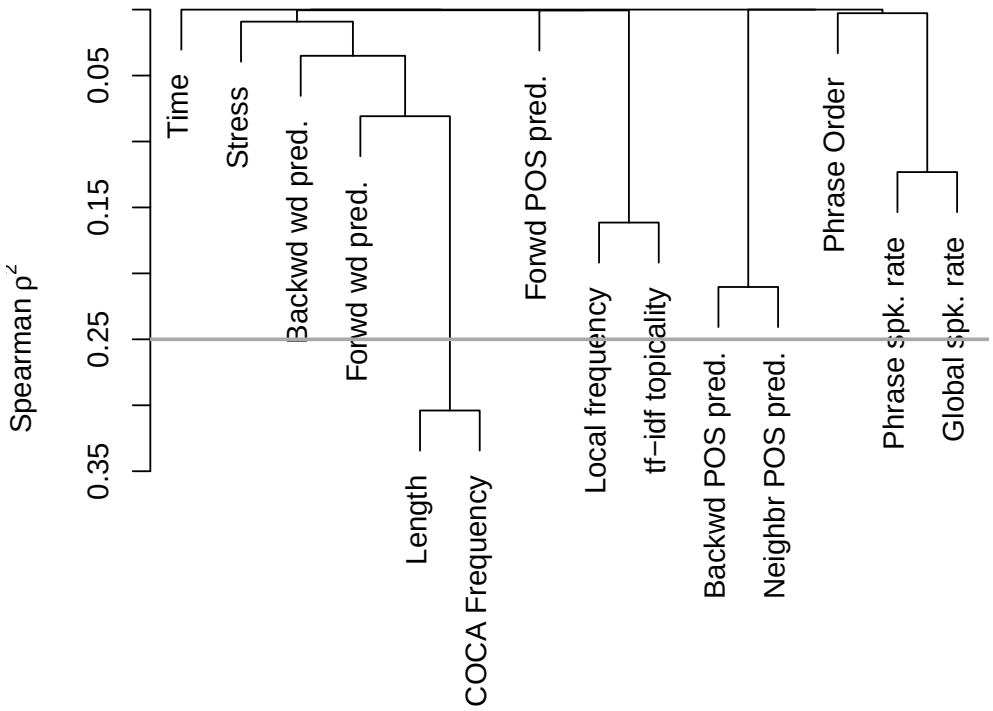

Figure 2.11: Cluster Analysis of Predictors Remaining after First Residualizations

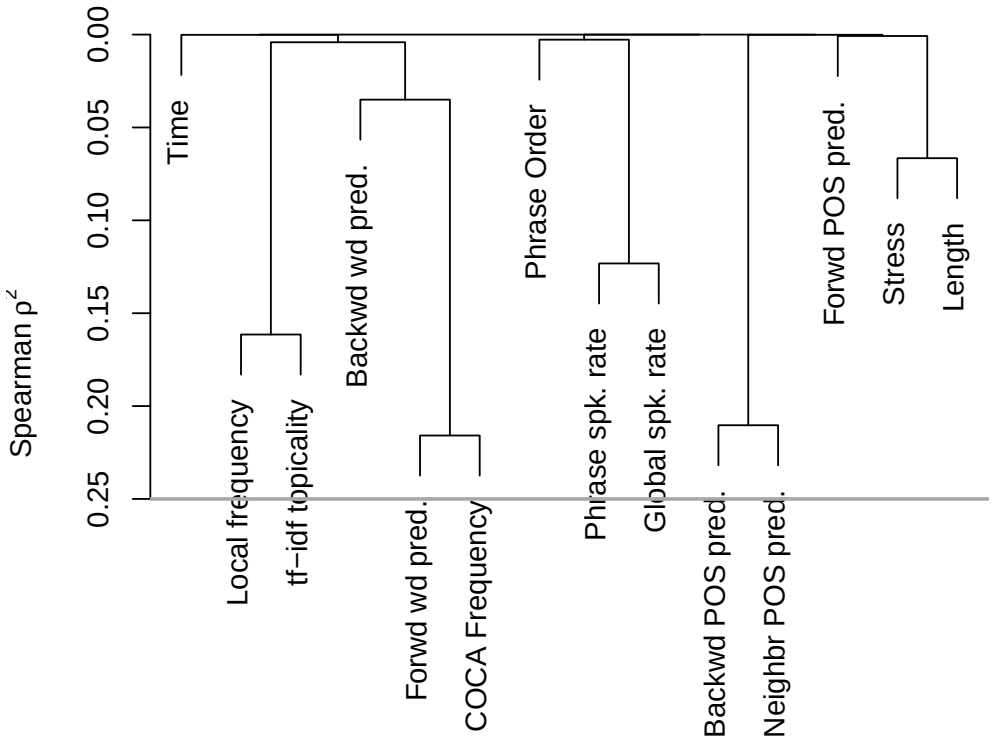

Figure 2.12: Cluster Analysis of Predictors Remaining after Final Residualization

The maximum variance among the remaining variables is $r^2 = 0.22$

One final pair of transformations is applied to each of the resulting numerical predictors: Each predictor is first shifted until its values are centered around zero. Next, each predictor is scaled by dividing each value by the standard deviation of the values of that predictor. Before scaling and centering, the numerical predictors have a condition number of 56.5. After scaling and centering, the condition number is reduced to 2.9.

Scaling and centering are not applied in order to reduce collinearity. Indeed, they can not do so (Belsley 1984). The centering simplifies the interpretation of correlation between random slopes and random intercepts in mixed effects models (Baayen 2008). This scaling increases the likelihood that mixed-effects models containing these predictors will converge, and decreases the potential for accuracy loss that comes from numerical computation over data with vastly different scales.

2.6 Modelling Techniques

Two modelling techniques are applied to the prediction of each dependent variable. The first technique, linear mixed-effects regression (LMER) modelling is described in Section 2.6.1 below. The second technique, Random Forest (RF) modelling, is described in Section 2.6.2 below. The results of these modelling techniques are combined to produce an LMER model informed by the Random Forest model (the *RF-Informed* model), as described in Section 2.6.3 below.

2.6.1 Linear Mixed-Effects Regression Modelling

In the word-based models described here, the 18 variables described in Table 2.2 are considered potential fixed-effect predictors. Two group-level predictors, word form type and speaker identity, form the basis of the random-effects structure of these models.

For each measure of word reduction, a series of linear mixed-effects regression models are constructed. To begin the process, a simple model is created with main effects for each predictor described in Table 2.2 and random intercepts for word form and speaker. This model serves as a baseline to which information is gradually added or refined until an optimal model is reached.

The next model is identical to the baseline model above, but with the insignificant main-effect predictors removed. (Details regarding significance testing are outlined below.)

The next modelling step involves searching for interactive effects between all pairs of the remaining fixed-effects predictors. The analysis in this step is consistent with an exploratory approach, in which no interaction is considered more likely *a priori*, and strong evidence is required before an interaction is selected for inclusion in the final model. This process involves several stages of model construction. First, a model is constructed with every possible two-way interaction between these predictors, along with main effects for each individual predictor and random intercepts for word and speaker. Next, an initial trimming removes any insignificant predictors from further consideration. For each of the remaining interactions, a model was created with that interaction, all significant main effects, and random intercepts for Word and Speaker. If adding this interaction did not significantly

Description	Brief Description
Age of speaker ('young' or 'old')	Age
Time when token appears in conversation	Time
Length of word (Resid.)	Length
Part of Speech	POS
Gender of interviewer	Interviewer Gender
COCA Frequency	COCA Frequency
Conditional prob. given following word	Backwd wd pred.
Conditional prob. given previous word	Forwd wd pred.
Conditional prob. of POS given following POS	Backwd POS pred.
Conditional prob. of POS given previous POS	Forwd POS pred.
Cond. prob. of POS given surrounding POSes (Resid.)	Neighbr POS pred.
Local frequency (Resid.)	Local frequency
tf-idf topicality (Resid.)	tf-idf topicality
Number of stressed syllables	Stress
Word's (ordinal) position in this phrase	Phrase Order
Gender of speaker	Gender
Average speaking rate for this speaker	Global spk. rate
Speaking rate for this phrase	Phrase spk. rate

Table 2.2: Predictor Variable Descriptions. Brief Descriptions are often used in Tables to Conserve Space

improve the model over the baseline model, it was eliminated from further consideration. Finally, a model was constructed with all significant main-effect predictors, the interactions found to significantly improve the model, and random intercepts for Word and Speaker.

The next step in the current model selection procedure is to explore the random-effects structure of the data more fully. Each of the preceding models included only random intercepts for word and speaker. At this stage, random slopes for each of the 18 original predictors, by both word and speaker, are evaluated.

Not every combination of predictor and random effect is sensible, of course: Random slopes allow each member of a group to react to a predictor variable differently. When the value of a predictor remains constant for each item within each member of a group, however, including separate slopes for each member of that group is not sensible. Word length, for example, remains constant within word (i.e., a word has the same (predicted) number of segments throughout the corpus). With only one value of length for each word type, a plot of length vs. reduction for a single word type would appear as a vertical series of dots. The line of best fit for such a plot would be a vertical line, with an undefined slope, and thus any attempt to fit a regression model to such a data set would fail. Thus, random slopes were only fitted for fixed-effect predictors that vary within-group for each of word form type and speaker.

Table 2.3 lists which variables have potential random slopes for each group level predictor. All such potential random slopes are tested, as described further below.

To determine the ideal random effects structure of a model, random slopes are added to the model and their usefulness in predicting reduction is evaluated. Random slopes are included for each fixed-effect predictor that varies within-group for each of word form type and speaker.

Predictor	Possible Slopes
Age	Word
Interviewer Gender	Word
Gender	Word
Global spk. rate	Word
Length	Speaker
POS	Speaker
COCA Frequency	Speaker
Local frequency	Speaker
Stress	Speaker
Time	Both
Backwd wd pred.	Both
Forwd wd pred.	Both
Backwd POS pred.	Both
Forwd POS pred.	Both
Neighbr POS pred.	Both
tf-idf topicality	Both
Phrase Order	Both
Phrase spk. rate	Both

Table 2.3: Set of Random Slopes for each Predictor

To determine which random slopes significantly improve the model, each random slope in Table 2.3 was added separately to a baseline model, and the improvement in model fit was measured. The baseline model selected contained random intercepts for word and speaker, as well as the significant main-effect predictors selected during the first step of the modelling process. Interactions were left out of this baseline to facilitate faster computation.

The addition of significantly helpful random slopes to the model with random intercepts and significant interactions and main effects completes the model selection process. The resulting "full" model will then be put through the model criticism process described below.

Two additional models are created to allow for a broader comparison between LMER models and random forests. The first model contains only the random-effects structure included in the final, "full" model described above, no fixed-effect predictors. This model shows how much of the LMER model's effectiveness comes from the predictive powers of word form and speaker. These two predictors are relatively easy to consider in an LMER model but nearly impossible to include in an RF model: Categorical variables with a large number of factor levels exponentially increase the problem space through which random forest models must search for an optimal fit.

The second model contains only the fixed-effects structure of the "full" model. This model can be used to estimate how much of the predictive power of the "full" model comes from its fixed-effects, including both main and interactive-effects. This model can also be thought of as a linear model including only those predictors available to the random forest models described in Section 2.6.2. This allows for a more direct comparison of linear mixed-effects models and random forest models. The amount of information available will not exactly match that of the random forest model, however. Random forest models implicitly consider interactions at arbitrary levels of complexity, and the number of variables allowed to interact can be easily increased well beyond 2 by a standard tuning parameter. However, n-

way interactions must be specified in advance for linear modelling, and increasing n increases the number of terms in the model (and the likelihood that model fitting will not converge) exponentially. The number of three-way interactions possible between the 14 variables in the present study, and thus the number of additional terms that would need to be added to explore three-way interactions, is $\binom{14}{3}$, or 364. An LMER fit to this many terms is unlikely to converge.

The models outlined in this section will be compared to each other in three ways: With log-likelihood ratio testing, Akaike Information Criterion (AIC - (Akaike 1974)) comparison, and overall model fit, as measured by the proportion of variance in the data explained by the model. All three tests consider how well the model fits the data, and AIC considers how parsimoniously a model does so.

2.6.1.1 Significance Testing

Each step in the modelling process may require a slightly different way of estimating significance. A somewhat conservative approach is adopted here, due to both the large number of data points involved and to the exploratory nature of the model selection process.

Where appropriate, two tests are applied to determine which predictors are correlated significantly with duration change in each model. For the main-effects and interaction structure of the duration model, both t-values and p-values can often be calculated. The absolute t-value is used as a measure of how far and how consistently each fixed effect in the model differs from zero. Absolute t-values below 2 reflect a low level of confidence that a fixed-effect predictor is contributing significantly to the model (Baayen *et al.* 2008), and predictors that fall below this threshold are excluded from future models. Due to the large number of potential interactions considered and the somewhat exploratory nature of the present study, a stricter threshold is applied to interactive effects: In the initial interaction-trimming stage, interactions with $|t| < 3$ (rather than $|t| < 2$) are removed from further analysis.

Estimated p-values can also be calculated for many of the duration models using the Monte Carlo simulation methods provided by the **pvals.fnc** function in the **languageR** package (Baayen 2011). If a predictor's p-value is greater than $p = 0.05$ in these simulations, it is excluded from further models. To be included in the final model, then, a predictor must pass both the t-value and p-value tests where both are available. This increases confidence that this predictor is linked to changes in word duration in speech. Unfortunately, **pvals.fnc** can not yet calculate p-values for models with parameters for correlation between random slopes. Such models will not be subjected to p-value testing, and t-values will be used as the primary test for significance.

The models attempting to predict deletions in Chapter 4 are fit as Poisson models rather than simple linear models. For Poisson models, the **lmer** function provides a true p-value based on a z-test. This p-value alone is used as the significance test for such models.

The next step of the interaction modelling process, in which a series of models are constructed with a single interaction added, must be thought of as a set of post-hoc tests for significance. Some post-hoc testing correction is appropriate, and in this case a simple Bonferroni correction (Bonferroni 1935) was used, in which the significant p-value ($p =$

0.05 here) is divided by the number of post-hoc tests performed. Each interaction must improve the model by a significant amount to pass through to the next stage of modelling. Improvement was calculated using log-likelihood estimation, as implemented by the **anova** function in **R**. The threshold significance level for these tests will be reduced by Bonferroni correction to well below 0.05.

The search for descriptively useful random slopes also involves multiple post-hoc testing. As with interactions testing, random slopes are tested for significance using multiple rounds of log-likelihood estimation. Bonferroni correction was again applied to determine a more appropriate significance threshold than $p < 0.05$ for these slopes.

2.6.1.2 Model Criticism

During the model criticism stage, an attempt is made to determine whether the model is being unduly influenced by a set of unusual data points. The procedure for model criticism applied here is derived from the process used in Baayen (2008).

The difference between the predicted and actual reductions for a production is known as a residual value. (i.e., a residual is the amount of reduction "left over" after predicted reduction is accounted for.) These residuals are normalized (i.e., transformed into z-scores) for the purpose of model criticism. Data points with very high or low residuals reflect words for which the model's prediction are poor. Their duration reduction or number of deletions is well beyond what is expected given the prediction of the model.

High- and low-residual points are considered potential outliers - word productions that may be unduly influencing the results of model fitting. At this stage, these potential outliers (here, tokens with reduction values more than 2.5 standard deviations away from their predicted values) are temporarily removed from the data set and an identical model is refit. This new, trimmed model is then compared to the untrimmed model. In particular, if the trimmed model shows qualitatively different results - predictors that dramatically change in their modelled effect on reduction, for example - the trimmed model is considered preferable, and used in the remaining analysis. If the trimmed model does not differ dramatically from the untrimmed model, however, the full data set and untrimmed model are used in the analysis that follows.

The trimmed model is preferred only if a dramatic, qualitative change in the structure of the model takes place. This untrimmed model is preferred by default. This is due to the fact that trimming data points for which the model makes poor predictions is anti-conservative. That is, trimming poorly-fit data points may artificially inflate the quality of the model fit, without adding to the accuracy of the model's prediction of the reduction process.

2.6.2 Random Forest Modelling

Random Forest modelling (Breiman 2001) is an ensemble method in which the results of several simple models are combined to make predictions about the data as a whole. In Random Forests, the simpler models are regression trees (Breiman *et al.* 1984). Individuals trees are constructed by selecting a series of "splits" - values of a predictor variable at which the data is divided into two groups. Splitting continues recursively until the data are partitioned into a set of "leaf" nodes based on their values of the dependent variables at

each split. The regression value of each data point in a leaf node is taken as the average value of the response variable within that node. Taken together, these values provide a piece-wise-constant regression curve, with each constant piece of the curve defined for a set of values of the predictor variables.

At each step of this process, a splitting value must be chosen from among all possible values of each predictor variable in the model. A given splitting value will divide the data into two groups, each with a different mean value of the response variable. The splitting variable and value that produce the smallest total variation around these mean response values is chosen at each step. This variation can be seen as the error term in the piece-wise-constant regression defined by the split: That is to say, the response values for each group are being modelled as the mean response value plus some error term. Splits are chosen to minimize this error at each step.

A sample regression tree is provided in Figure 2.13. The figure was generated using the **ctree** function in the R package **party** (Hothorn *et al.* 2006).

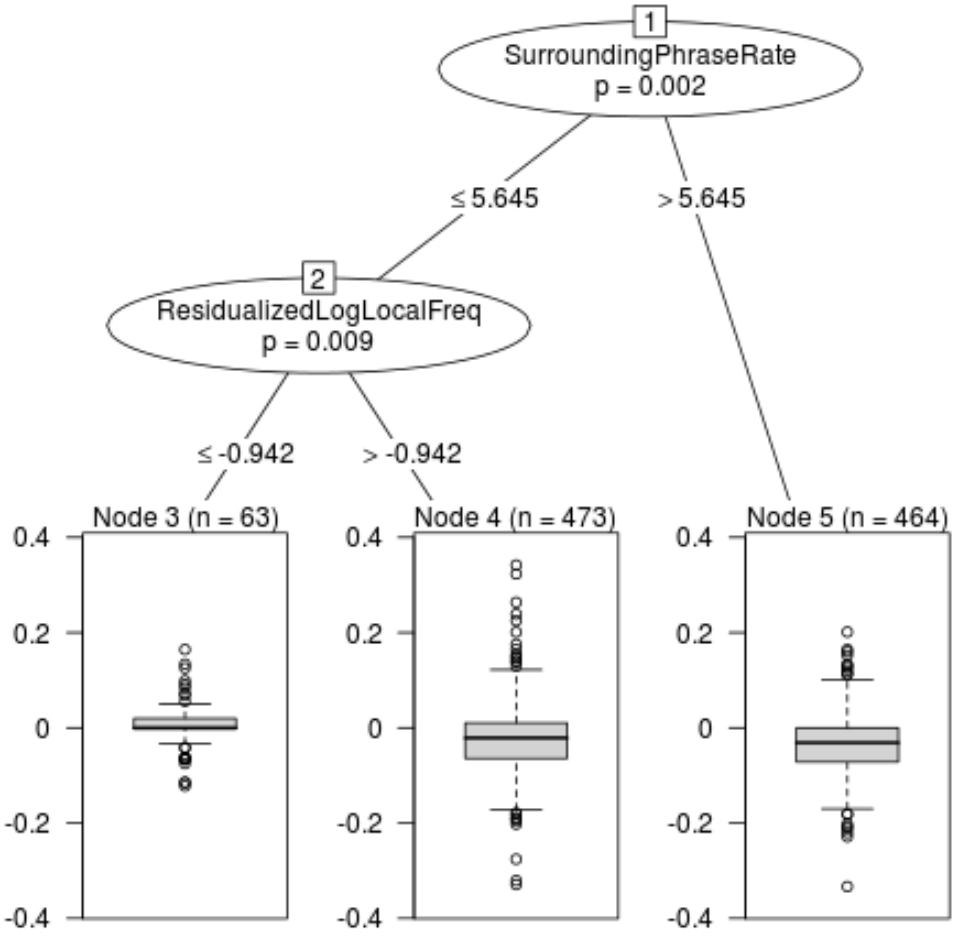

Figure 2.13: A Sample Regression Tree

In Figure 2.13, the first split divides words based on local speaking rate. Words in faster phrases (above about 5.6 syllables-per-second) are passed down to the right into (leaf) node 5. A box and whisker plot shows the distribution of the response value (in this case, change in word duration) for the word tokens in this group. The average duration change is slightly below zero, indicating that words in faster phrases are likely to have reduced durations. Tokens in slower phrases are further divided into two groups (at node 2) based on their relative frequency in the Buckeye corpus: Words with relatively high local frequencies will fall into leaf node 4, and are expected to show a slight decrease in duration on average. Words with lower local frequencies fall into node 3. These words are not expected to reduce, as the average duration reduction of words in node 3 is at or slightly above zero. In general, then, each word is passed down through the tree until it reaches a leaf node. The average duration change of words in that node is then taken as the predicted duration change for this word.

It is worth noting that these regression trees do not find a globally-optimal partitioning of the data. If these trees did provide a globally-optimal partitioning of the data, an ensemble of them would not be needed: One could simply construct a regression tree on the entire data set, confident that no improved partitioning could be found to model the data. Instead, the algorithm proceeds greedily at each step, selecting the partition that provides the optimal prediction at the local level. Global trends in the data are found by constructing a large number of these regression trees on randomly-chosen subsets of the data and predictor variables. The predictions of these trees are then aggregated to predict both the importance of each predictor variable to the modelling process, and the predicted response values for each data point in the model.

Each tree is constructed using a subset of the data, leaving several "Out-of-bag" (OOB) data points that can be used for cross-validation of the effectiveness of that tree. These OOB data points are key to the aggregation process. First, each OOB data point is passed through each tree to generate a predicted response value. The aggregate prediction of the overall forest model is taken by averaging the individual OOB predictions. A word token's response value in the regression model is thus taken as the average prediction over all trees for which it is out-of-bag.

To measure the importance of each predictor variable, OOB data points are passed through the regression trees twice for each splitting variable: In one pass, prediction proceeds as normal, and error in the prediction is noted. In the second pass, a token's value on the relevant predictor variable is permuted randomly, and the error of the resulting prediction is again noted. This second pass produces a baseline prediction. It can be seen as a kind of null hypothesis, in that it measures the predictive success of the tree when the values of a predictor variable are chosen at random and thus uncorrelated with the response variable. The effectiveness of a predictor can then be calculated by comparing the error rate of these two predictions: If randomly permuting the values of this predictor produces a large drop in accuracy, the "null hypothesis" can be rejected, and the variable is considered a useful predictor of the response. The overall importance of a predictor variable can then be calculated by averaging the (normalized) decrease in accuracy for this predictor across all trees for all OOB data points.

The use of OOB data points in calculating model predictions and variable importance scores precludes the need for independent cross-validation: In a sense, cross-validation is automatically performed for each tree in the forest, and the number of trees in the forest can be quite large.

The number of trees in a random forest (called *ntree*) can be seen as one of three tuning parameters: The second tuning parameter is the number of data points used to construct each tree, called *sampsize* in the randomForest package. The third tuning parameter is the number of predictor variables used to construct each tree, called *mtry*. In the present study, default values provided in the randomForest package are used to choose values for these parameters, as these default parameters represent best practices for RF modelling. One value is chosen for each of the first two parameters: An *ntree* value of 500 trees, and a *sampsize* value of 63.2% of the data. For *mtry*, tuning of the model is performed using the **tuneRF** function, again using the default values provided: First, a forest is constructed with $mtry = \lfloor \sqrt{p} \rfloor$, where p is the number of predictor variables in the model. (For the present study, the starting value of *mtry* is 4.) Next, forest models are constructed with a smaller or greater number of predictors by halving or doubling the value of *mtry*, respectively. The OOB error in prediction for each of these models is then compared to that of the original forest. The search for an optimal value of *mtry* continues recursively in each direction, terminating when a doubling or halving produces a higher OOB error than the previously-constructed model.

The optimal random forest model (i.e., the model with the optimal value of *mtry*) can then be compared to the LMER models described in the previous section. Comparisons are made in terms of three properties: First, the proportions of variance explained by each model are compared. Second, the importance of each predictor in the random forest model is compared to the effect size of that predictor in the optimal linear model. Third, partial-effects plots for random forest and linear models are compared.

2.6.3 Combining Modelling Techniques

The two modelling techniques applied above are likely to produce different inferences about how each predictor affects reduction. An attempt to integrate the results of these techniques is thus performed in order to produce a model of reduction that incorporates the results of each of the previously constructed models. The combination process itself necessarily depends on the results of those models, and thus cannot be summarized here. Instead, the specifics of the model combination process are described in each of the following two chapters, after the final results of initial modelling have been completed.

Chapter 3

Word Duration Reduction

3.1 Introduction

The present chapter describes the models that were constructed to describe reduction in word duration. Details of the modelling process are described in the previous chapter, but a brief overview is provided here. In Section 3.2, a series of linear mixed-effects regression models are constructed until a final, optimal model is discovered using some portion of the current set of predictors. (Fixed-effect predictors are listed in Table 3.1) The linear models are compared to each other in three ways: With log-likelihood ratio testing, Akaike Information Criterion comparison, and overall model fit, as measured by the proportion of variance in the data explained by the model (i.e., R-squared).

In Section 3.3, a random forest model is constructed using these predictors. The random forest model is compared to the final linear model. This comparison is made in terms of three properties: First, the proportions of variance explained by each model are compared. Second, the importances of each predictor in the random forest model is compared to the importance of these predictors in the optimal linear model. Third, partial-effects plots for both random forest and linear models are compared.

Section 3.4 describes an attempt to combine the results of LMER and RF modelling techniques.

Finally, Section 3.5 discusses the overall results, and their implications with regard to reduction.

Predictor
Age of speaker ('young' or 'old')
Time when token appears in conversation
Part of Speech
Gender of interviewer
COCA Frequency
Backwards word predictability
Forward word predictability
Backward POS predictability
Forward POS predictability
Number of stressed syllables
Word's (ordinal) position in this phrase
Length of word (res.)
Surrounding POS predictability (res.)
Local frequency (res.)
tf-idf topicality (res.)
Gender of speaker
Global speaking rate
Speaking rate for this phrase

Table 3.1: Predictor variable descriptions

3.2 Linear Mixed-Effects Regression Modelling

3.2.1 Baseline Model

The baseline model was fit with the 18 main-effect predictors listed in Table 3.1 along with random intercepts for word form and speaker. Table 3.2 shows the significant main effects (i.e., those that passed both the p- and t-value thresholds described in the previous chapter).

Description	Effect (ms)	Std.Err.	t.value	Pr(>|t|)
Speaking rate for this phrase	-13.0	0.3	-37.1	0.000
Backwards word predictability	-12.3	0.3	-35.9	0.000
Length of word (res.)	-8.4	0.5	-15.9	0.000
Forward word predictability	-4.9	0.4	-13.9	0.000
(Intercept)	-28.7	2.7	-10.8	0.000
Global speaking rate	5.5	0.8	6.6	0.000
Surrounding POS predictability (res.)	-2.2	0.4	-6.3	0.000
COCA Frequency	7.1	1.2	5.9	0.000
tf-idf topicality (res.)	-2.0	0.4	-5.8	0.000
Adverbs	-8.7	2.2	-4.0	0.000
Local frequency (res.)	-2.2	0.6	-3.5	0.000
Nouns	-5.6	1.7	-3.3	0.001
Verbs	-4.0	1.8	-2.2	0.027

Table 3.2: Main effects for baseline model. Effects with $|t| < 2$ or $p > 0.05$ are not shown

(A full description of the findings of the initial LMER modelling process, once a final linear model has been arrived at, is found in Section 3.2.6).

3.2.2 Removing Insignificant Predictors

The next model is identical to the baseline model above, but with all of the insignificant main-effect predictors (i.e., those with $|t| < 2$ or $p > 0.05$ in the previous model) removed en masse. Estimates for the fixed-effects parameters and variation in the random intercepts changed only minimally from the baseline model. The models are compared in Table 3.3 below.

	Model 1	Model 2
Degrees of Freedom	24	16
Proportion of Variance Explained (%)	11.9	11.9
log-likelihood ratio	88573	88569
log-likelihood improvement		-4.6
Log-likelihood improvement p-value		0.33
Akaike Information Criterion (AIC)	-177099	-177106
AIC Improvement		6.9

Table 3.3: Model Comparison: Baseline model (1) v. Model with Non-significant Main Effects Removed (2)

The smaller model is a weaker fit to the data than the baseline model, as shown by the slight decrease in log-likelihood ratio estimation. This difference does not reach significance, however. Moreover, the increase in AIC suggests that the smaller model is a more parsimonious description of the data. It is also worth noting that neither model is a strong fit for the data *in toto*, as both explain only 11.9% of the variance in duration reduction.

3.2.3 Exploring Interactive Effects

The evaluation of interactive effects on reduction in the present study is a multi-stage process. First, a model was constructed with every possible two-way interaction between all of the fixed-effect predictors that survived the trimming above. The model also contained main effects for the significant predictors and random intercepts for word and speaker. After the model was constructed, interactions with absolute t-values below a certain threshold were immediately removed from further consideration. (Due to the large number of possible interactions, and thus the increased likelihood of false positives, the threshold was raised from the main-effects level of $|t| > 2$ to the more stringent $|t| > 3$) This left a total of 10 interactions.

Next, a series of models were created to test each of the remaining interactions individually. These models provide a kind of double-check on the importance of each interaction, These interactions were each added individually to a separate model with only significant main effects and random intercepts for word form and speaker. Each such model was then compared to a nearly identical model with only the interaction under study absent. The two models were compared using log-likelihood ratio testing, establishing how the presence of the interaction under study improved the model fit. After Bonferroni correction (that is, after dividing the p-value threshold by the number of tests performed), the highest significant p-value dropped from 0.05 to ≈ 0.005. A total of 8 interactions remained after this process.

Finally, a model was constructed with the significant main effects, random intercepts for word and speaker, and the 8 interactions remaining after the preceding selection process.

	Model 1	Model 2
Degrees of Freedom	16	30
Proportion of Variance Explained (%)	11.9	12.3
log-likelihood ratio	88569	88818
log-likelihood improvement		249.4
Log-likelihood improvement p-value		0
Akaike Information Criterion (AIC)	-177106	-177577
AIC Improvement		470.7

Table 3.4: Model Comparison: Model without Interactions (1) v. Model with Selected Interactions (2)

Table 3.4 shows that adding these interactions led to strong improvements in both log-likelihood estimation and AIC. The increase in the proportion of variance in the data explained by the model is noticeable, but small ($\approx 0.4\%$)

3.2.4 Exploring Random Effects

The next step in the current model selection procedure is to explore the random-effects structure of the data more fully. Each of the preceding models included only random intercepts for word and speaker. Here, random slopes for each of the 18 original predictors, by word form or speaker or both, are evaluated. The set of random slopes under consideration in the present study are enumerated in Table 3.5, replicated from the previous chapter. To determine which random slopes significantly improve the model, each random slope in

Description	Random Slope By:
Age of speaker ('young' or 'old')	Word
Gender of interviewer	Word
Gender of speaker	Word
Global speaking rate	Word
Part of Speech	Speaker
Number of stressed syllables	Speaker
Length of word (res.)	Speaker
Local frequency (res.)	Speaker
Time when token appears in conversation	Both
COCA Frequency	Both
Backwards word predictability	Both
Forward word predictability	Both
Backward POS predictability	Both
Forward POS predictability	Both
Word's (ordinal) position in this phrase	Both
Surrounding POS predictability (res.)	Both
tf-idf topicality (res.)	Both
Speaking rate for this phrase	Both

Table 3.5: Random Effects Structure for LMER Modelling

Table 3.5 was added separately to a baseline model, and the improvement in model fit was measured. Improvement was calculated using log-likelihood estimation, as implemented by

the **anova** function in R. The baseline model selected contained random intercepts for word and speaker, as well as the significant main-effect predictors selected during the first step of the modelling process. (i.e., the predictors or levels listed in Table 3.2). Interactions were left out of this baseline to facilitate faster computation.

As this process consists of multiple post-hoc significance tests, Bonferroni correction was again applied to determine a more appropriate significance threshold than $p < 0.05$. As a result, the significance level ($p < 0.05$) was divided by the number of tests performed (27), leading to a significant p-value of ≈ 0.0019.

Once this process was completed, all of the random slopes that were found to contribute significantly to model fit were added to the model with significant main-effects predictors and random intercepts for word form and speaker. Table 3.6 describes the random-effects

Groups	Predictor	Std.Dev. (ms)	Corr		
Residual		70.387			
Word	(Intercept)	16.363			
Speaker	(Intercept)	4.272			
Word	Age (> 40)	4.900			
	Age (< 30)	5.355	0.05		
Word	Time	2.286			
Speaker	Time	2.245			
Speaker	Adjectives	3.249			
	Adverbs	5.931	0.52		
	Nouns	5.353	0.20	0.20	
	Verbs	2.825	-0.10	0.64	0.01
Word	Female Interviewer	0.001			
	Male Interviewer	9.531	0.00		
Speaker	COCA Frequency	3.828			
Word	Backwd wd pred.	8.944			
Speaker	Backwd wd pred.	3.574			
Word	Forwd wd pred.	6.453			
Word	Backwd POS pred.	10.399			
Word	Forwd POS pred.	5.352			
Speaker	Stress	3.132			
Speaker	Length	4.088			
Word	Neighbr POS pred.	5.828			
Word	tf-idf topicality	5.787			
Speaker	tf-idf topicality	1.516			
Word	Female Speaker	6.203			
	Male Speaker	6.160	0.07		
Word	Global spk. rate	3.127			
Speaker	Phrase spk. rate	5.549			

Table 3.6: Random Effects Found to Contribute Significantly to Model Fit

structure of this model. Table 3.7 compares the model fit of the random-slopes model to the previous model. Adding random slopes improved both AIC and log-likelihood scores, even with all interactions removed.

Table 3.7 also shows the most sizable jump in the proportion of variance explained by a model so far. Adding random slopes improved predictive power by 7.3 percentage points, an increase of about 59%

	Model 1	Model 2
Degrees of Freedom	30	50
Proportion of Variance Explained (%)	12.3	19.6
log-likelihood ratio	88818	89315
log-likelihood improvement		496.3
Log-likelihood improvement p-value		0
Akaike Information Criterion (AIC)	-177577	-178529
AIC Improvement		952.5

Table 3.7: Model Comparison: Main and Interactive Effects (1) v. Main and Random Effects (2)

Next, a more complete model was constructed, with all significant main effects, interactions, random intercepts, and random slopes.

Table 3.8 compares this full model to the best-fitting model constructed previously, the model with main effects and a full random-effects structure. The full model shows improvements in all measures of model fit, including a significant improvement in log-likelihood ratio.

	Model 1	Model 2
Degrees of Freedom	50	64
Proportion of Variance Explained (%)	19.6	19.2
log-likelihood ratio	89315	89401
log-likelihood improvement		86.5
Log-likelihood improvement p-value		0
Akaike Information Criterion (AIC)	-178529	-178674
AIC Improvement		145.1

Table 3.8: Model Comparison: Main and Random Effects (1) v. Main, Interactive, and Random Effects (2)

3.2.5 Model Criticism

As described in the Methods chapter, model criticism attempts to determine whether a model is being unduly influenced by a set of unusual data points. The model's standardized residuals are calculated, and used to determine potential outliers. Then, the effect of these potential outliers on the results of the model is explored.

Figure 3.1, generated using a modified version of the **mcp.fnc** function in Tremblay *et al.* (2013), shows several ways of visualizing the model residuals. The upper left panel shows the distribution of the standardized residuals of the model. The plot is narrower than a typical normal distribution, due to its longer tails.

This is shown more clearly in the upper-right panel, a quantile-quantile plot of the residuals. In this panel, residuals from the model ("Sample Quantiles") are ordered and compared to ordered samples from a normal distribution ("Theoretical Quantiles"). The straight line indicates the path that a normally distributed set of residuals would take. Deviation from this line is quite pronounced, especially for higher- and lower-valued residuals. The deviations represent a large number of data points for which the model over- or under-predicts reduction

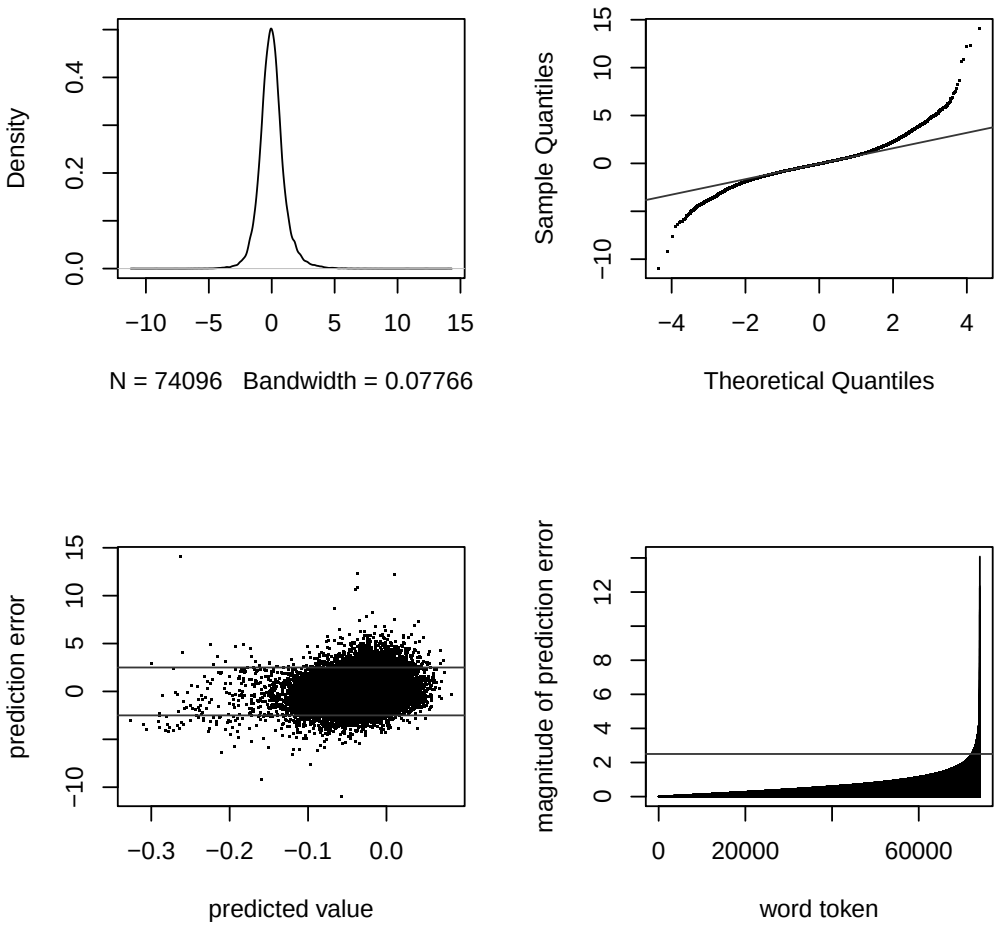

Figure 3.1: Model Criticism Plots

The scatter plot in the lower-left panel plots residuals against model predictions directly. Lines indicate ±2.5 standard deviations away from the mean residual value. The data points outside of these lines represent words for which model prediction is quite poor. The large number of these data points is further indication that a normal distribution is a poor fit for these residuals.

The bar chart in the lower-right panel shows the sorted absolute values of the standardized residuals. The chart shows most residuals falling below the horizontal line representing 2.5 standard deviations away from the mean residual. On the right of the chart, however, several data points show a dramatically higher residual, in some cases reaching 14.1 standard deviations beyond the mean residual.

The plots in Figure 3.1 suggest that some data points behave in a way that is dramatically different from the majority of the words in the model. This small set of words may be unduly influencing the results of the model fitting process. To determine whether or not this is the case, a new model was fit with these potential outliers removed. That is, words whose residuals in the model exceed 2.5 standard deviations away from the mean residual are removed, and a model with the same fixed- and random-effects structure as the previous model was fit. A total of 1,901 such data points, or 2.6% of the remaining words, were removed to calculate the new model.

Figure 3.2 shows the effect of removing these outliers: The distribution plot in the upper-left panel has much shorter tails than the commensurate plot in Figure 3.1, suggesting a closer approximation of a normal distribution.

The upper-right panel shows that the residuals in the trimmed model do not follow a normal distribution perfectly, with some high and low residuals deviating from the values expected in a normal distribution. Still, the deviation from normality is greatly reduced.

The lower-left panel shows fewer data points exceeding 2.5 standard deviations from the mean residual. Moreover, all of the most extreme data points (e.g., those that reach 10 or more standard deviations from the mean in Figure 3.1) have been removed.

The final panel in Figure 3.2 also shows a reduction in the maximum absolute value of the residuals. Some of the new residuals still exceed 2.5 standard deviations from the mean residual. The highest absolute residual in the new model is somewhat high, at 4.01. However, only 2% of the residuals are more than 2.5 standard deviations from the mean, and 89% of those residuals are within 3 standard deviations of the mean residual.

This improvement is not surprising, however. Removing data points that do not conform to a model naturally improves the quality of the fit of such a model. If the outliers were exerting undue influence on the model, however, the trimmed model may show qualitatively different effects of the predictors on the dependent variable. This possibility is examined in Tables 3.9 and 3.10.

Table 3.9 compares the effect sizes and t-values of the pre-trimming and post-trimming models. Table 3.9 shows that, overall, there is little qualitative change in the estimates of the fixed effects in the model.

No main effect or interaction loses or gains significance after trimming, and no effect changes from increasing reduction to decreasing reduction (nor vice versa). This lack of effect of trimming on the model's fixed-effects structure suggests that trimming is not necessary.

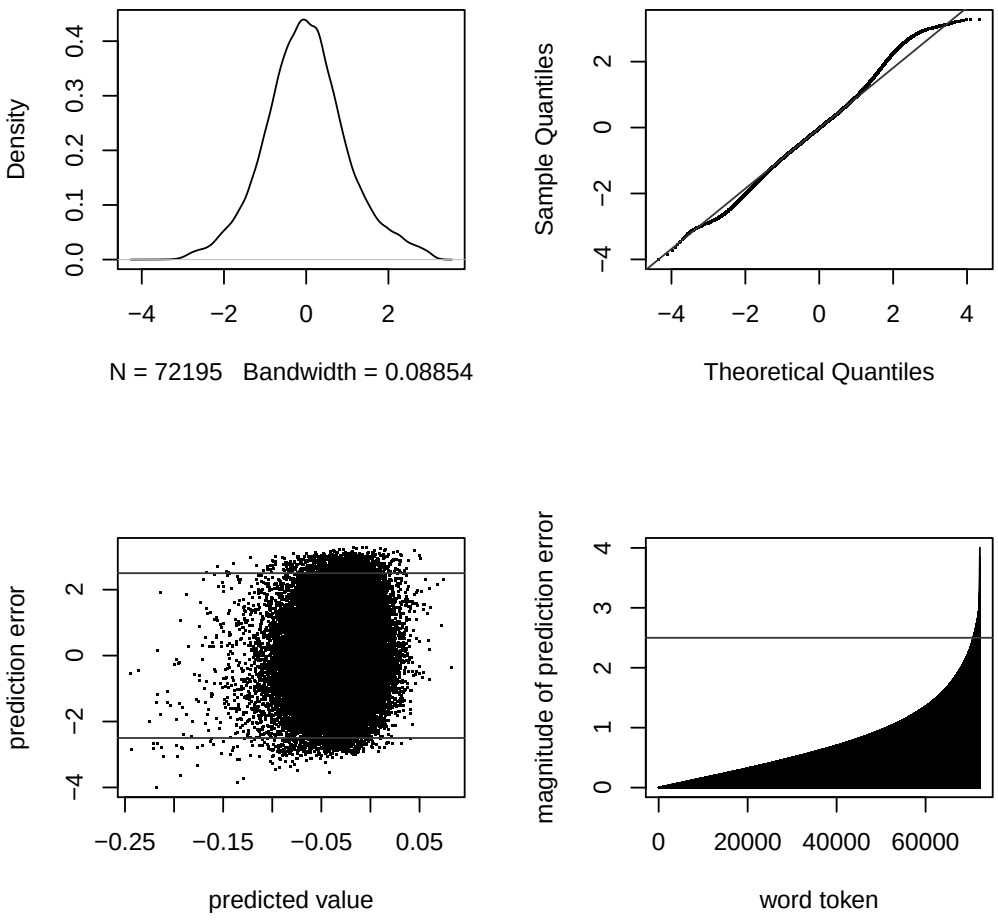

Figure 3.2: Post-Trimming Model Criticism Plots

Description	Est.1	Est.2	rank	t.val.1	t.val.2	t.val.chg
(Intercept)	-32.5	-33.1	0	-14.4	-16.9	-2.4
Backwards word predictability	-14.5	-11.8	0	-10.2	-9.4	0.8
Speaking rate for this phrase	-12.9	-11.4	0	-14.7	-16.3	-1.6
Length of word (res.)	-12.3	-10.5	0	-9.5	-9.6	-0.1
Adverbs	-8.6	-7.7	0	-3.5	-3.6	-0.1
Adverbs x Local frequency	-7.0	-6.8	0	-3.1	-3.5	-0.4
Nouns	-5.5	-6.2	+2	-2.6	-3.3	-0.7
Forward word predictability	-5.3	-6.0	+2	-7.8	-10.7	-2.9
Global speaking rate	5.8	5.7	-2	6.6	6.1	-0.5
Local frequency (res.)	-3.9	-4.9	+2	-3.4	-5.0	-1.6
COCA Frequency x Backwd wd pred.	-5.5	-4.6	-2	-7.6	-7.3	0.4
Nouns x Neighbr POS pred.	4.4	3.7	-1	3.6	3.5	-0.1
COCA Frequency x Length	-3.4	-3.6	+1	-4.3	-5.2	-1.0
tf-idf topicality (res.)	-2.3	-3.4	+5	-4.2	-6.9	-2.7
Nouns x Backwd wd pred.	-2.7	-3.0	+2	-2.1	-2.7	-0.6
Backwd wd pred. x Local frequency	-3.4	-2.9	-2	-8.3	-8.3	0.0
COCA Frequency x Forwd wd pred.	-1.9	-2.8	+5	-3.0	-5.1	-2.1
COCA Frequency	0.2	-2.2	+9	0.2	-1.6	-1.8
Surrounding POS predictability (res.)	-3.1	-2.1	-4	-3.0	-2.3	0.7
Adverbs x Backwd wd pred.	2.9	2.0	-4	1.7	1.4	-0.3
Global spk. rate x Phrase spk. rate	2.4	1.8	-3	2.9	2.7	-0.3
Verbs	-1.9	-1.7	0	-0.9	-0.9	-0.0
Adverbs x Neighbr POS pred.	2.0	1.5	-4	1.4	1.2	-0.2
Verbs x Backwd wd pred.	-0.7	-1.5	+1	-0.5	-1.3	-0.7
Verbs x Local frequency	0.7	1.3	0	0.5	1.1	0.6
Nouns x Local frequency	-0.9	-1.0	-4	-0.7	-0.9	-0.2
Verbs x Neighbr POS pred.	-0.4	-1.0	-1	-0.3	-0.9	-0.6

Table 3.9: Fixed-effects Before (1) and After (2) Trimming

Table 3.10 compares the standard deviations among random slopes for each of the random-effects predictors in the models. This table also shows no striking qualitative changes in the random-effects structure. Some slope variance coefficients appear, at first, to undergo dramatic changes. In particular, the variability among by-word slopes for female interviewers increases by more than 1.5 *million* percent. This apparently extreme increase is due to the extremely low variability among these slopes in the untrimmed model, however, and does not represent a dramatic change in absolute terms. The increase in standard standard deviation in the trimmed model amounts to less than 4 milliseconds. In this sense, both models agree: The variance among these slopes is very small. Furthermore, the fact that the variability *increases* when the data are trimmed can be seen as stronger evidence supporting the inclusion of by-word random slopes for interviewer gender. As these random slopes are already included in the untrimmed model, no change in model parameters is

Groups	Predictor	S.dev.1	S.dev.2	S.dev.chg	% chg.
Residual	NA	70.444	58.822	-11.622	-16
Word	(Intercept)	15.661	12.453	-3.208	-20
Speaker	(Intercept)	4.252	5.175	0.922	22
Word	Age (> 40)	5.994	5.207	-0.787	-13
	Age (< 30)	6.042	6.772	0.730	12
Word	Time	2.292	1.658	-0.635	-28
Speaker	Time	2.210	2.173	-0.038	-2
Speaker	Adjectives	3.542	2.245	-1.296	-37
	Adverbs	5.934	4.244	-1.690	-28
	Nouns	5.446	6.007	0.562	10
	Verbs	2.773	3.213	0.441	16
Word	Female Interviewer	0.000	3.559	3.559	1525598
	Male Interviewer	9.531	7.971	-1.561	-16
Speaker	COCA Frequency	3.733	4.088	0.355	9
Word	Backwd wd pred.	7.243	6.727	-0.516	-7
Speaker	Backwd wd pred.	3.463	3.175	-0.288	-8
Word	Forwd wd pred.	6.151	4.992	-1.159	-19
Word	Backwd POS pred.	9.468	9.414	-0.054	-1
Word	Forwd POS pred.	5.322	4.760	-0.562	-11
Speaker	Stress	3.175	3.637	0.462	15
Speaker	Length	4.102	3.319	-0.783	-19
Word	Neighbr POS pred.	5.572	5.174	-0.399	-7
Word	tf-idf topicality	5.794	5.320	-0.474	-8
Speaker	tf-idf topicality	1.529	1.431	-0.098	-6
Word	Female Speaker	6.692	4.847	-1.845	-28
	Male Speaker	5.781	5.032	-0.749	-13
Word	Global spk. rate	3.170	3.405	0.235	7
Speaker	Phrase spk. rate	5.017	3.938	-1.079	-22

Table 3.10: Random Effects Before (1) and After (2) Trimming

required to accommodate this difference.

Trimming outliers, then, has no strong qualitative effect on the structure of the model. Moreover, trimming data points that do not fit the model is anti-conservative, leading to inflated confidence in the quality of the model fit. For this reason, in the remainder of the analysis the full data set is used, and the model fit to the full data set is considered the final result of the current stage of the model selection procedure.

3.2.6 Results and Discussion

3.2.6.1 Main Effects

The main effects for the final model are included in Table 3.11, ordered by decreasing effect size (i.e., by decreasing slope coefficient). Not surprisingly, local speaking rate emerges as a strong predictor of reduction in word duration. (Effect size: -12.9msec, t-value: -14.7. Note that all numerical predictors are scaled and centered, so effect sizes reflect milliseconds in duration change over one standard-deviation of the values of the predictor.) When speaking more quickly, participants tended to shorten their words.

Global speaking rate also appears as a strong predictor of reduction, though not in the expected direction: The more quickly a participant spoke on average, the *less* likely that

Description	Effect (ms)	Std.Err.	t.value
(Intercept)	-32.5	2.3	-14.4
Backwards word predictability	-14.5	1.4	-10.2
Speaking rate for this phrase	-12.9	0.9	-14.7
Length of word (res.)	-12.3	1.3	-9.5
Adverbs	-8.6	2.5	-3.5
Adverbs x Local frequency	-7.0	2.3	-3.1
Global speaking rate	5.8	0.9	6.6
Nouns	-5.5	2.1	-2.6
COCA Frequency x Backwd wd pred.	-5.5	0.7	-7.6
Forward word predictability	-5.3	0.7	-7.8
Nouns x Neighbr POS pred.	4.4	1.2	3.6
Local frequency (res.)	-3.9	1.2	-3.4
COCA Frequency x Length	-3.4	0.8	-4.3
Backwd wd pred. x Local frequency	-3.4	0.4	-8.3
Surrounding POS predictability (res.)	-3.1	1.0	-3.0
Adverbs x Backwd wd pred.	2.9	1.7	1.7
Nouns x Backwd wd pred.	-2.7	1.3	-2.1
Global spk. rate x Phrase spk. rate	2.4	0.8	2.9
tf-idf topicality (res.)	-2.3	0.5	-4.2
Adverbs x Neighbr POS pred.	2.0	1.4	1.4
Verbs	-1.9	2.1	-0.9
COCA Frequency x Forwd wd pred.	-1.9	0.6	-3.0
Nouns x Local frequency	-0.9	1.3	-0.7
Verbs x Backwd wd pred.	-0.7	1.4	-0.5
Verbs x Local frequency	0.7	1.4	0.5
Verbs x Neighbr POS pred.	-0.4	1.3	-0.3
COCA Frequency	0.2	1.5	0.2

Table 3.11: Complete Fixed-effects Structure of Final LMER Model

participant was to shorten their word productions. This counter intuitive result is better explained when considered in tandem with its interaction with local speaking rate (Fig. 3.5). Such an explanation is provided in Section 3.2.6.2 below.

Predictability effects are also strongly supported in this model. A word's conditional probabilities, given either the previous or the following word, both emerged as strong predictors of duration change, each leading to a tendency for shortened words. In fact, a word's predictability given the following word ("Backwards Word Predictability") is the strongest predictor of reduction in the final model (-14.5msec, t-value: -10.2). Words whose part of speech is more predictable given the surrounding parts of speech are also more likely to be shortened, though the effect size is small (-3.1msec, t-value: -3.0).

Longer words and words more relevant to the conversation are also more likely to be shortened. Adverbs and nouns are also shown to be more likely to be shortened than adjectives. Adverbs and nouns represent the opposite extremes in type/token ratio, as illustrated in Figure 2.1. A study investigating the effects of type/token ratio on reduction may thus be an area of future interest.

Residualized local frequency remains a moderately strong and significant (-3.9msec, t-value: -3.4) predictor of reduction. That is, words that are more common in the Buckeye corpus than they are in the COCA are more likely to be shortened. COCA frequency,

however, appears to have no simple effect on reduction. Its main effect has the smallest effect size in Table 3.11, along with the lowest significance value. The disparity between the effects of these two frequency measures on reduction is addressed further in Section 3.2.6.2.

3.2.6.2 Interactive Effects

The interactions in this model are included in Table 3.11, and plotted below in Figures 3.3 through 3.7. Each line in a sub plot shows the effect of the predictor on the x-axis for a given value of the predictor with which it interacts. There are 7 lines in each sub plot, one for each of the quantiles into which the interacting predictor's values are divided. Only the central 99.9% of the range of values of the interacting predictor are included in these lines. This restriction is applied in order to avoid visually over representing the effects of the most extreme data points. (For the same reason, the x-axes span the central 95% of the data unless otherwise noted). The solid line in each plot represents the lowest (plotted) value of the interacting predictor. The dashed lines represent steadily higher values of the interacting predictor, increasing as line density decreases, until the dot-dashed line that represents the highest value of the interacting predictor in the corpus. The y-axes of the 8 sub plots are held to the same scale to allow for easy visual comparison of effect sizes. of the range of values of the interacting predictor are included in these lines.

Figure 3.3 shows two predictability interactions. The panel on the left shows an interaction between predictability given the following word ("Backwards word predictability") and COCA Frequency. This interaction sheds light on the question of frequency's inhibitory effect on reduction in the initial model. For low values of backward predictability, frequency appears to have an inhibitory effect on reduction, while for more predictable words frequency facilitates reduction. As a result, when a word is common and yet unpredictable from context a speaker is more likely to avoid reducing the word. Thus, this result suggests that predictability can be allowed to override frequency during production.

In terms of the underlying processes of speech production, two possible conclusions can be drawn from this result. The more specific conclusion is that the interaction suggests that predictability plays a stronger role than frequency in determining reduction in word duration. If the two factors are forced to compete, predictability appears to dominate. The relative effect sizes of frequency and predictability in the fixed-effects table in 3.11 (COCA frequency effect: 0.2msec, Backwards word predictability effect: -14.5msec) also suggests this. More broadly, the interaction implies that two potential systems of predictability are allowed to interact in the speaker's mind. One system of predictability is based on the frequency of a word in general speech. The other system of predictability is based on the frequency of a word in a very specific context (i.e., its relative frequency when given the following word). The strength of this interaction (Effect Size: -5.5msec, t-value: -7.6) suggests that these two predictability subsystems can not be considered in isolation, at least in the corpus under study.

This interpretation becomes more telling when the non-significant interactions are taken into consideration. Among the interactions that reach significance, none show an interaction between, e.g., forward and backward predictability. The current model, then, shows no evidence that predictability measures interact with each other. Since forward and backward

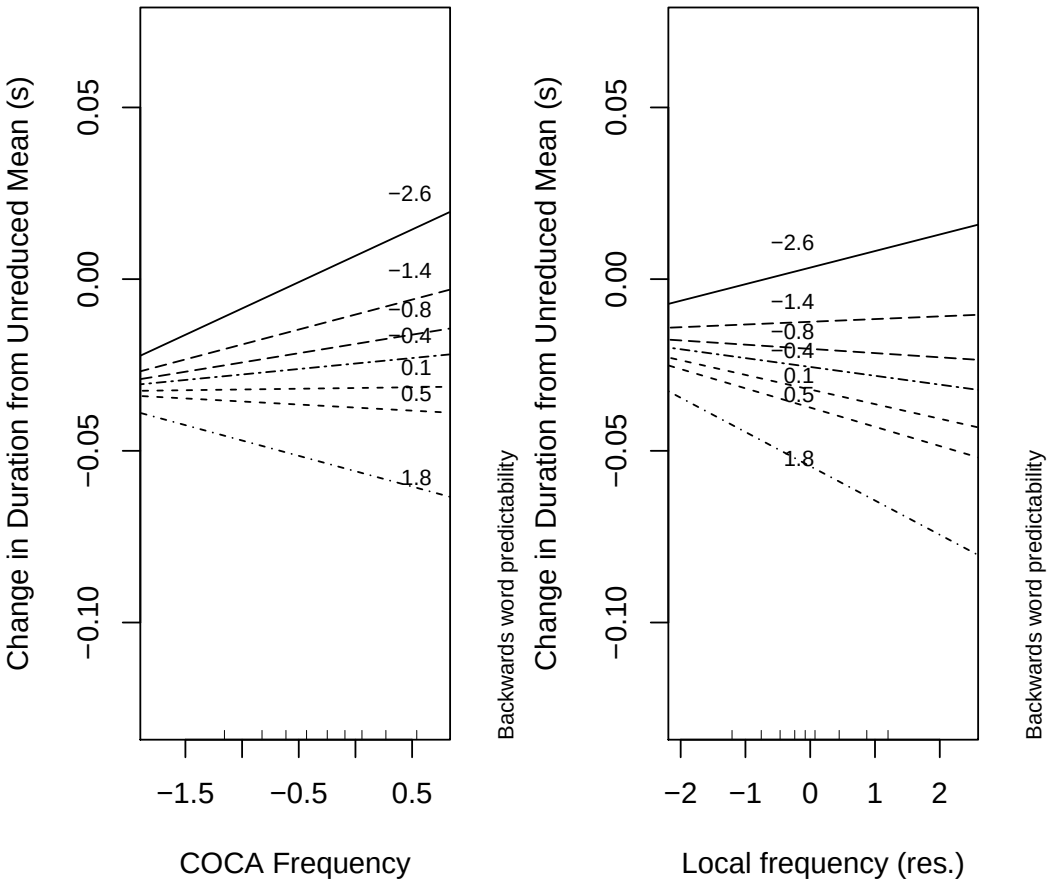

Figure 3.3: Backwards Word Predictability Interactions

predictability, for example, both lead to duration reduction in the model, the two measures may be taken as useful prediction systems in the speaker's mind. The lack of significant interaction between the two predictors shows no evidence that the two are part of the same prediction system. More precisely, information about a word's likelihood given preceding context does not appear to be used in concert with a word's likelihood given the following context when determining duration reduction.

It remains puzzling, however, that increased frequency should lead to shorter durations for unpredictable words. The model shows unpredictable low-frequency words shortening more than unpredictable high-frequency words. To interpret this result, it is worth recalling how predictability is measured in the current study. Backwards word predictability here represents the conditional probability of a word given the following word. This conditional probability is calculated by comparing the frequency of the pair of words together to the frequency of the target word in the corpus as a whole. Word productions with low conditional probability, then, appear relatively infrequently in the given context. As the frequency of these low-probability target words increases, however, the number of contexts in which the target words appear must also increase. This increased number of contexts for higher frequency words can be seen as an increase in the number of competing contexts that may be activated when a higher-frequency word is being produced. Lower-frequency unpredictable words, by contrast, will have fewer competing contexts that may be activated in a speaker's mind. Under this interpretation, the inhibitory nature of frequency in low predictability conditions becomes relatively straightforward: Speakers are shortening high-frequency words less because these words have more competing contexts than low-frequency words, and the competition is taxing some aspect of the word production system. Baayen (2010) found a similar inhibitory effect due to competition in a word's secondary family size. The more morphological-neighbours-of-morphological-neighbours a word has, the slower that word is processed.

The panel on the right of Figure 3.3 shows an interaction between backwards word predictability and (residualized) local frequency. This interaction is facilitatory. As predictability increases, the effect of local frequency on reduction also increases. Figure 3.3 shows that even relatively unpredictable words are more likely to be reduced if their frequency in the Buckeye corpus exceeds their frequency in the COCA. Only the lowest quantile of predictability shows an inhibitory effect of local frequency. The interaction suggests that a word's relative frequency in a local parlance combines with a word's predictability in its immediate context to encourage duration reduction. The inhibitory effect of local frequency in less predictable contexts can be understood in the same manner as the inhibitory effect of COCA frequency in similar contexts: High (local) frequency, low predictability words appear in more competing contexts than low (local) frequency words, and the increased competition leads to productions with longer word durations.

Comparing the two panels of Figure 3.3 leads to an interesting interpretation of frequency and reduction. In the left panel, COCA frequency is shown to have an inhibitory effect on reduction for several values of backwards word predictability. In the right panel, local frequency is shown to have an inhibitory effect for only the least predictable words in the corpus. Table 3.11 shows that local frequency is a stronger predictor of reduction than

general frequency. Figure 3.3 illustrates that the relative strength of these predictors remains the same under several predictability conditions: Local frequency usually leads to reduction, while COCA frequency leads to reduction only under specific predictability conditions. (i.e., when backwards word predictability exceeds a certain amount.)

The stronger effect of local frequency could be due to the difference in the type of data sampled for each frequency count. The general frequency measure samples from both written and spoke samples of American English, while the local frequency measure samples only spoken frequency, and samples only within a particular dialect region. In short, then, the panels in Figure 3.3 suggest that Buckeye frequency is a better predictor of reduction than COCA frequency, and that this is likely because the Buckeye corpus provides a better representation of the genre or dialect of the speakers under study than the COCA does. Indeed, since local frequency is residualized against COCA frequency, the residualized predictor may be seen as a measure of how appropriate a word type is to the genre under study. (Conversely, the COCA frequency predictor may be seen as a measure of how *in*appropriate a word is in the genre, explaining its weakness as a main effect and its tendency to correlate with longer productions in interactions.)

Figure 3.4 illustrates two interactions with COCA frequency. Both panels in Figure 3.4 again show that COCA frequency facilitates reduction only under certain conditions. The plot on the right shows an interaction similar to the one illustrated in the left plot of Figure 3.3. The interacting variable is predictability given the previous word, or "Forward predictability". Higher forward-predictable words are more likely to show a facilitatory frequency effect on reduction, while lower forward-predictable words show an inhibitory frequency effect. This coincides with the result for backward predictability described above. In both cases, predictability appears to override frequency, with high-frequency low-predictability words undergoing less shortening than low-frequency high-predictability words. As with the backward predictability interaction, the forward predictability interaction suggests that speakers combine frequency and predictability information when determining the length of their productions. And as with the previous interaction, it suggests that the increase in competing contexts that comes with increased word frequency in unpredictable contexts has an inhibitory effect on production.

The plot on the left of Figure 3.4 requires more care in its interpretation. One of the variables in the interaction (word length) has been residualized against the other (COCA frequency). As a result, the lines in the plot represent not frequency effects for short and long words, but frequency effects for words that are shorter or longer than expected given their COCA frequency. The plot reveals that unexpectedly short and unexpectedly long words behave very differently from words whose frequency and length are a better match for each other. Central values of residualized length, represented by the central 5 lines in the plot, show a behaviour similar to the predictability interactions described above. Among these central values, relatively long words show a facilitatory relationship between frequency and reduction, while relatively short words show an inhibitory relationship between frequency and reduction. The spread of these central values appears smaller than that found in the other plots, suggesting, at first, a relatively weak interaction between length and frequency. The extreme values of residualized length, however, show a great disparity in behaviour, and

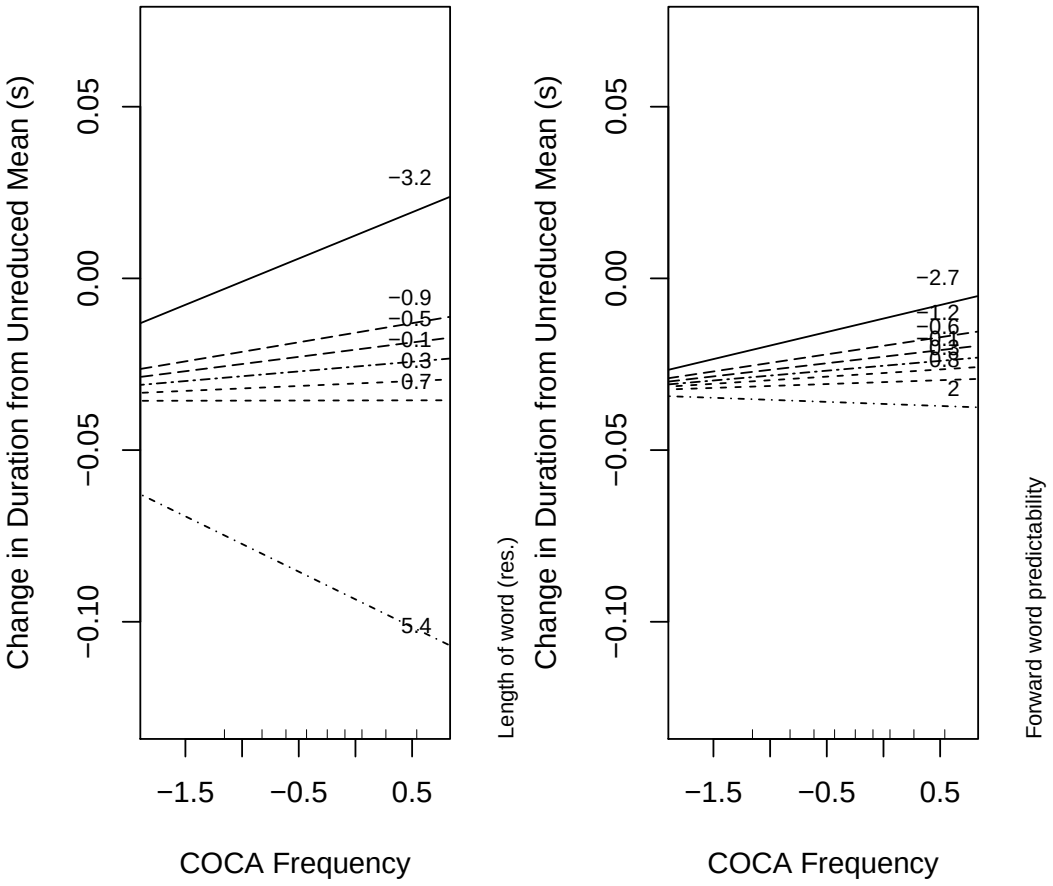

Figure 3.4: (COCA) Frequency Interactions

this interaction is the second strongest interaction in the final model (Effect Size: -3.4msec, t-value: -4.3) Unexpectedly short words show a strong inhibitory effect of frequency on reduction, while unexpectedly long words show a dramatic facilitatory effect of frequency.

Facilitatory co-operation between length and frequency is unsurprising, given that the two factors that have been shown to increase reduction in previous studies (e.g. Bell *et al.* (2009)). Inhibitory effects of frequency for shorter words, however, requires an explanation. One explanation can be found by considering phonological neighbourhood density. Neighbourhood density is positively correlated with frequency (Frauenfelder *et al.* 1993) and negatively correlated with word length (Pisoni *et al.* 1985). Short, high-frequency words, then, have more phonological neighbours than short, low-frequency words do on average. As a result, the short, high-frequency words have more competitors than their lower-frequency counterparts. The plot on the left of Figure 3.4 suggests that this competition inhibits duration reduction.

The inhibitory effect of frequency disappears for longer words in Figure 3.4 This suggests that as word length increases, competition from phonological neighbours becomes less of a factor in determining reduction.

If this result can be explained by phonological neighbourhood density, it can also be seen as arbitrating between speaker-oriented and listener-oriented models of reduction. Gahl *et al.* (2012) noted that words from dense phonological neighbourhoods are easy to produce but difficult to understand. The authors found that higher neighbourhood density led to increased reduction, providing support for a speaker-oriented model of reduction. The interpretation of Figure 3.4 provided here leads to the opposite conclusion: Speakers are resisting duration reduction as phonological neighbourhood density increases, suggesting that they are resisting reduction when such reduction would harm the listener.

The link between the frequency-length interaction and phonological neighbourhood density is still somewhat speculative, however. Frauenfelder *et al.* (1993) found that the correlation between word frequency and phonological neighbourhood density in the CELEX database (Baayen *et al.* 1993) is relatively weak. The interaction between length and frequency may indicate a relationship between reduction and a variable other than neighbourhood density. Age of acquisition, for example, also correlates with word frequency (Carroll & White 1973), and was omitted from the present analysis. It is also possible that COCA frequency is better described as a measure of a word's frequency in written English. Since local frequency is residualized against COCA frequency, the COCA frequency measure can be seen as accounting for the effects of written frequency, or relative unfamiliarity in spoken English, in the models presented here. The interaction between length and COCA frequency should thus be investigated more carefully in future work. In particular, Adding phonological neighbourhood density to the models presented here would likely provide invaluable insights.

Figure 3.5 shows an interaction between local and global speaking rates. Unsurprisingly, in all cases a faster local speaking rate correlates with more shortening of words.

Figure 3.5 helps to explain why higher average speaking rates correlate with lower rates of duration reduction, as illustrated by the main effect of global speaking rate in Table 3.11. The lowest line shows speakers who speak more slowly than average. When these people

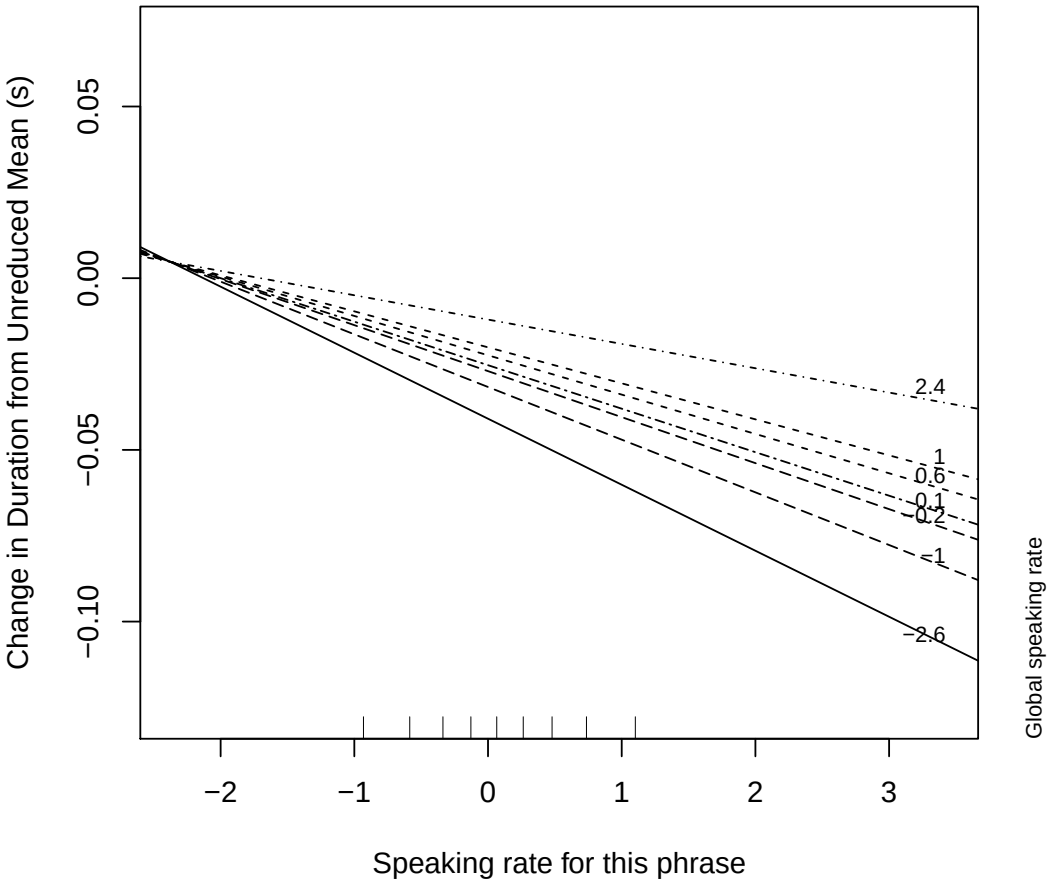

Figure 3.5: Global v. Local Speaking Rate Interaction. X-axis covers central 99.6% of data to accommodate plotting software

are speaking quickly, however, they reduce more severely than the average speaker. The top line shows speakers who speak more quickly on average. When *these* people show a higher than average local speaking rate, however, their reduction is less pronounced. The disparity increases as a speaker increases their (local) speaking rate, as shown by the fanning out of the lines towards the right edge of the plot.

The interaction in Figure 3.5, then, suggests that there are two types of speakers, or perhaps two poles on a continuum of speaker types: Speakers who speak quickly but uniformly throughout their dialogue, and speakers who speak slowly but with bursts of rapid, more reduced speech. The strength of the interaction suggests that the slower a speaker produces speech on average, the more likely they are to reduce words during bursts of more rapid speech. People with a higher than average global speaking rate, then, reduce less on average because their increased speed is spread out across the entire conversation, rather than occurring in short bursts of heavy reduction.

Figures 3.6 and 3.7 show three interactions involving part of speech. The left panel

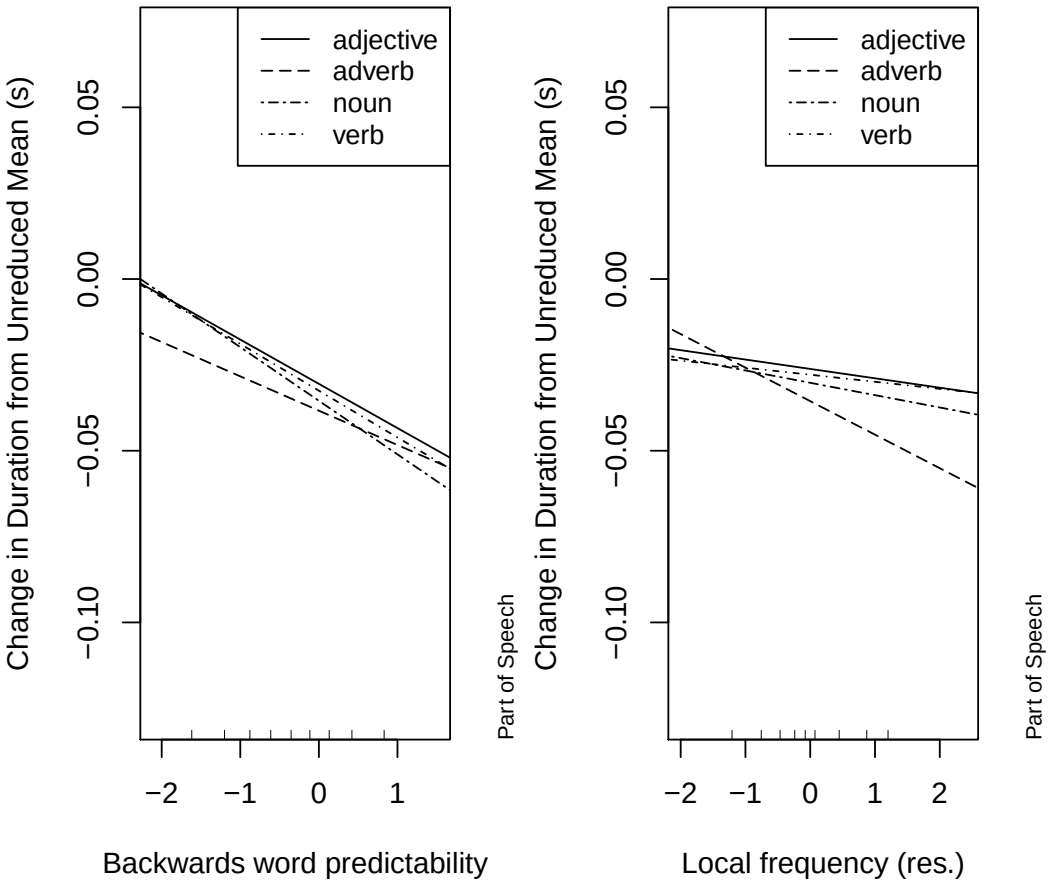

Figure 3.6: Part-of-Speech Interactions 1

of Figure 3.6 shows only a small difference between parts of speech: Adverbs appear less affected by predictability from following context. A close inspection of the left half panel suggests that this is partly due to the fact that adverbs show more reduction generally, including in less predictable contexts.

The right panel of Figure 3.6 shows a much starker difference in reduction between adverbs and other content words. Higher-frequency adverbs show much more reduction than higher-frequency nouns, adjectives, and verbs. This effect may also be derived from the much lower type/token ratio of adverbs when compared to other content words: The high-frequency adverb tokens are likely to come from a smaller set of types than high-frequency tokens of other parts of speech. The interaction appears to indicate that the lower number of types of high-frequency adverbs is matched by a lower variability in their reduction rates. Other parts of speech are likely to have more higher-frequency types, including more higher-frequency types that tend to resist reduction.

Figure 3.7 shows an interaction between part of speech and part-of-speech-based predictability. Verbs, adjectives, and adverbs show greater reduction as they become more likely given the parts-of-speech surrounding them. Nouns, however, show *less* reduction as they become more likely in structural context. This difference may be due to qualitatively differences in syntactic contexts in which nouns are more likely to be found. This possibility is explored further in Section 3.4.1.2 below.

3.2.6.3 Random Effects

The random effects structure of the final model is described in Table 3.12. The random slopes, in particular, help to solve a puzzle that appeared earlier in the modelling process. Namely, the fact that none of the sociolinguistic factors (age, gender, and interviewer gender) reached significance as main effects in the model. A lack of relationship between age or gender and reduction would be surprising given the importance that these factors are known to have in the description of conversational speech (Bell *et al.* 2009; Pluymaekers *et al.* 2005a). While not predictive of reduction in the larger model, allowing each word form type to vary separately in their relationship between these predictors and reduction leads to a better predictive model. This result can be interpreted as indicating that different age groups, for example, or different genders choose different words to reduce.

Similarly, random slopes by speaker for the number of stressed syllables in a word improve the model significantly, while not being a significant main-effects predictor of reduction in the model. The time at which a word appears in conversation is also not a significant main-effect predictor of reduction in the model, having been removed at the earliest stage of the modelling process. Allowing production time to vary, by both word and speaker, in its relationship to reduction leads to a stronger model, however. This suggests both that speakers respond differently over time in their reduction choices, and that different words are susceptible to reduction in differing ways as speakers proceed through their conversations.

Several predictors that do have significant main effects in the final model also show varying relationships with reduction by speaker or word form. Speakers respond differently from each other for different parts of speech, word length, topicality, predictability, COCA frequency, and local speaking rate. Different word forms also vary in how they respond to

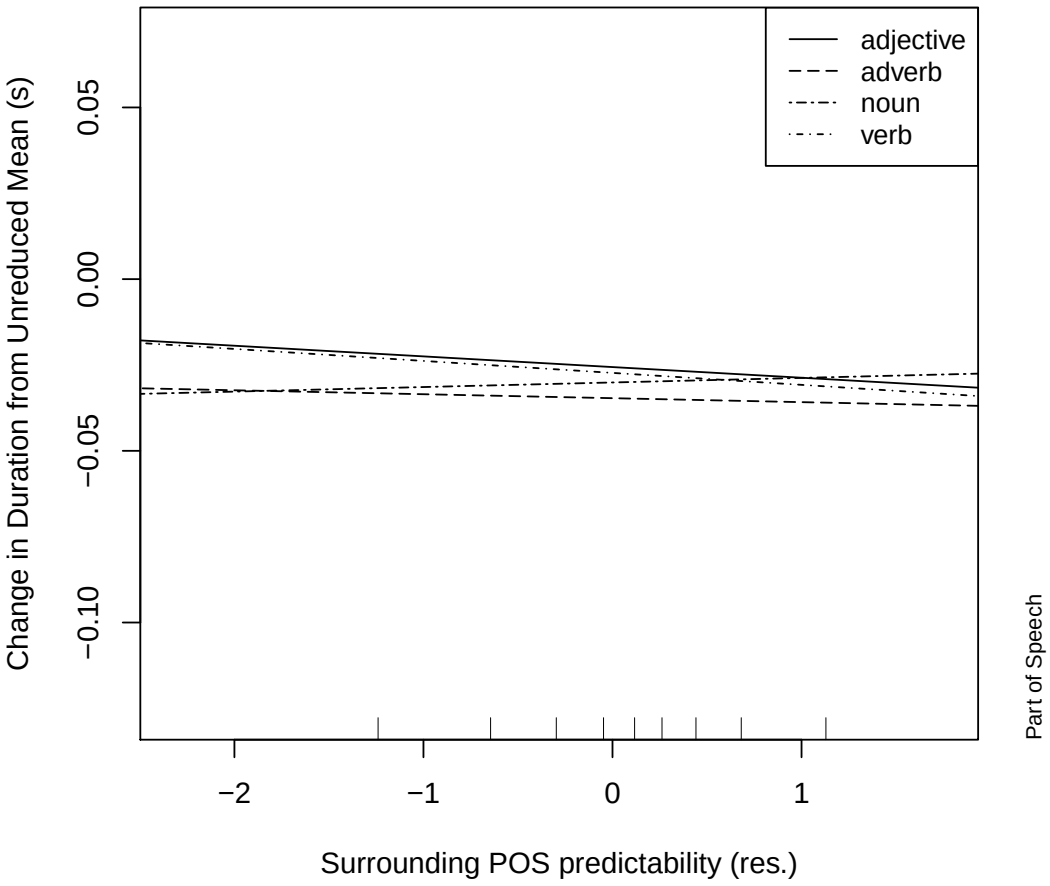

Figure 3.7: Part-of-Speech Interactions 2

Groups	Predictor	Std.Dev. (ms)	Corr		
Residual		70.444			
Word	(Intercept)	15.661			
Speaker	(Intercept)	4.252			
Word	Age (> 40)	5.994			
	Age (< 30)	6.042	0.32		
Word	Time	2.292			
Speaker	Time	2.210			
Speaker	Adjectives	3.542			
	Adverbs	5.934	0.50		
	Nouns	5.446	0.23	0.20	
	Verbs	2.773	-0.08	0.64	0.04
Word	Female Interviewer	0.000			
	Male Interviewer	9.531	0.00		
Speaker	COCA Frequency	3.733			
Word	Backwd wd pred.	7.243			
Speaker	Backwd wd pred.	3.463			
Word	Forwd wd pred.	6.151			
Word	Backwd POS pred.	9.468			
Word	Forwd POS pred.	5.322			
Speaker	Stress	3.175			
Speaker	Length	4.102			
Word	Neighbr POS pred.	5.572			
Word	tf-idf topicality	5.794			
Speaker	tf-idf topicality	1.529			
Word	Female Speaker	6.692			
	Male Speaker	5.781	0.09		
Word	Global spk. rate	3.170			
Speaker	Phrase spk. rate	5.017			

Table 3.12: Random-effects Structure of Final LMER Model

topicality, predictability, and global speaking rate. Allowing random slopes for each of these predictors improves the model fit significantly.

The specifics of how words and speakers respond differently to these predictors (which words are reduced more by women than men, for example, and which speakers respond differently to word length) is certainly a matter of interest to modellers of reduction. A more thorough investigation of these individual differences would be a worthwhile topic for future studies.

3.2.6.4 Comparing Random and Fixed Effects

The full mixed-effects model selected above combines both main- and random-effect predictors to attempt to predict reduction in the corpus. The final model predicts 19.2% of the variance in duration reduction. It is unclear from the model, however, how much of this variance is predicted by the random effect structure in the model, and how much is predicted by the fixed-effect structure in the model. One of the overall goals of the present study is to evaluate the relationship between LMER models and random forest models. The question of the relative predictive power of the fixed- and random-effects components is thus an important factor to investigate: The random-effects structure available to the LMER model

is unavailable to the random forest. Random forest models behave poorly when faced with predictors made up of large sets of unordered categorical factors. The number of possible splitting values for such predictors increases exponentially as the number of possible factors increases. Both random-effect predictors in the LMER models under study here, word form and speaker, represent precisely such a large set of unordered factors. As a result, the random effect structure of the LMER model must remain absent from the random forest models described below. Knowing how much of the LMER model's power comes from its random effects structure, then, is an integral part of comparing how well each modelling technique performs.

Two new linear models were created to evaluate the relative importance of the fixed- and random-effect structures of the model in predicting reduction. First, a model was constructed with all of the random slopes and intercepts listed in Table 3.12, with only a general intercept as a main-effect term. Second, a linear (non-mixed-effects) model was constructed with all main effects and interactions found in Table 3.11, with no random effects structure or predictors for word form or speaker. The proportion of variance explained by each model was then measured.

The model with only random effects is much more successful, predicting 20% of the variation in reduction in the model. The model with only fixed effects predicts only 5.7% of the variation in reduction in the model.

The latter model, containing fixed-effects alone, provides for the most direct comparison with the random forest model below. Adding word form and speaker as predictors in the random forest model is not computationally feasible. As a result, the fixed-effects-only model can be thought of as containing (a subset of) those predictors and interactions in the current study that are also available to Random Forest modelling.

3.3 Random Forest Modelling

3.3.1 Random Forest Model Fitting

In this section, a random forest model is fitted to the full data set of 74,096 data points used in the final LMER model. All 18 variables listed in Table 3.1 above are included as predictors. Tuning for this forest led to an *mtry* value of 3, meaning that 3 predictors are randomly selected to construct each tree in the forest.

3.3.2 Results and Discussion

3.3.2.1 Proportion of Variance Explained

The forest predicted 14.1% of the variance. In terms of overall model fit, this places the forest between the linear model without random slopes ($R^2 = 12.3\%$) and the model with random slopes added ($R^2 = 19.2\%$), as shown in Table 3.7. As mentioned above, however, the final LMER model had more information with which to make its predictions than the random forest models. Moreover, much of the predictive power of the LMER model appears to come from its random-effects structure. In fact, the model with only random effects outperforms the forest, predicting 20% of the variance in reduction. The model with only fixed-effects predictors, however, predicts just 5.7% of the variance. This fixed-effects-only

model performs much more poorly than the forest model, though both models are given the same set of predictor variables.

In terms of the amount of variance explained by each model, then, the present study finds mixed results. When given roughly the same amount of information, random forest models dramatically outperform linear models. In fact, the forest model fits the data better than the LMER model with random intercepts for word form and speaker added along with the complete fixed-effects structure from the optimal linear model. In this case, the random forest model performs better than an LMER model, even when the LMER model has more information to draw on in making predictions.

When the LMER modelling process is allowed to include random slopes, however, an LMER model can be constructed that fits the data better than the random forest model produced above. Indeed, an LMER model with *only* random slopes and intercepts, but with no fixed-effects predictors at all, predicts reduction more effectively than the current random forest model.

These mixed results make it difficult to decide whether random forests or LMERs are better tools for modelling duration reduction. Combining the two modelling tools by using a model selection process that draws on the strengths of each may provide the best results. Such a process is described in section 3.4.

3.3.2.2 Variable Importance Measures

The random forest model can be used to determine importance scores for each predictor variable. These importance scores are calculated using the **importance** function of the **randomForest** package (Liaw & Wiener 2002). This measure establishes how much the model fit suffers if a particular variable is left out of the analysis. Figure 3.8 compares this variable importance scores to the importance of each predictor in the final (full) linear model selected above. For the LMER model, a variable's importance to the model was calculated by comparing AIC values for two models: The full model and an identical model with the variable of interest removed. More specifically, any main effects, interactions, and random slopes for the target variable were removed from the model. The measures of variable importance for each model are scaled and centered in Figure 3.8 to allow for visual comparison.

The predictors in 3.8 are sorted in order of their importance in the random forest model. A clear discontinuity in importance appears between the topicality predictor and the gender predictor, indicated in Figure 3.8 by a vertical line separating the two. The simplest interpretation is that the predictors above this discontinuity are useful components of a model of duration reduction, while the predictors below the discontinuity are not. This interpretation coincides with the results of the LMER modelling process: All of the predictors below the discontinuity were trimmed during the first step of linear model selection.

There are differences between the results of LMER and random forest modelling, however. Two predictors in Figure 3.8 stand out as extremely useful in the LMER model: Predictability given the following word, and local speaking rate. Backwards word predictability also shows the highest importance score in forest model, providing evidence from multiple sources that a word's predictability is likely to coincide with duration reduction.

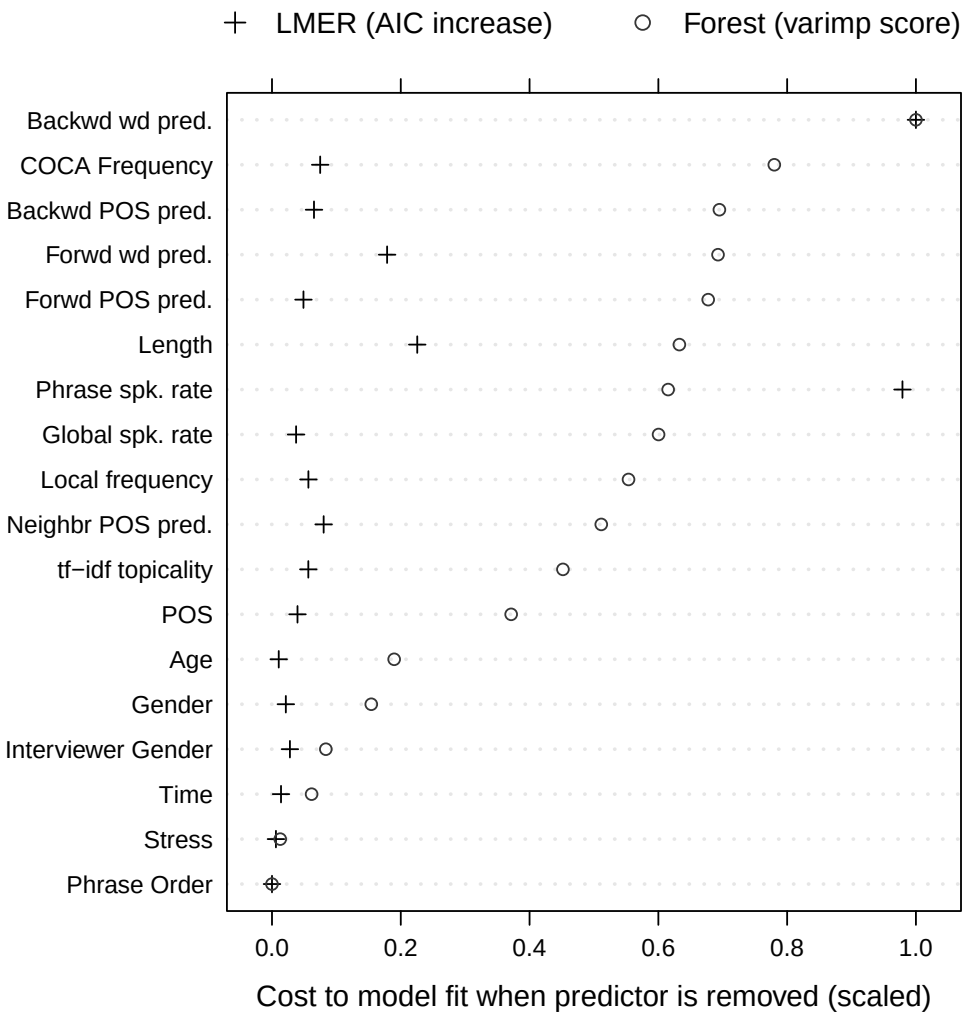

Figure 3.8: Comparison of Variable Importance across Model Types

Local speaking rate, however, is ranked quite differently in the LMER and forest models. In fact, it shows the largest difference in variable importance between the two models. Both models agree that speaking rate is important: The random forest model places speaking rate above the discontinuity described above, suggesting that speaking rate is a useful predictor. A link between local speaking rate and duration reduction is not surprising, implying that the faster a person is speaking the more likely they are to shorten their words. For this reason, it is surprising that speaking rate has such a low rank in the random forest models. The forests find speaking rate to be one of the less important predictors that can be considered useful, while the LMER model finds speaking rate to be one of the two most important predictors of reduction. The difference can be partly explained by the different roles that the speaking rate variable plays in the two classes of models. The LMER model selection process found that a random slope for speaking rate by speaker significantly improved model fit (see Table 3.12.) Without the ability to allow for random slopes, the random forest models cannot detect this source of variability in the data. As a result, the importance score for speaking rate in the forest models incorporates the loss of a main effect and (potential) interactive effects of speaking rate, while the importance score for speaking rate in the LMER model incorporates the loss of both a main effect and a random effect. Figure 3.8, then, shows that the random slope for speaking rate is much more useful in improving model fit than any potential interactions with speaking rate.

This difference in what type of effect is available to each class of models can also help to explain the very high score for backwards word predictability in the final LMER model. In the LMER model, backwards word predictability is included as a significant main effect, participates in two significant interactions, and is allowed to have a variable slope by speaker. The main and interactive effects are available to both classes of model, and indeed, both classes of model find backwards word predictability to be one of the strongest predictors of reduction. Still, the random slope for backwards word predictability in the LMER model pushes the importance of that variable beyond that found in the random forest models.

Some care should be taken when comparing the importance scores across random forest and LMER models. Each class of models uses a different measure of variable importance (percent increase in mean standard error for random forest models, and difference in AIC scores for LMER models), so comparisons between the two are necessarily imprecise and relative. The two predictors described above help to illustrate this problem: The predictors for speaking rate and backwards word predictability in the LMER model achieve importance scores so high that they dwarf the variation among importance values for the remaining predictors. To allow for a closer inspection of the relative importance of the remaining LMER predictors, the importance values for backwards word predictability and speaking rate in the LMER model were removed, and the remaining values were re-scaled. The result is shown in Figure 3.9

Many of the six variables at the bottom of Figure 3.9 show low importance scores in both random forest and LMER models. These variables were all trimmed from the main-effects portion of the LMER model at the earliest stage of model selection: Their effect sizes were too small to reach significance. The random forest model appears to agree that trimming these variables from the main-effects structure of the data was warranted. Their

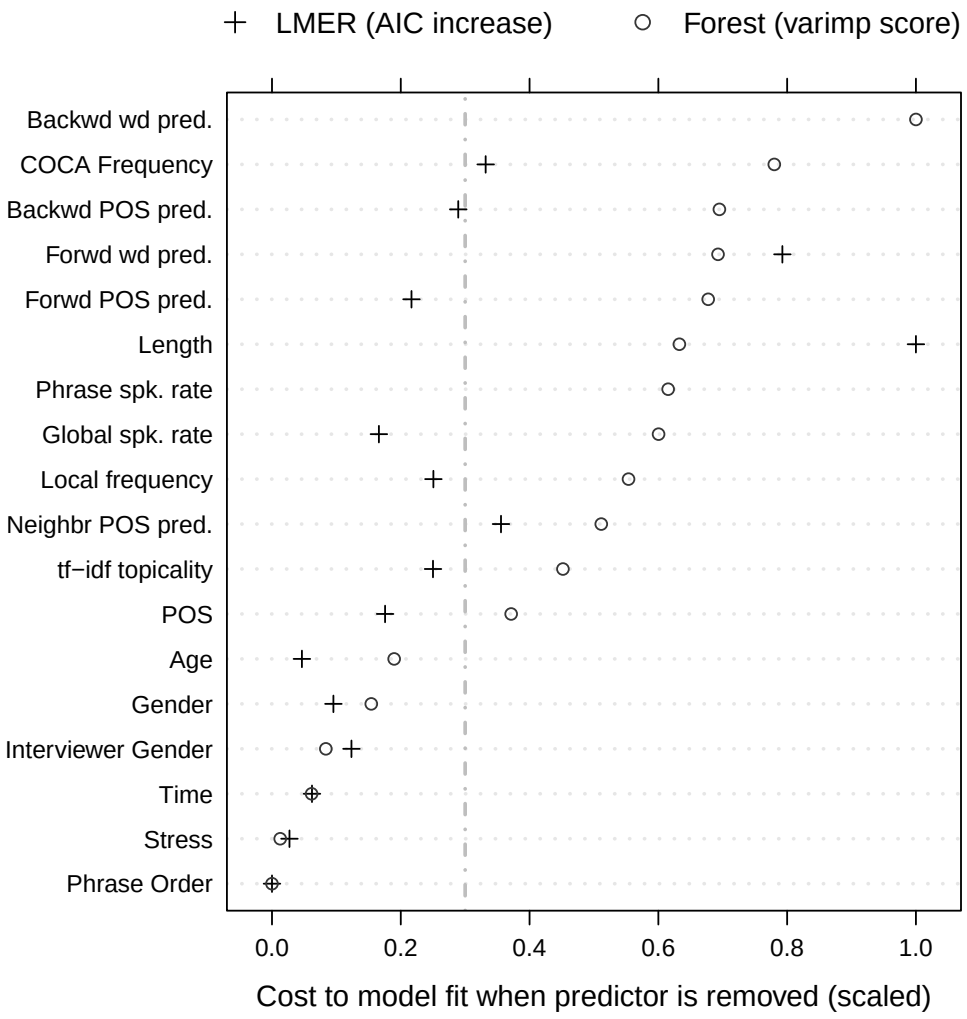

Figure 3.9: Comparison of Variable Importance across Model Types - LMER Values Re-scaled

low importance scores indicate that they contribute little to the fit of the forest model. Figure 3.9 appears to indicate that these variables are slightly more important in the LMER model than in the forest models, but this is likely in part an artifact of the scaling process. Phrase order, for example, does not enter into the LMER model in any way, so "removing" it has no effect on the AIC score of the final model. The LMER importance score for this variable, then, is zero, but the scaling and centering processes lead to it falling further to the right on the graph than it does in either forest model. For this reason, comparisons between the LMER and forest scores should be considered on a relative, rather than absolute basis.

Among these six low-importance predictors, interviewer gender has a relatively high importance score in the LMER model, as does speaker gender. Each variable scores higher in importance than roughly half of the other predictors in the LMER model. This is due to their presence in the random-effects structure of the model: The two variables show the highest variance in slopes among the random effects in the final LMER model, as shown in Table 3.12.

Two variables in Figure 3.9 stand out as having dramatically lower importance scores in the LMER model: Backwards and forwards part-of-speech predictability. Both predictors are ranked very highly by the forest model, but have the lowest possible importance score in the LMER model. These two variables were also trimmed during the initial stage of linear model selection (and thus precluded from interaction testing), and random slopes for the variables were not found to contribute significantly to model fit. As a result, these two variables do not appear in the final LMER model at all. The disparity between how highly these predictors are valued by the two classes of models can be partly explained by examining the nature of their partial-effects in the random forest model, plotted in Figure 3.10.

The sub plot on the left, for backwards POS predictability, shows a slight downward (reductionary) trend, followed by a sharp up-tick for a small number of extremely predictable words. In general, though, both sub plots in Figure 3.10 show that the main (partial) effects for these predictors are relatively flat across most words. The lack of strong partial main effects for these words contrasts with their high importance scores, illustrated in Figure 3.8. This contrast implies that these variables provide most of their predictive power through interactions with other variables. The weak main effect for these variables, shown now in both LMER and RF models, led them to be removed early in the LMER model selection process. Critically, these variables were removed before the step in which interactive effects were explored. As a result, no interactive effects for these variables were considered for inclusion in the final linear model.

3.3.2.3 Comparison of Partial Effects

This section compares plots of the effects of individual predictors on reduction. In particular, for each variable the partial effects predicted by the linear model are compared to the partial effects predicted by the random forest model.

The partial effects can be grouped into three broad categories: First, there are predictors for which the linear model and the forest model agree closely. Both models agree on the relative reduction levels expected for each part of speech, for example. Figure 3.11 shows

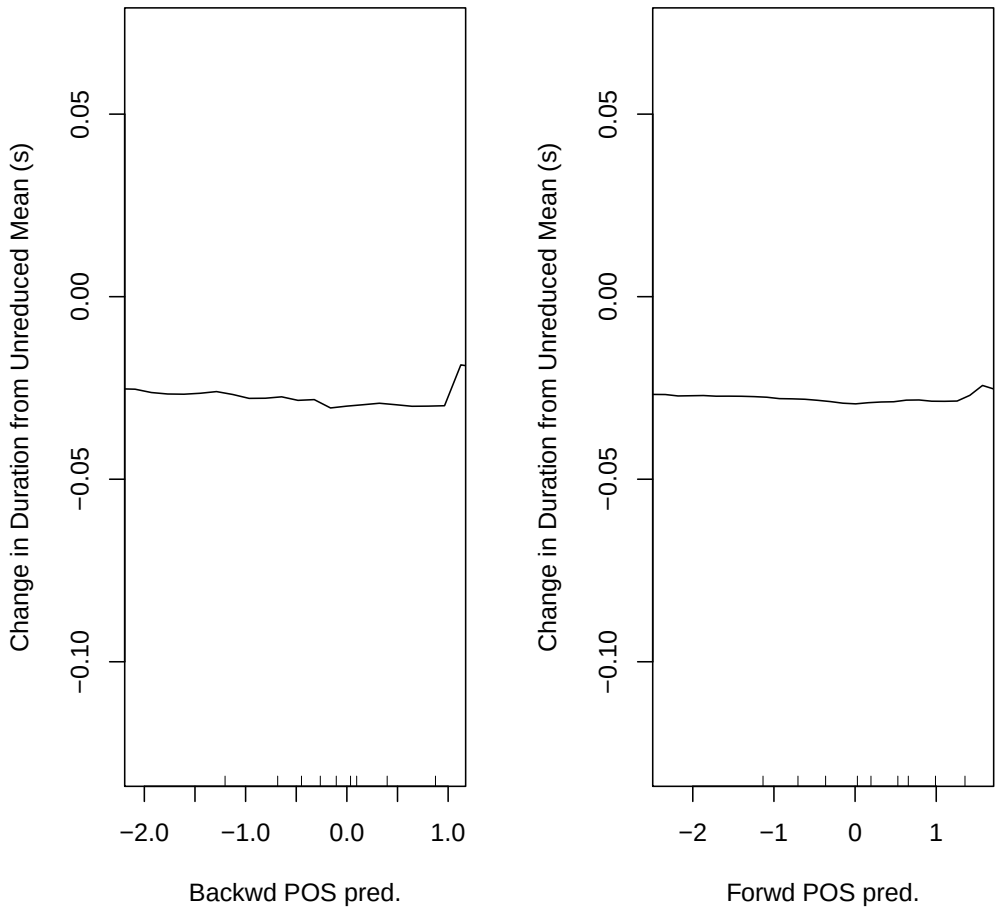

Figure 3.10: Partial Effects for Backward and Forward POS predictability

that both models find the most reduction for adverbs, followed by nouns and then adjectives and verbs.

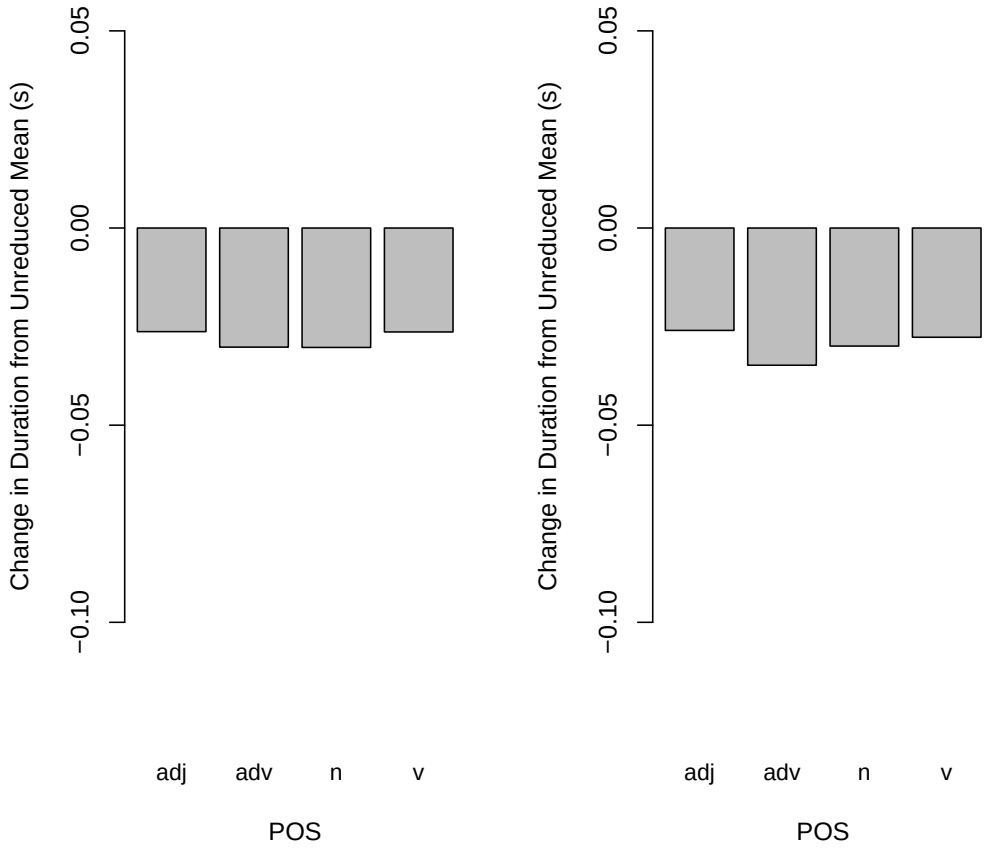

Figure 3.11: Partial Effect of Part of Speech in Random Forest (L) and LMER (R) Models

The models also agree on a roughly linear facilitatory effect of forward word predictability on reduction, as shown in Figure 3.12.

COCA frequency also falls into this category, with both models showing relatively little effect of COCA frequency on reduction (A complete set of partial effects plots can be found in Appendix A) The models also agree on a roughly linear facilitatory effect of local speaking rate on reduction, and a roughly linear inhibitory effect of global speaking rate on reduction.

The second category contains partial effects for which the forest model provides more detail than the linear model. More specifically, the random forest model shows a non-linear effect of certain predictors on reduction that the linear model is unable to capture.

Figure 3.13, for example, shows that both the linear model and the forest model find words becoming shorter as their predictability given the following word increases. In the forest model, however, the reductionary effect trails off as predictability increases beyond a

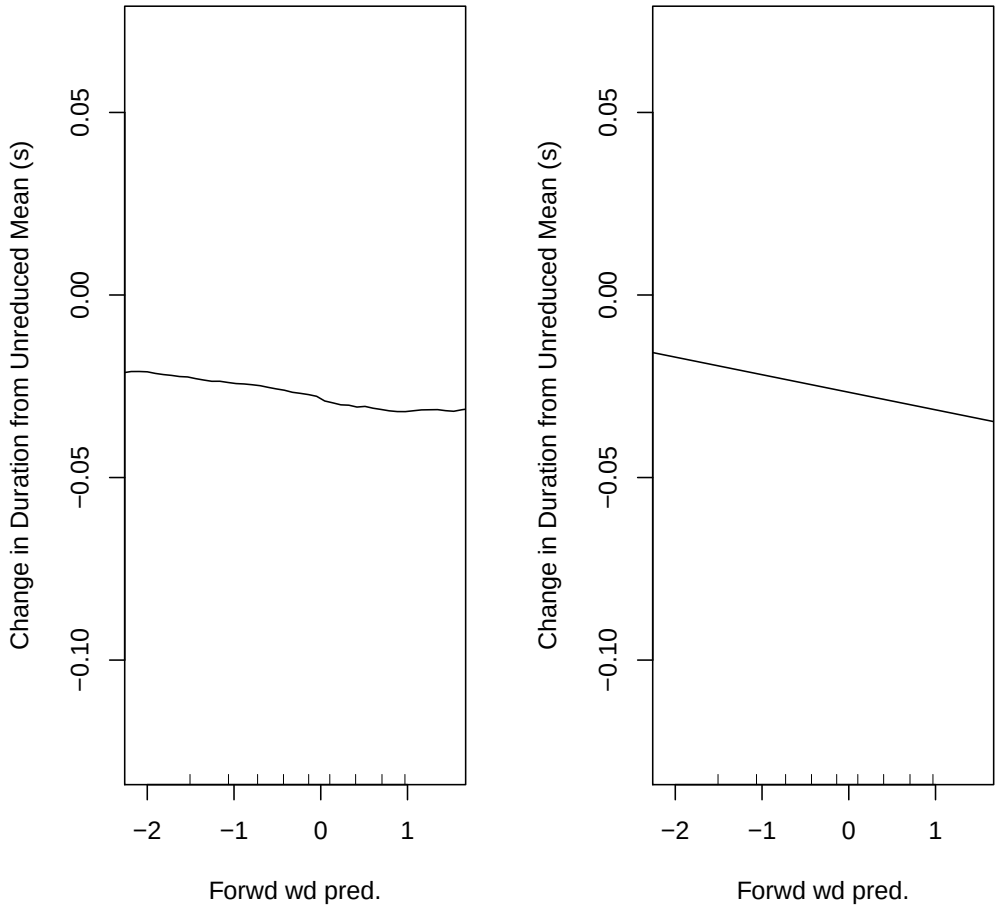

Figure 3.12: Partial Effect of Forward Word Predictability in Random Forest (L) and LMER (R) Models

certain level.

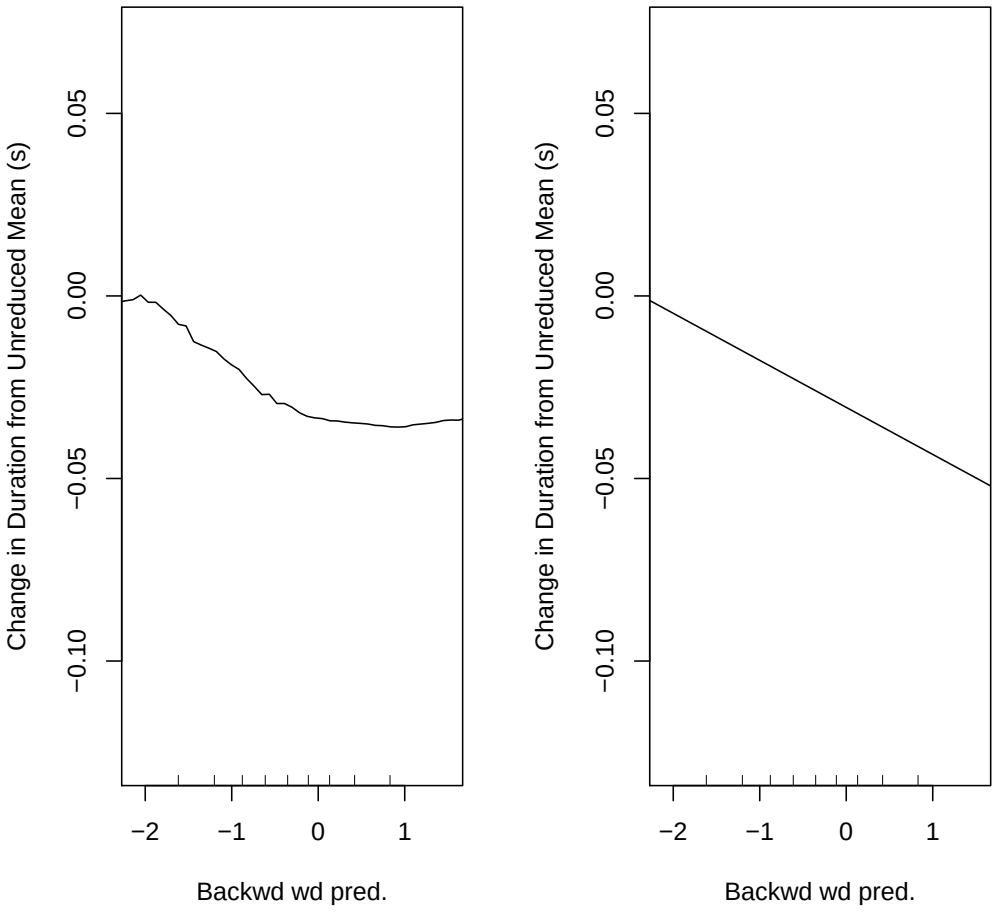

Figure 3.13: Partial Effect of Backward Word Predictability in Random Forest (L) and LMER (R) Models

Similarly, Figure 3.14 shows a reductionary effect of (residualized, citation) word length in both models. In the forest model, however, length appears to have little effect on the reduction of shorter words, and a stronger effect of length for the longest words. That is, reduction is less affected by the length of short words than it is by the length of longer words: The longest words are much more reduced than average-length words are. The shortest words, by contrast, are reduced at approximately the same rate as average-length words are.

This increase in effect size for higher values of a predictor is also found in (tf-idf) topicality and surrounding part-of-speech predictability, though the non-linearity (and the effect size) is less pronounced. For predictors in this category, the random forest model suggests that the linear model fit may be improved by allowing the predictors to vary with reduction in a non-linear fashion. Modelling these effects using splines, for example, or modelling the

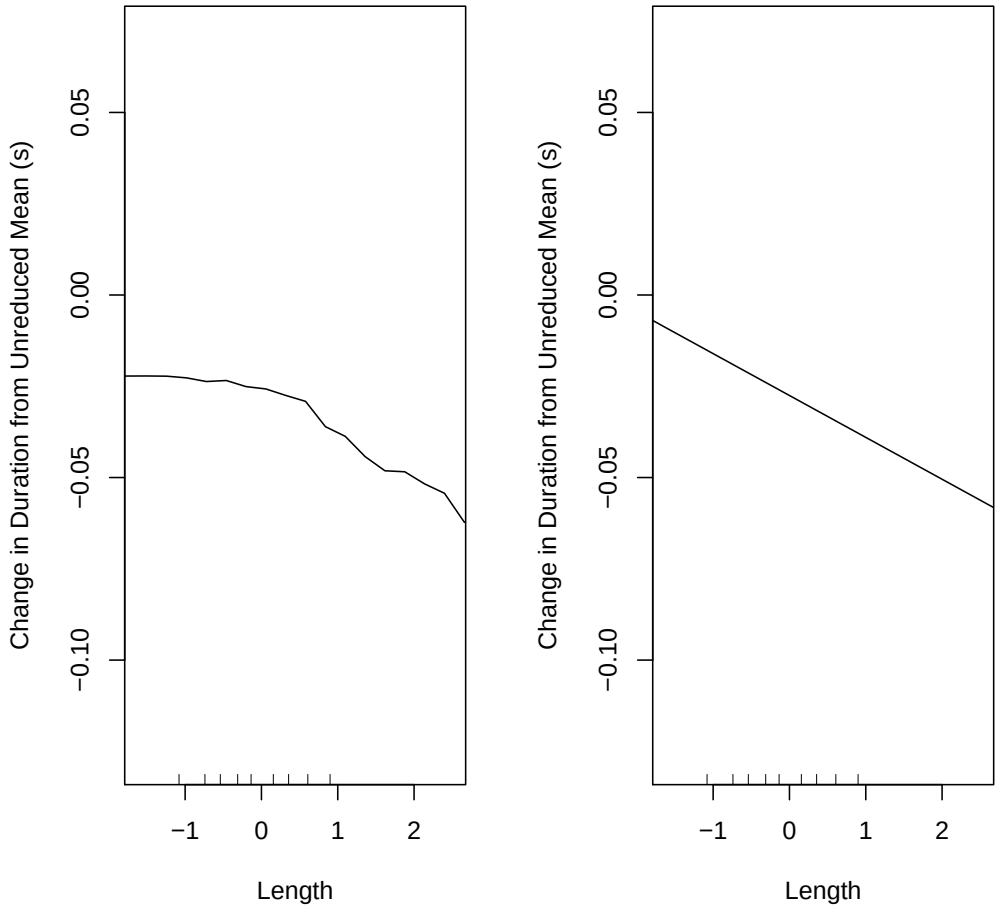

Figure 3.14: Partial Effect of (Residualized) Word Length in Random Forest (L) and LMER (R) Models

data using generalized additive modelling, may have a beneficial effect on the quality of the model produced.

The final category consists of predictors for which the two model classes appear to disagree entirely on the effect of a predictor on reduction.

Fortunately, this category contains only one predictor: Local (Buckeye) corpus frequency. Figure 3.15 shows that the forest model finds little effect of local frequency on reduction. The LMER model, however, finds a relatively strong facilitatory effect of local frequency on reduction.

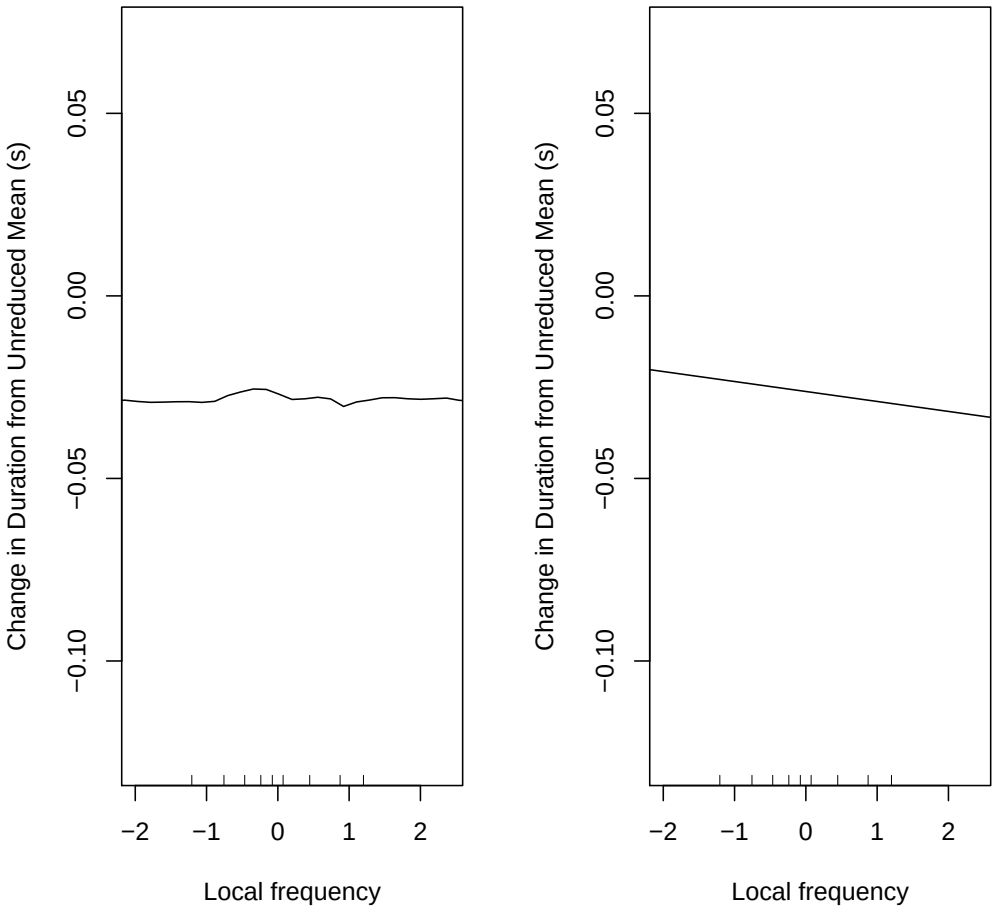

Figure 3.15: Partial Effect of (Residualized) Local Frequency in Random Forest (L) and LMER (R) Models

The cause of this disparity is difficult to determine with certainty. There are several possible explanations, however. The random forest model includes several predictors that the linear model does not include, and allows for arbitrarily complex interactions between the predictors. Similarly, the LMER model contains several random effects that the random forest model does not include. It is possible that, for example, the random forest model

finds no strong main effect of local frequency because the apparent effect found in the LMER is actually explained by some interaction between variables, or by some combination of the variables that are only found in the forest model. Conversely, it is possible that an effect of local frequency only becomes apparent when the variation of some predictor within groups of speakers or words is taken into account. The latter interpretation appears less likely, however, given that random slopes for local frequency by speaker were not found to significantly improve the LMER model. (See Table 3.12)

3.4 Combining Modelling Techniques

Both mixed-effects regression and random forest modelling appear to have clear strengths and weaknesses. As Section 3.3.2.2 above illustrates, each modelling tool is forced to consider only a subset of the possible sources of variation in the data. In both modelling techniques, this limitation is based in part on computing power. In random forest models, the inclusion of unordered factor variables with many levels increases the difficulty of the modelling task exponentially. Given that the random-effect variables in this study, word form and speaker, are unordered factor variables with many levels, the random forest model cannot be used to investigate the random-effects structure of the data.

Linear mixed-effects regression models also face a computational limitation. If a maximal, backward-fitting model selection process, like the one described in (Barr *et al.* 2013) were applied to the current data, the number of main effects, (two-way) interactive effects, random intercepts and random slopes adds up to nearly 200 potential predictors. Such a process was attempted for the current data, but even on a server with sixteen 2.9GHz processors and over 128GB of RAM, the process was abandoned after the first maximal model failed to converge after more than a week of processing. For large data sets and large numbers of predictors, then, current technology requires a forward-fitting approach for mixed-effects regression modelling. The modeller must make a long sequence of choices about which variables to exclude from each step of the modelling process. When modelling duration reduction here, for example, unpromising main-effects predictors were excluded in part as a way of limiting the number of interactions to be considered for inclusion in the model. As the analysis of the random forest model revealed, several interactions that increase the modeller's understanding of the data were likely excluded prematurely.

In short, random forest models suffer by omitting by-subject and by-item variance in the data, while forward-fitting LMER models suffer by omitting interactive effects whose component variables show little or no main effect by themselves.

One possible way to combine the strengths of the two modelling techniques is to use random forests as a quick way to determine which variables are likely to contribute to the model, on their own or as part of an interaction. Thus, random forest variable importance scores can be used to indicate to the linear modeller which variables are likely to contribute fruitful main effects and interactions, by acting as an initial filter through which variables are passed before linear modelling begins.

In the present study, for example, the six variables below the discontinuity in Figure 3.8 could safely be excluded from (fixed-effects) analysis before linear modelling begins.

Main- and interactive-effects for the remaining 12 variables can then be passed through the model-selection procedure applied above. This procedure represents a more sensible way of reducing the problem space to manageable levels: Interactions are not excluded based on the main-effect sizes of their components, but rather by the random forest model's determination that no significant interactive effect is likely to be found.

To test this method, the LMER model selection procedure described in the previous chapter was repeated, but modified to take the RF model's findings into account. The 12 variables above the discontinuity in Figure 3.8 were entered as main effects, and each unique pairing of these 12 variables was subjected to the interaction testing procedure used above.

3.4.1 Results and Discussion

The complete fixed-effects structure of the RF-Informed LMER model is shown in Table 3.13.

Description	Effect (ms)	Std.Err.	t.value
(Intercept)	-33.0	2.3	-14.5
Backwards word predictability	-14.8	0.9	-15.7
Speaking rate for this phrase	-12.7	0.9	-14.4
Length of word (res.)	-12.2	1.3	-9.3
Adverbs	-8.9	2.5	-3.5
Nouns x Backwd POS pred.	8.9	1.9	4.8
Verbs x Backwd POS pred.	-7.4	1.9	-3.8
Adverbs x Forwd POS pred.	-7.3	1.9	-3.9
Adverbs x Backwd POS pred.	6.7	2.1	3.2
Verbs	-5.8	2.1	-2.7
Global speaking rate	5.8	0.9	6.7
Nouns	-5.6	2.1	-2.7
Forward word predictability	-5.2	0.7	-7.7
COCA Frequency x Backwd wd pred.	-4.6	0.7	-6.9
Local frequency (res.)	-4.4	0.7	-6.4
COCA Frequency x Length	-3.4	0.8	-4.2
Backwd wd pred. x Local frequency	-3.2	0.4	-8.1
Global spk. rate x Phrase spk. rate	2.4	0.8	2.9
tf-idf topicality (res.)	-2.3	0.5	-4.2
Forward POS predictability	2.1	1.5	1.5
Nouns x Forwd POS pred.	-2.0	1.7	-1.2
COCA Frequency x Forwd wd pred.	-2.0	0.6	-3.1
Surrounding POS predictability (res.)	-1.5	0.5	-3.0
Backwd POS pred. x Forwd POS pred.	1.4	0.4	3.3
Backward POS predictability	-1.3	1.6	-0.8
Backwd POS pred. x Phrase spk. rate	1.3	0.4	3.1
COCA Frequency	0.7	1.5	0.5
Forwd wd pred. x Forwd POS pred.	-0.5	0.4	-1.4
Verbs x Forwd POS pred.	-0.2	1.6	-0.1

Table 3.13: Fixed-effects Structure of RF-Informed LMER Model

3.4.1.1 Main Effects

None of the main effects in the pre-random-forest model (see Table 3.11) are qualitatively different from their effects in the post-random-forest model described in Table 3.13.

3.4.1.2 Interactive Effects

In the post-random-forest model, the five interactions illustrated in Figures 3.3, 3.4 and 3.5 again survive the trimming process. The three part of speech interactions shown in Figures 3.6 and 3.6, however, were eliminated during the first round of interaction testing. These interactions were replaced by five new interactions, all including part-of-speech based conditional probability.

The new interactions are plotted in Figures 3.16, 3.17, and 3.18 below

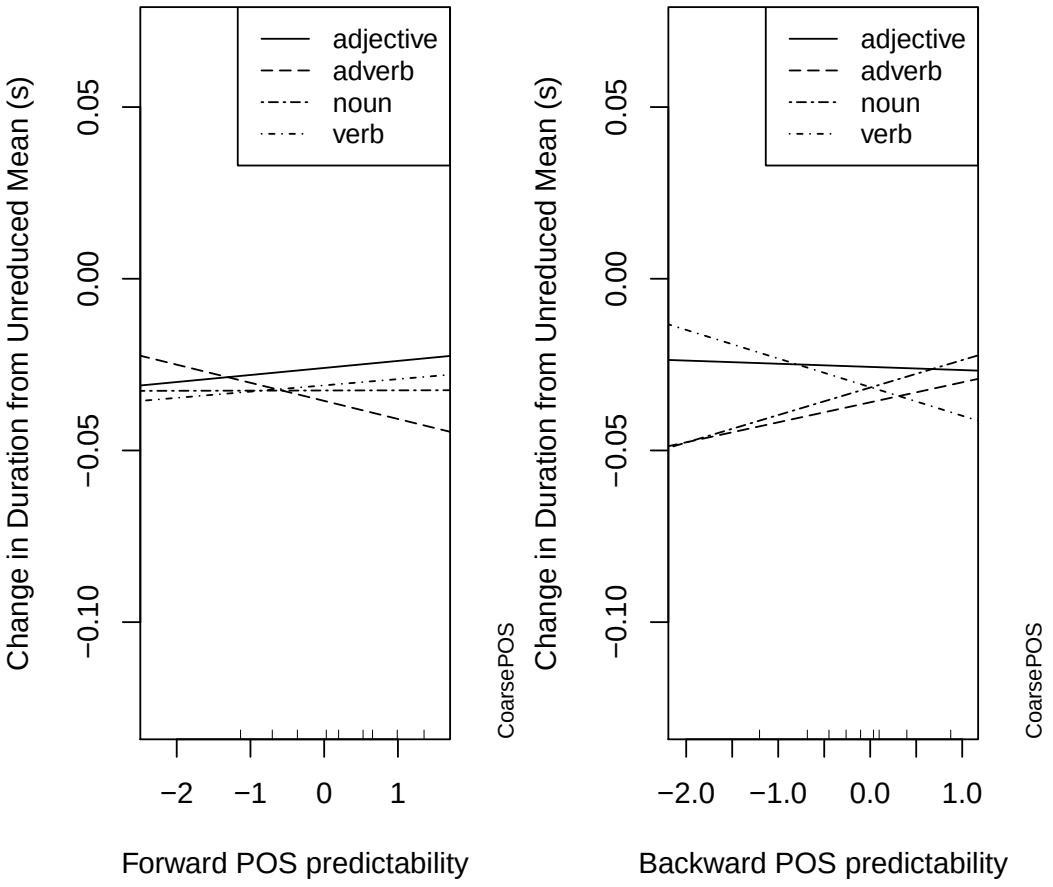

Figure 3.16: Part-of-Speech Interactions in RF-Informed LMER Model

Figure 3.16 illustrates the two interactions between part of speech and part-of-speech-based predictability.

In the left panel, a word's part of speech predictability given the preceding word's part of speech is illustrated. Nouns show little effect of this forward predictability, indicating that the duration of a noun is unlikely to change given the preceding word's part of speech. Indeed, the interaction between nouns and forward predictability does not meet the $|t| > 2$ significance threshold, as shown in Table 3.13, with an effect size of only -2.0 milliseconds. It is important to note that this significance level is calculated in comparison to the reference level of the part of speech predictor. In the present models adjectives are chosen as the reference level by default, since they fall first among POSes in alphabetical order. As a result, this result should be interpreted as indicating that nouns do not behave significantly differently from adjectives across the range of values of forward POS predictability. Verbs behave even less differently from adjectives across various levels of forward POS predictability: The verb-by-forward predictability component of the interaction in Table 3.13 has a t-value close to zero, and an effect size of only -0.2 milliseconds.

Adverbs, however, show a strong effect of forward predictability: They are more likely to be reduced as their predictability from preceding context increases. Since the predictability measure under consideration is based on part of speech, this result suggests that adverbs tend to be reduced when they are found in familiar syntactic contexts. This result has at least two possible interpretations in terms of speech production. In information-theoretical terms, the result suggests that when an adverb is expected, less phonetic information is required to convey which particular adverb the speaker is uttering. In speech processing terms, the result suggests that a speaker is able to produce (or habitually produces) an adverb more quickly when it is found in a syntactic position where adverbs are expected. More study would be required to arbitrate between these two interpretations.

The right panel of Figure 3.16 shows a much stronger interaction between part of speech and part-of-speech-based predictability. Adjectives are again chosen as the reference level, and Figure 3.16 suggests that this choice is particularly appropriate: Adjectives show little change in duration as they become more predictable from the following context. The effect sizes for the remaining interaction terms can thus be taken as how significantly part-of-speech predictability affects each part of speech.

The panel indicates that the duration of a verb depends highly on the part of speech that follows it, ranging from lengthening in unpredictable contexts to shortening in more predictable contexts. Adverbs and nouns, however, react to predictability in the opposite direction. Both adverbs and nouns undergo *less* duration reduction as they become more predictable from the following part of speech. All three of these interactions reach significance in Table 3.13, and have effect sizes that rank highly among the predictors in the model.

It is puzzling that predictability should lead to a decrease in reduction. Both forwards and backwards word predictability were shown to lead to strong reduction in word duration, as shown in Table 3.13. The reduction shown for verbs with high backwards part-of-speech predictability conforms to this pattern, showing that verbs become shorter as they become more predictable from following context. Adverbs and nouns, however, show the opposite tendency.

A similar result - longer processing times before more predictable words - has been

shown in eye movement studies by Kennedy and others (e.g., Kennedy (2000), Kennedy *et al.* (2002).) The studies found when readers fixated on a word followed by a high-frequency word, their fixation durations tended to be longer. The authors interpreted the result as indicating that during the fixation, the high-frequency word to the right was already undergoing processing. This parafoveal-on-foveal effect was supported by the fact that these high-frequency words tended to be skipped more often. A similar interpretation is possible here, for adverbs and nouns, at least: If an adverb or noun is followed by a highly predictable part of speech, preparation for the production of the following word may be taking place while the speaker is uttering the current word.

This interpretation may be best understood if backwards part-of-speech predictability is considered in terms of syntactic constituency. The number of part-of-speech types that can follow a word within a syntactic constituent is likely smaller than the number of part-of-speech types that can follow a word at the end of a syntactic constituent: Within a constituent, part-of-speech types are constrained by the set of possible variations of syntactic structure that such a constituent allows. At the edge of a constituent, however, the following part of speech can vary based on both what constituents can come next and what parts of speech can begin those constituents. Thus, a word is likely to have a low backwards POS predictability when it falls at the end of a syntactic constituent, and words with a higher backwards POS predictability are more likely to be found near the centre of, or within a highly routinized portion of, a syntactic constituent. Under these conditions, the following portions of the constituent may be being processed at the same time as the current portion. The present results suggest that this concurrent processing exists during speech production, putting an additional load on the speech production system and inhibiting reduction of the current word.

Backwards part-of-speech predictability has a different effect on verbs, however, suggesting that a different interpretation is necessary to explain verb behaviour. One such interpretation comes from the position verbs tend to occupy in English sentences. An English verb is likely to be followed by one of its arguments. These arguments usually represent a separate syntactic constituent, such as a noun phrase, prepositional phrase, or relative clause. A verb that is predictable from the following part of speech, then, is more likely to be found before the start of a new syntactic constituent. As a result, the constituent-processing explanation described above would not apply to verbs, and the more familiar facilitatory effect of predictability shown in the figure can be expected.

A more thorough investigation of the conditions under which part-of-speech predictability varies for each part of speech could further illuminate these results, but such an investigation is beyond the scope of the present study.

The left panel of Figure 3.17 shows an interaction between forward and backward part-of-speech predictability. The interaction shows a transition between facilitatory and inhibitory effects of forward POS predictability on word duration. Words whose part of speech is least predictable given the following part of speech tend to reduce more as their forwards POS predictability increases. These words, indicated by the downward-sloping line in the left panel of 3.17, are likely to fall at the edge of a syntactic constituent, for the reasons outlined in the discussion of the previous figure. Words more central to, or more predictable from,

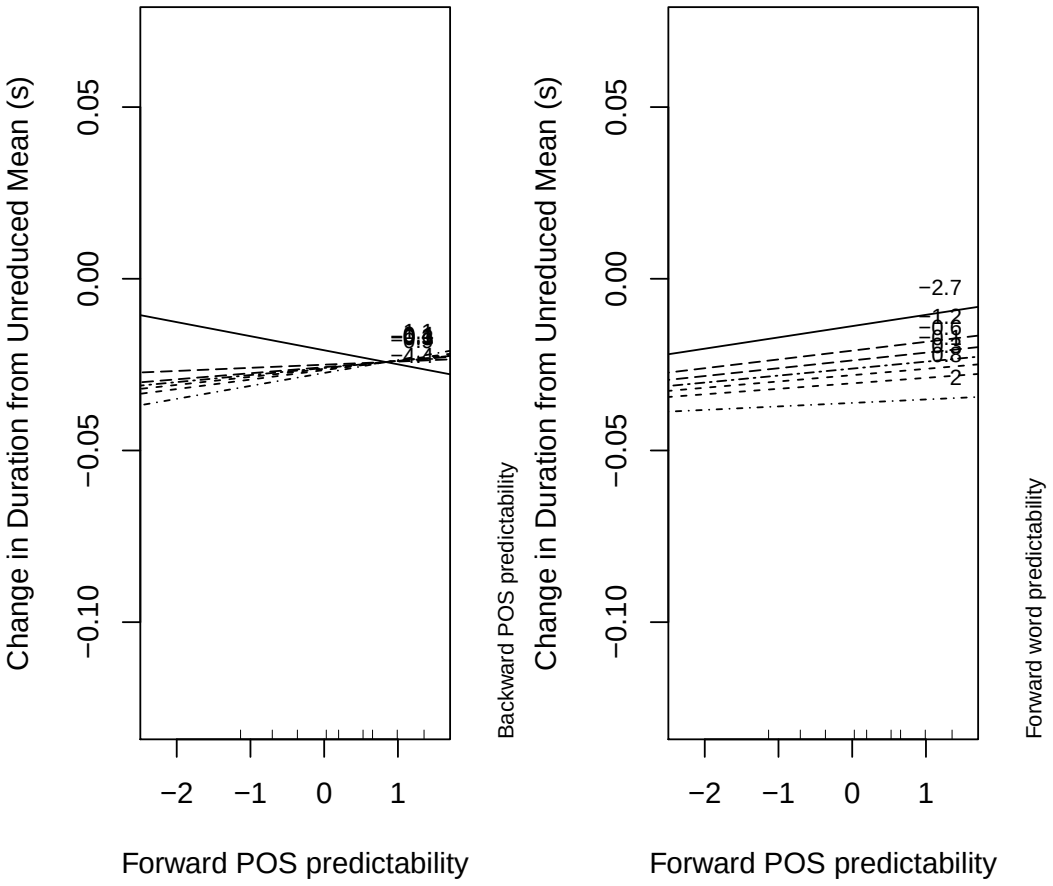

Figure 3.17: Forward Part of Speech Predictability Interactions in RF-Informed LMER Model

their syntactic constituency, appear to undergo less shortening. This effect is true for both forward and backward POS predictability: Most of the lines in the left panel of Figure 3.17 show less reduction as both forward and backward predictability increase.

This interaction also provides another potential explanation for why both measures fail to reach significance as main effects in the model described in Table 3.13. The effect of forward POS predictability can be either facilitatory or inhibitory, depending on the value of backwards POS predictability. If forward POS predictability is considered without taking this interaction into account, the inhibitory and facilitatory effects appear to cancel each other out, leading to the small, insignificant inhibitory effect described in Table 3.13.

The panel on the right of Figure 3.17 shows an inhibitory effect of forward POS predictability on reduction, and shows this inhibitory effect increasing as forward word predictability decreases. Words that are predictable given the previous word show little change in duration as their POS predictability given the previous POS increases. Words that are less predictable given the previous word, however, undergo less reduction as their part of speech's predictability increases.

This interaction could be seen as illustrating a type of prediction mismatch effect. If a word is in a predictable syntactic context, the speaker or listener may have prepared for a predictable word. When an unpredictable word is chosen instead, however, the decrease of information that comes with reduction appears to be dis-preferred. Thus, as less predictable words appear in more predictable syntactic contexts, the speaker is less likely to reduce the length of the word. In speaker-oriented terms, this lack of reduction could indicate a processing difficulty resulting from the mismatch in predictability. In listener-oriented terms, the lack of reduction could signal that the speaker provides more information to a listener receiving an unexpected mismatch in predictability.

Each line in Figure 3.18 shows that words are reduced more if the speaker is speaking more quickly. The interactive effect shows this reduction being mitigated by the predictability of a word's part of speech given the following part of speech. Unpredictable words show greater shortening than more predictable words do as speaking rate increases. Under the syntactic-constituency understanding of part-of-speech predictability, this result suggests that speakers are likely to shorten the latest words in a syntactic constituent when speaking quickly. Words that are more central to a syntactic constituent, or more predictable within a syntactic constituent, appear less likely to be shortened during fast speech.

Table 3.14 below compares the model with these five additional interactions to the model fitted earlier without them. Table 3.14 shows that the model with the additional interactions

	Model 1	Model 2
Degrees of Freedom	64	66
Proportion of Variance Explained (%)	19.2	19.2
log-likelihood ratio	89401	89457
log-likelihood improvement		56
Log-likelihood improvement p-value		0
Akaike Information Criterion (AIC)	-178674	-178782
AIC Improvement		107.9

Table 3.14: Model Comparison: Full Model (1) v. RF-Informed Model (2)

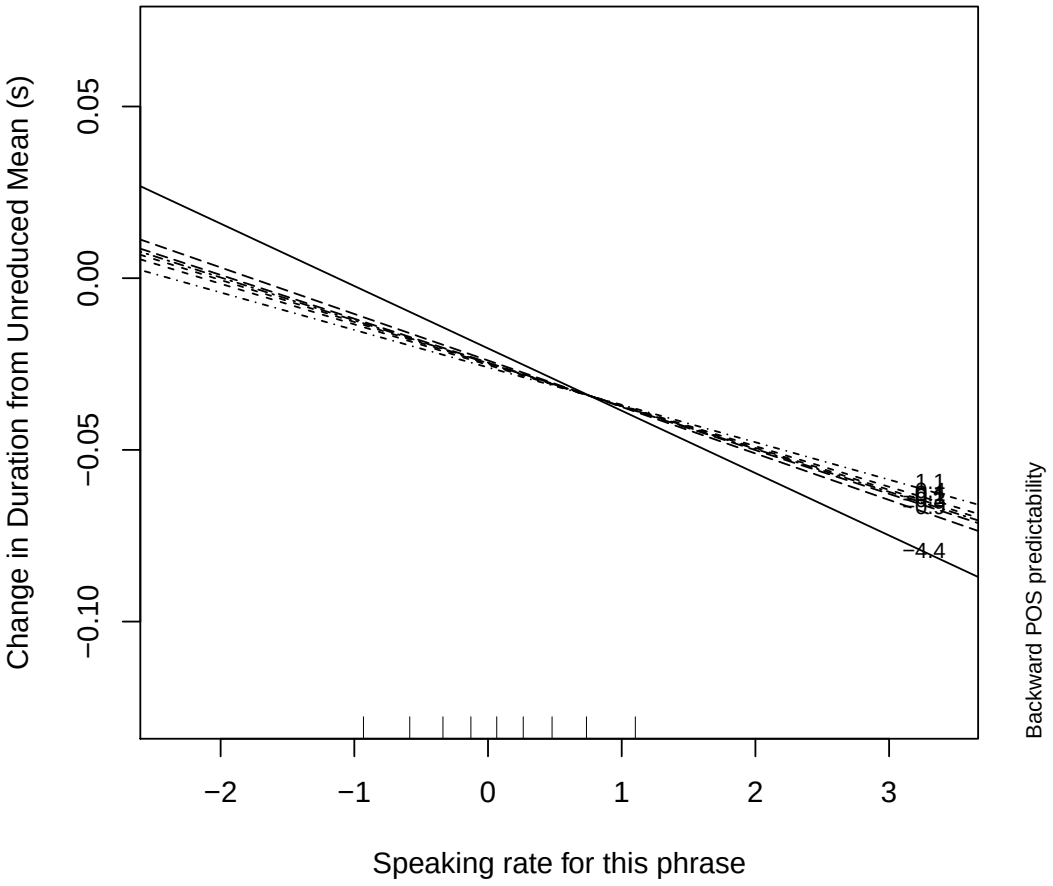

Figure 3.18: Speaking Rate Interaction in RF-Informed LMER Model. X-axis covers central 99.6% of data to accommodate plotting software

is a stronger and more parsimonious fit to the data than the earlier model. The proportion of variance explained increases slightly, from 19.2 to 19.2. Log-likelihood ratio testing finds the random-forest-filtered model is significantly ($p < 0$) more likely than the original complete LMER model. The addition of several additional parameters to the model informed by the random forests should, of course, improve model fit. The AIC values for the two models show that the improvement in model fit is not simply due to the increased number of parameters, as AIC scores are penalized for the addition of model parameters. The improvement in AIC scores, then, shows that the RF-Informed model is still the most efficient explanation of the data.

3.4.1.3 Random Effects

The random effects structure of the final model is described in Table 3.15.

Groups	Predictor	Std.Dev. (ms)	Corr		
Residual		70.438			
Word	(Intercept)	16.518			
Speaker	(Intercept)	4.240			
Word	Age (> 40)	4.635			
	Age (< 30)	5.601	0.01		
Word	Time	2.299			
Speaker	Time	2.175			
Speaker	Adjectives	3.378			
	Adverbs	5.906	0.50		
	Nouns	5.399	0.23	0.20	
	Verbs	2.835	-0.08	0.63	0.07
Word	Female Interviewer	0.001			
	Male Interviewer	9.663	0.00		
Speaker	COCA Frequency	3.779			
Word	Backwd wd pred.	7.317			
Speaker	Backwd wd pred.	3.427			
Word	Forwd wd pred.	6.107			
Word	Backwd POS pred.	6.486			
Word	Forwd POS pred.	4.965			
Speaker	Stress	3.152			
Speaker	Length	4.091			
Word	Neighbr POS pred.	5.489			
Word	tf-idf topicality	5.856			
Speaker	tf-idf topicality	1.496			
Word	Female Speaker	7.145			
	Male Speaker	5.281	0.08		
Word	Global spk. rate	3.162			
Speaker	Phrase spk. rate	5.000			

Table 3.15: Random-effects Structure of RF-Informed LMER Model

3.5 General Discussion

The findings of this chapter can be divided into two broad categories: First, there are the findings of the models themselves, which reveal several aspects of the nature of reduction in

the Buckeye corpus. Second, there are the findings about the modelling process itself, found by comparing and combining the LMER model selection and Random Forest modelling techniques. Each is discussed in a separate section below.

3.5.1 Model Findings

3.5.1.1 Main Effects

3.5.1.1.1 Demographic Predictors None of the demographic predictors included in the present study (Age, Speaker Gender, and Interviewer Gender) were found to contribute significantly to model fit as main or interactive effects. These predictors also have very low importance scores in the random forest model. Similar results have been found in other studies of the Buckeye Corpus. Gahl *et al.* (2012) and Yao (2011) found no significant effect of age or speaker gender on reduction. Bell *et al.* (2009), using the Switchboard corpus (Godfrey *et al.* 1992), found that younger speakers tended to reduce more than older speakers. The age measure used in the Switchboard corpus (years of age) is more gradient than the binary measure ("under 30" or "over 40") coded in the Buckeye corpus, including more than 500 speakers between 20 and 60 years of age. The difference in results between the Buckeye studies and the Switchboard study suggests that this gradient measure may be a better predictor of reduction.

Bell *et al.* (2009) also find an interactive effect between speaker gender and speech rate, with men speaking more quickly on average. No support for this interaction was found in the present study.

3.5.1.1.2 Phonological and Phonetic Predictors Faster speaking rates for the words surrounding the target led to more duration reduction in the present study, in both LMER and random forest models. Several studies have found that faster speaking rates correlate with decreased word durations. In some studies, speaking rate was measured separately for the words before and after the target word, and each of these measures was included as a predictor. Gahl *et al.* (2012) and Yao (2011) found facilitatory effects for both measures of local speaking rate, though the post-target speaking rate appeared to have a stronger effect on reduction. Indeed, in one study, Gahl (2008) found a significant effect on reduction *only* for post-target speaking rate, with no significant effect of pre-target speaking rate.

One study (Tily *et al.* 2009) found that speaking rate did not contribute significantly to the fit of a model of duration reduction. The study focused only on the duration of the function word 'to' in a very specific context, however. When placed in the context of the existing studies, this result suggests that local speech rate does tend to lead to reduction, though not in every condition.

Global speaking rate, calculated by speaker, has not been evaluated as a predictor of reduction in any of the studies cited here. The robust relationship between lower average speech rates and greater levels of word shortening shown in the present chapter is surprising. Independent studies of this effect would help to confirm or further explain it. The explanation provided above - that faster speakers reduce less during faster portions of a conversation than slower speakers - is supported by the current study, but further examination of this effect seems warranted.

The number of stressed syllables in a word had no effect on reduction in either the LMER or random forest models presented here. None of the studies of duration reduction cited here use this measure as a predictor, and the results of the present chapter suggest that this omission is warranted.

A word's expected length in phonemic segments, residualized here against COCA frequency, was found to correlate with reduction in both model types. Few studies have used this measure. Several studies examine the effect of orthographic length on duration. These studies find that this orthographic length has no significant effect on word duration (Bell *et al.* 2009; Gahl *et al.* 2012; Yao 2011). Other measures of word length did prove useful in modelling reduction, however: Bell *et al.* (2009) found that a word's average duration and number of syllables significantly improved a model of word duration. Gahl *et al.* (2012) found that a word's baseline duration, calculated as the sum of the corpus-wide mean duration of its segments, significantly improved model fit. Length results in the present study can not be compared directly to the length results found by these previous studies, however. Those studies take pure observed word duration as the dependent variable, and finds that it positively correlates with expected word duration: The longer a word was expected to be, the longer it tended to be. By contrast, the present study considers *reduction* from expected word duration as the dependent variable: A word form's expected duration is taken as the average duration of the productions with the same number of phonemic segments as the citation form. Duration reduction is then taken as the difference between this expected word form duration and the observed duration of the target word production. This reduction-based measure is found to *negatively* correlates with expected word length: The longer a word was expected to be given its citation form, the more likely it was to be shorter in duration than productions with citation-form length.

3.5.1.1.3 Predictability Local (Buckeye) frequency was found to be a useful predictor of reduction in the LMER model, though it did not meet the threshold for importance in the random forest model. Studies that used a frequency measure based on the corpus under study found that this "local frequency" measure corresponded to greater duration reduction (Bell *et al.* 2009; Gahl 2008).

Unlike the present study, however, several studies have found a strong relationship between an external measure of word frequency and duration reduction (Gahl *et al.* 2012; Baker *et al.* 2011; Aylett & Turk 2004; Baker & Bradlow 2009; Yao 2011). Recall that in the present study, COCA frequency did not contribute significantly to the model as a main effect, though the Random-Forest-filtered model found it to be a useful component of multiple interactive effects.

Both model types found backwards and forwards word predictability (i.e., the conditional probabilities of the token given the following or preceding words) to be significant predictors of duration reduction. Several existing studies confirm this facilitatory effect of predictability on reduction. Greater predictability given the previous word was found to lead to greater duration reduction in several studies (Gahl *et al.* 2012; Tily *et al.* 2009; Gahl 2008; Yao 2011), though some studies found no significant effect (Bell *et al.* 2009) Similarly, several studies of the relationship between conditional probability given the previous word and

duration reduction find a significant relationship (Gahl *et al.* 2012; Bell *et al.* 2009; Yao 2011), while a smaller number of studies found no relationship (Gahl 2008; Tily *et al.* 2009). On a related measure, Aylett & Turk (2004) find that predictability given the *two* preceding words leads to greater reduction in word duration.

3.5.1.1.4 Structural Constituency A word's position in its carrier phrase was not a significant predictor of reduction in either model type. This measure was not used in any of the studies cited here. Several studies have found an effect of phrase initial or phrase final status on reduction, but all such tokens were eliminated from the present analysis.

Part of speech was found to contribute significantly to the quality of both LMER and random forest models. In the present study, nouns were found to undergo the least reduction, followed by adjectives, adverbs, and verbs. This same ordering was attested to in two previous studies of duration in the Buckeye corpus (Gahl *et al.* 2012; Yao 2011).

Part-of-speech-based predictability measures were found to contribute significantly to the quality of model fit in the present study. Predictability given the surrounding parts of speech was found to facilitate reduction in both LMER and random forest models. Predictabilities given the previous or following part of speech were not found to improve LMER model fit initially. Random forest modelling, however, showed that these predictors were highly important in predicting reduction, and further analysis showed that these measures participate in interactive effects that significantly improve model fit. These measures are omitted from all of the studies cited here, suggesting that they constitute a relatively unexplored area in reduction research.

3.5.1.1.5 Topicality The present study uses term frequency-inverse document frequency (tf-idf - (Luhn 1958; Robertson & Jones 1976)) to measure the topicality of a particular word token. This measure was found to facilitate reduction in both LMER and random forest models. No study cited here uses this measure of topicality to predict reduction. Several studies, however, find an effect of a predictor that may strongly correlate with tf-idf: Previous mentions of the target word (Aylett & Turk 2004; Bard *et al.* 2000; Baker & Bradlow 2009; Bell *et al.* 2009; Fowler & Housum 1987; Lam & Watson 2010). A topical word, as measured by tf-idf, is one that appears unusually frequently in a particular stretch of conversation. Most tokens with high topicality are likely second mentions, not only within the conversation but within the roughly 10 minutes of conversation that form their carrier document.

Other studies have found an effect of some measure of givenness on duration. Kahn & Arnold (2012) find that a word preceded either by another token of that word or a picture of what that word signifies is likely to be shorter than a word that has not yet been introduced. Anderson & Howarth (2002) find that a word spoken previously by any party involved in the conversation is also likely to be shorter than an unmentioned word. These measures suggest that givenness or topicality effects on duration are widespread and multi-modal.

3.5.1.1.6 Time The time at which the target word appears in an interview was not found to have a significant effect on reduction in any models presented here. This measure

has not been used in existing studies of reduction, and the present study finds no evidence that it should be included in the main-effects structure of models of duration reduction.

3.5.1.2 Interactive Effects

After random-forest-based filtering, the final model found 10 interactions that contributed significantly to model fit. The details of these interactions, along with their implications, are discussed in Sections 3.2.6.2 and 3.4 above. The present section focuses on evidence for interactive effects in existing studies of reduction.

Few such studies use the interaction-selection procedure employed here, and no work cited here considers as many interactions as the present study. Gahl *et al.* (2012) tested two interactions, and found that they did not contribute significantly to model fit. Baker & Bradlow (2009) found a three-way interaction between second mention, frequency, and speech style. The interaction indicated that in plain speech, second mention and frequency contribute collaboratively to reduction. As the present corpus is limited to plain speech, the closest analogue to this interaction in the present study would be one between topicality and frequency. No such interaction was found to contribute significantly to any of the models produced in the present chapter.

Bell *et al.* (2009) found a significant interaction between speaker gender and speech rate, with men speaking faster than women. No such result was found in the present study, and the low importance of speaker gender in random forest modelling suggests that no such interaction is likely to appear in the Buckeye corpus.

Bell *et al.* (2009) also find a facilitatory interaction between local frequency and backwards word predictability. This interaction was also found to be facilitatory in the current model, as shown in Table 3.13 and illustrated in Figure 3.3

3.5.1.3 Random Effects

Several existing studies include random intercepts for word and speaker in their models of duration. (e.g., (Gahl *et al.* 2012; Kahn & Arnold 2012; Lam & Watson 2010; Yao 2011)) Few studies consider (or, at least, report) the effect of including random slopes in their models, however. Kahn & Arnold (2012) include random slopes in their model, but do not describe their effects in detail. Indeed, it is not entirely clear which random slopes were found to contribute significantly to the fit of their model. Lam & Watson (2010) tested several random slopes, but found that none significantly improved model fit. This may be due, in part, to the restriction that many studies place on their words of interest. If analysis is confined to monomorphemic, monosyllabic words, the amount of variability by word is likely to be reduced. In the present study, random intercepts and slopes are used to quantify the variability that is not accounted for by the main-effects structure of the model. With greater variability in the character of the words under study, the likelihood that random slopes capture this greater variability increases.

It is also possible that adding predictors to the current study would decrease the importance of random slopes. Still, given the number of random slopes found to significantly improve model fit, as well as the proportion of variance explained by the random effects, it

is surprising that so few studies of word duration have found an important contribution of random slopes.

3.5.2 Modelling Techniques

The modelling process described in the present chapter reveals that each of the modelling techniques used has clear strengths and weakness. In forward-fitting LMER modelling, at least with the procedure outlined here, some interactive effects were overlooked. When the results of the random forest model were taken into account, the number of interactions found to significantly improve model fit was doubled. Five interactions in total were eliminated from consideration too early. Each of these interactions sheds light on some effect of syntactic constituency on reduction, and the loss of this insight severely limits the understanding of the data. Once added to the model, the interactions also help to explain why they were eliminated from consideration in the first place. Figure 3.17 shows that the relevant predictors range from increasing with reduction rates to decreasing with reduction rates. The net result of this variation was that each predictor registered only a weak main effect, precluding their consideration in interactions. A modeller could address this problem by simply considering every possible two-way interaction between variables, but this approach proves computationally prohibitive for large data sets with large numbers of predictors.

Random forests can be used to provide a short cut to determining where informative interactions should be looked for. By implicitly considering arbitrarily complex interactions, random forest modelling provides a clear test of which variables are likely to contribute to the quality of model fit, as main effects and as members of interactive effects. Perhaps more importantly, random forest modelling can be used to decide which variables are *unlikely* to be present in useful interactions, reducing the problem space faced by the modeller.

Random forests also provide the modeller with a more precise understanding of the way in which each predictor affects the dependent variable. As illustrated in Section 3.3.2.3, the linear model constructed here oversimplifies the effects of predictability and word length, for example, on reduction. In each case, the random forest describes a non-linear relationship between the predictor and reduction at a level of detail impossible in a strictly linear model. Spline predictors can be used, of course, to describe arbitrarily complex curves in LMER models. The choice of which predictors are likely to have non-linear effects on the response variable, however, along with the complexity allowed for each of these non-linear predictors, are parameters that the modeller must tune by hand in an ad-hoc fashion after considering the variables in isolation. With random forest modelling, the modeller simply asks the forest to consider a set of predictors, and the modelling process reveals which effects may best be modelled as non-linear.

Random forest models were shown to have their own drawbacks in the analysis of linguistic data, however. In particular, the inability to consider predictors that represent large, unordered groups of values caused the models to overlook important insights about the data. Such predictors are integral to most forms of linguistic analysis. In the present study, the effects of word form and speaker proved impossible to model using random forests. These predictors were used in the LMER model as grouping variables, and random slopes and intercepts based on these groups were found to significantly improve model fit. Table 3.12

reveals the rich random-effects structure of the final LMER model. The models described in Section 3.2.6.4 revealed that this random-effects structure is almost as capable of fitting the data on its own as the full model is. These random-effects not only improve model fit, but also provide insight into the nature of reduction in the Buckeye corpus. Word form (COCA) frequency, gender, and age, for example, were not found to be useful predictors as part of the fixed-effects structure, despite good theoretical reasons to consider them important. The random-effects structure of the final model reveals that these predictors are indeed exerting influence on reduction, but this influence is expressed differently for different speakers or different words. Studies aimed at other levels of linguistic analysis are likely to find similar restrictions when using random forest modelling: Several key linguistic variables - phonemes, speakers' geographic locations, or categories of meaning, for example - are likely to contain relatively large unordered sets of values. Understanding the effect of these variables is likely to form a key component of such studies. Random forest models alone cannot provide this understanding.

Given the results of the present chapter, then, a combination of Random Forest modelling and Linear Mixed-Effects Modelling may provide the modeller with more insight into the structure within the data than either method can alone.

Chapter 4

Segment Deletion

4.1 Introduction

The present chapter describes the models that were constructed to describe the deletion of segments from words. The model selection process, and the way in which models are compared, is nearly identical to that described in the methods chapter and implemented in the duration reduction chapter. To maintain consistency, the separation of pre- and post-Random Forest LMER modelling is adopted here. In fact, in the present chapter this separation reveals that the process of combining modelling techniques requires some modification, as described in Section 4.3.2. This section summarizes the process of the previous chapter, and the modifications required to apply the process to segment deletions.

In brief, the process consists of forward-fitting linear mixed-effects regression models, adding main effects, interactions, and random slopes in successive stages. At each stage, the linear models are compared to each other by log-likelihood ratio testing, Akaike Information Criterion comparison, and a comparison of the overall proportion of variance explained by the model.

A random forest model is also created, and compared to both the final linear model and to linear models that contain an amount of information that reflects the information available to random forest modelling. The linear model is compared to the random forest model in terms of the proportions of variance explained, the importance each model places on each predictor, and the partial effects on deletion rates established for each predictor.

There is one important difference in the linear modelling process, however: Deletions are modelled here as a Poisson process. Each word is seen as having some unknown probability of having a segment deleted in each production of that word. By counting the number of deletions under certain conditions (i.e., given certain values of the predictors), the modelling process attempts to determine when a word is more or less likely to have a segment deleted.

The assumptions underlying the way in which the number of deletions is assessed are worth re-iterating here. As in the previous chapter, the question of what a production can be said to be reduced *from* must be operationalized. In both chapters, the citation form provided in the corpus is taken as the expected unreduced form. In the present chapter, the number of deletions is taken as the difference between the number of expected (citation-form) segments and the number of segments actually produced by the speaker. Assimilations

that result in fewer observed segments, then, are counted as deletions.

Poisson modelling requires that this number of deletions be positive. That is, word productions in which the observed number of segments exceeds the citation number of segments can not be included. As a result, words in the data that have epenthetic segments that would lead them to have to a negative deletion count must be excluded from the analysis. Fortunately, only 2 such words were excluded.

4.2 Linear Mixed-Effects Regression Modelling

4.2.1 Baseline Model

The baseline model was fit with the 18 main-effect predictors described and discussed in the previous chapters, along with random intercepts for wordform and speaker. The strengths of each predictor's effect on deletion in this baseline model are shown in Table 4.1 below.

Description	Effect	Std.Err.	z.value	Pr(>|z|)
Length of word (res.)	0.601	0.022	27.0	0.000
(Intercept)	-2.459	0.119	-20.7	0.000
Backwards word predictability	0.108	0.011	9.8	0.000
Speaking rate for this phrase	0.083	0.010	7.9	0.000
Forward word predictability	0.069	0.012	5.9	0.000
Local frequency (res.)	0.168	0.029	5.8	0.000
Number of stressed syllables	-0.119	0.032	-3.7	0.000
Surrounding POS predictability (res.)	0.032	0.010	3.1	0.002
tf-idf topicality (res.)	0.034	0.011	3.0	0.003
Male Speaker	0.196	0.078	2.5	0.012
Global speaking rate	0.092	0.037	2.5	0.013
Forward POS predictability	0.025	0.012	2.1	0.035
Nouns	-0.124	0.070	-1.8	0.077
Verbs	0.124	0.075	1.7	0.097
Male Interviewer	0.103	0.078	1.3	0.187
Adverbs	0.083	0.082	1.0	0.311
Time when token appears in conversation	-0.007	0.009	-0.8	0.440
Age (under 30)	-0.057	0.078	-0.7	0.470
Backward POS predictability	-0.010	0.019	-0.6	0.578
Word's (ordinal) position in this phrase	-0.004	0.009	-0.5	0.633
COCA Frequency	-0.012	0.055	-0.2	0.830

Table 4.1: Main effects for baseline model.

4.2.2 Removing Insignificant Predictors

The next model that was fit is identical to the baseline model above, but with the insignificant main-effect predictors (i.e., those with $p > 0.05$) removed. The models are compared in Table 4.2 below.

Unfortunately, Table 4.2 suggests that the trimming process was too aggressive: The model with all predictors is significantly ($p < 0.004$) more likely as a description of the data than the trimmed model. The trimmed model also scores more poorly on the Akaike

	Model 1	Model 2
Degrees of Freedom	23	14
Proportion of Variance Explained (%)	33.2	33.2
log-likelihood ratio	-18112	-18124
log-likelihood improvement		-12.1
Log-likelihood improvement p-value		0.004
Akaike Information Criterion (AIC)	36270	36277
AIC Improvement		-6.3

Table 4.2: Model Comparison: Baseline model (1) v. Model with Non-significant Main Effects Removed (2)

information criterion, providing strong evidence that the trimmed model should be less preferred: The trimmed model has 9 fewer degrees of freedom than the baseline model, but the AIC, which values more parsimonious models, still considers the trimmed model a poorer fit to the data. The choice of predictors with which to proceed must thus be selected and motivated by the modeller. The process by which these predictors are selected is described here.

Clearly, fewer predictors should be removed from the model at this stage. The results described in Table 4.1 suggest two particular candidates for inclusion in the next stage of modelling: Part of speech and interviewer gender. Table 4.1 reveals that the part of speech predictors - nouns and verbs in particular - approach the $p < 0.05$ significance level, with $p \approx 0.077$ and $p \approx 0.097$, respectively. Moreover, the (absolute) effect sizes for nouns and verbs are higher than the effect sizes of most of the more significant predictors in Table 4.1.

Interviewer gender's z-value places it farther from the traditional significance threshold, at $p \approx 0.19$, but its effect size ranks it above half of the effect sizes of the predictors that reached the $p < 0.05$ significance level. The five remaining predictors Table 4.1 all display z-values below 1, and the likelihood that each is affecting deletion rates is at most 56%. These five insignificant predictors are not included in subsequent models.

A new model was fit, then, with main effects for all predictors in Table 4.1 that have $z \geq 1$, and random intercepts for word and speaker. The main-effects structure of this less conservative model is shown in Table 4.3, and the model is compared to the baseline model in Table 4.4 below.

The log-likelihood ratio testing illustrated in Table 4.4 shows that the new model is a poorer, though not significantly poorer, model of the data than the baseline model. The slight improvement in the AIC score for the smaller model suggests that it is a more efficient description of the data. As a result, the model described in Table 4.3 was used as the basis of the remaining model selection process.

Description	Effect	Std.Err.	z.value	Pr(>|z|)
Length of word (res.)	0.602	0.022	27.2	0.000
(Intercept)	-2.473	0.094	-26.2	0.000
Backwards word predictability	0.108	0.011	9.8	0.000
Speaking rate for this phrase	0.083	0.010	7.9	0.000
Local frequency (res.)	0.170	0.028	6.1	0.000
Forward word predictability	0.069	0.012	5.9	0.000
Number of stressed syllables	-0.118	0.032	-3.7	0.000
Surrounding POS predictability (res.)	0.029	0.008	3.5	0.000
tf-idf topicality (res.)	0.034	0.011	3.0	0.003
Male Speaker	0.197	0.079	2.5	0.012
Global speaking rate	0.088	0.037	2.4	0.016
Forward POS predictability	0.024	0.012	2.0	0.040
Nouns	-0.127	0.070	-1.8	0.070
Verbs	0.128	0.075	1.7	0.086
Male Interviewer	0.102	0.078	1.3	0.189
Adverbs	0.081	0.081	1.0	0.320

Table 4.3: Main effects for reduced model.

	Model 1	Model 2
Degrees of Freedom	23	18
Proportion of Variance Explained (%)	33.2	33.2
log-likelihood ratio	-18112	-18113
log-likelihood improvement		-0.9
Log-likelihood improvement p-value		0.888
Akaike Information Criterion (AIC)	36270	36262
AIC Improvement		8.3

Table 4.4: Model Comparison: Baseline model (1) v. Model with Main Effects with $z < 1$ Removed (2)

4.2.3 Exploring Interactive Effects

Potential interactive effects were selected using the same multi-stage process described in the previous chapter. First, a large model was constructed with the main effects predictors shown in Table 4.3, along with every possible two-way interaction between them and random intercepts for word and speaker. The model failed to converge after the maximum number of default iterations (300) of the numerical fitting method used in the **lmer** function implemented in the **lme4** R package. Still, the final iteration of the model provides an approximation of the strength of the interactions under consideration. Interactions with absolute z-values below 3 in the final iteration of the model that were immediately removed, leaving 5 interactions of interest. (The use of parameters from an iteration of a model that has not converged to inform model simplification here is modelled after a similar technique used by Barr *et al.* (2013)).

A series of smaller models were created to test the effect of each of these remaining interactions individually.

For each of the 5 interactions, a model was created with that interaction, all remaining main effects, and random intercepts for Word and Speaker. Log-likelihood ratio tests were then conducted to compare each of these models to an identical model with the interactive-

effect term removed. If an interaction did not significantly improve the model, it was eliminated from further consideration.

This process must be thought of as a series of post-hoc tests for significance, and some post-hoc testing correction is appropriate. In this case, a simple Bonferroni correction (i.e., dividing the highest significant p-value by the number of tests performed) was used, reducing the target p-value from 0.05 to ≈ 0.01.

During the model fitting process, one model had a Cholesky matrix that was not positive definite, indicating a violation of one of the modelling assumptions. This was likely due to the fact that the interaction term in the model was between highly correlated predictors. The number of stressed syllables in a word, and (residualized) word length are correlated at a rate of ($r \approx -0.733$). As a result, the number of data points for which the interactive term and the main effect term differ may not be sufficient to allow for a model in which the two are described as varying independently. That is, a linear model containing both terms should be considered unreliable. As a result, this interaction was excluded from further consideration.

A total of 3 interactions remained after the process described above. A model was thus constructed with all significant main-effect predictors, the 3 remaining interactions, and random intercepts for Word and Speaker. The fixed-effects portion of this model is described in Table 4.5.

Description	Effect	Std.Err.	z.value	Pr(>|z|)
Length of word (res.)	0.600	0.022	27.0	0.000
(Intercept)	-2.478	0.095	-26.2	0.000
Backwards word predictability	0.108	0.011	9.8	0.000
Speaking rate for this phrase	0.083	0.010	7.9	0.000
Forward word predictability	0.072	0.012	6.1	0.000
Local frequency (res.)	0.166	0.028	5.9	0.000
tf-idf topicality x Male Speaker	0.063	0.016	3.9	0.000
Forward POS predictability	0.048	0.013	3.7	0.000
Number of stressed syllables	-0.112	0.032	-3.5	0.000
Surrounding POS predictability (res.)	0.029	0.008	3.5	0.000
Forwd wd pred. x Forwd POS pred.	0.033	0.010	3.3	0.001
Forwd POS pred. x Stress	-0.042	0.015	-2.8	0.005
Male Speaker	0.205	0.079	2.6	0.009
Global speaking rate	0.089	0.037	2.4	0.016
Nouns	-0.126	0.070	-1.8	0.072
Verbs	0.119	0.075	1.6	0.111
Male Interviewer	0.101	0.078	1.3	0.194
Adverbs	0.083	0.081	1.0	0.310
tf-idf topicality (res.)	-0.006	0.015	-0.4	0.676

Table 4.5: Fixed-effects Structure of Model with Significant Main and Interactive Effects

Table 4.6 shows that adding these interactions led to strong improvements in both log-likelihood estimation and AIC. The increase in the proportion of variance in the data explained by the model is noticeable, but quite small ($\approx 0.1\%$)

	Model 1	Model 2
Degrees of Freedom	18	21
Proportion of Variance Explained (%)	33.2	33.3
log-likelihood ratio	-18113	-18095
log-likelihood improvement		17.9
Log-likelihood improvement p-value		0
Akaike Information Criterion (AIC)	36262	36232
AIC Improvement		29.8

Table 4.6: Model Comparison: Model without Interactions (1) v. Model with Selected Interactions (2)

4.2.4 Exploring Random Effects

In this section, the effect of allowing predictors to vary in slope by wordform or speaker is investigated. These random slopes were again fitted only for fixed-effect predictors that vary within-group for each of wordform type and speaker. Each random slope in Table 3.5 was again added separately to a baseline model, and the improvement in model fit was measured. Improvement was calculated using log-likelihood estimation, as implemented by the **anova** function in R. The baseline model selected contained random intercepts for word and speaker, as well as the 13 significant main-effect predictors selected during the first step of the modelling process. (i.e., the predictors listed in Table 4.3). Interactions were again left out of this baseline to facilitate faster computation. As this process consists of multiple post-hoc significance tests, Bonferroni correction was again applied to determine a more appropriate significance threshold than $p < 0.05$. More precisely, with 27 group covariates, Bonferroni correction leads to a significant p-value of ≈ 0.0019.

Once this process was completed, the random slopes found to significantly improve model fit were all added to the baseline model to form a new model with a complete random-effects structure. The random-effects structure of this model is shown in Table 4.7. Table 4.8 compares the previous model to the new model with all random slopes added. The previous model contained the interactive-effect terms, while the random-slopes model does not. Still, the random slopes model shows an improvement in AIC score, and a significant improvement in log-likelihood. Table 4.8 shows that adding random slopes also improves the proportion of variance explained by the model, as it did in the model of duration reduction. Adding random slopes improved predictive power by 4.7 percentage points, an increase of about 14%

Next, the interactions found to contribute significantly to model fit were added to the model with random slopes. Table 4.9 compares the model with both interactions and random slopes to the model with random slopes but no interactions. The proportion of variance explained by the models is roughly the same, but log-likelihood ratio testing and the Akaike Information Criterion both suggest that the full model is the best fit to the data so far.

Groups	Predictor	Std.Dev.	Corr		
Speaker	(Intercept)	0.201			
Word	(Intercept)	0.680			
Word	Age (> 40)	0.421			
	Age (< 30)	0.262	1.00		
Speaker	Time	0.101			
Speaker	Adjectives	0.228			
	Adverbs	0.161	0.61		
	Nouns	0.194	0.92	0.68	
	Verbs	0.086	0.77	0.18	0.84
Word	Female Interviewer	0.095			
	Male Interviewer	0.187	0.09		
Speaker	COCA Frequency	0.075			
Word	Forwd wd pred.	0.144			
Word	Forwd POS pred.	0.142			
Speaker	Length	0.071			
Word	tf-idf topicality	0.079			
Speaker	tf-idf topicality	0.033			
Word	Female Speaker	0.788			
	Male Speaker	0.494	0.99		
Word	Global spk. rate	0.106			
Word	Phrase spk. rate	0.028			

Table 4.7: Random Effects Found to Contribute Significantly to Model Fit

	Model 1	Model 2
Degrees of Freedom	21	46
Proportion of Variance Explained (%)	33.3	38
log-likelihood ratio	-18095	-17855
log-likelihood improvement		240.5
Log-likelihood improvement p-value		0
Akaike Information Criterion (AIC)	36232	35801
AIC Improvement		431

Table 4.8: Model Comparison: Main and Interactive Effects (1) v. Main and Random Effects (2)

	Model 1	Model 2
Degrees of Freedom	46	49
Proportion of Variance Explained (%)	38	38
log-likelihood ratio	-17855	-17850
log-likelihood improvement		4.6
Log-likelihood improvement p-value		0.026
Akaike Information Criterion (AIC)	35801	35798
AIC Improvement		3.2

Table 4.9: Model Comparison: Main and Random Effects (1) v. Main, Interactive, and Random Effects (2)

4.2.5 Model Criticism

Model criticism is again applied here, in order to identify potentially unduly influential outliers and mitigate their effects on the conclusions drawn from the model. The search for outliers again begins with an examination of the standardized residuals (i.e., prediction errors) of the model. Figure 4.1 shows several ways of visualizing the error in model prediction. The upper-left panel of Figure 4.1 shows the distribution of these standardized

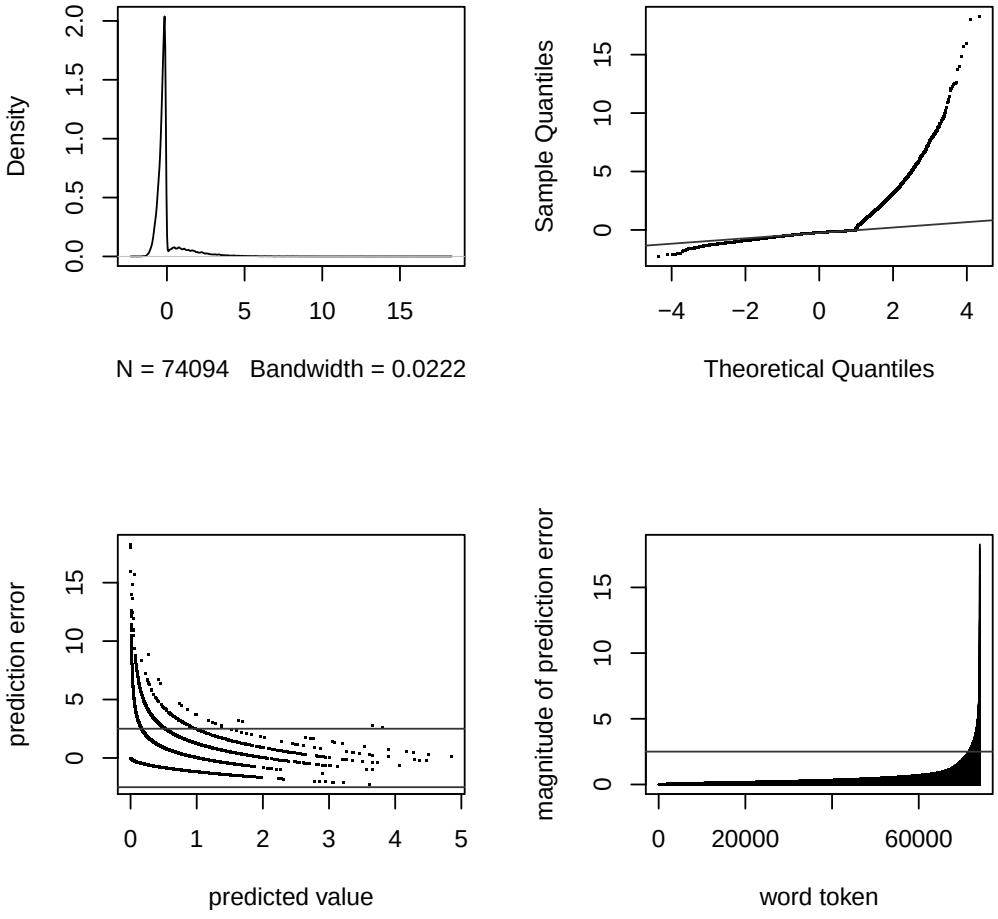

Figure 4.1: Model Criticism Plots

residuals of the model. The plot is heavily skewed, with a much longer right tail than would be predicted from a normally distributed variable.

This right skew is also illustrated in the upper-right panel, a quantile-quantile plot of the residuals. The quantile-quantile plot deviates from normality very sharply at the right side of the graph, suggesting that the model severely underestimates the chances of deletion for several words.

It is worth noting at this stage that the scale of the residuals does not directly correspond

to the number of deletions. That is, there are no words in the corpus with 15 segments deleted. Rather, the scale of the residuals represents the fact that the size of the residual value in a Poisson model can increase exponentially as the prediction deviates from the actual number of deletions. (That is, the residuals of a Poisson model are not expected to be normally distributed.) This is illustrated in Figure 4.2, which plots residual values against the difference between the model predictions and the actual number of deletions for each token.

In Poisson models, residuals are calculated differently for each value of the dependent variable. This fact, too, is illustrated in Figure 4.2. Each number of (observed) deletions is plotted with a different character in Figure 4.2, and each number of observed deletions clearly follows a different curve. In short, words with fewer observed deletions are penalized more severely for underestimating the number of deletions.

This fact helps to explain the striated nature of the scatter plot in the lower-left panel of Figure 4.1. The panel plots residuals against model predictions directly.

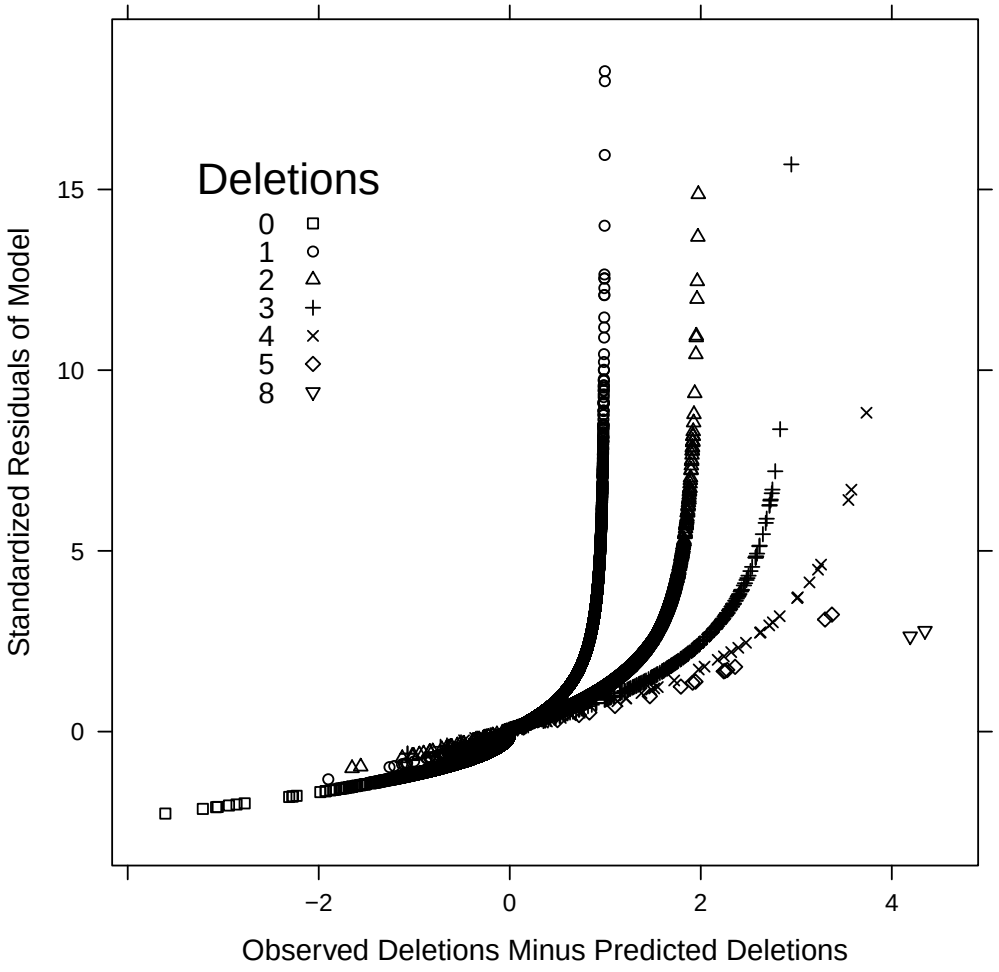

Figure 4.2: Residual Values v. Prediction Error, Grouped by Number of Deletions

In the lower-left panel of Figure 4.1, then, each apparently distinct curve likely represents the residuals for a certain number of observed deletions. Several of these residuals are quite high, showing data points for which the model underestimates the number of deletions. (Lines indicate ±2.5 standard deviations away from the mean residual value.)

The bar chart in the lower-right panel of Figure 4.1 shows the sorted absolute standardized residuals. The horizontal line indicates 2.5 standard deviations away from the mean residual. Of the panels in Figure 4.1, this panel most clearly illustrates the proportion of data points for which residuals exceed this 2.5 standard deviation threshold. Most bars fall below the threshold, and only a relatively small proportion exceed the threshold. Still, those data points that do exceed the threshold do so by a large amount: The highest residual value reaches 18.3.

These data points with high-valued residuals may be unduly influencing the model. To investigate this possibility, all data points with absolute residuals above 2.5 were removed from the data set, and a new model was fit. A total of 2609 data points, or 3.5% of the remaining words, were trimmed to create this model.

Unfortunately, while the model fit to the trimmed data set converged, it failed to meet the convergence conditions for a generalized linear model. That is, the model represents what the **lmer** method calls a 'false convergence'. Examination of the intermediate results of the model-fitting algorithm suggests that the false convergence is caused by an extremely low variance of the random slopes for female interviewers by word. This random slope was removed, and the model was refit to the trimmed data set.

Figure 4.3 shows the model criticism plots of the trimmed model with the random slope for interviewer gender removed. This figure looks remarkably similar to Figure 4.1, suggesting that the model that was fit to the trimmed data suffered prediction problems similar to those found in the model fit to the full data set.

Table 4.10 compares the standard deviations of the random-effects in the untrimmed and trimmed models. Table 4.10 shows no significant qualitative difference between the random effects structures of the untrimmed and trimmed models. The largest change appears in the random slopes by word for each gender of speaker. Qualitatively speaking, however, both models agree that these random slopes are among the most variable random effects in the model. The random slope for verbs by speaker also nearly doubles in variation in the trimmed model, but both models agree that the variation for this random slope is relatively small.

Table 4.11 compares the effect sizes and t-values of the fixed-effects in the pre-trimming and post-trimming models. Three effects in Table 4.11 show qualitative differences in the model that was fit to the trimmed data. All three involve forward part-of-speech-based predictability: The main effect of the predictor, along with both of its interactive effects (with the number of stressed syllables, and with forward word-based predictability), are significant predictors of deletion in the trimmed model, but fail to reach significance in the untrimmed model.

As Table 4.5 shows, all three of these effects were highly significant before random slopes were added to the model, and Table 4.7 shows that by-word slopes for forward POS predictability show a fair amount of variance. This suggests that these by-word random

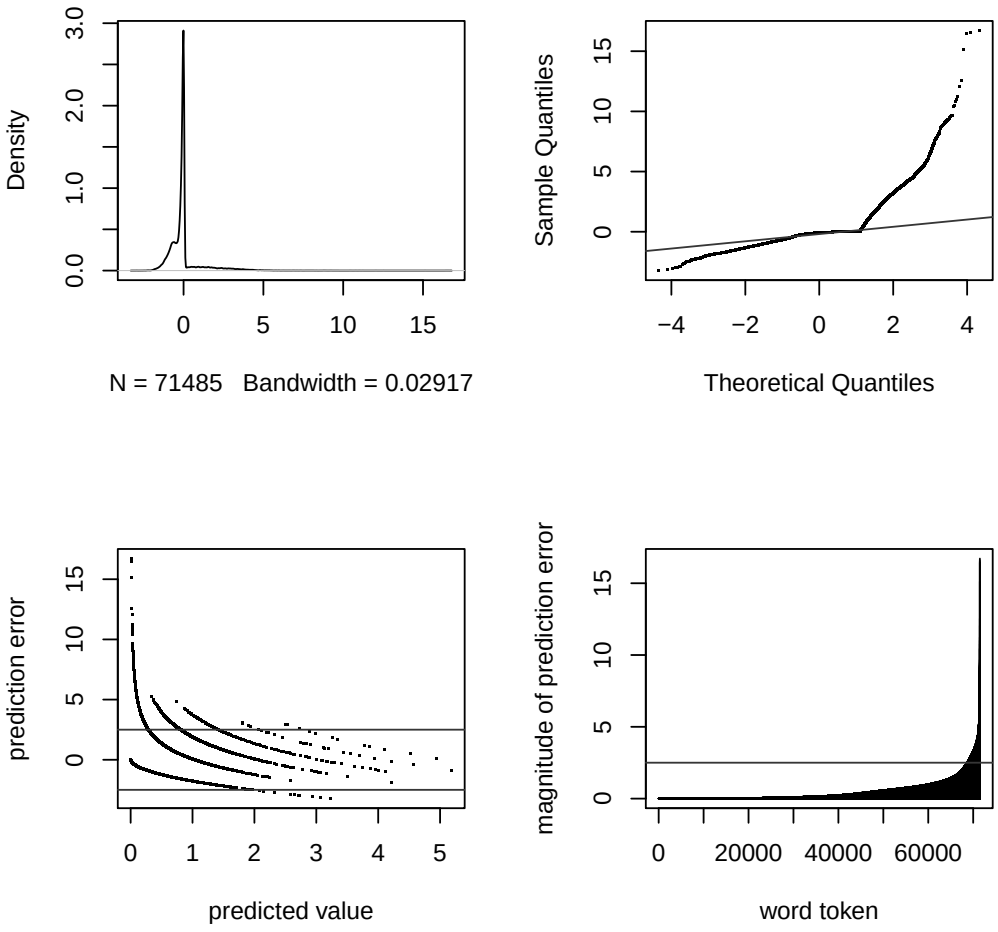

Figure 4.3: Model Criticism Plots

Groups	Predictor	S.dev.1	S.dev.2	S.dev.chg	% chg.
Speaker	(Intercept)	0.201	0.187	-0.013	-7
Word	(Intercept)	0.677	0.665	-0.013	-2
Word	Age (> 40)	0.423	0.723	0.299	71
	Age (< 30)	0.265	0.520	0.255	96
Speaker	Time	0.101	0.119	0.018	18
Speaker	Adjectives	0.229	0.273	0.044	19
	Adverbs	0.161	0.246	0.085	53
	Nouns	0.195	0.296	0.101	52
	Verbs	0.087	0.197	0.110	126
Word	Female Interviewer	0.107	NA	NA	NA
	Male Interviewer	0.182	NA	NA	NA
Speaker	COCA Frequency	0.075	0.135	0.061	81
Word	Forwd wd pred.	0.143	0.128	-0.014	-10
Word	Forwd POS pred.	0.137	0.148	0.011	8
Speaker	Length	0.072	0.107	0.036	50
Word	tf-idf topicality	0.078	0.091	0.013	17
Speaker	tf-idf topicality	0.029	0.043	0.014	49
Word	Female Speaker	0.795	1.932	1.137	143
	Male Speaker	0.500	1.203	0.703	141
Word	Global spk. rate	0.106	0.172	0.066	63
Word	Phrase spk. rate	0.028	0.034	0.006	24

Table 4.10: Random Effects Before (1) and After (2) Trimming

slopes mediate the undue influence on model fit caused by the extreme data points.

In fact, an argument can be made that *any* by-word random slopes are likely to produce over-fitting in models of the current data set. Six of the seven predictors whose significance level dropped noticeably in Table 4.11 contain by-word random slopes in the full model. The potential for over-fitting derives from the large number of levels of the random effect: There are a total of 3,015 unique wordforms among the 74,094 data points modelled here. Adding by-word random slopes thus allows a predictor to have 3,015 different possible relationships to the dependent variable.

A partial test of this argument was performed by removing by-word random slopes for forward POS predictability from the model parameters and re-fitting the full data set using LMER. The results of this process are shown in Table 4.12. As in the trimmed model, the three effects related to forward POS predictability all reach significance. In fact, all three have higher p-values in the trimmed model than they do in the model with by-word slopes removed.

Avoiding all by-word random slopes would require extensive re-modelling, and is beyond the scope of the present study. Instead, the full model fit to the trimmed data set is taken as the final result of this stage of model selection. (Note that this means that by-word random slopes for interviewer gender are excluded from the final model, for the reasons described above.)

Description	Est.1	Est.2	rank	p.val.1	p.val.2	p.val.chg
(Intercept)	-2.650	-4.503	0	0.000	0.000	0.000
Male Speaker	0.355	1.092	+1	0.000	0.000	0.000
Length of word (res.)	0.627	1.021	-1	0.000	0.000	0.000
Local frequency (res.)	0.168	0.252	+1	0.000	0.000	0.000
Verbs	0.188	0.224	-1	0.021	0.047	0.026
Number of stressed syllables	-0.124	-0.220	0	0.000	0.000	-0.000
Male Interviewer	0.062	0.166	+6	0.426	0.054	-0.373
Global speaking rate	0.118	0.138	0	0.002	0.004	0.003
Backwards word predictability	0.107	0.134	0	0.000	0.000	0.000
Speaking rate for this phrase	0.088	0.115	0	0.000	0.000	0.000
Nouns	-0.076	-0.112	0	0.297	0.276	-0.021
Forward word predictability	0.063	0.076	0	0.000	0.000	0.000
Adverbs	0.120	0.055	-6	0.174	0.616	0.443
Forward POS predictability	0.014	0.052	+4	0.472	0.027	-0.444
Forwd POS pred. x Stress	-0.030	-0.043	0	0.080	0.035	-0.045
tf-idf topicality x Male Speaker	0.038	0.039	-2	0.105	0.200	0.095
Forwd wd pred. x Forwd POS pred.	0.021	0.035	0	0.070	0.013	-0.057
Surrounding POS predictability (res.)	0.024	0.033	-2	0.004	0.000	-0.004
tf-idf topicality (res.)	0.002	0.015	0	0.940	0.571	-0.369

Table 4.11: Fixed-effects Before (1) and After (2) Trimming

Description	Est.1	Est.2	rank	p.val.1	p.val.2	p.val.chg
(Intercept)	-2.650	-2.620	0	0.000	0.000	0.000
Length of word (res.)	0.627	0.626	0	0.000	0.000	0.000
Male Speaker	0.355	0.359	0	0.000	0.000	0.000
Local frequency (res.)	0.168	0.167	+1	0.000	0.000	0.000
Verbs	0.188	0.149	-1	0.021	0.062	0.041
Number of stressed syllables	-0.124	-0.123	0	0.000	0.000	0.000
Global speaking rate	0.118	0.118	+1	0.002	0.002	0.000
Nouns	-0.076	-0.111	+3	0.297	0.123	-0.174
Backwards word predictability	0.107	0.108	0	0.000	0.000	0.000
Adverbs	0.120	0.105	-3	0.174	0.230	0.056
Speaking rate for this phrase	0.088	0.090	-1	0.000	0.000	0.000
Male Interviewer	0.062	0.067	+1	0.426	0.396	-0.030
Forward word predictability	0.063	0.065	-1	0.000	0.000	0.000
Forward POS predictability	0.014	0.044	+4	0.472	0.001	-0.471
tf-idf topicality x Male Speaker	0.038	0.038	-1	0.105	0.099	-0.006
Forwd POS pred. x Stress	-0.030	-0.038	-1	0.080	0.013	-0.067
Forwd wd pred. x Forwd POS pred.	0.021	0.034	0	0.070	0.002	-0.068
Surrounding POS predictability (res.)	0.024	0.026	-2	0.004	0.002	-0.003
tf-idf topicality (res.)	0.002	0.001	0	0.940	0.975	0.035

Table 4.12: Fixed-effecs With (1) and Without (2) Forward POS Predictability Slopes. Both models fitted to untrimmed data set.

4.2.6 Results and Discussion

4.2.6.1 Main Effects

Description	Effect	Std.Err.	z.value	Pr(>|z|)
(Intercept)	-4.503	0.147	-30.7	0.000
Male Speaker	1.092	0.106	10.3	0.000
Length of word (res.)	1.021	0.042	24.3	0.000
Local frequency (res.)	0.252	0.044	5.8	0.000
Verbs	0.224	0.113	2.0	0.047
Number of stressed syllables	-0.220	0.048	-4.6	0.000
Male Interviewer	0.166	0.086	1.9	0.054
Global speaking rate	0.138	0.048	2.9	0.004
Backwards word predictability	0.134	0.013	10.2	0.000
Speaking rate for this phrase	0.115	0.013	8.7	0.000
Nouns	-0.112	0.102	-1.1	0.276
Forward word predictability	0.076	0.018	4.3	0.000
Adverbs	0.055	0.110	0.5	0.616
Forward POS predictability	0.052	0.024	2.2	0.027
Forwd POS pred. x Stress	-0.043	0.020	-2.1	0.035
tf-idf topicality x Male Speaker	0.039	0.030	1.3	0.200
Forwd wd pred. x Forwd POS pred.	0.035	0.014	2.5	0.013
Surrounding POS predictability (res.)	0.033	0.010	3.5	0.000
tf-idf topicality (res.)	0.015	0.026	0.6	0.571

Table 4.13: Fixed-effects Structure of Final Model

The main effects for the final model are included in Table 4.13, ordered by decreasing effect size.

Residualized word length is the most highly significant predictor of deletion in the model, though its effect size is lower than that of speaker gender. Longer words are more likely to be produced with deleted segments than shorter words. This result is unsurprising: Longer words contain more information, and should thus be more able to withstand deletion while still providing enough identifying information to the listener.

Male speakers were also found to be more likely than female speakers to delete segments. This result was the strongest in the current model, with an effect size of 1.092 (p-value: 7.1e-25).

Words with higher relative frequencies in the Buckeye Corpus than in the COCA corpus also showed a greater tendency towards deletion. The fact that frequency leads to deletion is unsurprising. More surprising is the lack of a significant main effect of COCA frequency on deletion. As in the duration reduction models, COCA frequency alone does not appear to relate to deletion in a straightforward way. This general frequency measure does contribute to the model, however: Table 4.14 shows that different speakers react differently to COCA frequency. Moreover, in Section 4.3.2 below, the results of LMER and random forest models are again combined to search for overlooked interactions. During this process, COCA frequency is shown to contribute to model fit as a member of two significant interactions, as illustrated in Figure 4.14.

Part of speech effects also reach significance in the final model, though only verbs showed significant differences in reduction when compared to adjectives. The lack of significant

effects for nouns and adverbs may be the result of the choice of adjectives as the reference level against which they are compared. (Recall that adjectives were chosen as the reference level by default, since they fall first among parts of speech in alphabetical order.) In the final model, verbs and adverbs show a tendency towards more deletions, while nouns show a (non-significant) tendency towards fewer deletions. Adjectives, then, fall near the centre of the range of part of speech effects. These trends suggest, at least, that nouns are less likely to undergo deletion. The trend may be seen as implying that nouns are more important, in some informational sense, than verbs. Or rather, that the importance or difficulty of identifying a noun correctly tends to preclude segment deletion.

The number of stressed syllables in a word is one of the few measures under consideration that inhibits deletion. Words with more stressed syllables are strongly (Effect size: -0.220) and very significantly ($p = 3.7e-06$) less likely to undergo segment deletion than words with fewer stressed syllables. By contrast, recall that the number of stressed syllables in a word had no effect on duration reduction in the previous chapter. The present result is intuitive, suggesting that segments are less likely to be deleted from stressed syllables than they are from unstressed syllables.

Faster global and local speaking rates both lead to a greater number of deletions, with global speaking rate having the stronger effect. This suggests that faster speakers tend to delete more segments, and that all speakers tend to delete more segments when speaking quickly.

Several predictability measures also lead to increased deletion. Predictability given a word's preceding or following word both show significant effects on deletion. Predictability given a word's preceding or surrounding parts of speech also show significant increases in deletion rates. These results imply that increased predictability, of word or of structure, allows for less phonetic information to be transmitted to the listener.

4.2.6.2 Interactive Effects

The interactions in this model are included in Table 4.13, and plotted below in Figures 4.4 and 4.5.

Figure 4.4 shows an interaction between topicality and speaker gender. Overall, topicality appears to have no significant effect on deletion rates ($p = 0.57$). Figure 4.4, however, suggests that topicality does lead to an increase in deletions, but this reductionary effect is more pronounced for male speakers than it is for female speakers.

Figure 4.5 shows two interactions with forward POS predictability. The left panel of the figure shows POS predictability shifting from inhibiting deletion to facilitating deletion depending on how predictable the target word is from the preceding context. Words with high word predictabilities tend to show more deletions as their POS predictabilities increase. This suggests that the two types of predictability are collaborating, and that an increase in both leads to an increase in deletions.

Words with low word predictability, however, show fewer deletions as they become more syntactically predictable. The models in the previous chapter showed a similar interaction between these two variables, with less predictable words becoming less reduced as their POS predictability increased. In the previous chapter, this was explained in terms of a

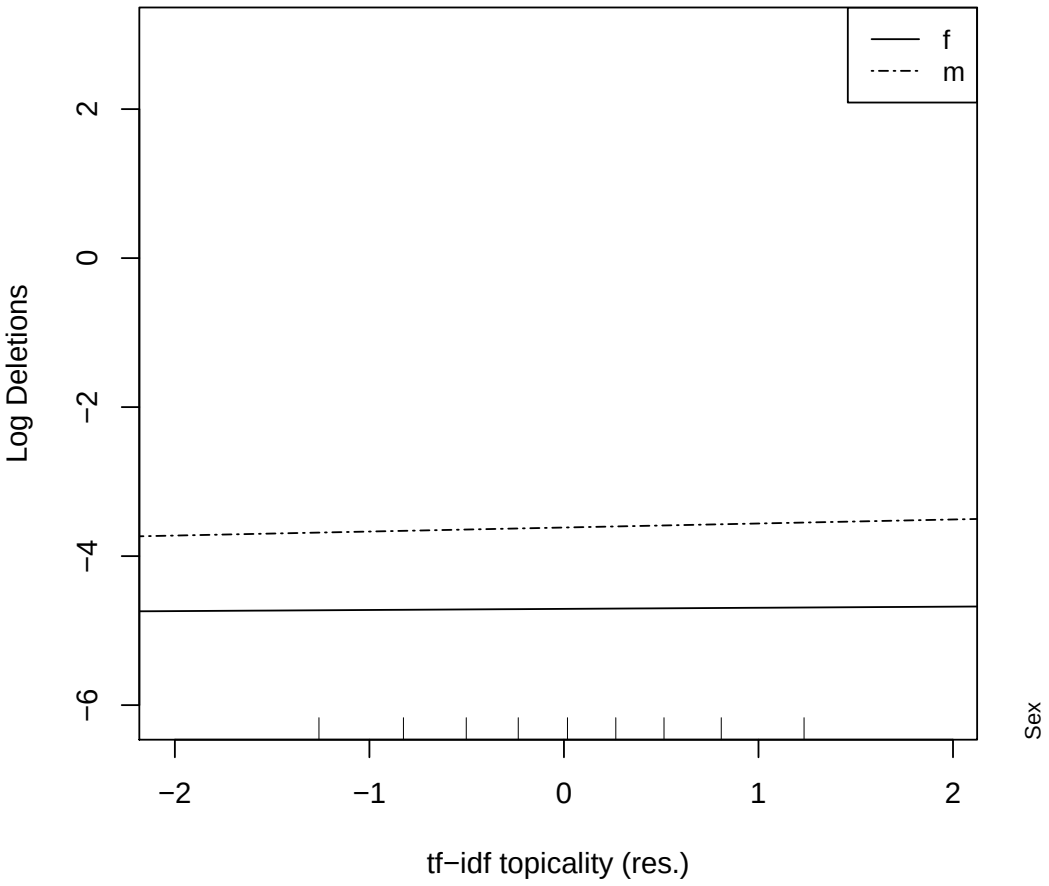

Figure 4.4: Interaction between Speaker Gender and Topicality

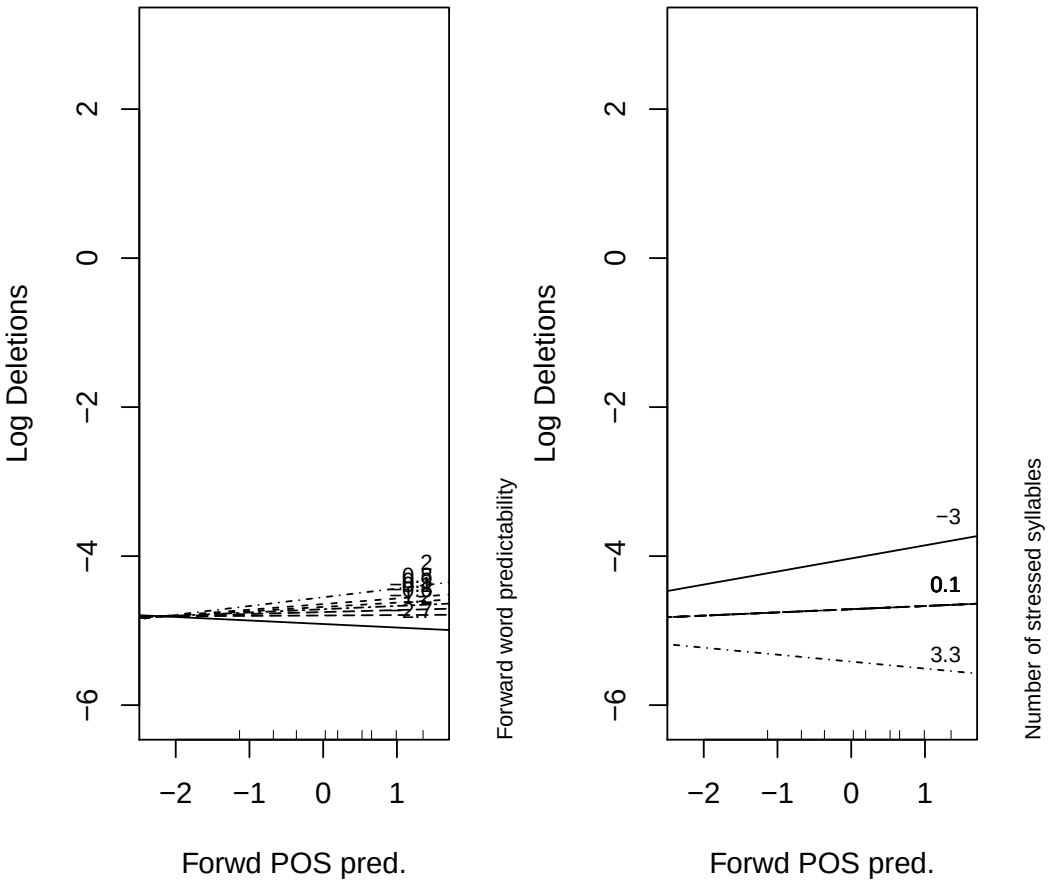

Figure 4.5: Forward Part-of-Speech-Based Predictability Interactions

prediction mismatch effect: If a less predictable word appears in a more routinized syntactic context, deletion is less likely to occur in that word.. It is also possible that words with high-structural but low-lexical predictability are more likely to be infrequent words. If this is the case, the present interaction could represent a simple inhibitory effect of lower lexical frequency.

The right panel of Figure 4.5 illustrates another interaction with POS predictability. In this case, words with many stressed syllables show less deletion when syntactically predictable, while words with few stressed syllables show more deletions when syntactically predictable. This interaction only appears for the most extreme numbers of stressed syllables, however. The central 5 out of 7 quantiles overlap in the figure, suggesting no difference in the way words with moderate numbers of syllables undergo deletion as predictability increases. In particular, words with an average number of stressed syllables appear to show a facilitatory effect of forward POS predictability on deletion: More predictable words are more likely to have deleted segments. The more extreme numbers of stressed syllables show a collaboration between predictability and unstressed syllables: Predictable words and words with few stressed syllables are shown in Table 4.13 to be more likely to undergo segment deletion. The panel on the right of Figure 4.5 shows that these two measures combine to increase (or decrease) the likelihood of segment deletion more than would be predicted by either measure alone.

4.2.6.3 Random Effects

Groups	Predictor	Std.Dev.	Corr		
Word	(Intercept)	0.665			
Speaker	(Intercept)	0.187			
Word	Age (> 40)	0.723			
	Age (< 30)	0.520	1.00		
Speaker	Time	0.119			
Speaker	Adjectives	0.273			
	Adverbs	0.246	0.72		
	Nouns	0.296	0.93	0.81	
	Verbs	0.197	0.83	0.51	0.90
Speaker	COCA Frequency	0.135			
Word	Forwd wd pred.	0.128			
Word	Forwd POS pred.	0.148			
Speaker	Length	0.107			
Speaker	tf-idf topicality	0.043			
Word	tf-idf topicality	0.091			
Word	Female Speaker	1.932			
	Male Speaker	1.203	1.00		
Word	Global spk. rate	0.172			
Word	Phrase spk. rate	0.034			

Table 4.14: Random Effects Structure of Final Model

The random effects structure of the final model is summarized in Table 4.14. As in the previous chapter, the random effects structure reveals that some factors with no main-effect in the model nevertheless have an effect on deletion rates for some words and speakers. In

some cases, the effects are dramatic. Age group, for example, was not found to have a significant main effect on segment deletion. Random slopes for each age group by word, however, show very high variability in the way each age group reacts to particular words.

Speaker gender proves to be a strong contributor to both the fixed and random effects structure of the model. Indeed, the amount of variability that female speakers show in deletion rates by word exceeds the amount of variability in any of the remaining random effects. Moreover, the amount of variability exceeds the effect size of all of the fixed-effects predictors in the model.

Global and local speaking rate also vary by wordform in their effect on deletion. This suggests that speakers choose different words to reduce when speaking quickly, or that the number of deletions in a quickly-produced word depends in some way on the identity of that word. The effect also suggests that fast and slow speakers differ in the words from which they tend to delete segments.

The random effects structure described in Table 4.14 also shows that some predictors lead to different rates of deletion depending on the speaker under study. Some of these predictors (Part of speech, citation length, and topicality) also play a role in the fixed-effects structure of the model, though some are only predictive of deletion as members of interactions. Others played no role in the fixed-effects structure of the current model: COCA frequency has no direct connection to deletion rates in the current model, for example. Its presence in the random-effects structure suggests that COCA frequency does have some effect on deletion, but that this effect varies by speaker. Similarly, the time at which a word appears in a conversation has no fixed effect on deletion rates in the current model. The random slope for time by speaker suggests that different speakers respond differently to how far they have proceeded in a conversation.

As with random effects on duration, further investigation into the variation in deletion rates within these grouping variables would be a profitable area of future study.

4.2.6.4 Comparing Random and Fixed Effects

The final linear model can again be compared to LMER models generated with only fixed-effect, or only random-effect, predictors. The model with only random effects is extremely successful, predicting 38% of the variation in deletions in the model. That is, a model with only random effects predicts deletion *exactly as well* as the final mixed effects model. The model with only fixed effects performs poorly, predicting only 8.8% of the variation in deletion counts in the model.

Taken together, these results imply that much more of the variability in deletion rates is due to variation in word and speaker behaviour than is due to the particular predictors under study. A model with several main and interactive effects offers a relatively poor prediction of the deletion rates found in the data. Models that allow each word and speaker to behave differently in response to several predictors, on the other hand, predict the variation in deletion rates quite well. Moreover, this prediction accuracy is strong *whether or not the more direct effects of those predictors are included in the model.* This fact should be interpreted with caution however. It may indicate further support for the idea that allowing slopes to vary over a large number of values of a factor (wordforms, in the present study,)

leads to over fitting, and thus to overconfidence in the ability to predict reduction.

Alternatively, this result may be seen as the result of the present full-corpus approach. The present study attempts to model deletion in as many different words as possible. This necessarily entails greater variation in the character of the words under study than an investigation of a more limited set of words. Still, these results show that the predictive power of the fixed-effects structure is impoverished. This suggests that some sources of variability have been overlooked, and additional predictors should be included in future models of segment deletion. This suggestion is revisited in Section 4.3.1 below.

4.3 Random Forest Modelling

The random forest modelling of deletions is essentially identical to the random forest modelling of duration change in the previous chapter. Since random forest modelling is a non-parametric modelling technique, no assumption of normality in the dependent variable is made. As a result, no change is required to indicate that deletions should be modelled as the result of a Poisson process. Deletions are simply treated as an ordered variable, and modelled using piece-wise-constant regression as duration change was in the previous chapter. Tuning for this forest led to an *mtry* value of 6, meaning that 6 predictors are randomly selected to construct each tree in the forest.

4.3.1 Results and Discussion

4.3.1.1 Proportion of Variance Explained

The forest predicted 31.5% of the variance in deletions. This under performs all but one of the linear models considered above. Even the baseline model, with main effects and random intercepts for wordform and speaker only, predicts 33.2% of the variance in deletion rates found in the corpus.

The LMER model that corresponds best to the random forest model, however, predicted only 8.8% of the variation in the data. In the context of the analysis in Section 4.2.6.4, this result is reassuring. Random forests are unable to account for variation by word or speaker in the current study. In Section 4.2.6.4 it is shown that random slopes and intercepts for these variables can be used to predict deletion with or without the addition of any fixed-effect information. This result suggests that the relationship between the predictors under study and deletion rates is not strong, or at least not direct.

When the same set of predictors was included in a random forest, rather than LMER model, however, they were able to predict a much larger proportion of the variance in the data. The random forest, then, suggests that these predictors do indeed have a sizable effect on deletion rates. The reduced explanatory power of these predictors as fixed effects in the LMER models is likely due to the present modelling procedure. Recall that there are two sources of variation, in principle, that random forests can account for but the LMER model selection process used here does not. First, LMER main effects in the present study are all linearly (or log-linearly) related to the dependent variable in the model. The random forest assumes no such linear relationship, and predicts the dependent variable as a piece-wise-continuous function of each predictor. Second, the interactions between predictors in a

random forest can be arbitrarily complex, while the number and complexity of interactions considered in an LMER model is limited by current computing power.

It is likely that both of these sources of variation explain the jump in proportion of variance explained by the random forest model. Non-linear relationships between predictors and deletion are indeed present in the random forest model, as shown by Figure 4.12 and discussed in Section 4.3.1.3 below. Interactions that are overlooked by the current LMER model are again found after consulting the random forest model, as described in Section 4.3.2 below.

4.3.1.2 Variable Importance Measures

The measures of variable importance for each model are displayed in Figure 4.6. (Importance measures are scaled and centered to allow for visual comparison.)

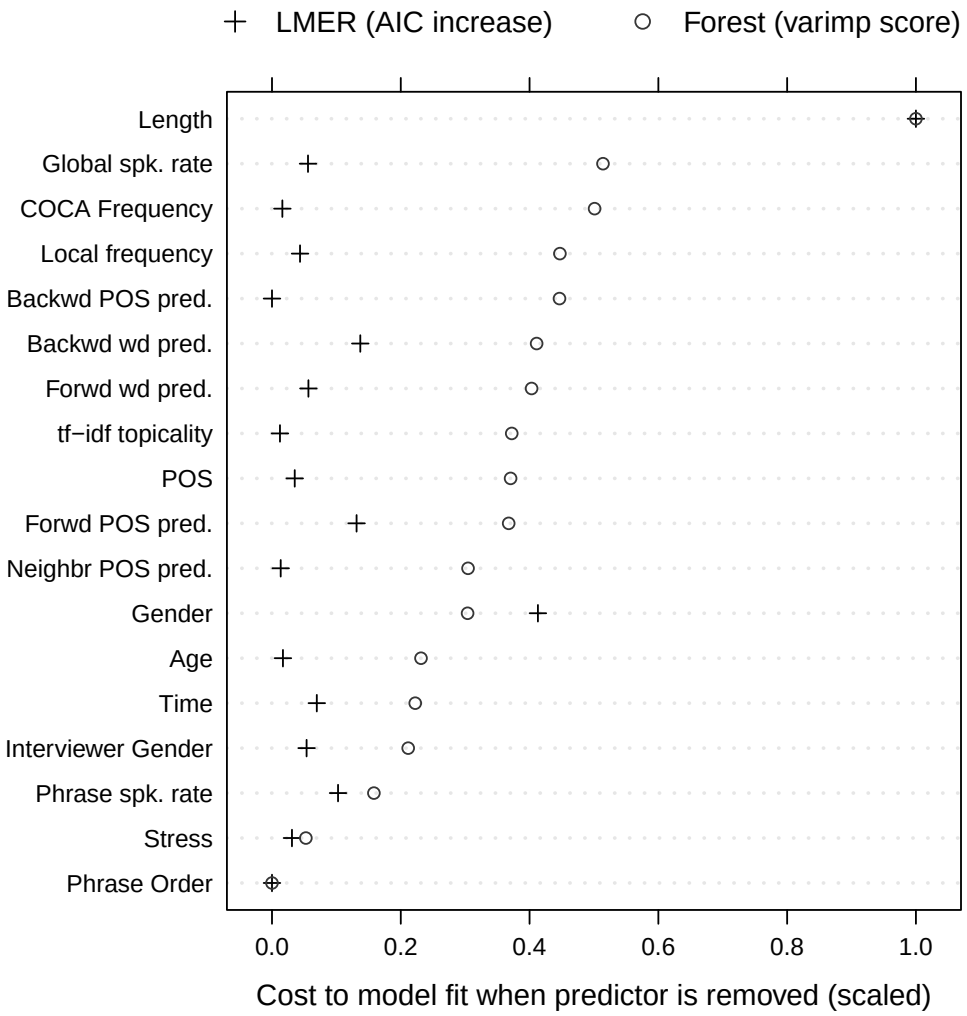

Figure 4.6: Comparison of Variable Importance across Model Types

As in the previous chapter, variable importance in the final LMER model was calculated by removing all main, interactive, and random effects of a predictor and measuring the

change in AIC score. A variable's importance is thus a measure of how much its presence in the model improves AIC score.

The importance of one predictor (word length) dominates the plot, masking variability among the remaining predictors. A second plot, with word length removed and the remaining importance values re-scaled, is shown in Figure 4.7.

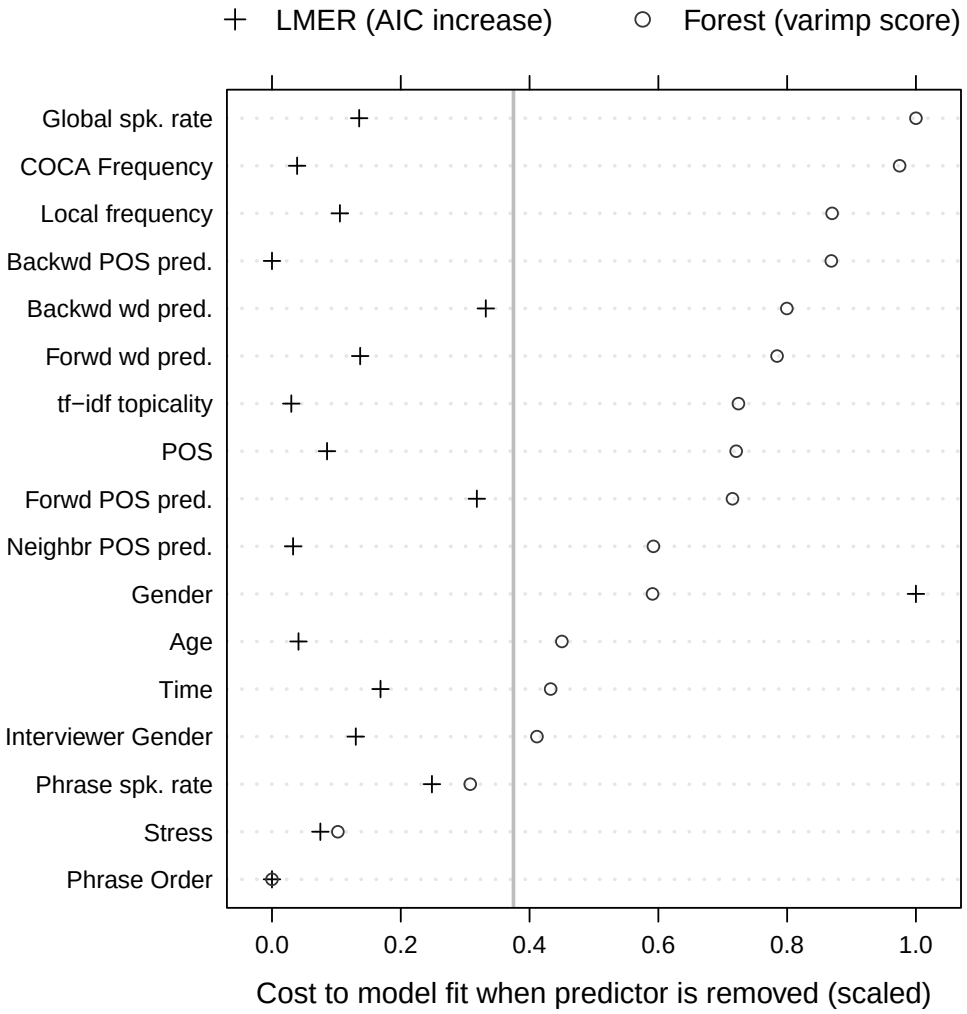

Figure 4.7: Comparison of Variable Importance across Model Types - all Values Re-scaled

Figure 4.7 again reveals predictors whose importance is undervalued in the LMER model due to early exclusion from interaction modelling. In particular, COCA frequency (the second strongest remaining predictor in the forest model), backwards part-of-speech-based predictability, and speaker age group appear much more important to the random forest than they do to the linear regression model. All three of these predictors were eliminated from the LMER model before interaction testing began.

Two additional predictors were removed from LMER model selection before interaction testing: Each of these predictors measures a word's position in the conversation, one marking its ordinal position in the phrase and one marking the time in the conversation at which

the word was uttered. Both models agree that a word's position in its phrase (for phrase-medial words) is not a useful predictor of deletion. The LMER model appears to consider time of utterance within a conversation as a highly useful predictor, as illustrated in Figure 4.7. This can only be due to the importance of time-of-utterance to the random-effects structure of the model, as described in Table 4.14, since time-of-utterance is excluded from the fixed-effects portion of the model.

Unlike the results found in the previous chapter, the random forest model puts a low importance value on some predictors that the LMER finds highly useful. At first glance, the most natural discontinuity at which to cut the variables in the random forest model falls between interviewer gender and local speaking rate, indicated by the vertical line in Figure 4.7. Three predictor variables fall below this continuity. Two of these variables, the number of stressed syllables and the local speaking rate, showed sufficient importance during LMER model selection to remain through the interaction-selection procedure. Indeed, local speaking rate has one of the strongest significance levels ($p \approx 3e-18$) among the predictors of deletion in the model. It is unclear why this variable should have so low an importance score in the random forest model. Similarly, the number of stressed syllables in a word was a strong (-0.220) and significant ($p = 0.0000$) predictor of deletion rates in the LMER model, while ranking second-lowest in importance in the random forest model.

These contradictory results raise an important question about Random Forest variable importance scores. Namely, how high an importance score is required before a variable is considered important to the model. Existing studies that use random forest models (e.g., Strobl *et al.* (2009)) assume that variables with no true effect on the dependent variable will be distributed randomly around an absolute variable importance of zero. Using this definition, "unimportant" variables can be defined as those that fall closer to zero than the most negative importance score. This approach proved impossible in the present study, however: Due to the large number of data points, no variables received negative importance scores. In the previous chapter, a natural discontinuity appeared between variables that were important and variables that were unimportant and this discontinuity matched the results of the LMER modelling process. In the present chapter, however, a more careful choice of discontinuity must be made. While there is a large difference between the importance of interviewer gender and local speaking rate, there is also a large difference between the importance of local speaking rate and stress. Using either discontinuity as the basis for a threshold, however, would prevent stress from being considered important, a conclusion that the LMER model strongly disputes.

Worse still, the results of the present study suggest that absolute importance scores can not be used to determine an appropriate importance threshold. The absolute random forest importance scores for speaking rate and stress are smaller than the absolute importance scores for several variables that were considered *un*important in the random forest model of duration in the previous chapter.

The process of combining random forest and LMER modelling results, then, requires some modification when compared to the process applied in the previous chapter. The new process is described in Section 4.3.2 below.

4.3.1.3 Comparison of Partial Effects

In this section, the partial effects of each predictor are compared across modelling techniques. There are again a few broad categories into which these partial-effects comparisons fall.

First, there are predictors whose behaviour is largely agreed upon by the two models. Each part of speech, for example, has the same relative effect on deletion rates in both models, as illustrated in Figure 4.8. Both models also agree that men show a higher chance of deleting a segment than women. The remaining predictors agreed on by both techniques

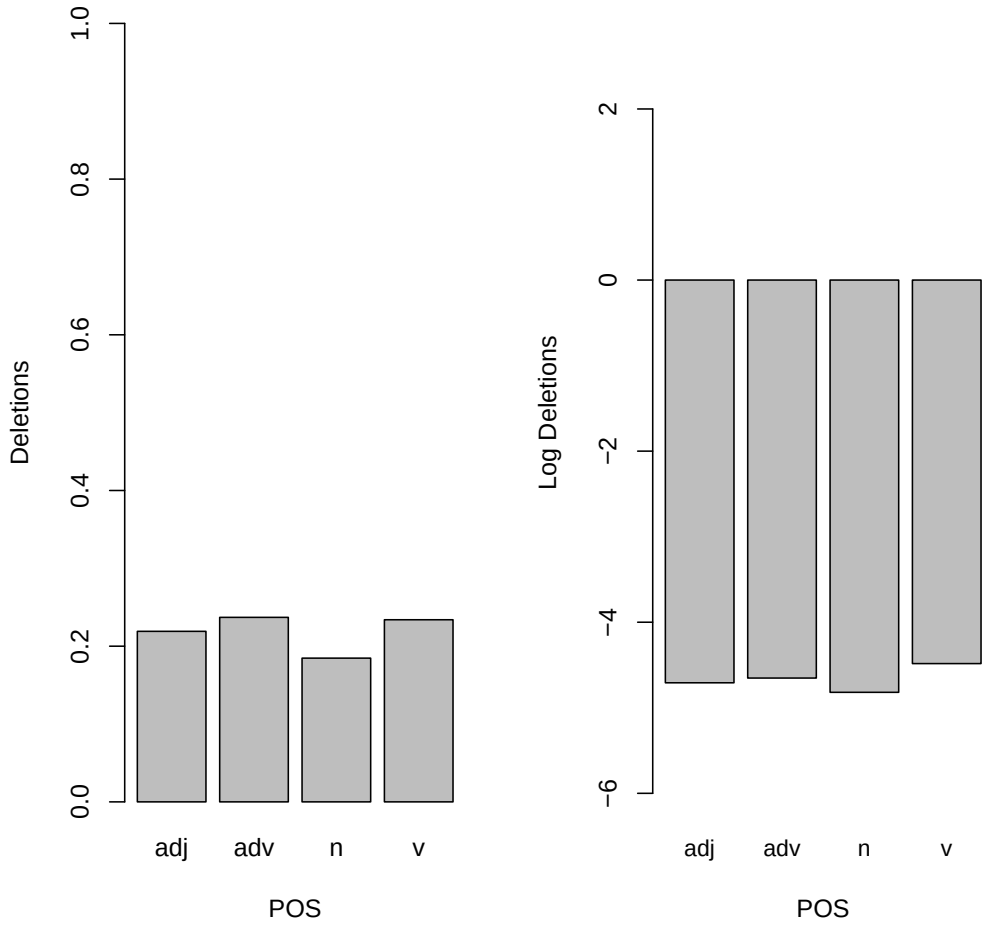

Figure 4.8: Partial Effect of Part of Speech in Random Forest (L) and LMER (R) Models

are those for which the predictor has little effect of its own on deletion rates. Tf-idf-based topicality shows relatively little effect on deletion in both models, as illustrated in Figure 4.9. Other predictors with only weak partial effects in both models include Interviewer Gender and Forward part-of-speech predictability.

A second category of partial-effects comparisons is made up of predictors for which the

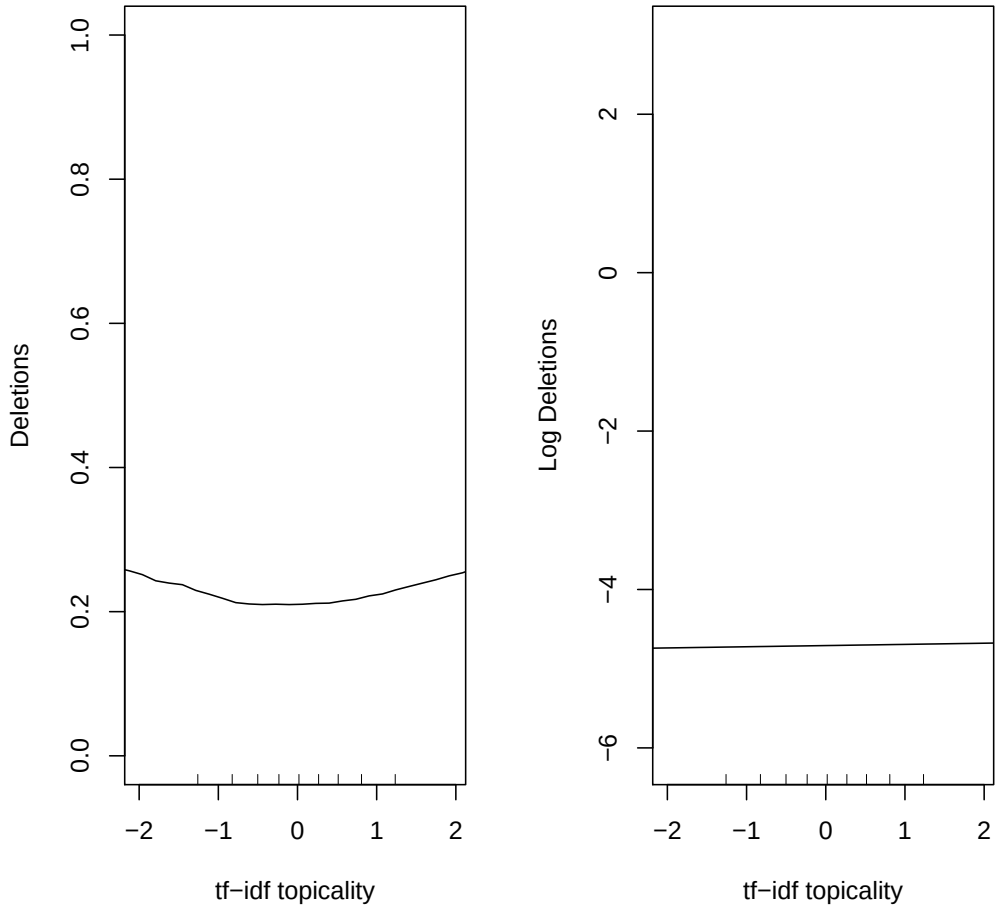

Figure 4.9: Partial Effect of (Residualized) TF-IDF topicality in Random Forest (L) and LMER (R) Models

random forest model finds little effect, while the LMER model finds significant facilitation of deletion. Figure 4.10 illustrates one member of this category, (residualized) local frequency. Despite the flat partial effect of local frequency, the random forest model assigns a relatively

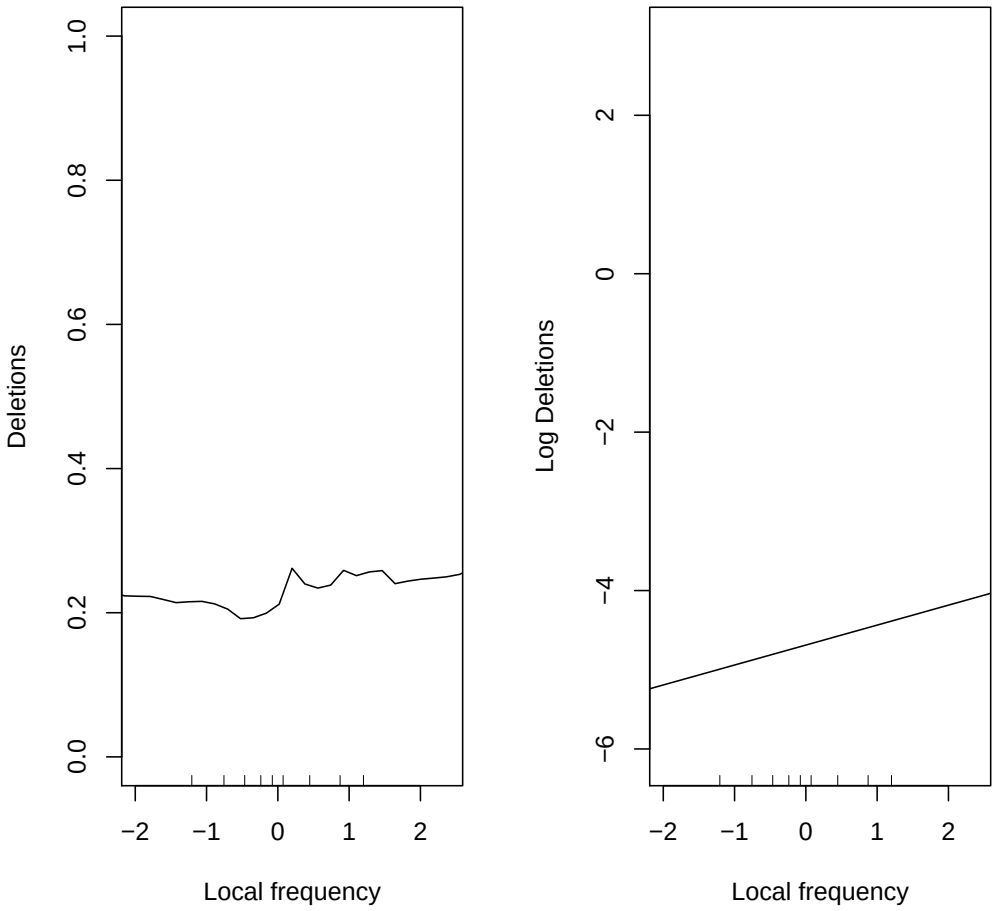

Figure 4.10: Partial Effect of Local Frequency in Random Forest (L) and LMER (R) Models

high importance score to the variable. Not all predictors in this category fit this description. Local speaking rate, one of the strongest predictors of deletion in the LMER model, shows little partial-effect and has one of the lowest variable importance scores in the random forest model. Figure 4.11 illustrates the partial effects for local speaking rate in the two models

The remaining predictors that show little partial effect in the forest model but reductionary partial effects in the LMER model are conditional probability given the previous or following word, and global speaking rate.

The two remaining predictors for which partial effects are available in both models do not fit neatly into either of the above categories. Residualized word citation-form length is clearly facilitatory of deletion in both models. As Figure 4.12 illustrates, however, the two models differ in how they characterize that deletion. In the random forest model the

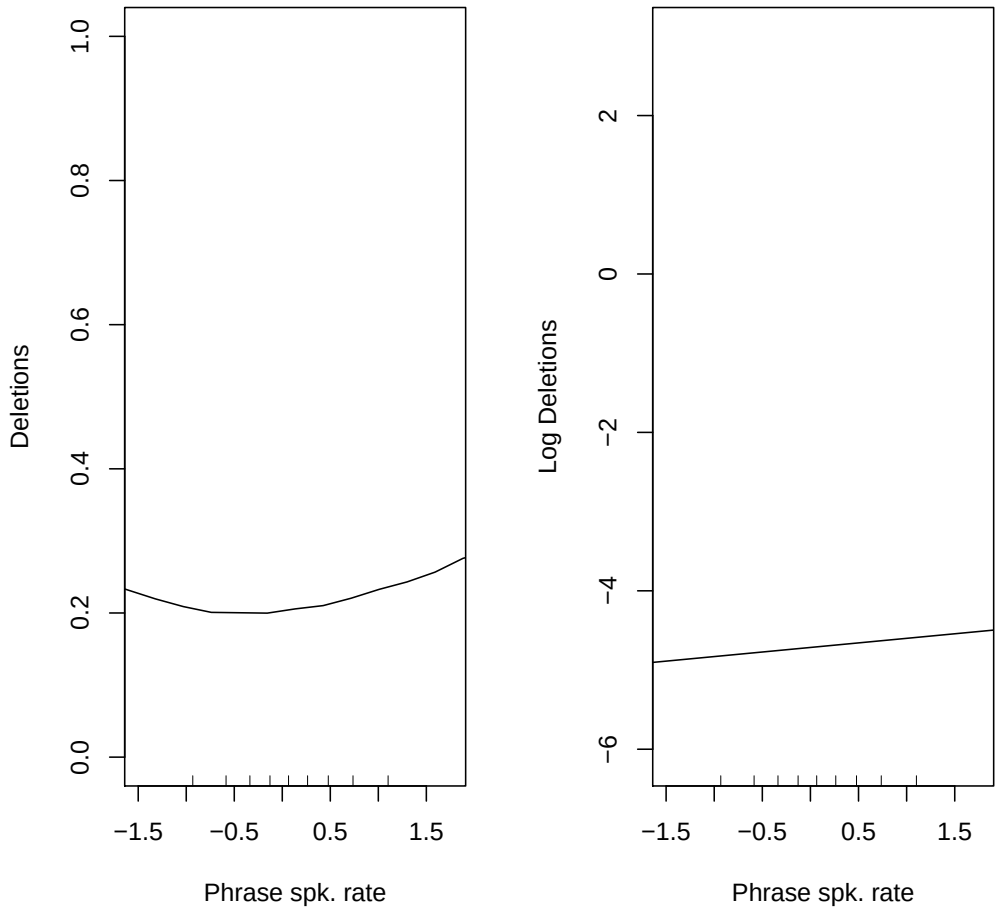

Figure 4.11: Partial Effect of Local Speaking Rate in Random Forest (L) and LMER (R) Models

effect of word length on deletion, displayed in the left panel, increases gradually and non-linearly as word length increases. In the LMER model, as displayed in the right panel, word length appears to have a more dramatic effect on deletion. (This effect is linear, of course, due to the nature of the model.) The LMER partial effect plotted in Figure 4.12 shows

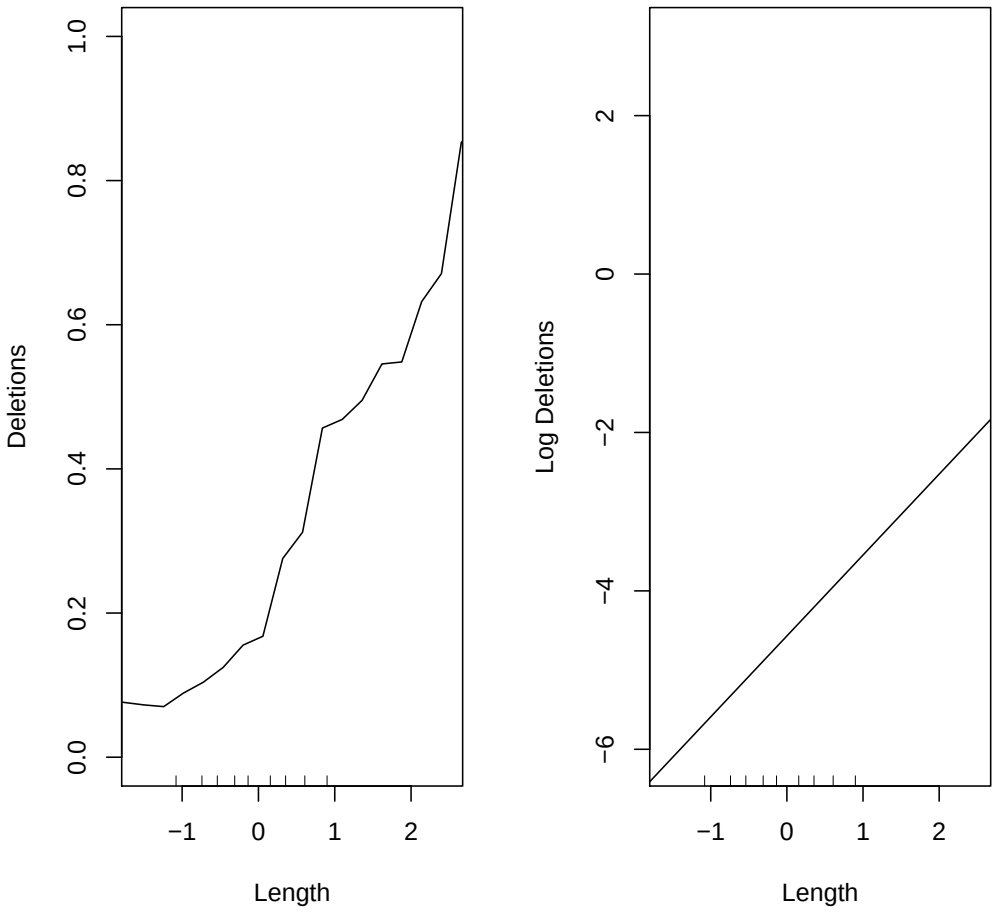

Figure 4.12: Partial Effect of Word Length in Random Forest (L) and LMER (R) Models

deletion rates increasing as word length increases. By contrast, the random forest partial effect illustrates that this effect is stronger for long words than it is for short words. That is, words not only show more deletions as their length increases, they also show an increase in *the effect* of length on deletion as their length increases.

The final predictor under consideration is the number of stressed syllables in the word. This predictor behaves quite differently in the two classes of model. In the LMER model, the number of stressed syllables in a word inhibits deletion, with one of the largest effect sizes and highest significance levels of all the predictors in the model. In the random forest model, however, the number of stressed syllables has the second-lowest variable importance score, suggesting little effect of stressed syllables on deletion rates in the current data. The

partial-effects plot in 4.13 appears to show that an increase in stressed syllables actually leads to an *increase* in deletion in the random forest model. The fact that this predictor has

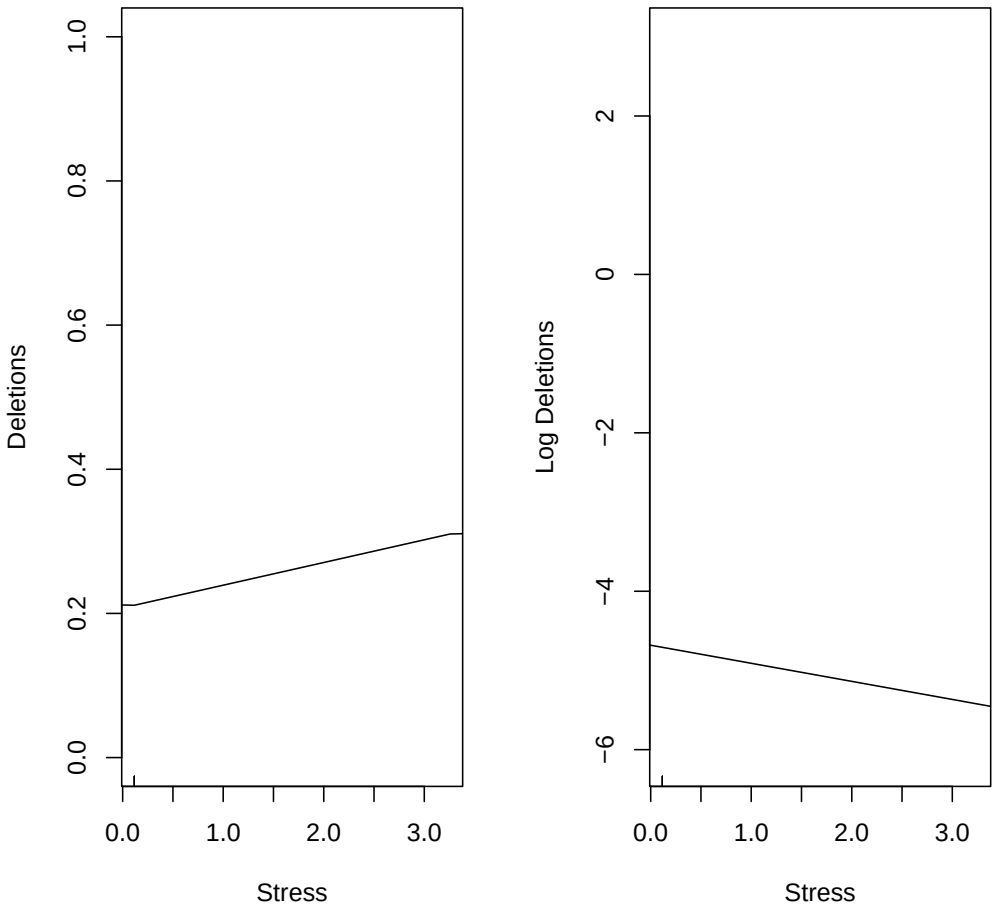

Figure 4.13: Partial Effect of the Number of Stressed Syllables in Random Forest (L) and LMER (R) Models

a low importance score suggests that this slight trend need not be considered with much weight. Still, the two models make qualitatively different predictions about the effect of stress on deletion rates. It is unclear why this should be the case. The result is discussed further in Section 4.4.1.1.2 below.

4.3.2 Combining Modelling Techniques

As mentioned in the previous section, the random forest assigns low importance values to some predictors that LMER modelling finds highly important. As a result, for the present section, any variable found to be useful in predicting deletion *by either model* will be considered. That is, any variable surviving either one of the trimming criteria described above, in LMER modelling or random forest modelling, is included in interaction testing.

As a result, in the modelling below, all of the predictors in the current study except a word's order in its carrier phrase are tested for interactions with each other.

During the initial stage of interaction testing, in which all possible two-way interactions are included in a single model, the model-fitting algorithm again failed to converge. This is unsurprising: The model constructed for interaction testing earlier in the chapter also failed to converge. The first model likely failed to converge due to the large number of variables it was being asked to fit, and the number of variables in the new model has increased. As in the interaction testing above, the final model iteration was used to obtain approximations of the strength of each interaction.

Interestingly, one of the interactions found to be significant during the initial model selection fails to reach significance in the final iteration of the RF-informed model. In fact, the interaction, between forward word predictability and forward POS predictability, is the strongest interaction falling below the significance threshold. The interaction falls just below the previously chosen threshold of $|t| > 3$, with $t = 2.8$. Adding this interaction (and only this interaction) to the set of interactions under study, then, is equivalent to a lowering of the threshold for inclusion from $|t| > 3$ to $|t| > 2.8$, a small change. Due to the strong evidence shown above (see e.g. Table 4.13 and Figure 4.5) that this interaction is a useful component of the model, it was included in the next stage of interaction testing.

In this second interaction testing stage, a set of pairs of models were constructed, one with one of the remaining interactions as a predictor and one without it. Log-likelihood ratio testing was then used to determine whether the interaction contributed significantly to model fit. A total of 10 interactions were submitted to this process.

4.3.2.1 Results and Discussion

The complete fixed-effects structure of the RF-Informed LMER model is shown in Table 4.15.

4.3.2.2 Main Effects

No dramatic qualitative change is found in the random forest-informed model. Forward POS predictability is not significant in the new model, though this is likely due to the use of the full, untrimmed data set in RF-Informed model selection. (Recall that this predictor was non-significant in the previous LMER model before trimming took place.) This result suggests that the trimmed data set should have been used to perform the RF-Informed modelling process.

4.3.2.3 Interactive Effects

All of the interactions illustrated in Figures 4.4 and 4.5 again survive the trimming process. Four new interactions also survived the trimming process. Each of these interactions included one of the variables excluded from interaction testing in the initial model. Two interactions involving (COCA) frequency are illustrated in Figure 4.14. The two remaining interactions, involving the time at which a word appears in the conversation, are illustrated in Figure 4.15.

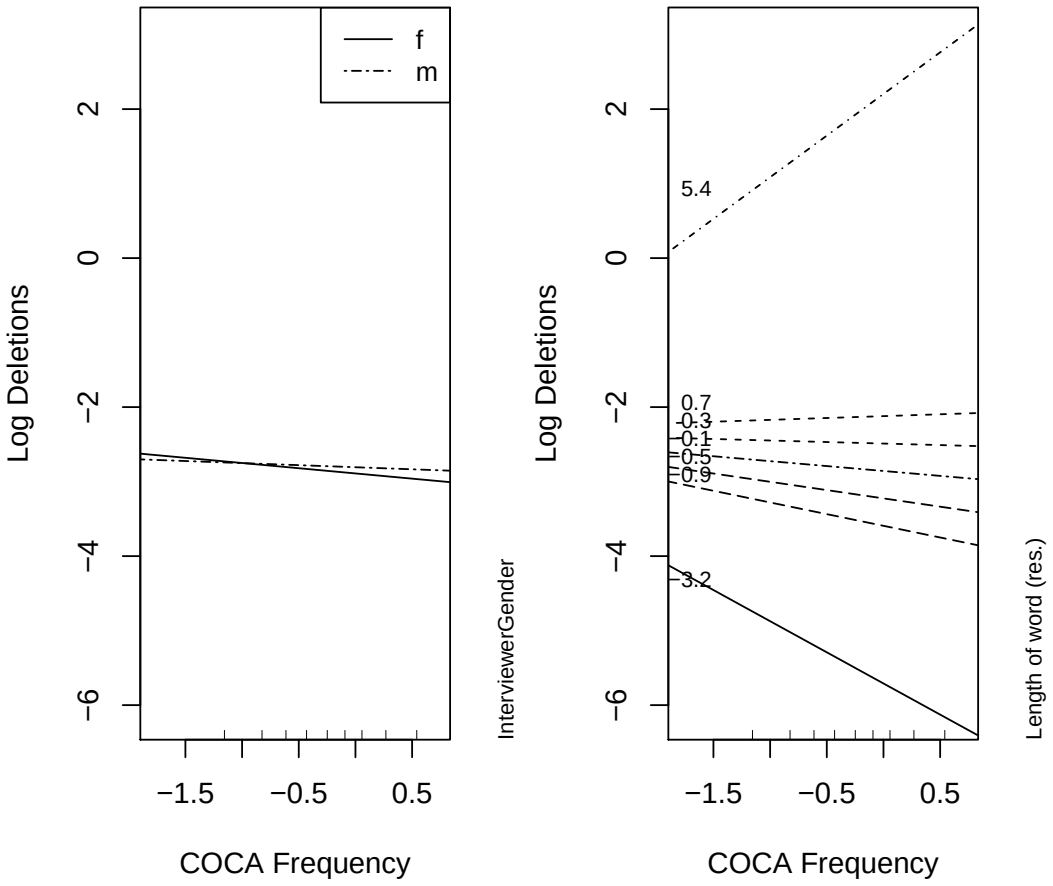

Figure 4.14: COCA Frequency Interactions

Description	Effect	Std.Err.	z.value	Pr(>|z|)
(Intercept)	-2.724	0.125	-21.7	0.000
Length of word (res.)	0.920	0.047	19.7	0.000
Male Speaker	0.354	0.080	4.4	0.000
COCA Frequency x Length	0.227	0.031	7.4	0.000
Verbs	0.200	0.081	2.5	0.014
Local frequency (res.)	0.175	0.028	6.2	0.000
Adverbs	0.123	0.088	1.4	0.163
Global speaking rate	0.117	0.038	3.1	0.002
COCA Frequency	-0.109	0.059	-1.8	0.066
Backwards word predictability	0.107	0.011	9.5	0.000
Nouns	-0.098	0.073	-1.4	0.176
Male Interviewer	0.093	0.078	1.2	0.233
Number of stressed syllables	-0.087	0.032	-2.7	0.006
Speaking rate for this phrase	0.087	0.011	7.8	0.000
Male Interviewer x COCA Frequency	0.084	0.048	1.8	0.076
Time x Male Interviewer	0.063	0.033	1.9	0.054
Forward word predictability	0.060	0.015	4.0	0.000
Time x Global spk. rate	-0.048	0.016	-3.0	0.003
tf-idf topicality x Male Speaker	0.038	0.024	1.6	0.111
Forwd POS pred. x Stress	-0.035	0.017	-2.1	0.039
Surrounding POS predictability (res.)	0.023	0.010	2.3	0.024
Forwd wd pred. x Forwd POS pred.	0.021	0.012	1.8	0.075
Age (under 30)	-0.018	0.077	-0.2	0.812
Time when token appears in conversation	-0.015	0.023	-0.6	0.521
Forward POS predictability	0.014	0.019	0.8	0.449
Backward POS predictability	0.002	0.019	0.1	0.935
tf-idf topicality (res.)	-0.001	0.021	-0.1	0.949

Table 4.15: Fixed-effects Structure of RF-Informed LMER Model

The left panel of Figure 4.14 shows speakers responding differently to word frequency depending on whether they had a male or female interviewer. Both lines in the plot show a decreased likelihood of deleted segments as a word's frequency increases. Indeed, Table 4.15 shows that this frequency effect is stronger when the interviewer is female: Low-frequency words are more likely to show deletion, and high-frequency words are less likely to show deletion, when the interviewer is female. Unlike the inhibitory main effect of COCA frequency, this effect cannot be explained by word length: People speaking to women produced words with the same citation length (given their frequency) as people speaking to men did on average.

People speaking to male interviewers tended to delete more segments in general, as described in Table 4.15. The interaction in the left panel of Figure 4.14, then, suggests that one of the conditions in which people delete less when faced with a female interviewer is when a word is higher in frequency.

The right panel of Figure 4.14 shows an extremely strong interaction between word frequency and word length. Recall that the word length variable has been residualized against the frequency variable. Words with high scores on the residualized length measure are thus longer than would be expected given their (COCA) frequency. Figure 4.14 shows that these longer-than-expected words exhibit a strong facilitatory effect of their frequency

on deletion. By contrast, words that are shorter than expected, or even slightly longer than expected, show an inhibitory effect of frequency on deletion. The more frequent these shorter words are, the less likely they are to undergo segment deletion.

This result parallels an effect found in the previous chapter. The simplest explanation for this effect is that length overwhelms frequency in predicting deletion. Shorter-than-expected words have fewer segments to delete, and as their length decreases the amount of information lost to a potential deletion increases commensurately.

The residualized nature of the length predictor can help to explain why shorter words show an inhibitory frequency effect. A word that is shorter than expected given its frequency is a poor candidate for deletion. In fact, Table 4.15 shows that residualized length is the strongest predictor of deletion in the final LMER model.

Each line in the right panel of Figure 4.14 represents a set of words for which unexpected-shortness is held constant as COCA frequency is allowed to vary. An information-theoretic explanation of this result can be constructed: Shorter-than-expected words should contain less identifying information than words whose length is expected given their frequency, while longer-than-expected words contain more identifying information than may be necessary. The diachronic trend towards the shortening of frequent words suggests that these longer words are prime candidates for deletion, and Figure 4.14 bears out this prediction. Shorter-than-expected words, on the other hand, likely have a higher information density than longer words. Moreover, as described in the previous chapter, these short words may come from dense phonological neighbourhoods, suggesting that high phonological neighbourhood density inhibits deletion, providing further support for a listener-oriented model of reduction. As explained in the previous chapter, however, this result is somewhat speculative, and phonological neighbourhood density should be included in future models of reduction. (Neighbourhood density was not included in the models presented here because it had not been linked to reduction by the time this study was designed.)

The left panel of Figure 4.15 shows an interaction between the gender of the interviewer and the time when the current word token appears in the conversation. When the interviewer is male, speakers appear to gradually increase the number of deletions they make as they proceed through a conversation. When the interviewer is female, however, speakers actually decrease the number of deletions they make as time goes on.

The right panel of Figure 4.15 shows an interaction between a speaker's global speaking rate and the time when the word appears in the conversation. The interaction appears to show faster speakers gradually reducing the number of deletions they make as time passes. Slower speakers appear instead to increase the number of deletions they make, until the two groups are deleting segments at roughly the same rate.

Table 4.16 below compares the model with these five additional interactions to the model fitted earlier without them. (Note that the model fit to the untrimmed data set was used, since log-likelihood and AIC comparisons are not valid between models fit to different data sets.) Table 4.16 shows that the post-RF model shows improvement in log-likelihood ratios and AIC, suggesting that it is a more parsimonious fit to the model, with only a slight decrease in the proportion of variance explained. Random Forest filtering has thus improved the quality of the model.

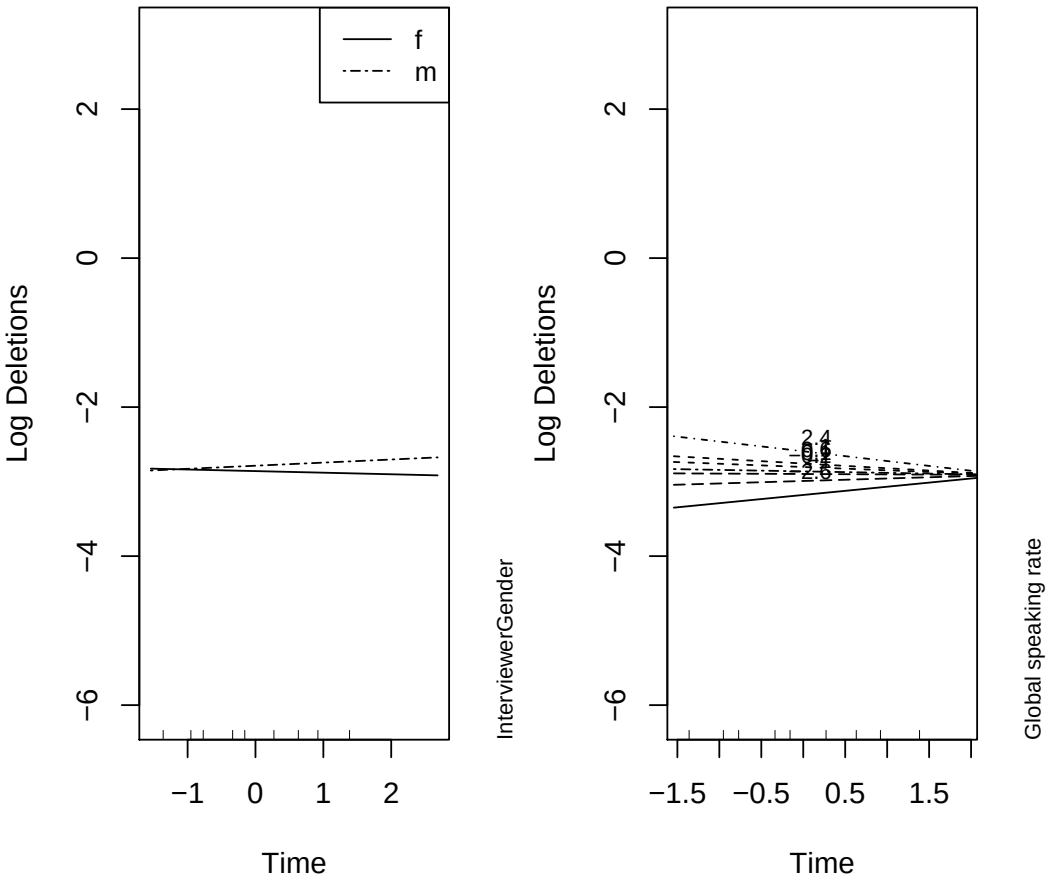

Figure 4.15: Time Interactions

	Model 1	Model 2
Degrees of Freedom	49	57
Proportion of Variance Explained (%)	38	37.9
log-likelihood ratio	-17850	-17815
log-likelihood improvement		35.1
Log-likelihood improvement p-value		0
Akaike Information Criterion (AIC)	35798	35744
AIC Improvement		54.2

Table 4.16: Model Comparison: Full (untrimmed) Model (1) v. RF-Informed Model (2)

4.4 General Discussion

The findings of this chapter can be divided into two broad categories: First, there are the findings of the models themselves, which reveal several aspects of the nature of deletion in the Buckeye corpus. Second, there are the findings about the modelling process itself, found by comparing and combining the LMER model selection and Random Forest modelling techniques. Each is discussed in a separate section below.

4.4.1 Model Findings

4.4.1.1 Main Effects

4.4.1.1.1 Demographic Predictors Speaker gender was found to have a significant effect on deletion rates in both LMER and random forest models in the current study. In both model types, men were more likely to delete segments than women.

Several studies have failed to find an effect of speaker gender on segment deletion rates in general (Patterson *et al.* 2003; Pluymaekers *et al.* 2005a; Raymond *et al.* 2006; Strik *et al.* 2008; Zimmerer 2009). Each of these studies, however, consider deletions in contexts that do not match the present study perfectly. Three of these studies consider deletion rates in languages other than English (Pluymaekers *et al.* 2005a; Strik *et al.* 2008; Zimmerer 2009), and most consider deletion in more restricted contexts than the present study, limiting their study to coronal (Raymond *et al.* 2006; Strik *et al.* 2008; Zimmerer 2009) or schwa (Patterson *et al.* 2003) deletions. Some of these studies found sub-contexts in which deletion rates differ by gender. Raymond *et al.* (2006) found that women are more likely than men to delete word-medial English coronal segments in environments when flapping is possible. Zimmerer (2009) found that women delete fewer word-final *t*'s in verbs, though this effect disappeared once other factors were taken into consideration. Gender has an effect on deletion rates when considered broadly, then, but not in the specific contexts studied in the above-cited literature.

Speaker age was not found to affect deletion rates in the models presented here. Two studies have found that older speakers are less likely to delete segments than younger speakers: Raymond *et al.* (2006) found an age effect in the Buckeye Corpus, though only for coronal stops in onset position. Strik *et al.* (2008) found an age effect in Dutch, though their study divided speaker age into three categories rather than two. At least two explanations can be found for the discrepancy between these studies and the present work. The results of Raymond *et al.* (2006) suggest that speaker age may affect deletion rates, but only in

certain phonotactic contexts. Indeed, Table 4.14 shows that the relationship between age and deletion rate varies depending on the wordform being produced. The results of Strik *et al.* (2008), on the other hand, suggest that the age categories used in the Buckeye corpus are too coarse to reveal an effect of age. It seems likely that one or both of these factors can be used to explain the lack of significance of the age effect in the present models.

Interviewer gender was not found to be a significant predictor of deletion rates in the present models, nor in any of the work cited here. It does participate in important interactions, however, as described in Section 4.4.1.2 below.

4.4.1.1.2 Phonological and Phonetic Predictors The speaking rate for the words surrounding the target was found to be a strong predictor of deletion, though its importance as a predictor was less pronounced in random forest models. Most studies of deletion have found (or taken as granted) that faster speaking rates lead to an increase in deletion (Fosler-Lussier & Morgan 1999; Raymond *et al.* 2006; Guy *et al.* 2008; Bürki *et al.* 2011). Two studies have failed to find an effect of speaking rate: Patterson *et al.* (2003) found no effect of speaking rate on deletion. Pluymaekers *et al.* (2005a) found that faster speaking rates led to duration reduction, but not deletion. In both of these studies, deletion rates were analyzed for a specific subset of the corpora only: Patterson *et al.* (2003) limit their analysis to schwa deletion, and Pluymaekers *et al.* (2005a) study only the seven most frequent words with a particular suffix. The discrepancy between their results and the present study may be a result of this narrower focus of inquiry.

The effect of average speaking rate on deletions is relatively understudied in the works cited here. Raymond *et al.* (2006) find that the ratio between global and local speaking rates can be used to predict segment deletion. Speakers exceeding their average speaking rate were more likely to delete segments. This result parallels the finding, illustrated in Figure 3.5 in the previous chapter, that global and local speaking rates interact to predict duration reduction, though no similar interaction was found for deletion rates in the present study.

The present study has led to conflicting predictions about the relationship between stressed syllables and deletion. Words with large numbers of stressed syllables are strongly linked to lower deletion rates in the LMER model, while weakly (if at all) linked to *higher* deletion rates in the random forest model. (See Figure 4.13.) The results of the LMER model are better supported by existing literature: In several studies, segments in stressed syllables were found to be less likely to undergo deletion than segments in unstressed syllables (Greenberg 1999; Pluymaekers *et al.* 2005a; Raymond *et al.* 2006; Van Bael *et al.* 2007; Zimmerer 2009) Figure 4.6 shows that the random forest model finds the number of stressed syllables to be one of the weakest predictors in the model. It thus seems likely that the 'facilitation' displayed in Figure 4.13 does not represent a truly significant effect. Still, *some* effect of stress on deletion rates is expected given the works cited here. It appears as if this effect is being subsumed in the random forest model, possibly by a correlated predictor such as word length.

Longer words are more likely to undergo deletion in the two model classes described here. This result mirrors the results found in other studies of deletion, in which words

with more expected segments (Van Bael *et al.* 2007) or syllables (Patterson *et al.* 2003; Raymond *et al.* 2006) undergo more deletion than shorter words.

4.4.1.1.3 Predictability A word's frequency in the corpus under investigation has reliably predicted deletion in several studies (Fosler-Lussier & Morgan 1999; Jurafsky *et al.* 2001; Guy *et al.* 2008; Zimmerer 2009; Meunier & Espesser 2011), though Priva (2008) found no effect of local frequency. Local frequency is predictive of deletion in the models developed here, though its partial effect in the random forest model appears relatively flat (See Figure 4.10).

One of the more surprising results in the present study is the finding that COCA frequency does not appear to facilitate deletion. Indeed, the LMER model produced here finds that an increase in COCA frequency actually *decreases* the likelihood that a segment will be deleted. This inhibitory effect of frequency is not replicated in other studies of deletion, though some studies have found no link between externally-measured frequency and deletion rates (Raymond *et al.* 2006; Guy *et al.* 2008; Schuppler *et al.* 2009). Guy *et al.* (2008) attribute this lack of effect to genre differences between the two corpora used in their study: The training corpus (CELEX) was composed of modern English, while the test corpus was composed of New Zealand English from the 1940's. The Buckeye and COCA corpora also represent different genres in several important ways. The COCA contains language sampled from a larger collection of people, taken over a wider period of time. Moreover, the COCA is composed largely of written language, and its spoken sub genres consist largely of scripted speech for television. One or more of these genre differences provide the most likely explanation for the lack of a facilitatory COCA frequency effect in the current study.

Backward and Forward word predictability were both found to lead to more deletions in the LMER model, though their partial effects were not strongly associated with deletion in the random forest model. Existing studies of predictability have found mixed results. The effect of predictability given the following word appears to be more robust. Several studies show an effect of this backward predictability on deletion (Pluymaekers *et al.* 2005a; Raymond *et al.* 2006; Schuppler *et al.* 2009), though a few studies fail to find such an effect (Jurafsky *et al.* 2001; Bürki *et al.* 2011)

Each of the studies listed above failed to find an effect of predictability given the previous word. Fosler-Lussier & Morgan (1999) found a predictability effect, though only when the two preceding words are taken into account. The present results suggest that such a predictor should not be rejected from reduction modelling studies: Predictability given the previous word does appear to have an effect on deletion rates.

4.4.1.1.4 Structural Constituency A word's position in its carrier phrase was not found to be a significant predictor of deletion in either of the model types presented here. Existing studies have similarly found that phrase position has little effect, with two notable exceptions: Words at the beginning or end of an utterance have been shown to undergo reduction or lengthening in several studies. These words, however, are excluded from the present analysis. The current results, then, suggest that words that are central in an utterance do not behave differently from each other in a consistent way with respect to deletion.

Both LMER and random forest models find that parts of speech differ in their deletion rates. Nouns are least likely to contain deletions, followed by adjectives, adverbs, and verbs. This gradient of increasing likelihood of deletion nearly matches the gradient of type/token ratios for each part of speech (nouns >adjectives >verbs >adverbs), suggesting that this ratio, or possibly an entropy measure, can be used to predict this result. None of the studies cited here consider the relationship between part of speech and deletion rates.

Part-of-speech based predictability measures had varying effects on deletion rates. Increased predictability given the preceding or surrounding parts of speech led to higher deletion rates in the LMER model, though no strong (partial) effects were found for these predictors in the random forest model. Predictability given the following part of speech had no significant effect on deletion rates in either model type. Existing studies do not model the effect of part-of-speech based predictability on deletion. The present results suggest that some consideration of these factors may benefit future studies of reduction, though a more sophisticated measure of syntactic constituency may prove more illuminating.

4.4.1.1.5 Topicality The topicality measure in the present study, term frequency-inverse document frequency, was not a significant predictor of deletion in any of the models presented here. Topicality did, however, participate in an interaction with speaker gender, with men tending to delete more segments as topicality increased. The tf-idf measure was not considered by the studies cited here.

4.4.1.1.6 Time The time at which the target word appears in an interview had no main effect on deletion rates, in neither the present study nor existing studies of deletion. Time of utterance did contribute to two interactive effects, however, as described below.

4.4.1.2 Interactive Effects

After incorporating the information gleaned from the random forest model, seven interactions were revealed to significantly improve the fit of the LMER model. None of these interactions have been directly attested to in existing studies of deletion rates.

The interaction between speaker gender and topicality has a straightforward interpretation: Men are more responsive to topicality than women when it comes to deletion rates.

Two interactions with Forward POS predictability illustrate the conditional effect this variable has on deletion. Table 4.15 indicates that higher predictability given the previous part of speech leads to an increased rate of deletion. The interactions show that this facilitatory effect does not hold true in two specific conditions. First, when the part-of-speech-based predictability and word-based predictability are mismatched, deletion is less likely to occur. In particular, words with low word-based predictability are less likely to be reduced as their POS-based predictability increases. Second, when a word has a large number of stressed syllables, it is less likely to undergo deletion as its POS-predictability increases. Why long, prosodically heavy words should show less deletion in predictable syntactic contexts is unclear, though Figure 4.5 suggests that this effect only appears for words with an extremely high number of stressed syllables.

The effect of COCA-based frequency on deletion rates is also conditional. This external

frequency has a reliable inhibitory main effect on deletion. An interaction with interviewer gender shows that this effect is stronger when the interviewer is female. An interaction with word length (residualized against COCA frequency) shows that words that are longer than expected given their COCA frequencies actually show *more* deletions as their frequency increases. COCA Frequency, then, has a (very strong) facilitatory effect on deletions, but only for words that are longer than expected.

The time at which a word token appears in the conversation has no reliable main effect on deletion, but it does modulate two other predictors. Speakers respond differently to different interviewer genders, for example. When speaking to a male interviewer, speakers gradually increased their deletion rates, while female interviewers saw their participants gradually lower their deletion rates. This interaction could indicate accommodation in deletion rates, as male speakers delete more segments on average than female speakers do (see Table 4.15).

Time of utterance also modulates the effect of global speaking rate on deletions. As the conversations proceeded, faster talkers tended to produce fewer deletions, while slower talkers tended to produce more. By the end of the conversation, deletion rates converge until there is little difference between faster and slower speakers. (This interaction is illustrated in the right panel of Figure 4.15).

4.4.1.3 Random Effects

Few of the deletion studies cited here incorporate random effects into their models. Some studies include speaker or word as a main effect rather than as a random effect (Patterson *et al.* 2003; Pluymaekers *et al.* 2005a; Zimmerer 2009). Bürki *et al.* (2011) include random intercepts in their model, but for variables at the phonemic level that are not considered here. Schuppler *et al.* (2009) include random intercepts for word and speaker, though no random slopes. They also incorporate following word as a random intercept, finding it to improve model fit significantly.

4.4.2 Modelling Techniques

The present chapter revealed a complication to the RF-Informed LMER modelling procedure described in the previous chapter. Namely, the difficulty in determining which variables in a random forest are absolutely, rather than relatively, important. In the models created in the previous chapter, the random forest model indicated that some variables were overlooked during the LMER modelling process. In the models created in the present chapter, however, each model type found significant predictors that the other model overlooked. In response, the criterion used to decide whether a variable be included in a RF-Informed LMER model was re-framed: Any predictor variable that *either* model type considers useful is included in the LMER interaction testing process.

This criterion was implicitly followed in the previous chapter, though all predictors that the random forest found important also significantly improved LMER model fit. Following this criterion proved quite successful, with five new interactions found in the previous chapter and four new interactions found in the present chapter. This suggests that combining random forests and LMER models, using the method described here, is useful in discovering interactions that either model alone fails to find useful.

Chapter 5

Conclusion

This chapter is divided into three sections. Section 5.1 outlines the implications of the results of the modelling process. In particular, it is focuses on what the findings of the preceding chapters regarding the conditions under which phonetic reduction takes place. Section 5.2 outlines the findings regarding the modelling process itself, and outlines how these findings can be used to improve the statistical modelling of reduction. (Each of these sections contains a point-form summary for easy reference.) The final section describes the broader implications of the study for the nature of reduction in general.

5.1 Model Findings

This section describes what the modelling process revealed about the nature of reduction. Each subsection focuses on a particular set of levels of linguistic analysis.

5.1.1 Demographic Predictors

Speaker age, speaker gender, and interviewer gender, are considered in this section. Table 5.1 shows that very few of the models showed a main effect for any of these predictor variables. The only exception to this trend is that men are more likely to delete segments from words than women. When combined with the lack of difference in *duration* reduction by gender this leads to an interesting result: Men and Women appear to reduce to the same degree, but differ in the reduction strategies that they apply.

The interaction between time and interviewer gender shown in Table 5.2 suggests that speakers gradually become more like their interviewer in reduction strategy: Speakers tend to delete more segments the longer they talk to males, and fewer segments the longer they talk to females. This sensitivity to the listener's reduction behaviour provides evidence for a listener-oriented aspect of reduction. (For a more extensive study of socially-mediated phonetic accommodation, see Babel (2009).)

The interviewer's gender also affects the way in which speakers respond to (COCA) frequency, though only with respect to deletion rates. In Chapter 4, higher COCA Frequency was shown to induce a non-significant trend towards a decreased rate of deletion. The interaction between interviewer gender and COCA frequency does reach significance, however: Speakers with a female interviewer showed a stronger inhibitory effect of frequency

	Duration			Deletion		
	LM1	RF	LM2	LM1	RF	LM2
Gender				*	*	*
Interviewer Gender					*	
Age					*	
Phrase speaking rate	+	+	+	+	+	+
Global speaking rate	-	-	-	+	+	+
Stress				-		-
Length	+	+	+	+	+	+
Local frequency	+	*	+	+	+	+
COCA Frequency		*			*	
Forwd wd pred.	+	+	+	+	+	+
Backwd wd pred.	+	+	+	+	+	+
Forwd POS pred.		*		+	+	
Backwd POS pred.		*			*	
Neighbr POS pred.	+	+	+	+	+	+
Phrase Order						
POS	*	*	*	*	*	*
tf-idf topicality	+	+	+		*	
Time					*	

Table 5.1: Summary of Results for Main Effects. '+' indicates increasing reduction, '-' indicates decreasing reduction, and '*' indicates some other type of significant effect. The 'RF' column indicates how predictors contributed to Random Forest models. The remaining columns indicate how predictors contributed to the most complete linear models before ('LM1') and after ('LM2') random forest modelling.

on deletion rates than speakers with a male interviewer. For more frequent words, that is, speakers are less likely to delete segments if their interviewer is female. This may be taken as further evidence of accommodation to the interviewer's gender.

Male and female speakers also respond differently to a word's topicality, as shown in Table 5.2. Figure 4.4 illustrates the nature of this interaction: Female speakers show little or no effect of topicality, while male speakers appear more likely to delete segments as topicality increases.

Taken together, these results suggest that when gender has an effect on reduction, it is more visible in measures of segment deletion than it is in measures of word duration. None of the above effects are visible in the duration models.

5.1.2 Phonetic Predictors

The predictors considered in this section are speaking rate (both by-speaker and by-phrase), the number of stressed syllables, and the length of the word form in phones.

5.1.2.1 Speaking Rates

Unsurprisingly, participants were more likely to both shorten words and delete segments from words when speaking more quickly. Participants who spoke faster on average also showed more reduction than slower speakers, though only in the the sense that they were more likely to delete phones from words. Indeed, participants with fast average rates of

	Duration		Deletion	
	LM1	LM2	LM1	LM2
COCA Frequency x Backwd wd pred.	*	*		
COCA Frequency x Length	*	*		*
COCA Frequency x Forwd wd pred.	*	*		
Local frequency x Backwd wd pred.	*	*		
Global speaking rate x Phrase speaking rate	*	*		
POS x Local frequency	*			
POS x Neighbr POS pred.	*			
POS x Backwd wd pred.	*			
POS x Backwd POS pred.		*		
POS x Forwd POS pred.		*		
Backwd POS pred. x Forwd POS pred.		*		
Phrase speaking rate x Backwd POS pred.		*		
Forwd wd pred. x Forwd POS pred.		*	*	*
tf-idf topicality x Gender			*	*
Forwd POS pred. x Stress			*	*
COCA Frequency x Interviewer Gender				*
Time x Interviewer Gender				*
Time x Global speaking rate				*

Table 5.2: Summary of Results for Interactive Effects. '*' indicates that adding the interaction significantly improved model fit during model selection. 'LM1' and 'LM2' are used as in the previous table.

	Duration		Deletion	
	Word	Speaker	Word	Speaker
Gender	*		*	
Interviewer Gender	*			
Age	*		*	
Phrase speaking rate		*	*	
Global speaking rate	*		*	
Stress		*		
Length		*		*
Local frequency				
COCA Frequency		*		*
Forwd wd pred.	*		*	
Backwd wd pred.	*	*		
Forwd POS pred.	*		*	
Backwd POS pred.	*			
Neighbr POS pred.				
Phrase Order				
POS		*		*
tf-idf topicality	*	*	*	*
Time	*	*		*

Table 5.3: Summary of Results for Random Effects. '*' indicates that adding the random slope significantly improved model fit.

speech were *less* likely to shorten the duration of their words than participants who spoke more slowly. Table 5.1 shows that these results were found across all model types.

A possible explanation for this result is found in the interaction, illustrated in Figure 3.5, between global and local speaking rates. This interaction suggests that faster speakers shorten their words more evenly across a conversation than slower speakers: Slower speakers, by contrast, shorten their words much more when speaking quickly than faster speakers do under the same circumstances.

No similar effect is found in the models of deletion: No interaction between global and local speaking rates was found to contribute to the model, and slower speakers delete fewer segments on average than faster speakers do. This difference decreased as speakers proceeded through their conversations, however. Faster and slower speakers tended to converge in their deletion rates as they proceed through a conversation. This is illustrated in the right panel of Figure 4.15, in which nearly identical deletion rates are found for each group near the end of the conversation. This may represent a kind of regression to the mean: Fast and slow speakers both begin with a deletion rate that fits their speaking style, then gradually adjust to fit some average deletion rate. This average rate may represent a kind of optimal deletion rate given a certain length of conversation, or it may represent another type of accommodation to the interviewers' deletion rates. More study would be required before either of these conclusions could be confirmed.

One additional interaction involving speaking rate was discovered after random forest filtering. The interaction, illustrated in Figure 3.18, shows the effect of local speaking being mitigated as a word's part of speech becomes more predictable from syntactic context. This result is explored in more detail in Section 5.1.4 below.

Adding random slopes for speaking rate variables improved the models in most cases, as illustrated in Table 5.3. For the reasons described in Section 2.6.1 (see Table 2.3), random slopes by speaker for global speaking rate cannot be calculated in either model. This leaves six possible random slopes by speaking rate, of which four were found to significantly improve model fit.

Random slopes by word for global speaking rate improved both deletion and duration models. The main effects structure of the model indicates that faster and slower speakers show different patterns of reduction. The random effects structure indicates that the magnitude of this difference depends on which word forms are being reduced.

Adding random slopes by word for local speaking rate improved the deletion model as well. This indicates that different word forms react differently to the speed at which a person is speaking during a phrase. Some word forms show a greater resistance to deletion than others do. (And conversely, some word types are easier to delete from than others.) The degree of variation is small, however. Indeed, Table 4.7 shows that the effect of local (phrase) speaking rate is less variable by word than any other useful random slope in the deletion model. Moreover, Table 5.3 shows that the effect of local speaking rate on *duration* reduction is not variable enough to merit including a by-word random slope for it in the final model. These results suggest that in fast speech all word types tend to be reduced to a similar degree. Or, more accurately, variation between word types in the effect of speech rate on reduction is well accounted for by the predictor variables included in the present

study.

Adding random slopes by *speaker* for local speaking rate, however, *did* significantly improve the quality of the word duration model. Table 3.12 shows that this variation is relatively broad, with a standard deviation of about 5 milliseconds among the 40 speakers (and 70,000+ data points studied here) in the corpus. This indicates that speakers vary in how much they shorten their words when speaking quickly. An analogous type of variation is described above: The interaction between global and local speaking rates shows that fast speakers and slow speakers shorten their words differently when speaking quickly. The variation in random slope by speaker shows that a speaker's idiolect leads to a similar difference in their response to speaking rate.

In the case of speaker age, however, some of the variation captured by the random slope may not be due to idiolectal variation. As described in Section 3.5.1.1.1, the gradient measure of speaker age in the Switchboard corpus has been found to predict reduction, while the binary measure used in the Buckeye corpus has not. A more gradient measure of age, then, may reduce the variation in by-speaker random slopes.

5.1.2.2 Word Length

Recall that the word length variable is measured in terms of the number of phones present in a word's citation form. This citation length is then residualized against (COCA) frequency in all models presented here. As a result, a word's 'length' incorporates two forms of expectation: Expected number of segments given citation form, and expected number of segments given the frequency of the word form in COCA. Values of the length predictor, then, reflect how much longer or shorter a word's citation form is than other words with similar frequencies.

Table 5.1 shows that as this length increases, words become shorter in duration and more likely to undergo deletion. This result is unsurprising, as it can be explained by at least two straightforward factors, possibly working in concert with each other. One factor is that, in general, longer words have a lower information density and a lower phonological neighbourhood density than shorter words. Shortening a long word is easier, then, because the resulting reduced form is still relatively easy to identify by the listener: The shortened form still contains a fair amount of distinguishing information, and still has relatively few phonological neighbours to be confused with. (The link between phonological density, word length, and COCA frequency is somewhat speculative, however. See Section 3.2.6.2 for details.) Another factor is diachronic in nature: More frequently used words tend to become shorter over time. Words that are longer than their frequency predicts are thus especially likely to undergo reduction over time. In fact, it is possible that these words have actually already undergone this reduction in the speakers' mental lexicons, and the corpus citation forms are simply out of date.

Each of these explanations in turn predict that word length and reduction should have a non-linear relationship. From a diachronic perspective, while frequent words are likely to lose segments over time, infrequent words are not expected to gain segments over time. Shorter-than-expected words are thus less likely to represent outdated citation forms, and less likely to show a strong effect on reduction, than longer words. This prediction is

examined further in Section 5.1.3 below.

The non-linear partial effects plots produced by the Random Forest models show that the non-linearity described above is indeed the case. The left panels of Figures 3.14 and 4.12 show reduction (in duration and number of deleted phones, respectively,) increasing as length increases. In both cases, this reduction effect accelerates at or near a residualized length of zero, the point at which words become longer than expected.

Word length participates in only one significant interaction, with COCA frequency, but this interaction is the strongest interaction in the present study (Figure 4.14), and one of only two interactions found to significantly improve both duration and deletion models (See Table 5.1). Implications of the interaction are described in Section 5.1.3 below.

5.1.2.3 Stress

None of the models of word duration found the number of stressed syllables in a word to be a useful predictor of word shortening. Both mixed-effects models of deletion, however, found that words with more stressed syllables tended to have fewer deletions. Thus, stressed syllables resist deletion but not shortening. When pressed to provide a reduced form, then, the speech production system 'prefers' to maintain the segmental content of a stressed syllable while shortening the duration of those segments (or the surrounding segments) instead.

5.1.3 Predictability

The role of predictability in reduction is shown here to be both complex and illuminating.

5.1.3.1 Frequency

The simplest result is that local frequency is a better predictor of reduction in the Buckeye Corpus than COCA frequency is. This implies that there is an important difference between the two corpora. There are indeed several notable differences between the two corpora, each providing a possible explanation for the difference in importance of the frequency measures. The differences in sample population may make the Buckeye corpus more representative of the dialect under study. The differences in genre provide multiple possible explanations: The extremely high prevalence of written or scripted language in COCA may make the Buckeye corpus more representative of spontaneous, conversational speech in general. The use of only a single genre may mean that the Buckeye frequency measure captures a usage pattern specific in pragmatics and semantics to that particular context. In order to best predict reduction in conversational speech, then, frequency counts should be drawn from as representative a sample of the speech under study as possible.

The interaction between COCA frequency and (relative) length has more complex implications. This interaction is robust; It is one of only two interactions found to improve both deletion and duration models. The interaction can be described as indicating that common, longer-than-expected words are very easy to shorten and delete from. This interaction suggests that these longer-than-expected words may be likely to lose some of their segments permanently. Indeed, it is possible that these words already have lost segments from their true citation forms, but this change has not yet been applied to the citation forms in the

dictionaries used in the present study. In short, these may be words for which the dictionary is out of date, at least in Central Ohio.

A more subtle implication comes from consideration of the cases in which higher frequency leads to *longer* productions. These cases provide indirect evidence for a listener-oriented reduction process. As Gahl *et al.* (2012) indicated, words from dense phonological neighbourhoods are easier for a speaker to produce, but more difficult for a listener to identify. The frequency-by-length interaction can be seen as an indirect measure of phonological neighbourhood density: Shorter-than-expected high frequency words should come from dense phonological neighbourhoods, since word frequency is inversely correlated with neighbourhood density. Such words show less shortening (Fig. 3.4) and fewer deletions (Fig. 4.14) than lower-frequency words of the same relative length, which should come from less dense phonological neighbourhoods. Hence, high phonological neighbourhood density leads to less reduction, conforming to the prediction made by listener-oriented models of production. Further research into the frequency-length interaction is required to determine whether this is truly an effect of phonological neighbourhood density, however, for reasons outlined in Section 3.2.6.2.

5.1.3.2 Predictability from Lexical Context

Both measures of frequency are shown to be less important to predicting duration reduction than predictability from lexical context. This is illustrated in Figure 3.3: frequent but unpredictable words are less likely to be reduced than infrequent but predictable words. This predictability-override is stronger for COCA frequency than it is for local frequency, in two ways: First, COCA frequency is more easily overridden than the effect of relative local frequency Figure 3.3 demonstrates this: More plot lines slope upwards in the left panel than in the right, indicating that higher COCA frequency leads to less duration reduction in a larger range of predictability conditions. Second, COCA frequency is overridden in the model by both forwards *and* backwards predictability (See Figure 3.4), but local frequency is only overridden by low predictability given the following word. The fact that frequency and predictability interact also implies that frequency-based prediction processes likely interact with context-based prediction processes.

Predictability from lexical context can itself be overridden by predictability from structural context. This is illustrated by the interaction between predictability given the following word and predictability (of POS) given the following POS. This interaction is found in both duration and deletion models, as indicated in Table 5.2 and illustrated in Figures 3.17 and 4.5. The interaction shows that lexically-predictable words are less likely to be reduced in structurally-unpredictable contexts than structurally-predictable contexts. That is, lexical-context predictability can itself be overridden by structural-context predictability.

5.1.4 Structural Constituency

Section 3.4.1.2 describes interactions between POS and POS-based predictability in models of duration reduction. These interactions have important implications regarding the role of structural constituency in reduction.

The left panel of Figure 3.16 illustrates that adverbs shorten more after parts-of-speech that usually precede adverbs. This result can be explained in either information-theoretic or processing-oriented terms. In terms of information, the result suggests that when an adverb is expected given the previous POS, less phonetic information is required to convey which particular adverb the speaker is uttering. In speech processing terms, the result suggests that a speaker is able to produce (or habitually produces) an adverb more quickly when it is found in a syntactic position where adverbs are expected.

Both of these explanations should predict similar behaviour for the other three POS types, but no such behaviour is observed. The interaction, then, has another implication: Adverbs are qualitatively different from nouns, verbs, and adjectives, and this difference leads to differing reduction behaviour. Figure 2.1 illustrates one possible difference between adverbs and other parts of speech: Adverbs have an extremely low type/token ratio in the Buckeye corpus. It is possible that type/token ratio itself plays a role reduction, and is thus a possible subject of future research.

The right panel of Figure 3.16 suggests that a part of speech's predictability from following syntactic context is an indirect marker of distance to a syntactic phrase boundary. The interaction indicates that this predictability measure marks these boundaries differently for different parts of speech. Detailed reasoning for this is provided in Section 3.4.1.2., but the implications are summarized here. Nouns and adverbs in Figure 3.16 are longer if they're more commonly found before the following part of speech. This suggests that the production of these two parts of speech is so routinized that the two relevant words are being processed simultaneously. As a result, the first word in the sequence takes longer to produce since the following word is being prepared for production concurrently. The behaviour of nouns and adverbs, then, is consistent with the interpretation of higher backward POS predictability as indicating that the target word is farther from a phrase boundary.

Verbs respond differently to this predictability, however, decreasing in duration as they become more predictable from the following part of speech. This result may be due to the fact that English verbs are most likely to come before one of their arguments. Thus, if a part of speech is most commonly found after verbs, it can be said to be most commonly found at the start of an argument. These arguments are potentially-long sequences of words that may be seen as complete syntactic (sub-)constituents in themselves. For verbs, then, increased predictability from following POS indicates increased likelihood that the verb falls at the end of a syntactic constituent, or at least before a coherent sub-constituent. Unlike nouns and adjectives, then, verbs with higher backward POS predictability are *more* likely to fall at a structural boundary. Concurrent processing with the following word becomes less likely, and thus reduction becomes more likely, and the result illustrated in Figure 3.6 is predicted.

Both of these results can thus be seen as indicating more shortening at the end of a syntactic constituent than in the middle of one. This has two important implications: First, that structural boundaries *within* phrases lead to reduction, analogous to the reduction found at the boundaries of the phrases themselves described in other studies. (See, e.g., Fougeron & Keating (1997).) Second, it suggests that not only common word sequences but also common part-of-speech sequences can become routinized in a way that affects their

production.

The qualitative differences between the behaviour of each part-of-speech type under various predictability conditions has a further implication for future modelling: When POS-predictability appears to interact with another variable, POS itself should also be added to the interaction. Such three-way interactions are beyond the scope of the present study.

However, when nouns and adverbs are taken together, they are more numerous in both types and tokens than verbs, as shown in Figure 2.1. Further interactions involving backwards POS predictability, then, are treated as though higher backwards POS predictability indicates that a POS is more tightly connected with the following POS, and that the two words are likely members of the same syntactic constituent.

The interaction illustrated in Figure 3.18, then, should be interpreted as further evidence that words within a syntactic constituent are being prepared concurrently during production, inhibiting duration reduction. This interaction shows the reduction-inducing effect of speech rate being mitigated by this reduction inhibiting process. Or, alternately, that words near the end of a syntactic constituent are more likely to be shortened during fast speech than the words that precede them.

Broadly speaking, the results found in this section suggest that syntactic structure within phrases has a strong and complex effect on reduction. A more thorough study, focused on disentangling this complexity by considering syntactic structure in greater detail, is likely to yield interesting results pertaining to phonetic reduction.

5.1.5 Topicality

Topicality is operationalized here as term-frequency-inverse-document-frequency (tf-idf) residualized against frequency. This measure was shown to be a significant, if weak, predictor of duration reduction (See Table 3.13). Topicality did not reach significance as a predictor of deletion, but one useful interaction suggests that this may be due to gender differences: Figure 4.4 suggests that men are more likely to delete segments as topicality increases, but women show little response to topicality.

Both results suggest that topicality is playing some role in phonetic reduction. Further study would be required to determine how the operationalization of topicality here compares to other measures of given-ness, such as a simple count of previous mentions. The presence of topicality effects on both deletion and duration reduction suggests that such study is warranted.

5.1.6 Time

Speakers changed deletion rates as they proceeded through a conversation, though only under certain conditions. The interactions that illustrate this are found in Figure 4.15. The linguistic implications of these interactions are described above.

5.1.7 Random Effects

The present study suggests that by-word and by-speaker variation are highly useful in modelling reduction. That is, allowing each word form and speaker to respond to the

predictor variables differently can be used to predict reduction very effectively.

The random-effects structure of the LMER models, summarized in Table 5.3, can be used to predict variation nearly as well as the best LMER models that combine both fixed and random effects. The random-effects structure of the duration reduction model can predict 14.8% of the variation in reduction, while the full, final LMER model predicts 15.7% of the variation. The random-effects structure of the deletion model predicts 36.8% of the variance in deletion, while the most efficient full LMER model predicts 36.7% of deletion rates.

This does not mean that the fixed-effects predictors themselves do not truly predict reduction, of course. Random forest models, which cannot at present account for by-word or by-speaker variation, still manage to predict duration reduction ($R^2 \approx 13.8\%$) and deletion rates ($R^2 \approx 31.6\%$) to a degree similar to that of the LMER models.

What it does mean is that variation between words and speakers is an important area of future studies of reduction. Examining this variation in detail may reveal how speakers can be grouped by reduction strategies, and how word forms can be grouped according to their resilience to these strategies.

5.1.8 Comparison of Duration and Deletion Models

The variable importance scores for the LMER models of duration reduction and deletions are illustrated in Figure 5.1. Four variables in particular stand out as having drastically different levels of importance in predicting the two reduction measures. Predictability given the following word and local speaking rate are much more crucial to models of duration reduction than to models of deletion rates. Word form citation length and speaker gender are much more crucial to models of deletion rates than to models of duration reduction. A similar pattern of results can be observed in the variable importance scores of the RF models, illustrated in Figure 5.2. The same four predictors show the greatest importance difference between models, and the differences are in the same directions.

Two of these results appear to be a result of how closely related the predictors in question are to the reduction measures themselves: Long words have the most segments to lose, so it is unsurprising that length is the strongest predictor of deletion rates. And words in fast speech are by definition shorter in duration than words in slow speech, so it is unsurprising that the speaking rate of the surrounding words is a strong predictor of the duration reduction found in a target word.

The other two results seem to illustrate precisely why it is important to consider multiple measures of phonetic reduction when studying the process: Several interesting effects were found to affect only one of the two reduction measures under study.

Speaker gender played almost no role in any model of duration reduction presented here. In these models, gender had no main effect, did not play a part in any useful interactions, and fell below the importance threshold in random forest modelling. Considering only this one measure of reduction, then, may lead researchers to falsely conclude that gender has no effect on phonetic reduction. Deletion modelling, however, shows clearly that this is not the case. Both random forests and LMERs showed that men are more likely to delete segments from words than women. Indeed, gender is the second-strongest predictor of deletion in the final LMER model. The interaction between time and interviewer gender also illustrates

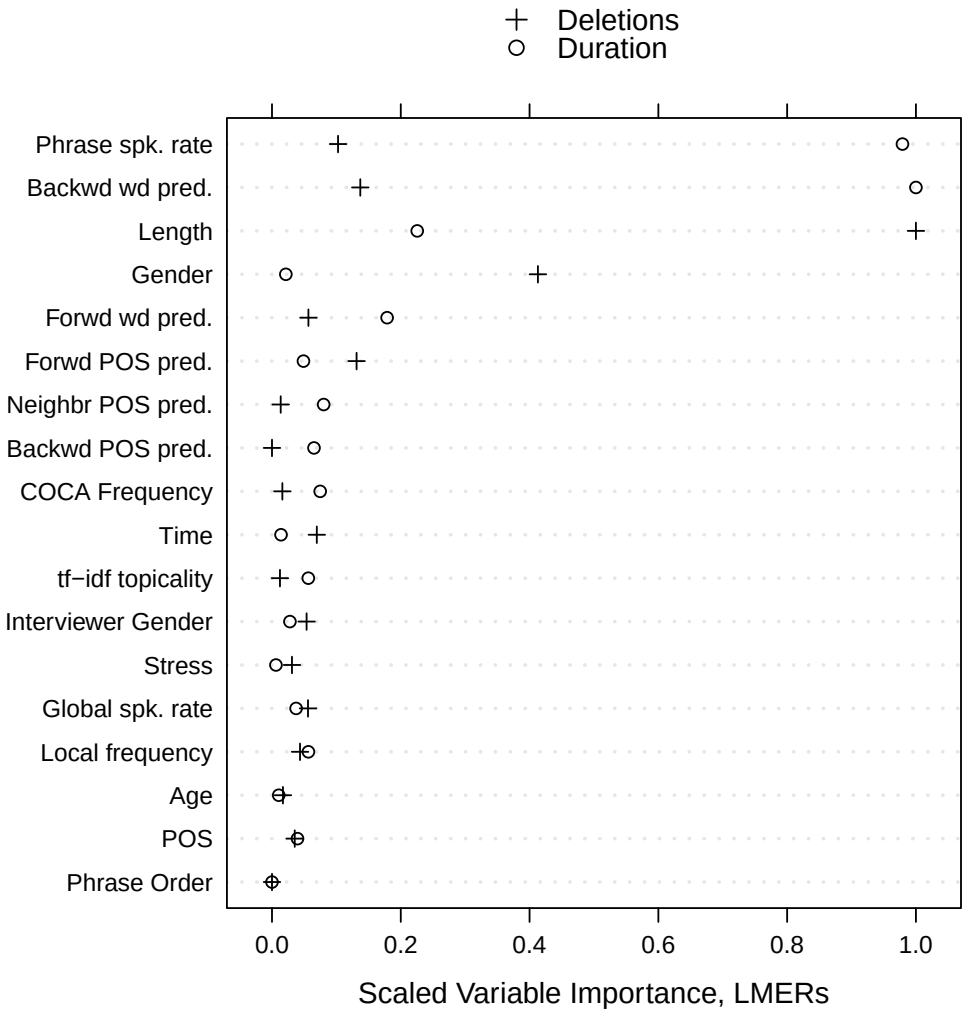

Figure 5.1: Comparison of Variable Importance in LMER models across Dependent Variables, in Descending Order of Difference Magnitude

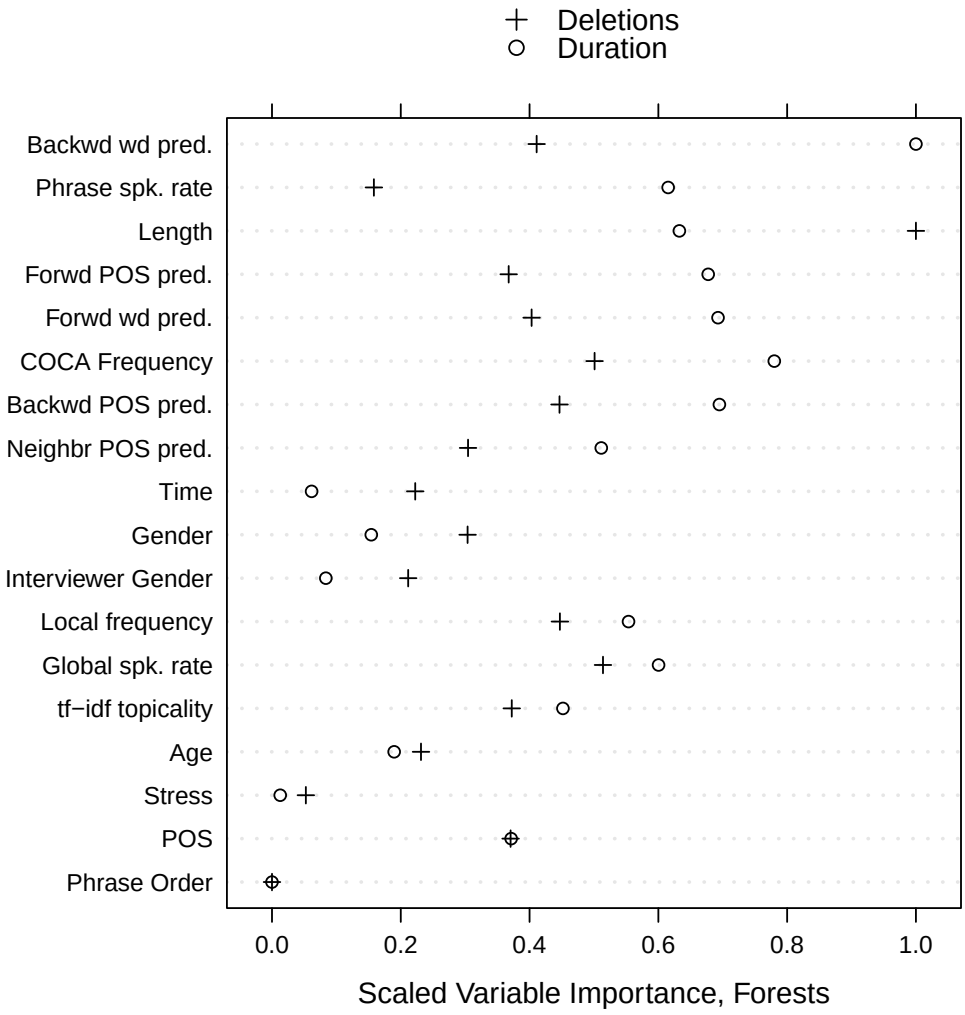

Figure 5.2: Comparison of Variable Importance in Random Forests across Dependent Variables, in Descending Order of Difference Magnitude

indirectly how this difference can affect a speaker's reduction behaviour: As Figure 4.15 shows, the longer someone is interviewed by a man, the more likely they are to delete segments from their speech, while the longer someone is interviewed by a woman, the less likely they are to delete segments. The gender of both the speaker and the interviewer are affecting phonetic reduction, then, but only in deletion rates.

Conversely, a word's predictability given the following word shows only a simple, moderately strong main effect on deletion rates. A researcher studying only deletion rates, then, may simply conclude that this predictability measure has a straightforward effect on deletion rates and move on. The duration reduction models, however, reveal much more about how backwards predictability affects reduction. This measure is, in fact, the strongest and most highly significant main-effects predictor of duration reduction, stronger even than speaking rate (Table 3.13). Figure 3.13 shows that the relationship between predictability and reduction is non-linear, clearly revealing that the increase in reduction for predictable words occurs mostly for words that are less predictable from following context than average.

Interactions with backwards predictability, shown in 3.3, help explain why frequency appeared to have a relatively weak effect on reduction, by showing that the effect of frequency varies depends on this predictability measure. This in turn points to a factor affecting reduction that the present study had overlooked. Namely, competition from context: A decrease in a frequent word's predictability from context entails a larger number of possible contexts for that word, to a greater degree than a decrease in predictability of an infrequent word. 3.3 illustrates that this number of competing contexts can overwhelm the facilitation of lexical access by frequency long assumed to be a key component of reduction.

The non-linearity and interactions described above are not present in models of deletion, and the insights into reduction that they provide would thus be lost to a researcher modelling only deletion rates.

Whole word duration and segment deletion are themselves relatively crude measures of reduction, and other measures of reduction may reveal still more about the properties of spontaneous speech. Warner (2011b), for example, describes studies that operationalize reduction in terms of modifications of individual segments or decreases in vowel space, tonal space, or spectral properties. Kondrak (2000) defined a gradient method for comparing the overall phonetic difference between pairs of words (*ALINE*). Such a method could be used to incorporate both fusion and deletion into a measurement of reduction. Future statistical modelling studies of reduction would likely benefit from the use and comparison of one or more of these measures.

5.1.9 Summary of Findings Regarding Reduction

This subsection summarizes the above findings in point form for easy reference.

- Women and Men use different reduction strategies (Sec. 5.1.1)
- Speakers gradually adopt interviewer-like deletion behaviours (Sec. 5.1.1)
  - Suggests that deletion is a listener-oriented (or listener-mirroring) reduction process.

- Speakers prefer shortening to deletion in stressed syllables (Sec. 5.1.2)
- Slow speakers shorten words more in fast speech than fast speakers do (Sec. 5.1.2)
- A gradient measure of age is preferable to a binary measure of age (Sec. 5.1.2)
- Citation forms in dictionaries may be out of date, or may not reflect Central Ohioan (Sec. 5.1.3)
- Corpora used to calculate frequency should be as representative of the speech in the target corpus as possible (Sec. 5.1.3)
- Shorter words become less likely to reduce as their written frequency increases (Sec. 5.1.3)
  - May be due to increased phonological neighbourhood density, suggesting a listener-oriented reduction process
  - Further investigation of this effect is warranted
- Effects of predictability from lexical context override (and communicate with) frequency effects (Sec. 5.1.3)
- Effects of predictability from structural context override (and communicate with) Effects of predictability from lexical context (Sec. 5.1.3)
- Different parts of speech respond in qualitatively different ways to structural predictability (Sec. 5.1.4)
  - Suggests that effect of, e.g., type/token ratio should be further explored
- Reduction is more likely at the end of a syntactic (sub-)constituent (Sec. 5.1.4)
  - Implies concurrent preparation of pairs of words in *syntactically* common sequences
- Syntactic structure affects reduction in ways too complex for this study to fully explain, and should be explored further (Sec. 5.1.4)
- Topicality effects on reduction can be found using tf-idf; this measure should be compared to others (Sec. 5.1.5)
- Variation between words and between speakers are crucial to predicting reduction, and should be explored in more detail (Sec. 5.1.7)
- Modellers should use more than one measure of reduction (Sec. 5.1.8)

5.2 Modelling Techniques

The present study reveals that Random Forest and LMER modelling techniques can be combined to improve a modeller's understanding of the structure of the data. The most useful way to combine the two models may be by first constructing a random forest model of the data. This random forest model can then be used as both filter and guide to the construction of an optimal linear model. The variable importance measure can be used to determine which variables are unlikely to contribute significantly, as either main effects or interactive effects, to the fixed-effects structure of the model. In the present study, for example, a total of 9 interactions (5 in the duration model and 4 in the deletion model) were initially overlooked due to the weak main effects of one of their component predictors. When these interactions were added to each LMER model, model fit improved significantly.

In some cases, these interactions helped to explain why these predictors were so weak in initial modelling. Figure 4.14 provides the most dramatic illustration of this, revealing a very strong frequency effect on deletion, but one that varies from inhibition to facilitation depending on the length of word under study. When interpreted as a result of inhibition due to phonological neighbourhood density, this interaction also provides evidence that reduction processes have a listener-oriented aspect. Without random forest modelling this interaction would have gone unnoticed, and important evidence regarding the processes underlying reduction would have been overlooked.

The partial effects in the forest can be used to determine which variables are best modelled as non-linear predictors of the response variable. Indeed, if several variables appear to have a non-linear effect on the dependent variable, the modeller may choose to forego linear models altogether in favour of a more flexible technique such as generalized additive modelling.

Random forest models also offer at least two benefits that are not fully explored here: The ability to handle highly correlated predictors, and the ability to handle predictors that are not normally distributed. For the sake of comparison, in the present study both types of models were fed the same set of sometimes de-correlated, sometimes log-transformed predictors. In future studies, however, the random forest filtering described above may also be performed on untransformed predictors. Such a model may provide a more easily interpretable understanding of the relationship between the independent and dependent variables. In the present study, for example, the effect of word length on reduction can not be described in a simple fashion. Its correlation with frequency forces the modeller to apply some transformation to word length in order to include it in a linear model. In the present study, length was residualized against frequency, forcing the interpretation of this variable to be based on how long a word is relative to how long it is expected to be given its frequency. It is difficult to consider such a measure a true predictor of the effect of word length on speech processing.

Moreover, correlations such as the one between frequency and length force the modeller to make a set of choices, selecting a correlation-reducing transformation to apply, and choosing how to apply that transformations. Random Forest modelling can inform this process. In the present study, for example, local frequency and length were both residualized against COCA frequency as a result of two assumptions: First, that the generally powerful

and broadly applicable effect of frequency on most aspects of psycholinguistic processing indicates that frequency is likely the more fundamental determiner of phonetic reduction. Second, that the larger sample of language found in the COCA makes it a better measure of a word's true frequency than the smaller Buckeye corpus. Constructing a Random Forest with untransformed predictors would allow the modeller to replace these assumptions with evidence. The Random Forest model's variable importance scores indicate which of the correlated predictors is most useful in predicting reduction. When choosing which predictor of a correlated predictor set to residualize against, then, the modeller can choose the most RF-important variable in the set.

Random Forests have limitations, of course. The limitation most relevant to the study of reduction is the inability to handle large sets of unordered factors. This limitation prevents word form and speaker, both examples of such factors, from being included in Random Forest models As described in Section 5.1.7 above, these factors are powerful tools for predicting reduction. Random Forests can not be used to explore these factors in greater detail, let alone to include them as control variables. A model type that can include random intercepts and slopes for these factors, such as LMER models, is thus a necessary component in the study of phonetic reduction.

In some cases, however, this limitation may actually be a strength. Section 4.2.5 revealed a potential danger inherent in random-effects modelling: When a factor is composed of *too* large a collection of levels, including random slopes for that factor may subdivide the corpus into portions that are quite small. As a result, over fitting becomes more likely, and the relationship between a predictor and the response variable may be overshadowed by the behaviour of small groups of data points. In the present study, the use of by-word random slopes in a corpus of linguistic data may have caused both of these problems. The problem is likely exacerbated for predictors for which random slopes by both word *and* speaker are included, dividing the corpus into even smaller units from which broad generalizations are derived. Further study is thus suggested into the ratio of the number of levels in a grouping variable and the number of data points in the corpus under study.

5.2.1 Summary of Findings Regarding Modelling Techniques

This subsection summarizes the above findings in point form for easy reference.

- An initial Random Forest model can improve the forward-fitting LMER modelling process; interactions, non-linear relationships, and dominant predictors in correlated predictor sets can all be discovered or suggested by RF modelling.
- RF models cannot incorporate the by-word and by-speaker variation that is crucial to predicting reduction, so they should be used in combination with model types that can, such as LMERs.
- The use of random slopes for factors with many levels may lead to over fitting. Further study into how many such levels can be included without limiting the generalizability of results is warranted.

5.3 Broader Implications

Perhaps the broadest implications of the present study are that phonetic reduction is widespread in spontaneous speech, and that some amount of this reduction can be predicted using linguistic properties.

Moreover, the factors linked to reduction span several levels of linguistic analysis. Every level of linguistic analysis studied here was shown to have some link to phonetic reduction. Demographic/social properties like gender, phonetic properties like stress, predictability properties like frequency and conditional probability, structural/syntactic properties like part of speech predictability, semantic properties like topicality, and pragmatic properties like time in conversation, were all used to improve predictions of when reduction took place.

As this is an observational study, (rather than an experiment,) the results can only show what *predicts* reduction, rather than what *causes* it. Nonetheless, the breadth of predictive properties implies that reduction should be considered in the construction of psycholinguistic models of nearly any level of language production.

While the wide range of factors found here to affect reduction suggest that it plays an important role in human language processing, the definite conclusions that can be drawn are necessarily limited. The conclusions of the present study can only be assumed with certainty to apply to Central Ohioan English speakers undergoing sociolinguistic-style interviews. To better understand the role and nature of phonetic reduction, then, broader and deeper investigations into the phenomenon are required: Studies of reduction in several genres taken from speakers of several languages are required to evaluate whether the results found here apply to human language processing in general. Studies that aim to incorporate reduction into psycholinguistic models of, for example, syntactic processing, require a more thorough and subtle analysis of syntactic properties than this study is able to present.

In short, the findings presented here indicate that phonetic reduction is likely an essential property of human language. As a result, the number of studies into this essential property should only increase.

Bibliography

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19.716–723.

Anderson, A.H., & B. Howarth. 2002. Referential form and word duration in video-mediated and face-to-face dialogues. In *Proceedings of the sixth workshop on the semantics and pragmatics of dialogue (EDILOG 2002), Edinburgh, UK*, 4–6. Citeseer.

Archer, K.J., & R.V. Kimes. 2008. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis* 52.2249–2260.

Argamon, S., M. Koppel, & G. Avneri. 1998. Routing documents according to style. In *First international workshop on innovative information systems*. Citeseer.

Aylett, M., & A. Turk. 2004. The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47.31–56.

——, & ——. 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America* 119.3048.

Baayen, R. H. 2008. *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge, U.K.: Cambridge University Press.

Baayen, R. H., 2011. *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"*.. R package version 1.4.

Baayen, R. H, D. J Davidson, & D. M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* p. in press.

——, R. Piepenbrock, & H. van Rijn. 1993. *The CELEX lexical database (CD-ROM)*. Philadelphia: Linguistic Data Consortium.

Baayen, R.H. 2010. The directed compound graph of english. an exploration of lexical connectivity and its processing consequences. *New impulses in word-formation (Linguistische Berichte Sonderheft 17). Hamburg: Buske* 383–402.

Babel, Molly Elizabeth, 2009. *Phonetic and social selectivity in speech accommodation*. UNIVERSITY OF CALIFORNIA dissertation.

Baker, R.E., M. Baese-Berk, L. Bonnasse-Gahot, M. Kim, K.J. Van Engen, & A.R. Bradlow. 2011. Word durations in non-native english. *Journal of phonetics* 39.1–17.

——, & A.R. Bradlow. 2009. Variability in word duration as a function of probability, speech style, and prosody. *Language and speech* 52.391–413.

Bard, E.G., A.H. Anderson, C. Sotillo, M. Aylett, G. Doherty-Sneddon, & A. Newlands. 2000. Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language* 42.1–22.

Barr, Dale J, Roger Levy, Christoph Scheepers, & Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68.255–278.

Bates, Douglas, Martin Maechler, & Ben Bolker, 2011. *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-42.

Bell, A., J.M. Brenier, M. Gregory, C. Girand, & D. Jurafsky. 2009. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language* 60.92–111.

——, D. Jurafsky, E. Fosler-Lussier, C. Girand, & D. Gildea. 1999. Forms of english function words–Effects of disfluencies, turn position, age and sex, and predictability. In *Proceedings of ICPhS*, volume 99, p. 395–398. Citeseer.

——, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, & D. Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *Journal of the Acoustical Society of America* 113.1001–1024.

Belsley, D.A. 1984. Demeaning conditioning diagnostics through centering. *The American Statistician* 38.73–77.

Bonferroni, C. E. 1935. Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni* 13.

Breiman, L. 2001. Random forests. *Machine learning* 45.5–32.

——, J. H Friedman, R. Olshen, & C. J Stone. 1984. *Classification and Regression Trees*. Belmont, California: Wadsworth International Group.

Bürki, A., M. Ernestus, C. Gendrot, C. Fougeron, & U.H. Frauenfelder. 2011. What affects the presence versus absence of schwa and its duration: A corpus analysis of french connected speech. *The Journal of the Acoustical Society of America* 130.3980–3991.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 711–733.

Carroll, J. B, & M. N White. 1973. Word frequency and age of acquisition as determiners of Picture-Naming latency. *Quaterly Journal of Experimental Psychology* 25.85–95.

Curran, James, Stephen Clark, & Johan Bos. 2007. Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 33–36, Prague, Czech Republic. Association for Computational Linguistics.

Davies, M. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14.159–190.

del Prado Martín, Fermín Moscoso. 2003. *Paradigmatic Effects in Morphological Processing: Computational and cross-linguistic experimental studies*. MPI Series in Psycholinguistics. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics.

Díaz-Uriarte, R., & S.A. De Andres. 2006. Gene selection and classification of microarray data using random forest. *BMC bioinformatics* 7.3.

Fosler-Lussier, E., & N. Morgan. 1999. Effects of speaking rate and word frequency on pronunciations in convertional speech. *Speech Communication* 29.137–158.

Fougeron, C., & P. Keating. 1997. Articulatory strengthening at the edges of prosodic domains. *Journal of the Acoustical Society of America* 101(6).3728–3740.

Fowler, C. A, & J. Housum. 1987. Talkers' signalling of "New" and "Old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language* 26.489–504.

Frauenfelder, U. H, R.H. Baayen, & F. M. Hellwig. 1993. Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language* 32.781–804.

Gahl, S. 2008. "Time" and "thyme" are not homophones: Lemma frequency and word durations in a corpus of spontaneous speech. *Language* 84.

——, Y. Yao, & K. Johnson. 2012. Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of memory and language* .

Gamon, M. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of COLING*, volume 4, p. 611.

Gelman, A., & J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*, volume 648. Cambridge University Press New York.

Godfrey, J. J., E. C. Holliman, & J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92.*, volume 1.

Greenberg, S. 1999. Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29.159–176.

Gregory, M. L, W. D Raymond, A. Bell, E. Fosler-Lussier, & D. Jurafsky. 1999. The effects of collocational strength and contextual predictability in lexical production. *CLS* 35.151–166.

Guy, G.R., J. Hay, & A. Walker. 2008. Phonological, lexical, and frequency factors in coronal stop deletion in early new zealand english. *Laboratory Phonology* 11.

Harrell, Jr, Frank E, with contributions from Charles Dupont, & many others., 2013. *Hmisc: Harrell Miscellaneous*. R package version 3.12-2.

Hothorn, Torsten, Kurt Hornik, & Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15.

Jaeger, Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* X.in press.

Johnson, K. 2004. Massive reduction in conversational american english. In *Spontaneous speech: data and analysis. Proceedings of the 1st session of the 10th international symposium*, 29–54, Tokyo, Japan.

Jurafsky, D., A. Bell, E. Fosler-Lussier, C. Girand, & W. Raymond. 1998. Reduction of english function words in switchboard. In *Fifth International Conference on Spoken Language Processing*. ISCA.

——, A. Bell, M. Gregory, & W. D Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In *Frequency and the emergence of linguistic structure*, ed. by J. L Bybee & P. Hopper, 229–254. Amsterdam: Benjamins.

Kahn, J.M., & J.E. Arnold. 2012. A processing-centered look at the contribution of givenness to durational reduction. *Journal of Memory and Language* .

Kennedy, A. 2000. Parafoveal processing in word recognition. *The Quarterly Journal of Experimental Psychology: Section A* 53.429–455.

——, J. Pynte, & S. Ducrot. 2002. Parafoveal-on-foveal interactions in word recognition. *The Quarterly Journal of Experimental Psychology: Section A* 55.1307–1337.

Kiesling, S., L. Dilley, & W.D. Raymond. 2006. The variation in conversation (vic) project: Creation of the buckeye corpus of conversational speech. *Dept. of Psychology, Ohio State University, available at www. buckeyecorpus. osu. edu* .

Köhler, Reinhard. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Kondrak, G. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, p. 288–295.

Koppel, M., N. Akiva, & I. Dagan. 2003. A corpus-independent feature set for style-based text categorization. In *IJCAI-2003 Workshop on Computational Approaches to Text Style and Synthesis, Acapulco, Mexico*. Citeseer.

Lam, T.Q., & D.G. Watson. 2010. Repetition is easy: Why repeated referents have reduced prominence. *Memory & cognition* 38.1137–1146.

Levelt, W. J.M, A. Roelofs, & A. S Meyer. 1999. A theory of lexical access in speech production. *Behavioral and brain sciences* 22.1–38.

Liaw, A., & M. Wiener. 2002. Classification and regression by randomforest. *R news* 2.18–22.

Luhn, H.P. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development* 2.159–165.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, & Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.* 19.313–330.

Meunier, C., & R. Espesser. 2011. Vowel reduction in conversational speech in french: The role of lexical factors. *Journal of Phonetics* 39.271–278.

Newmeyer, F. J. 2006. On gahl and garnsey on grammar and usage. *Language* 82.399–404.

Nicodemus, K., J. Malley, C. Strobl, & A. Ziegler. 2010. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC bioinformatics* 11.110.

Ohala, J. J. 1993. Sound change as nature's speech perception experiment. *Speech Communication* 13.161.

Patterson, D., P.C. LoCasto, & C.M. Connine. 2003. Corpora analyses of frequency of schwa deletion in conversational american english. *Phonetica* 60.45–69.

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

Pisoni, D. B, H. C Nusbaum, P. A Luce, & L. M Slowiaczek. 1985. Speech perception, word recognition and the structure of the lexicon. *Speech Communication* 4.75–95.

Pitt, M. A, K. Johnson, E. Hume, S. Kiesling, & W. Raymond. 2005. The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication* 45.89–95.

Pluymaekers, M., M. Ernestus, & R. H Baayen. 2005a. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62.146–159.

——, ——, & ——. 2005b. Frequency and acoustic length: the case of derivational affixes in dutch. *Journal of the Acoustical Society of America* 118.2561–2569.

——, M. Ernestus, & R.H. Baayen. 2005c. Lexical frequency and acoustic reduction in spoken dutch. *The Journal of the Acoustical Society of America* 118.2561.

Priva, U.C. 2008. Using information content to predict phone deletion. In *Proceedings of the 27th West Coast Conference on Formal Linguistics*, 90–98. Cascadilla Proceedings Project.

Raymond, W.D., R. Dautricourt, & E. Hume. 2006. Word-internal/t, d/deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change* 18.55.

Robertson, S.E., & K.S. Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information science* 27.129–146.

——, & K. Spärck Jones. 1994. Simple, proven approaches to text retrieval. Technical Report UCAM-CL-TR-356, University of Cambridge, Computer Laboratory.

Sarle, WS. 1990. The varclus procedure. *SAS/STAT User's Guide,* .

Scheibman, Joanna, & Joan Bybee. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in english. *Linguistics* 37.575–596.

Schuppler, B., W. Van Dommelen, J. Koreman, & M. Ernestus. 2009. Word-final [t]-deletion: An analysis on the segmental and sub-segmental level. In *Proceedings of interspeech*, 2275–2278.

Strik, H., J. van Doremalen, & C. Cucchiarini. 2008. Pronunciation reduction: how it relates to speech style, gender, and age. In *Proceedings of Interspeech*, 1477–1480.

Strobl, C., A.L. Boulesteix, T. Kneib, T. Augustin, & A. Zeileis. 2008. Conditional variable importance for random forests. *BMC bioinformatics* 9.307.

Strobl, Carolin, James Malley, & Gerhard Tutz. 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods* 14.323.

Tagliamonte, S.A., & R.H. Baayen. 2010. Models, forests and trees of york english: Was/were variation as a case study for statistical practice. *Manuscript submitted for publication* .

Tily, H., S. Gahl, I. Arnon, N. Snider, A. Kothari, & J. Bresnan. 2009. Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition* 1.147–165.

Tree, J. E Fox, & H. H Clark. 1997. Pronouncing 'the' as 'thee' to signal problems in speaking. *Cognition* 62.151–167.

Tremblay, Antoine, Dalhousie University, Johannes Ransijn, & University of Copenhagen, 2013. *LMERConvenienceFunctions: A suite of functions to back-fit fixed effects and forward-fit random effects, as well as other miscellaneous functions.*. R package version 2.0.

Van Bael, C., RH Baayen, & H. Strik. 2007. Segment deletion in spontaneous speech: A corpus study using mixed effects models with crossed random effects. In *Proceedings of interspeech*, 2741–2744.

Warner, N. 2011a. Methods for studying spontaneous speech. In *The Oxford handbook of laboratory phonology*, ed. by A. Cohn, C. Fougeron, & M. Huffman, 621–633. Oxford University Press.

—— 2011b. Reduction. In *The Blackwell Companion to Phonology: Phonological Processes*, ed. by Marc van Oostendorp, Colin J Ewen, Elizabeth Hume, & Keren Rice, 1866–1891. John Wiley & Sons.

Yao, Y., 2011. *The effects of phonological neighborhoods on pronunciation variation in conversational speech*. University of California dissertation.

Zimmerer, F., 2009. *Reduction in natural speech*. University of Frankfurt dissertation.

Zipf, G. K. 1935. *The Psycho-Biology of Language*. Boston: Houghton Mifflin.

Appendices

Appendix A

Partial Effects Plots

Throughout this Appendix, The abbreviations 'RF' and 'LME' are used to refer to Random Forest and Linear Mixed-Effects Models, respectively. The notation (Resid.) indicates that the predictor was residualized against another predictor during decorrelation.

A.1 Word Duration Reduction Models

A.1.1 Demographic Predictors

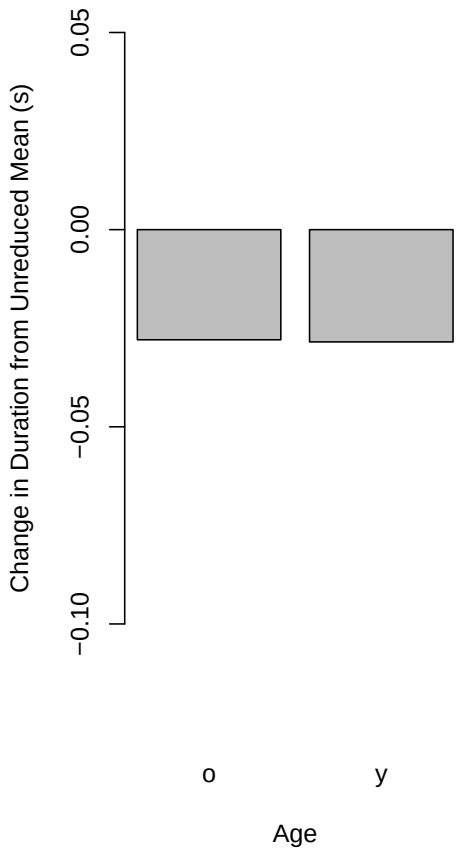

Figure A.1: Partial Effect of Speaker Age in RF Duration Model

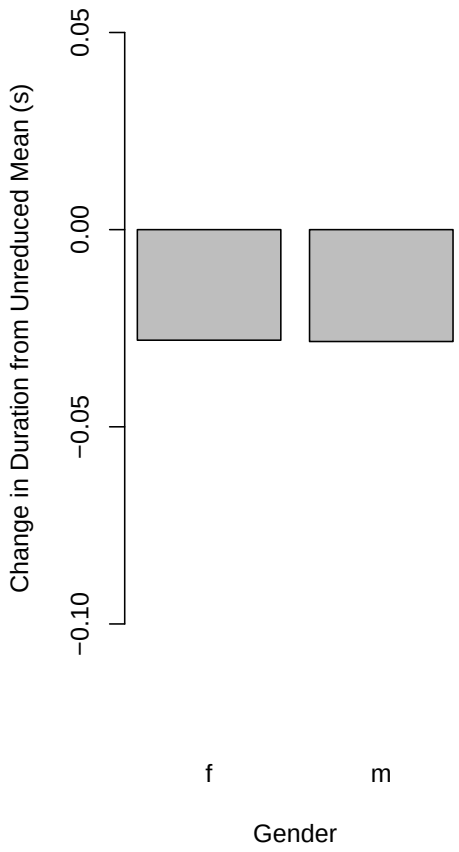

Figure A.2: Partial Effect of Speaker Gender in RF Duration Model

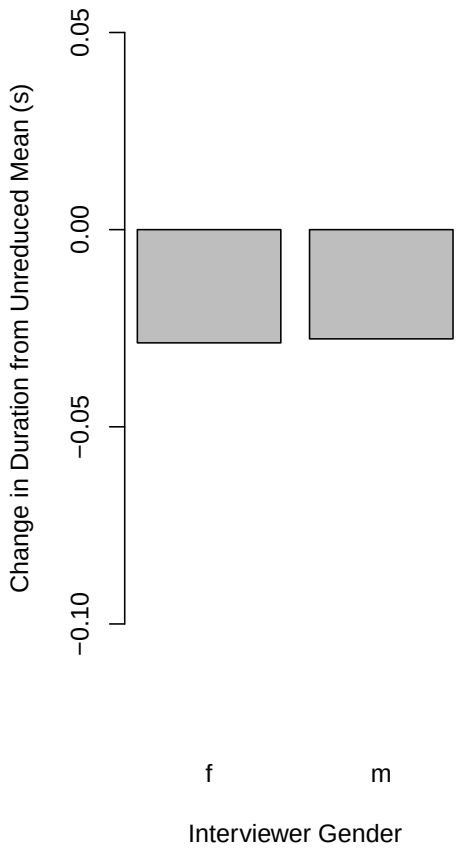

Figure A.3: Partial Effect of Interviewer Gender in RF Duration Model

A.1.2 Phonological and Phonetic Predictors

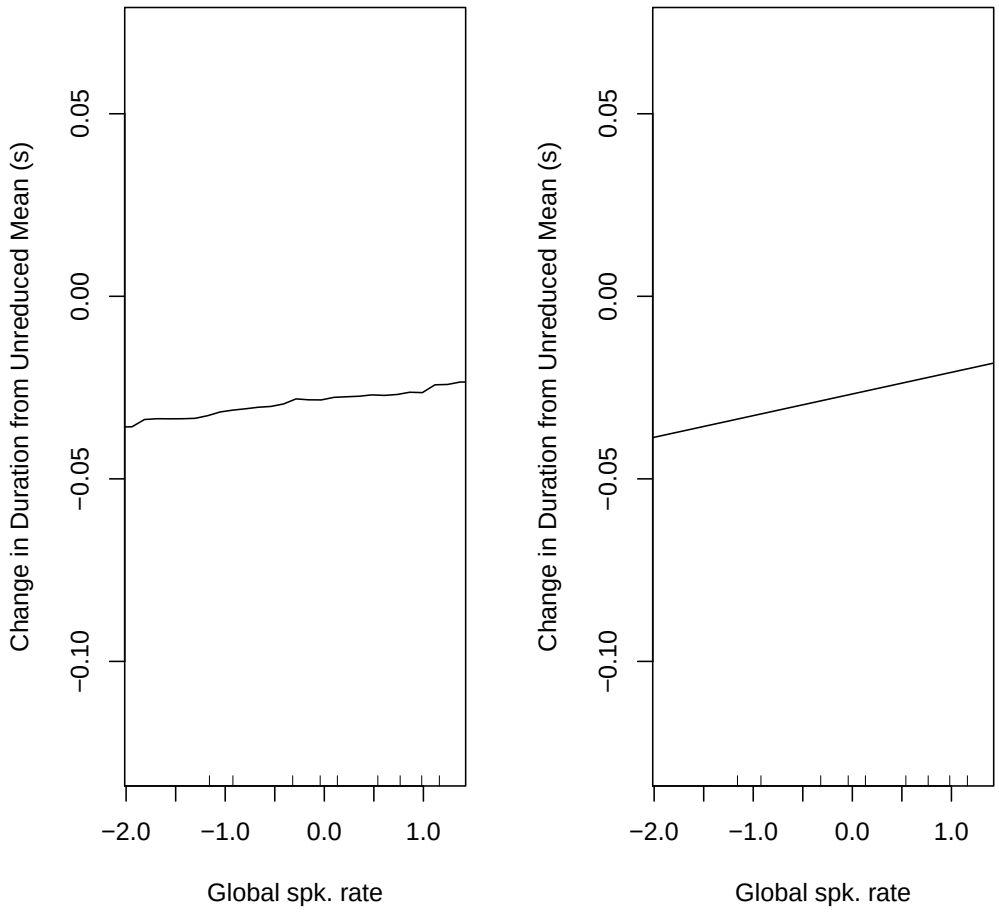

Figure A.4: Partial Effect of Average Speech Rate in RF (L) and LME (R) Duration Models

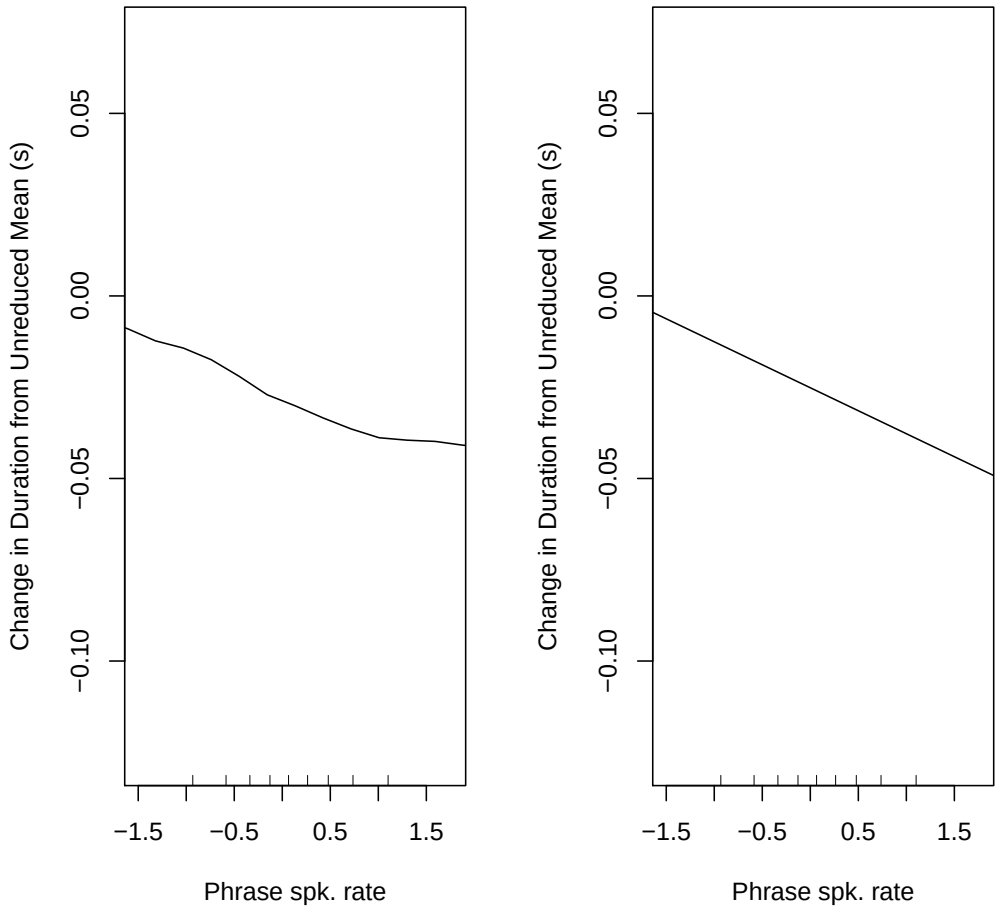

Figure A.5: Partial Effect of Local Speech Rate in RF (L) and LME (R) Duration Models

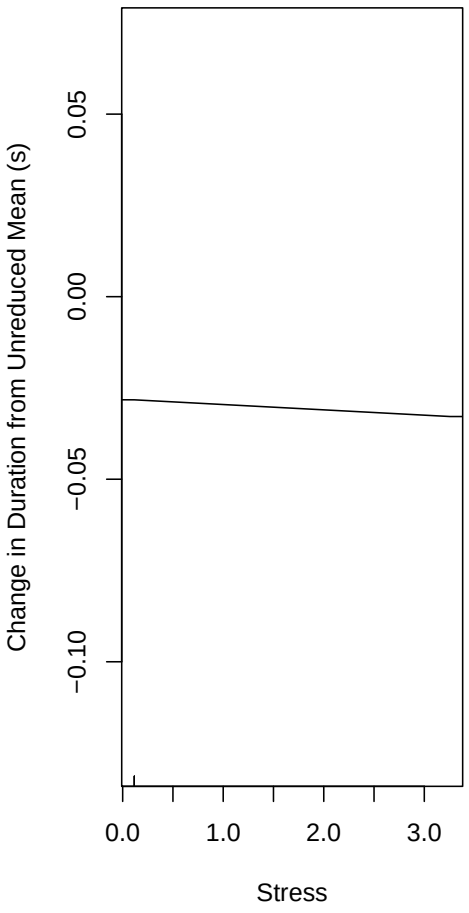

Figure A.6: Partial Effect of Stressed Syllables in RF Duration Model

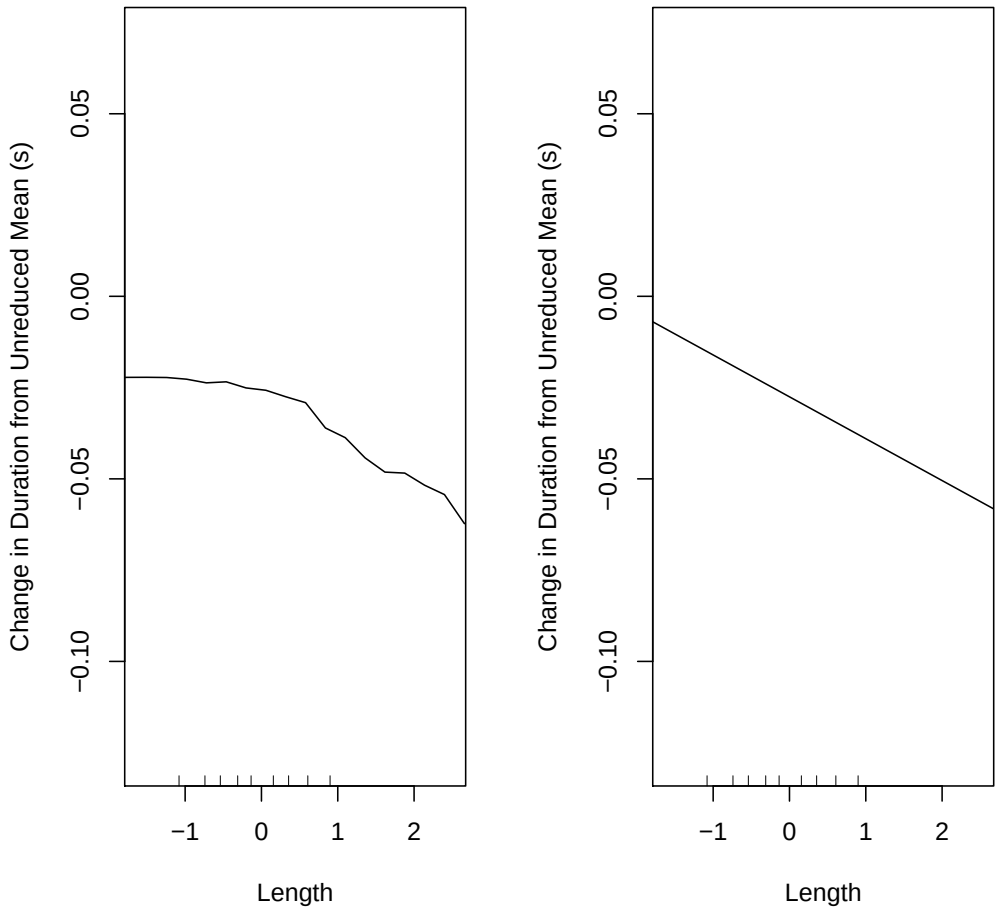

Figure A.7: Partial Effect of Word Length (Resid.) in RF (L) and LME (R) Duration Models

A.1.3 Predictability

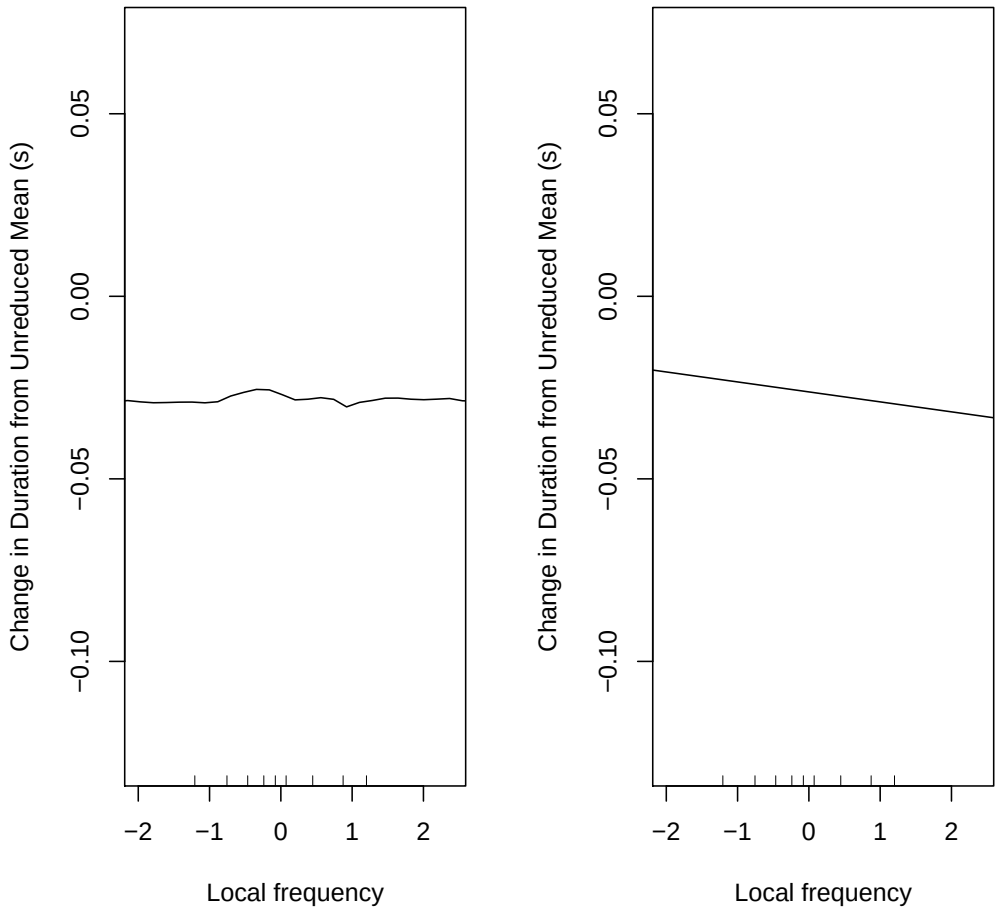

Figure A.8: Partial Effect of Buckeye Frequency (Resid.) in RF (L) and LME (R) Duration Models

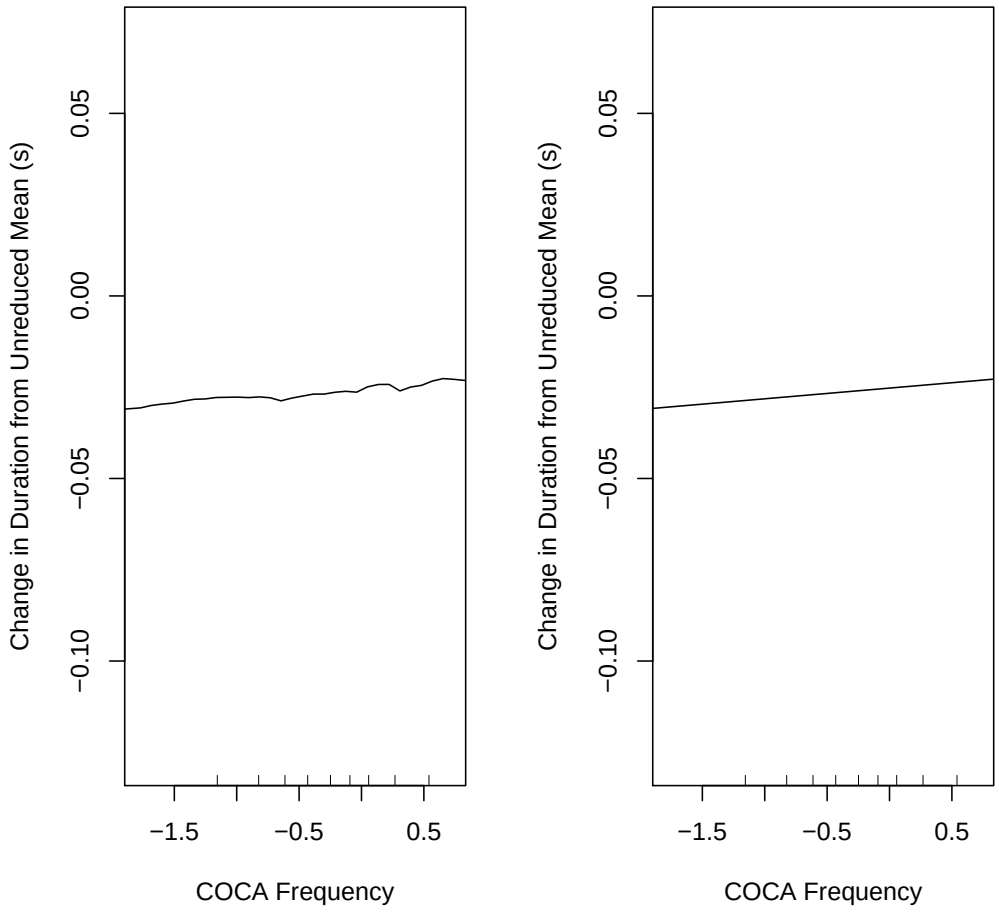

Figure A.9: Partial Effect of COCA Frequency in RF (L) and LME (R) Duration Models

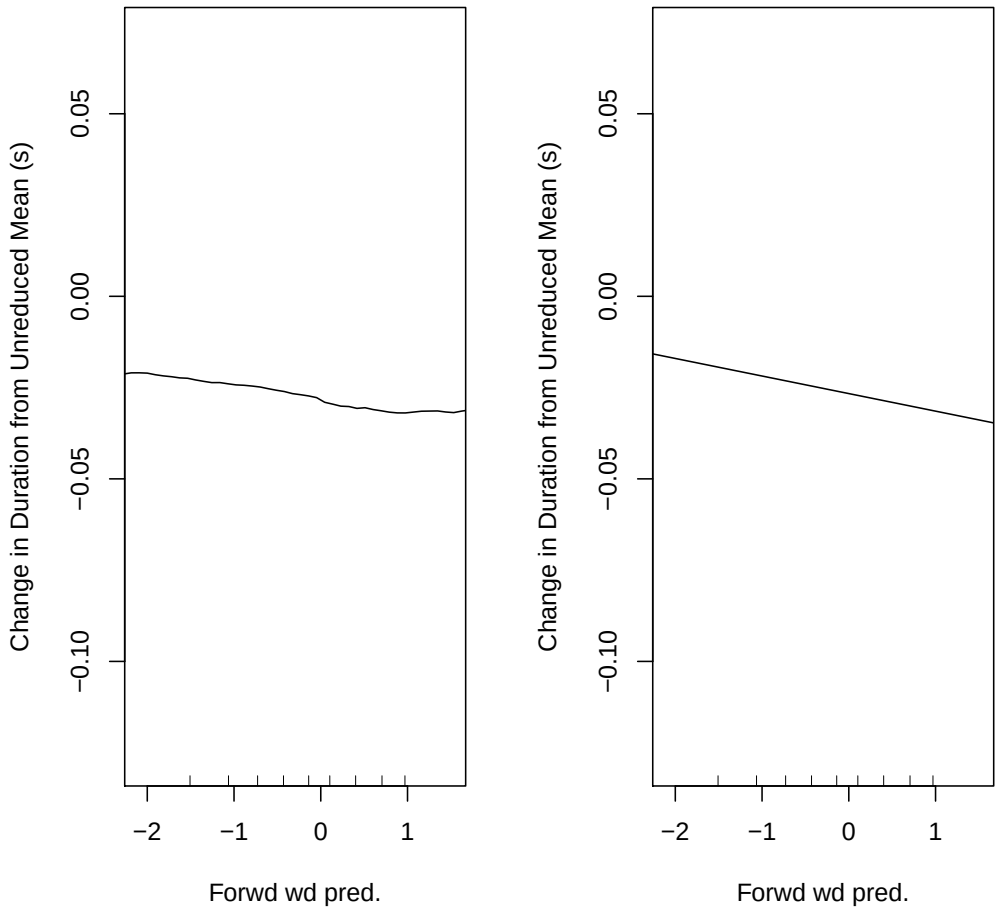

Figure A.10: Partial Effect of Forward Word Predictability in RF (L) and LME (R) Duration Models

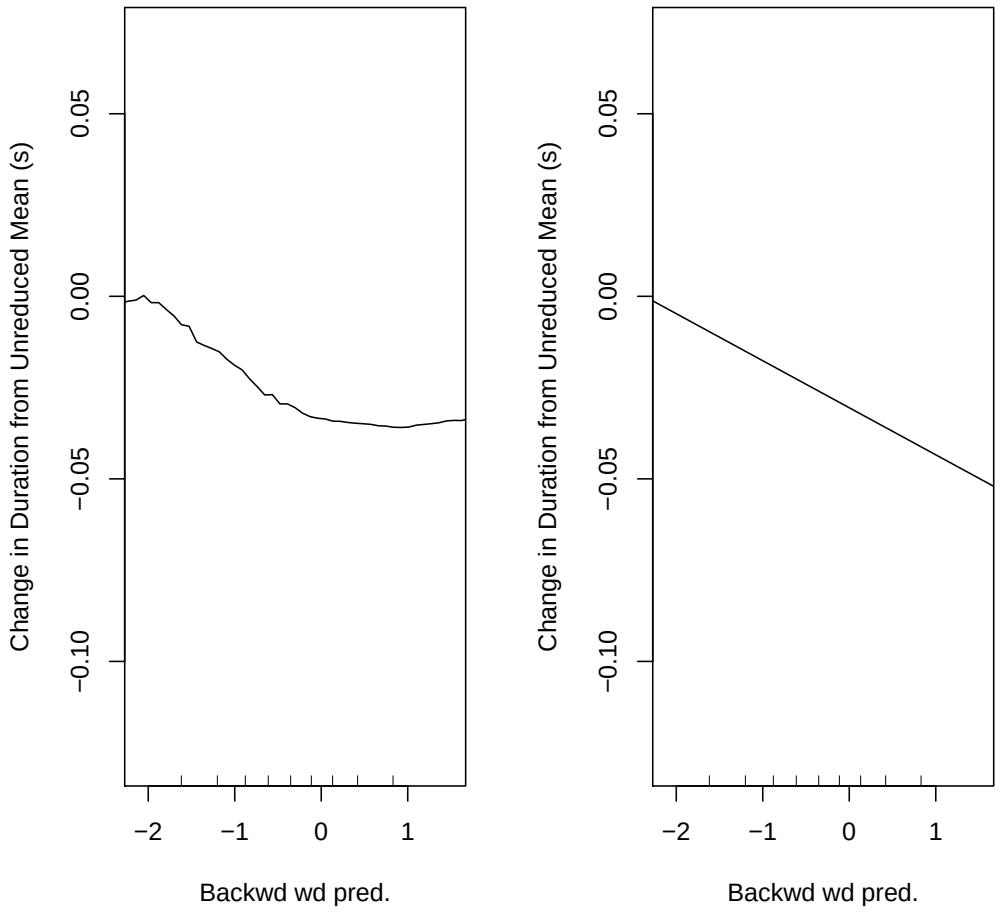

Figure A.11: Partial Effect of Backward Word Predictability in RF (L) and LME (R) Duration Models

A.1.4 Structural Constituency

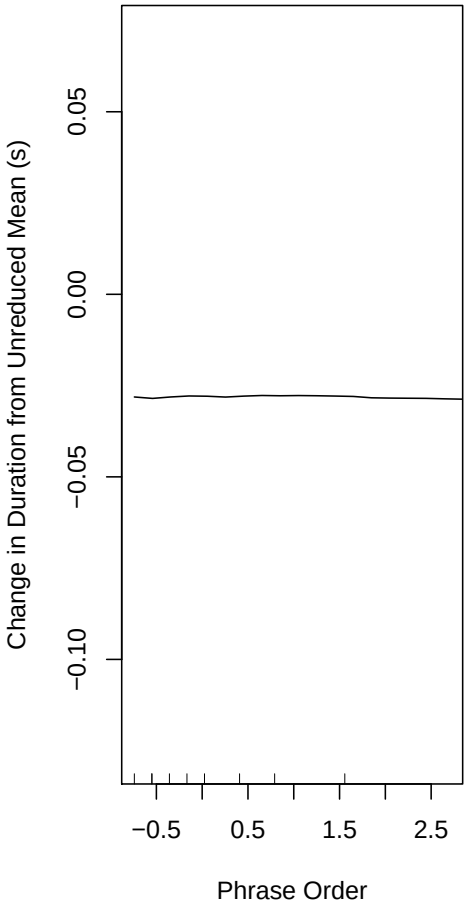

Figure A.12: Partial Effect of Phrase Order in RF Duration Model

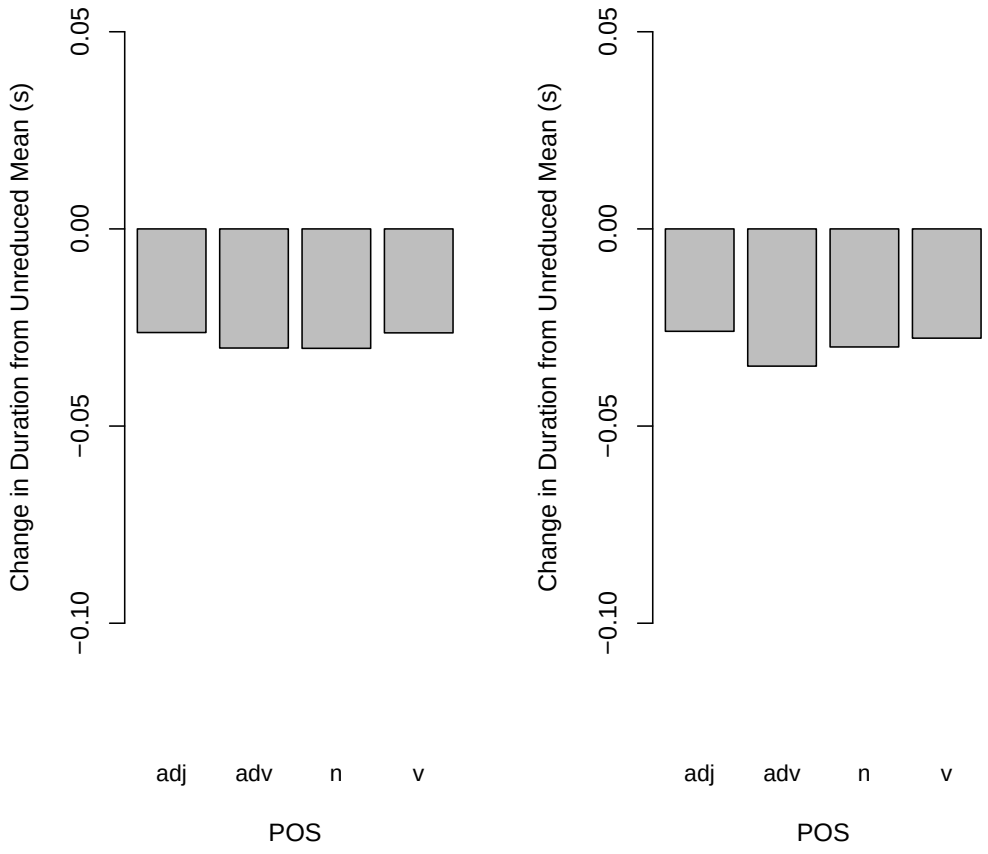

Figure A.13: Partial Effect of Part of Speech in RF (L) and LME (R) Duration Models

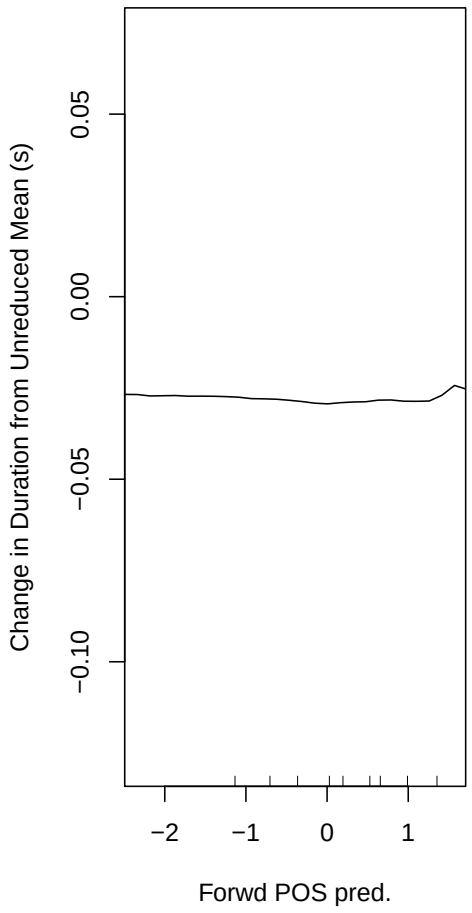

Figure A.14: Partial Effect of Forward POS Predictability in RF Duration Model

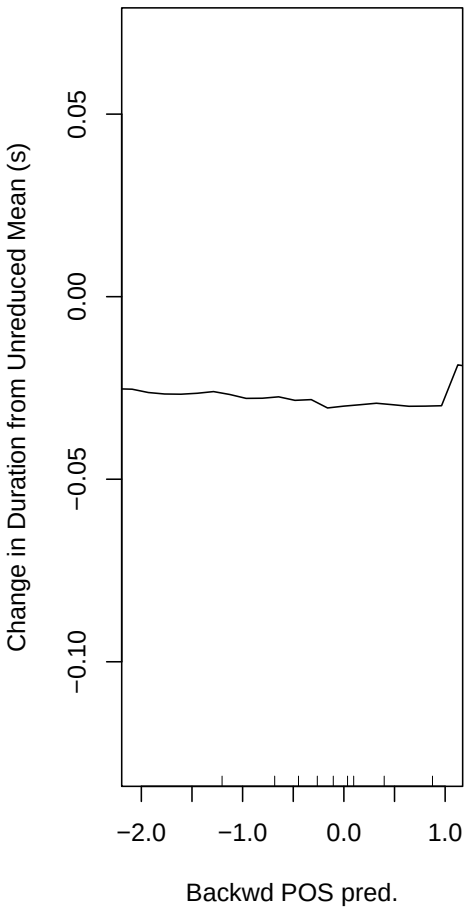

Figure A.15: Partial Effect of Backwards POS Predictability in RF Duration Model

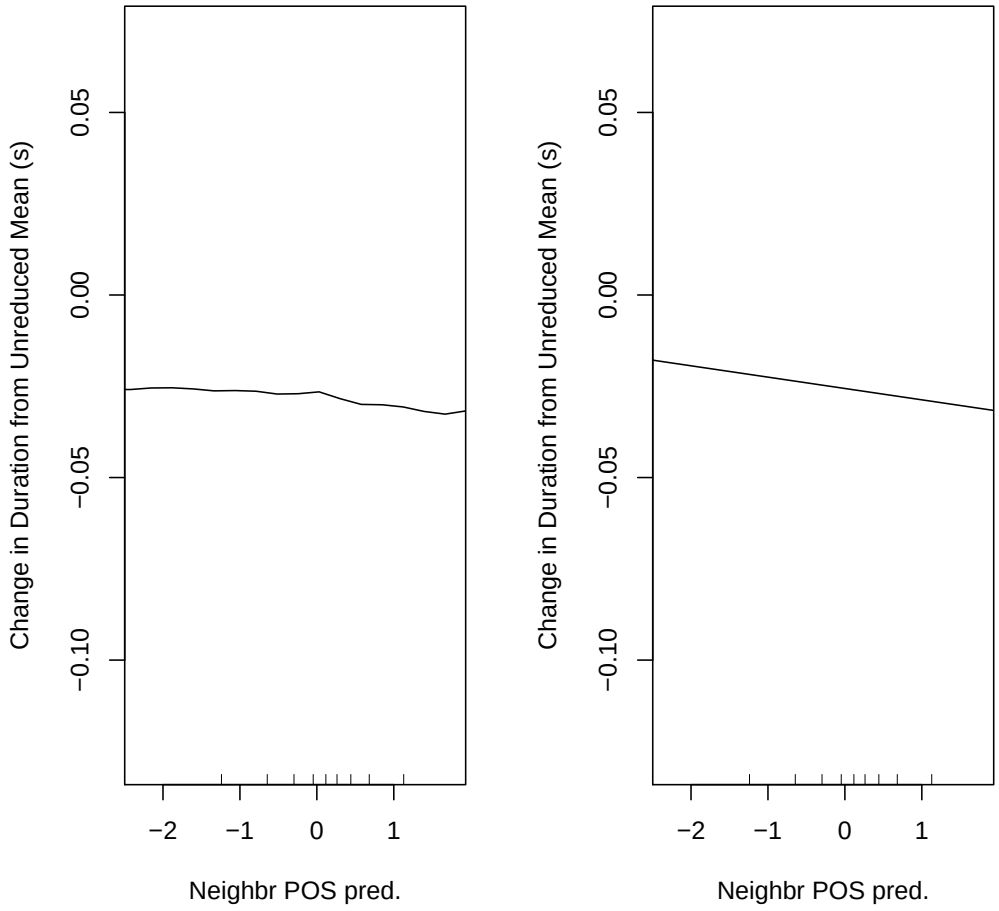

Figure A.16: Partial Effect of Surrounding POS Predictability (Resid.) in RF (L) and LME (R) Duration Models

A.1.5 Topicality

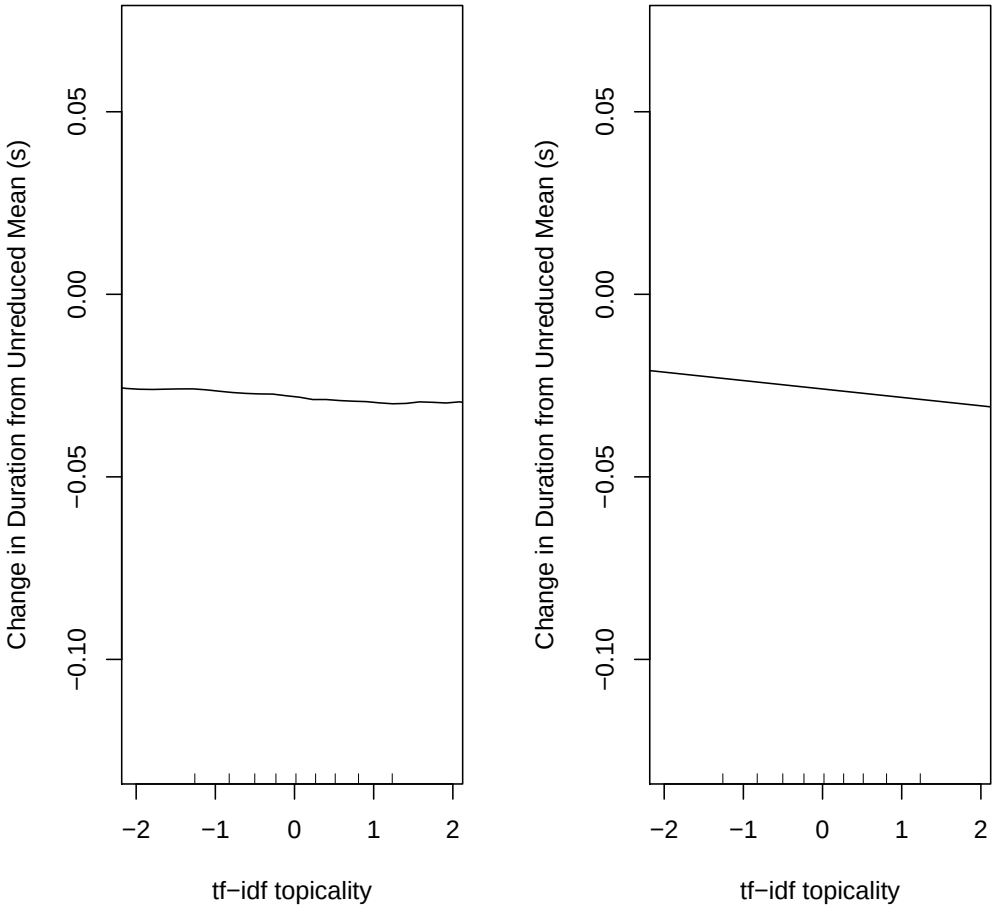

Figure A.17: Partial Effect of TF-IDF Topicality (Resid.) in RF (L) and LME (R) Duration Models

A.1.6 Time

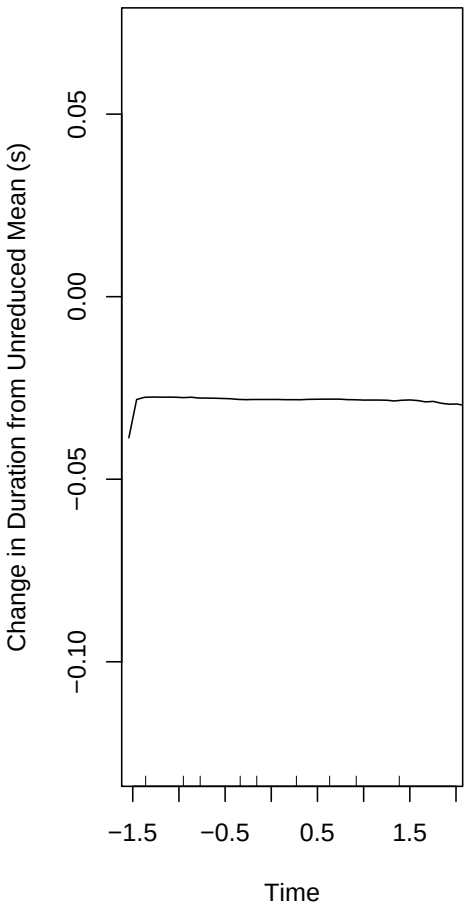

Figure A.18: Partial Effect of Time in Conversation in RF Duration Model

A.2 Segment Deletion Models

A.2.1 Demographic Predictors

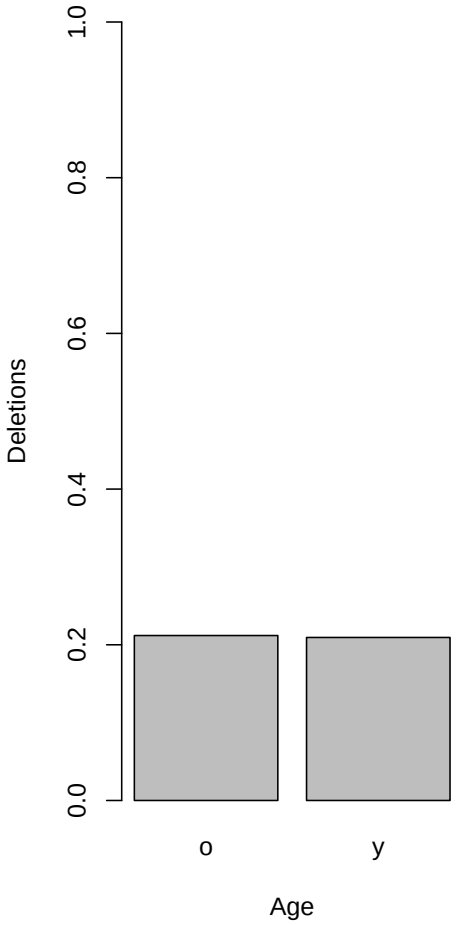

Figure A.19: Partial Effect of Speaker Age in RF Deletion Model

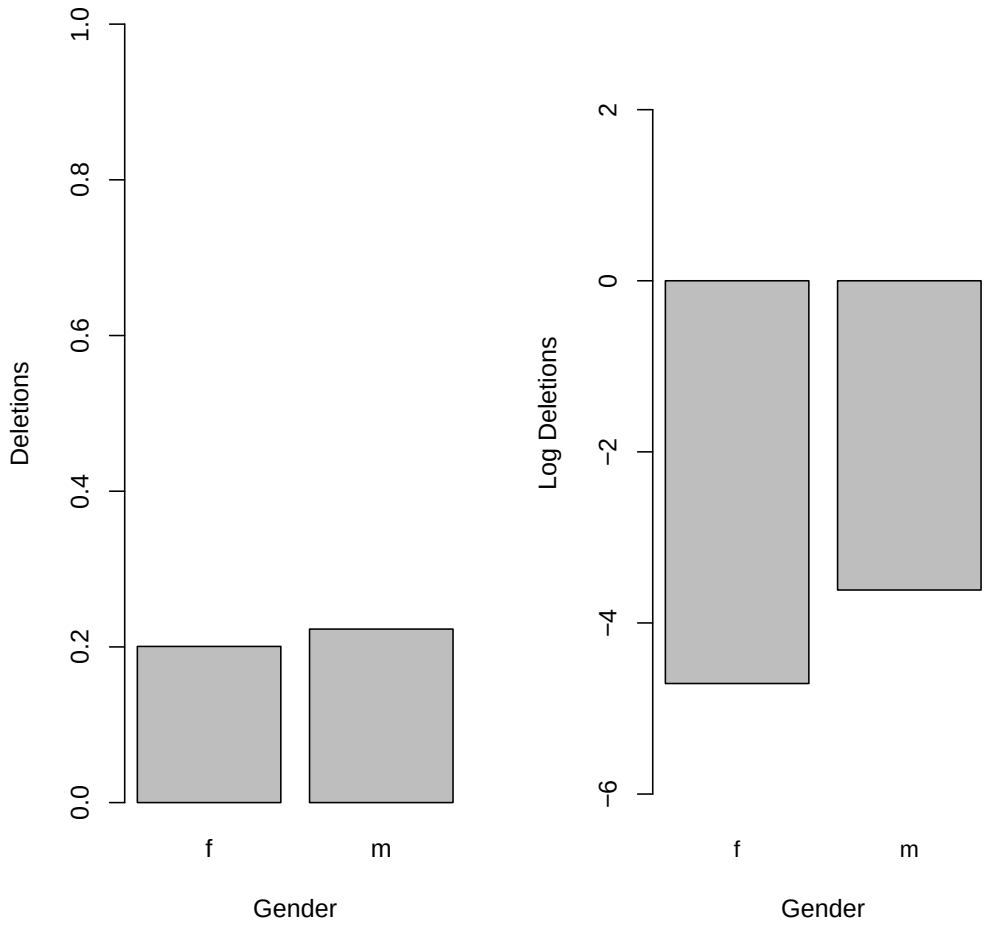

Figure A.20: Partial Effect of Speaker Gender in RF (L) and LME (R) Deletion Models

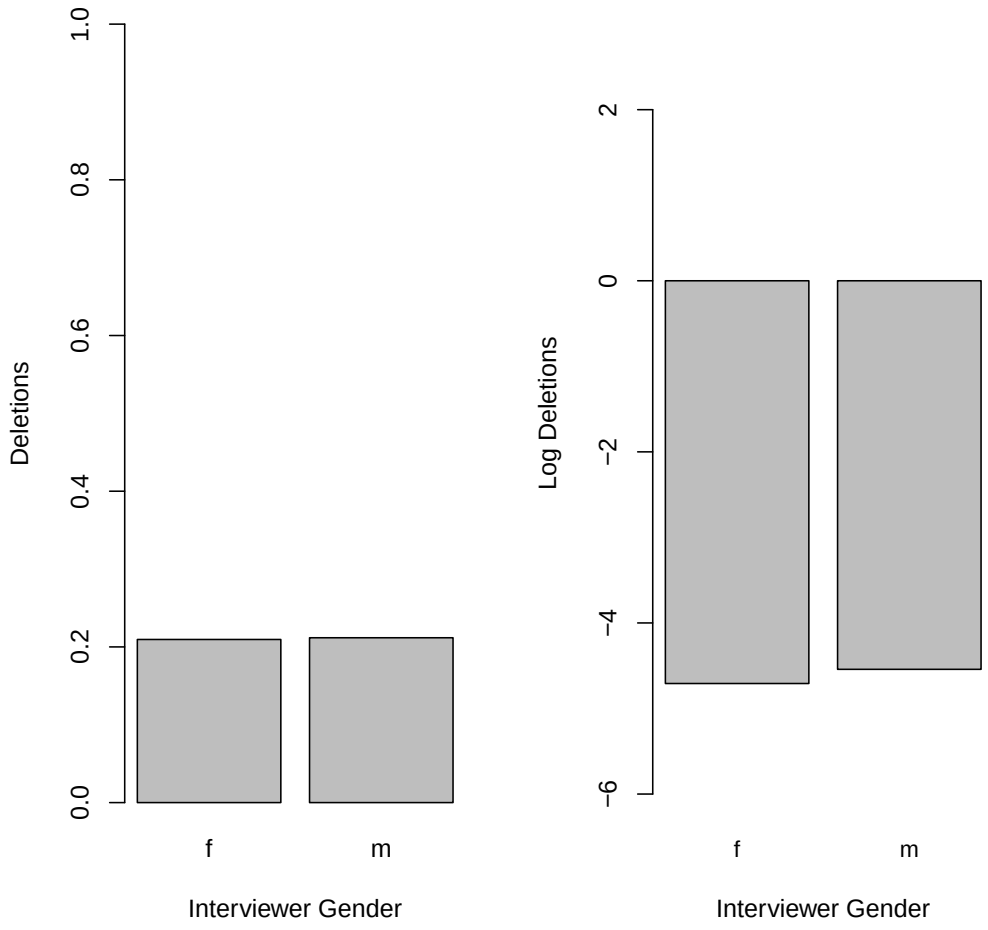

Figure A.21: Partial Effect of Interviewer Gender in RF (L) and LME (R) Deletion Models

A.2.2 Phonological and Phonetic Predictors

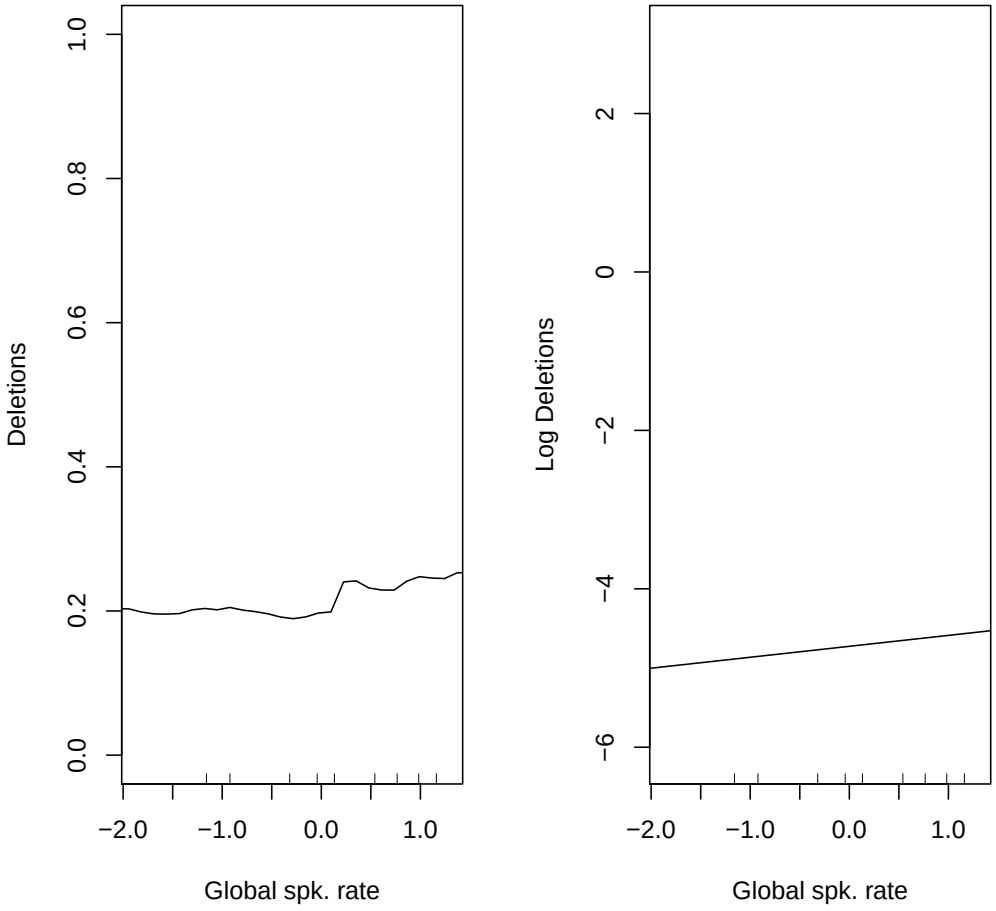

Figure A.22: Partial Effect of Average Speech Rate in RF (L) and LME (R) Deletion Models

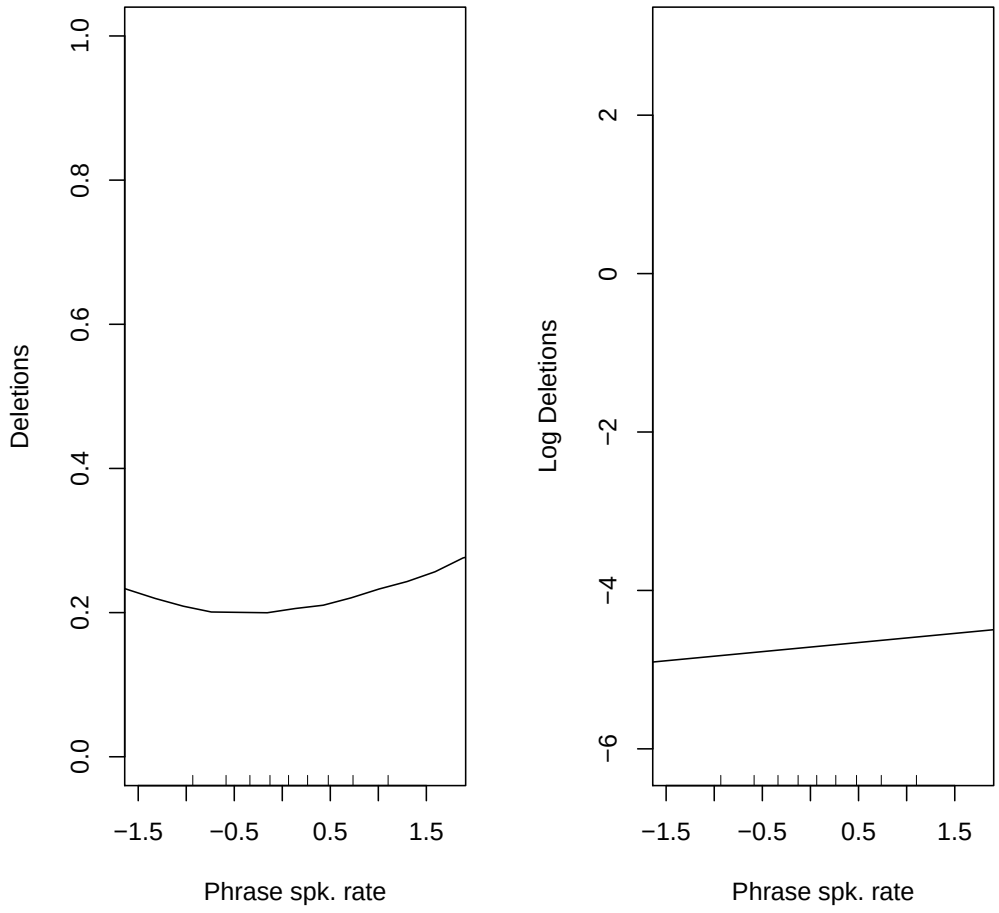

Figure A.23: Partial Effect of Local Speech Rate in RF (L) and LME (R) Deletion Models

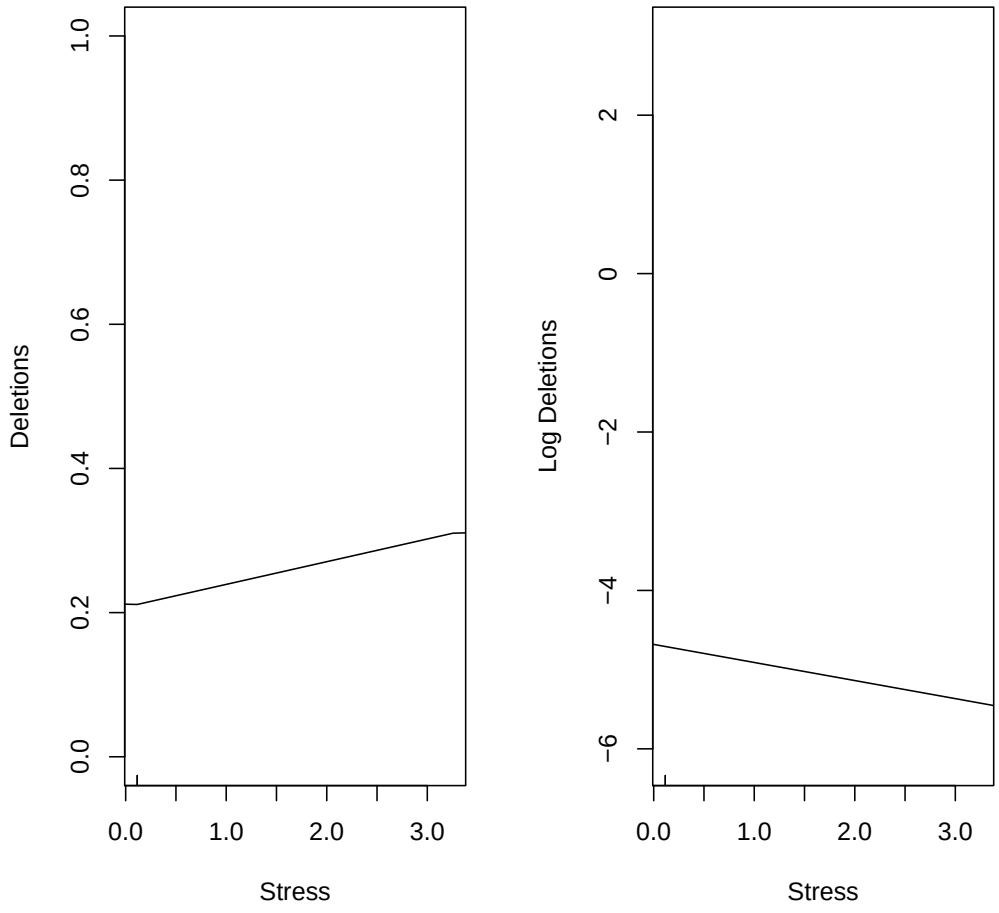

Figure A.24: Partial Effect of Stressed Syllables in RF (L) and LME (R) Deletion Models

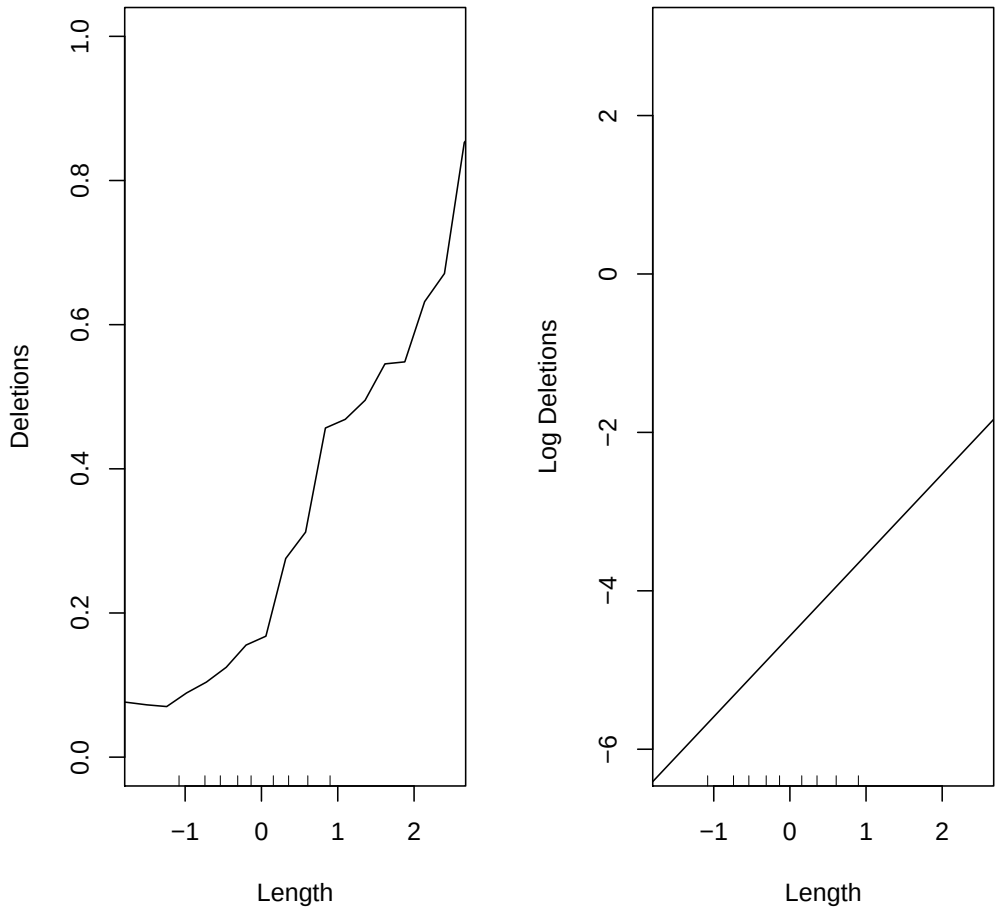

Figure A.25: Partial Effect of Word Length (Resid.) in RF (L) and LME (R) Deletion Models

A.2.3 Predictability

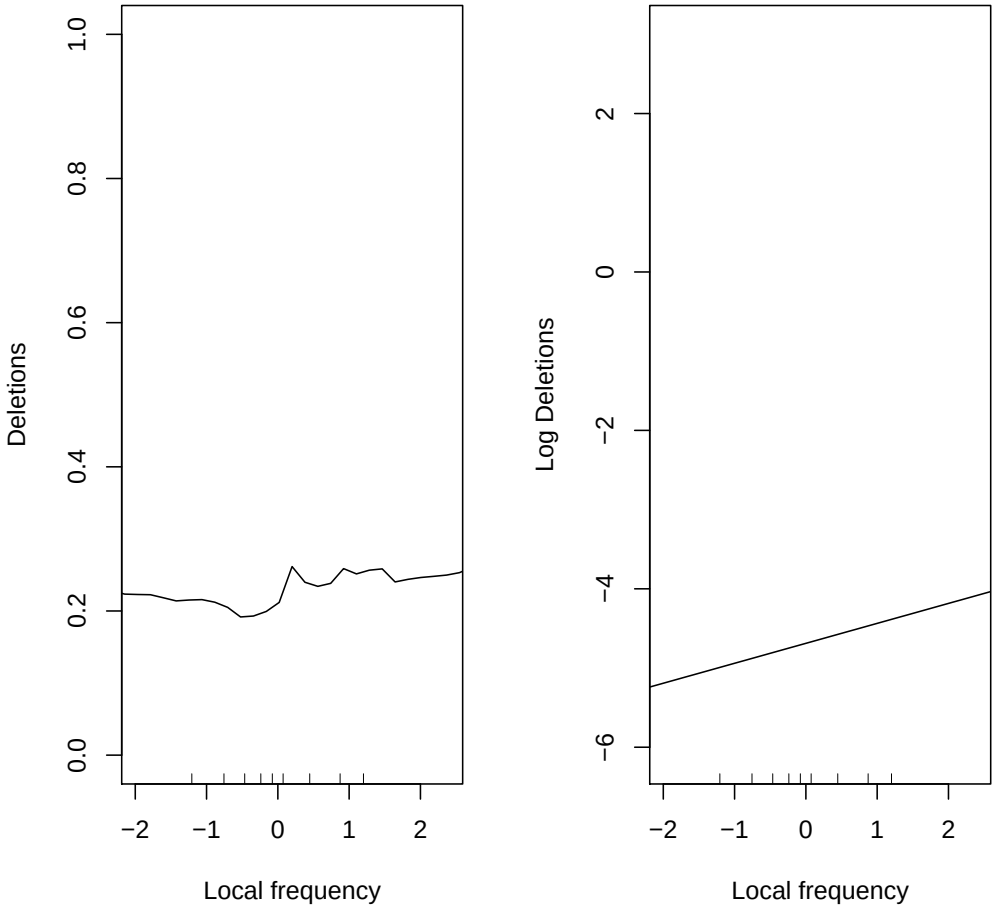

Figure A.26: Partial Effect of Buckeye Frequency (Resid.) in RF (L) and LME (R) Deletion Models

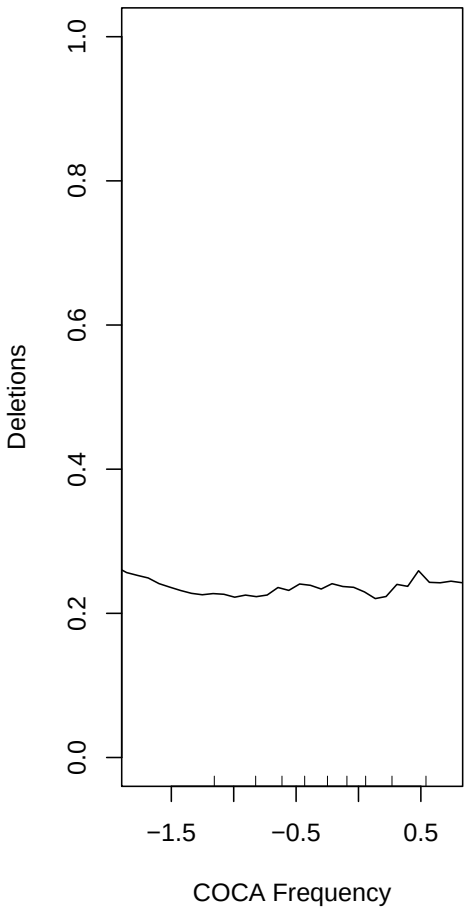

Figure A.27: Partial Effect of COCA Frequency in RF Deletion Model

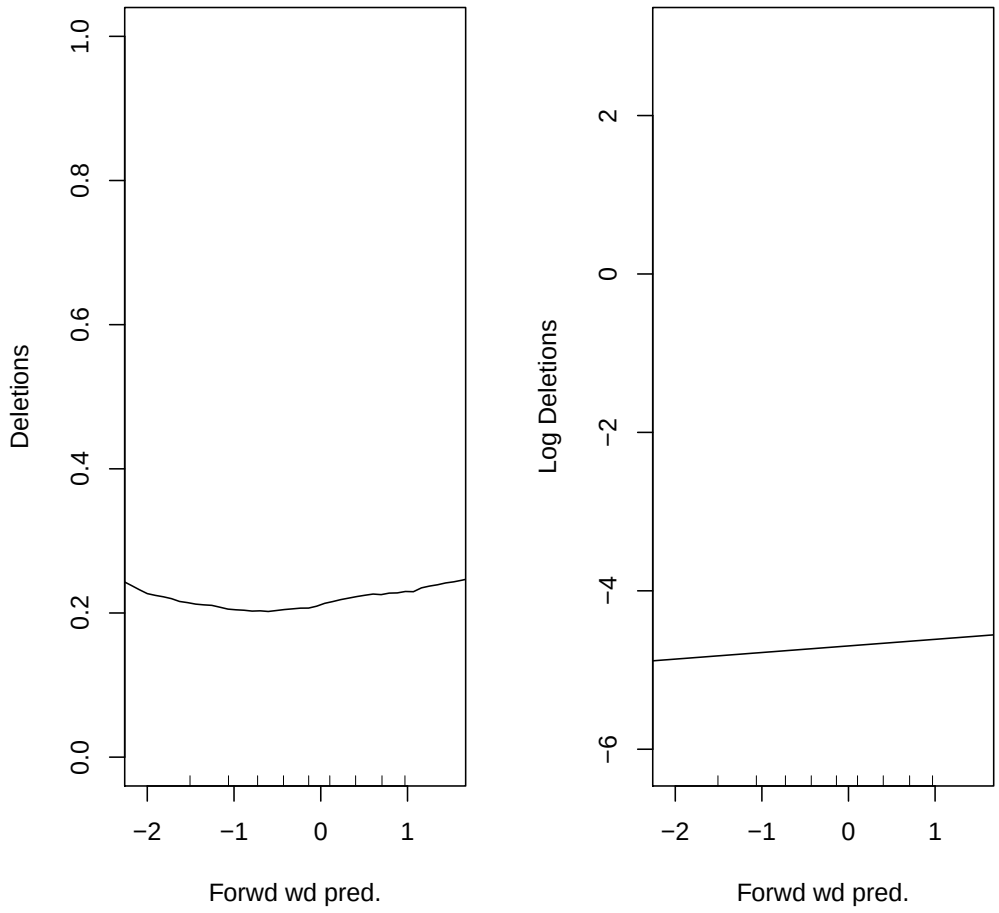

Figure A.28: Partial Effect of Forward Word Predictability in RF (L) and LME (R) Deletion Models

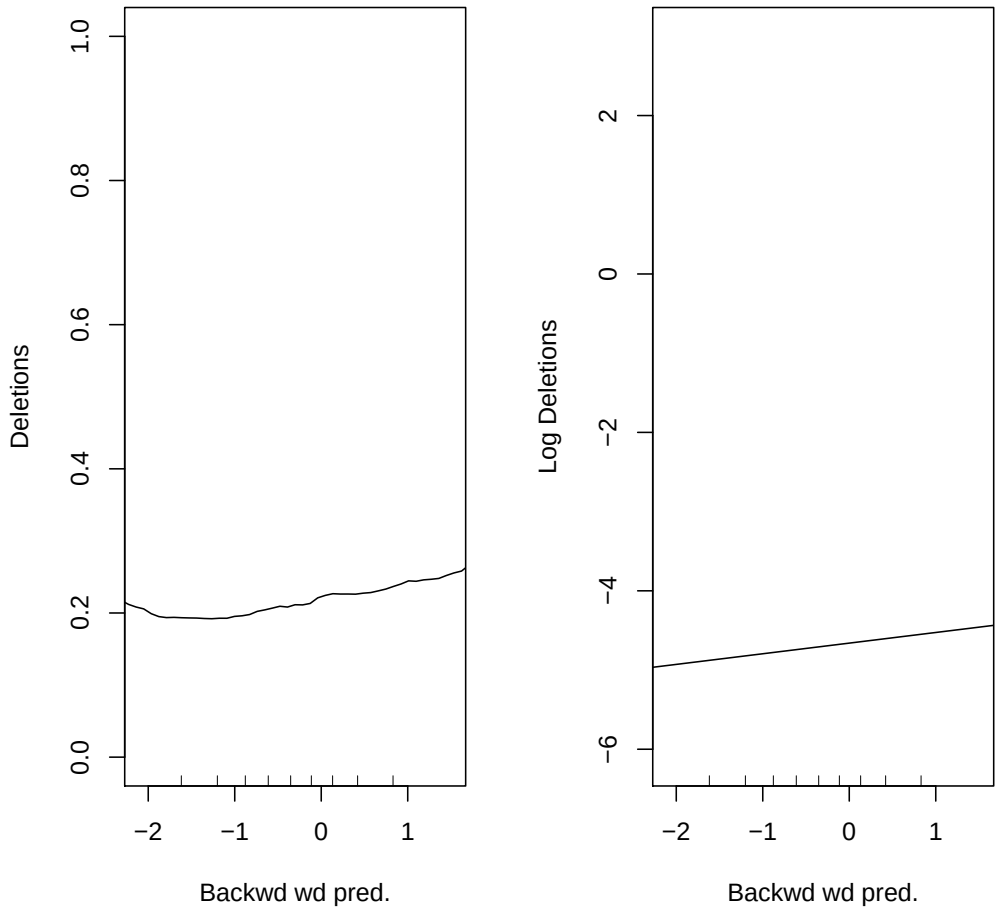

Figure A.29: Partial Effect of Backward Word Predictability in RF (L) and LME (R) Deletion Models

A.2.4 Structural Constituency

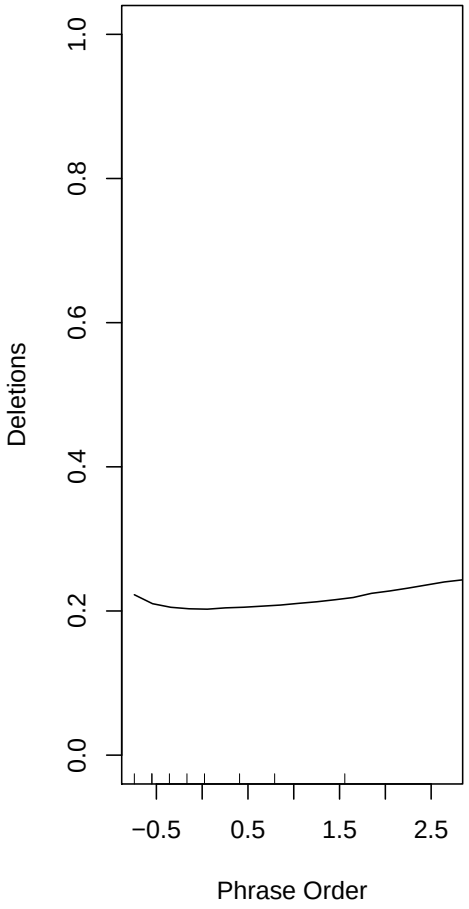

Figure A.30: Partial Effect of Phrase Order in RF Deletion Model

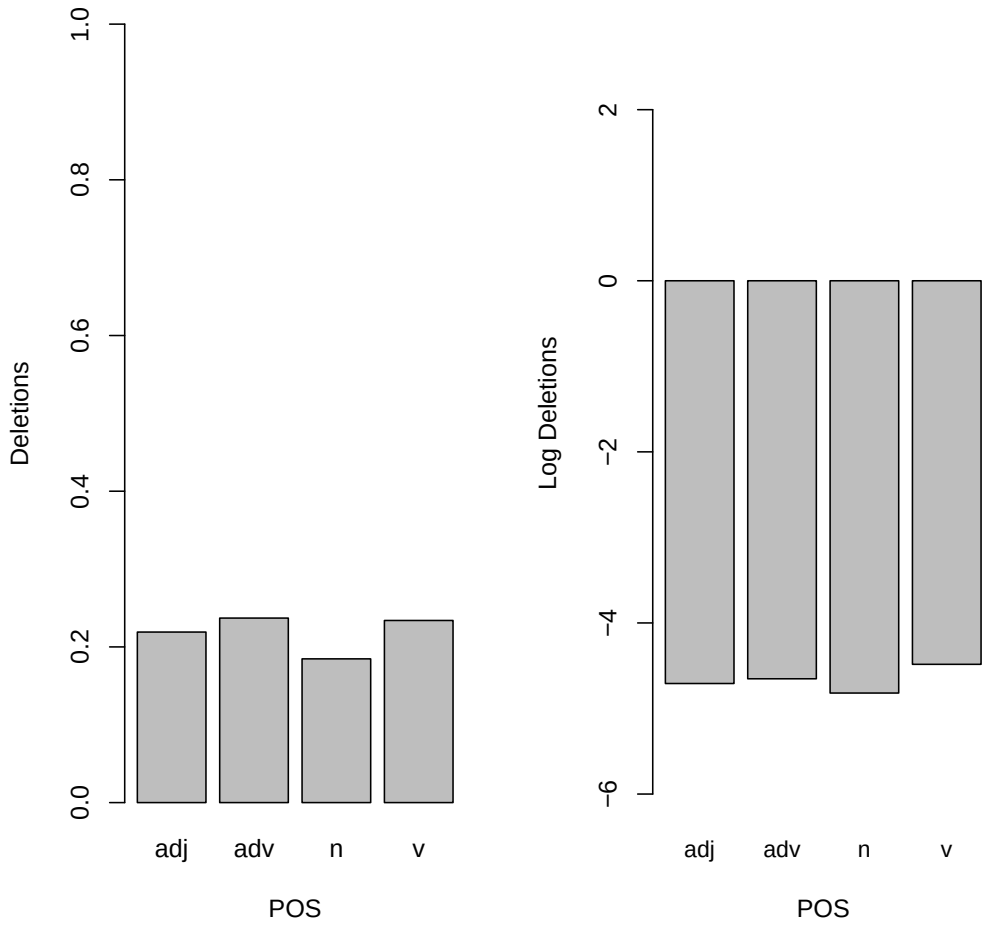

Figure A.31: Partial Effect of Part of Speech in RF (L) and LME (R) Deletion Models

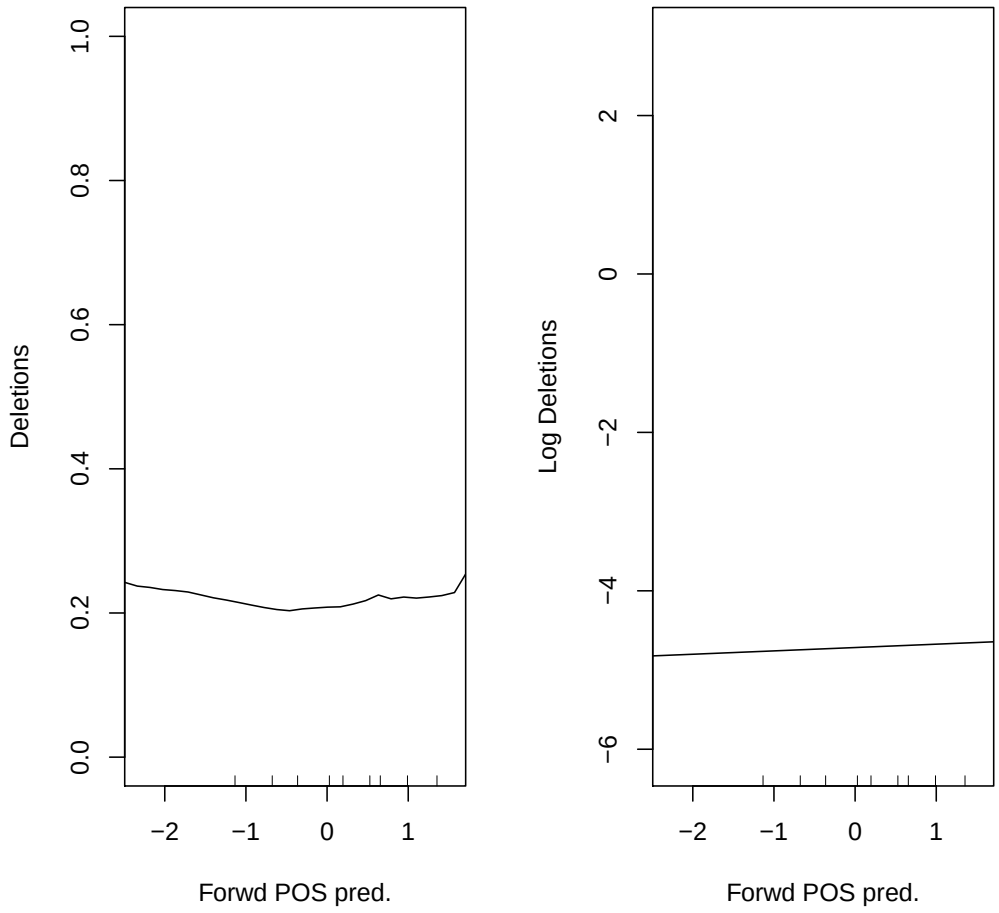

Figure A.32: Partial Effect of Forward POS Predictability in RF (L) and LME (R) Deletion Models

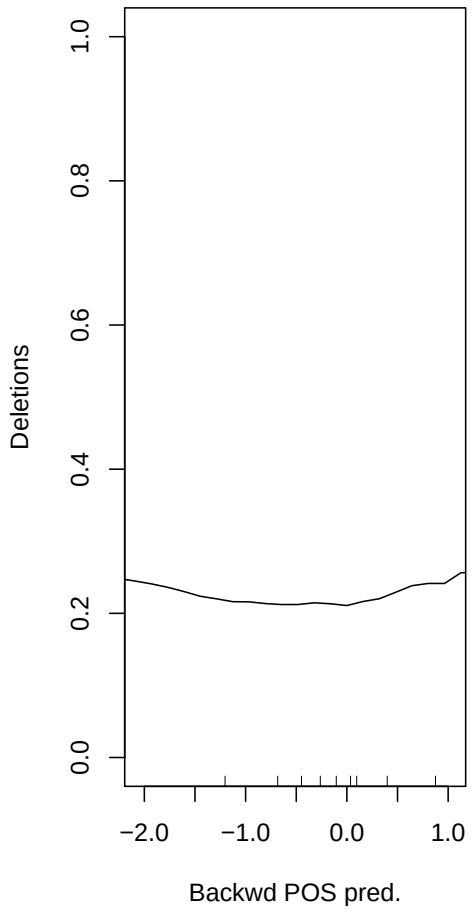

Figure A.33: Partial Effect of Backwards POS Predictability in RF Deletion Model

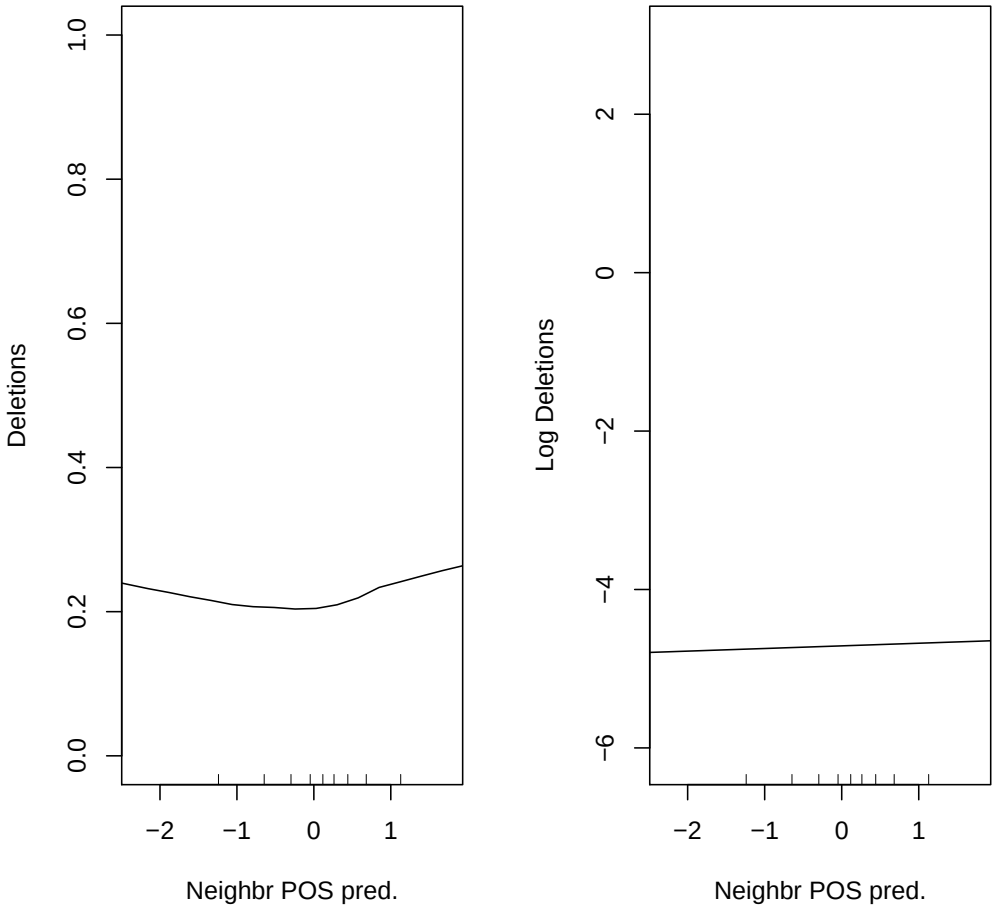

Figure A.34: Partial Effect of Surrounding POS Predictability (Resid.) in RF (L) and LME (R) Deletion Models

A.2.5 Topicality

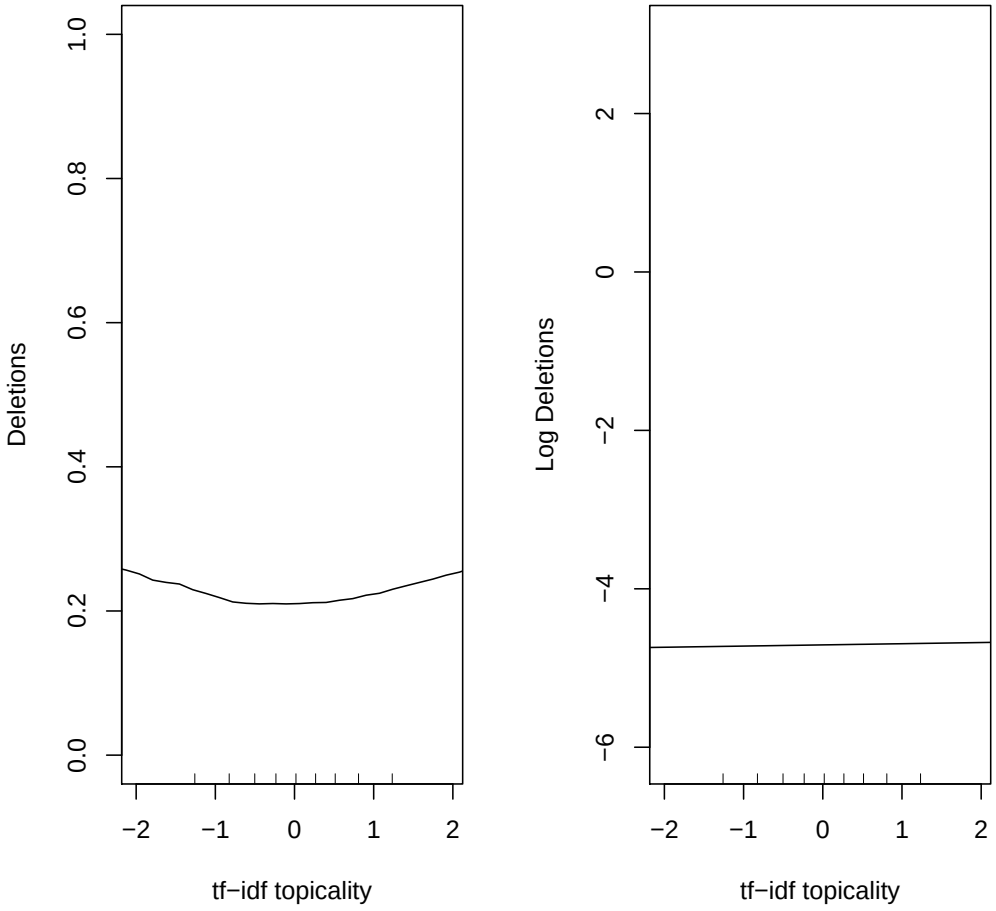

Figure A.35: Partial Effect of TF-IDF Topicality (Resid.) in RF (L) and LME (R) Deletion Models

A.2.6 Time

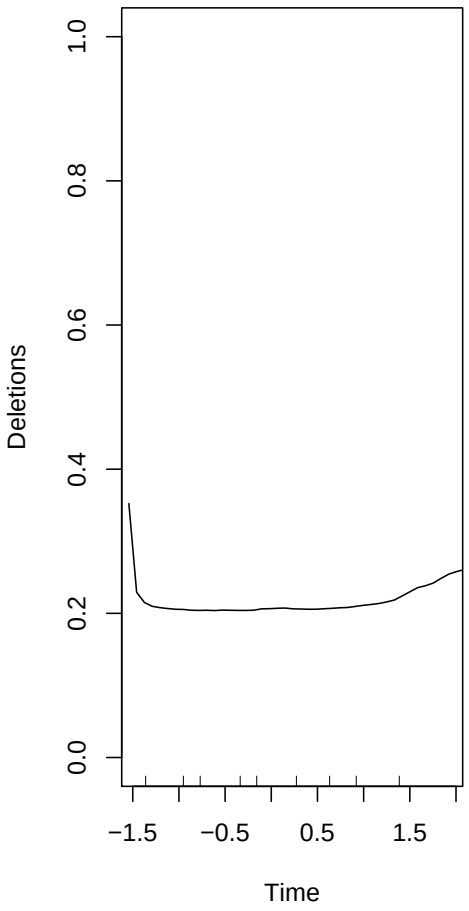

Figure A.36: Partial Effect of Time in Conversation in RF Deletion Model