Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation

Scott Seyfarth

Language-users reduce words in predictable contexts. Previous research indicates that reduction may be stored in lexical representation if a word is often reduced. Because representation influences production regardless of context, production should be biased by how often each word has been reduced in the speaker's prior experience. This study investigates whether speakers have a context-independent bias to reduce low-informativity words, which are usually predictable and therefore usually reduced. Content word durations were extracted from the Buckeye and Switchboard speech corpora, and analyzed for probabilistic reduction effects using a language model based on spontaneous speech in the Fisher corpus. The analysis supported the hypothesis: low-informativity words have shorter durations, even when the effects of local contextual predictability, frequency, speech rate, and several other variables are controlled for. Additional models that compared word types against only other words of the same segmental length further supported this conclusion. Words that usually appear in predictable contexts are reduced in all contexts, even those in which they are unpredictable. The result supports representational models in which reduction is stored, and where sufficiently frequent reduction biases later production. The finding provides new evidence that probabilistic reduction interacts with lexical representation.

1 Introduction

1.1 Probabilistic reduction

In speech production, language-users reduce words when they are predictable in the local context, as well as when they are frequent overall (Lieberman, 1963; Whalen, 1991; Gahl, 2008). This reduction manifests as a broad array of articulatory and acoustic effects, including differences in word and syllable duration, vowel dispersion and quality, plosive voice onset time, syllable deletion, and language-specific segmental deletion, among others (Bell et al., 2003; Aylett and Turk, 2006; Baker and Bradlow, 2009; Hooper, 1976; Bybee, 2002; Jurafsky et al., 2001; Everett et al., 2011; Clopper and Pierrehumbert, 2008; Yao, 2009; Gahl and Garnsey, 2004; Bybee, 2006; Tily et al., 2009; Kuperman and Bresnan, 2012; Demberg et al., 2012; Moore-Cantwell, 2013). These phenomena have been known for over a century (see Bell et al., 2009, for a review), and are usually described together as the *probabilistic reduction hypothesis*—words with higher probability are articulatorily reduced, for a variety of local and global probabilistic measures.

The cause of probabilistic reduction is not fully understood, although it can be accounted for in several different (and compatible) models of speech production. For

example, such reduction may indicate that speakers actively manage their productions to balance audience-design considerations with articulatory efficiency (Lindblom, 1990). Under this theory, speakers hyper-articulate unpredictable words in order to improve listeners' chances of parsing words that they have low expectations for. They hypo-articulate words that listeners can easily predict based on the context, in order to save on articulatory effort. Smooth-signal or uniform-information-density versions of this theory frame this behavior as speakers' preference for keeping a constant rate of information transfer (Aylett and Turk, 2004; Pluymaekers et al., 2005; Levy and Jaeger, 2007). Speakers spend more time on unpredictable words, which are informative, and relatively little time on predictable words, which provide less new information.

An alternative account for probabilistic reduction is based in speaker-internal processing factors (Bard et al., 2000; Munson, 2007; Bell et al., 2009). Under this theory, words are activated more strongly by their phonological, semantic, and syntactic associates. This facilitates retrieval and speeds production (Gahl et al. 2012, cf. Baese-Berk and Goldrick 2009). For example, Kahn and Arnold (2012) show that a linguistic prime causes speakers to reduce a word target even when audience-design factors are controlled for, while a non-linguistic prime does not trigger reduction.

1.2 Is reduction stored in lexical representation?

An important question is whether probabilistic reduction is exclusively an online effect, or whether it is also represented offline in the lexicon. It is generally argued that unreduced citation forms have a privileged representational status (Ernestus et al., 2002; Kemps et al., 2004; Ranbom and Connine, 2007). However, there is evidence that reduced forms are also represented. Lavoie (2002) and Johnson (2007) show that words with homophonous citation forms can have very dissimilar distributions of reduction variants in conversational speech, and each word may in fact have special reduced variants that are unattested for its homophone. For example, [fɻ] and [fə] are attested variants of *for* but not of *four*. This suggests that reduced forms are to some extent word-specific, and therefore associated with lexical representation, rather than created exclusively online during production.

Furthermore, language-users have a processing advantage for common reduced forms of a word. This advantage is relative to how often the word is reduced (Connine and Pinnow, 2006; Connine et al., 2008). For example, French *genou* [ʒənu] is often realized in a reduced form [ʒnu], which lacks an audible schwa. On the other hand, *querelle* [kəʀɛl] is more often realized with a full schwa in the first syllable. In isolated word production, speakers are faster to produce forms like [ʒnu] than [kʀɛl], all else held equal, where [ʒnu] but not [kʀɛl] is a common word-specific reduction (Racine and Grosjean, 2005; Bürki et al., 2010). In lexical decision experiments, Ranbom and Connine (2007) and Pitt et al. (2011) show that listeners are faster to classify reduced forms like English *gentle* [dʒɛɾ̃l], with a nasal flap, than [dʒɛnʔl], where the flap but not the glottal stop is a usual reduction of [t] in words like *gentle*.

These findings indicate that reduced variants, when they are typical realizations of a word, are likely stored in representation (Pitt, 2009; Ernestus, 2014).

There are at least three ways this storage might be implemented. First, storage of reduction might involve multiple phonologically-abstract, categorical variants, which include both unreduced and reduced forms of a word (as described above). Second, individual productions of reduced words might be stored as exemplars with fine-grained phonetic detail, including acoustic reduction (Pierrehumbert, 2002; Johnson, 2007). Third, reduction might be represented indirectly via changes to articulatory timing relations that are lexically specified (Browman and Goldstein, 1990; Byrd, 1996; Lavoie, 2002).

Is probabilistic reduction stored in lexical representation? Reduction associated with high contextual probability is standardly treated as an online phenomenon, such as a kind of priming or else active management of information density, as in 1.1. The evidence discussed here suggests that reduction is stored when it occurs often enough. Therefore, if a word is very often reduced because it typically occurs in high-probability contexts, language-users may store this reduction in lexical representation as well.

1.3 Informativity

In usage, some words almost always occur in predictable contexts, whereas others are unlikely in each of the contexts that they occur in, even though they might be relatively frequent overall. For example, the word *current* usually occurs in the context of *current events* or *the current situation*, and is therefore usually predictable in context. On the other hand, the word *nowadays* has roughly the same log-frequency overall as *current*, but *nowadays* occurs in a wide variety of contexts (see figure 1). Thus, on average, *nowadays* is more unpredictable in each of its contexts.

The average predictability of a word in context is its *informativity* (Cohen Priva, 2008; Piantadosi et al., 2011). Word informativity is formally defined as:

$$-\sum_{c} P(C = c \mid W = w) \log P(W = w \mid C = c) \tag{1}$$

In equation 1, c is a context and w is a word type. Context is usually operationalized simply as the n preceding or following words in an utterance. The informativity of a word type is the averaged probability with which a word will occur given each of the contexts that it can occur in. This average is weighted by the frequency with which the word occurs in each context. Usually-predictable words (like *current*) have low informativity, because they tend to provide less new information in actual communicative use. Usually-unpredictable words (*nowadays*) have high informativity, because in actual use they tend to be surprising and informative.

Because low-informativity words are usually predictable, they are also usually reduced. On the other hand, high-informativity words are rarely reduced. The experiments described in 1.2 demonstrate that reduced forms of a word are more

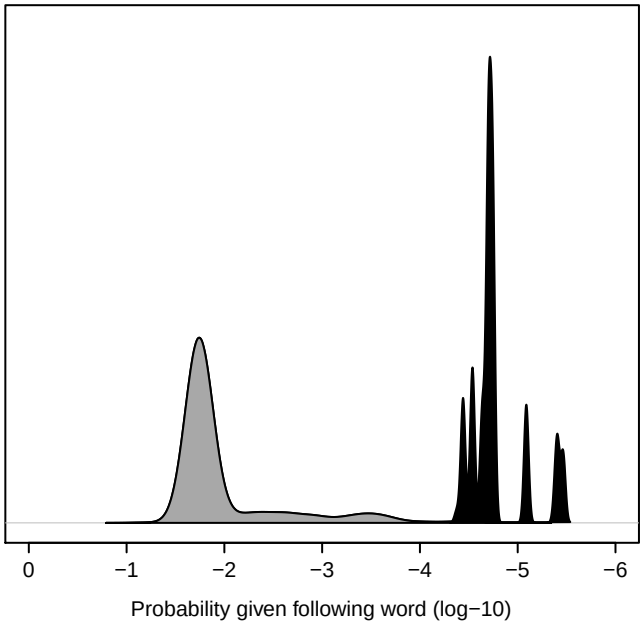

Figure 1: Density plot showing the (log-10) probability of tokens of *current* (gray) and *nowadays* (black) given the following word as context; predictability is higher to the left. Tokens of *current* usually occur in predictable contexts, and so *current* has more density on the left (low informativity); tokens of *nowadays* are usually unpredictable, and so *nowadays* has more density on the right (high informativity). Tokens taken from Fisher; probabilities from COCA.

accessible if a reduced form is a typical realization of that word. If probabilistic reduction is stored, reduction of low-informativity words should be more accessible than reduction of high-informativity words. In a model that assumes abstract variants, reduced variants of low-informativity words should be retrieved and produced in spontaneous speech relatively more often than could otherwise be explained. In general terms, the prediction is that speakers should have a stronger bias for reducing a word if that word is usually reduced and usually predictable elsewhere. The current study evaluates this hypothesis through an analysis of conversational speech corpora.

1.4 Evidence for linguistic informativity effects

1.4.1 Sub-lexical informativity

There is evidence from language-specific phonetics that average contextual predictability does affect sub-lexical representation and processing. Using speech production data from English and Dutch, Aylett and Turk (2004, 2006) and van Son and Pols (2003) show that greater within-word predictability is associated with shorter segment and syllable token duration, which is consistent with the larger pic-

ture of probabilistic reduction. Building on this research, Cohen Priva (2008, 2012) demonstrates that certain features of English segment types are best modeled by the average contextual predictability of each type within words across the lexicon. Consonants that are on average less predictable in context have longer durations and are deleted less often than consonants that are more predictable, even when they do occur in highly predictable contexts. This paper extends these findings from segment realization to word realization.

There is related within-word evidence from perception for the claim that language-users are biased by average contextual probabilities, in addition to local ones. Lee and Goldrick (2008) show that the errors that English and Korean speakers make in recalling nonce words are dependent not only on immediate segment transition probabilities, but also on the typical reliability of different intra-syllable boundaries in each of the two languages. This suggests that the representation of sub-syllabic units (the body and rime) is shaped by the average predictability within each unit.

Together, these results provide some evidence that language-users represent average contextual trends—not just local relationships—and that this representation influences both production and perception below the word level. However, similar processing effects have yet to be shown at the word level.

1.4.2 Word lengths

Piantadosi et al. (2011) and Mahowald et al. (2013) show that word informativity is correlated with word lengths. Words that are usually predictable in context have fewer letters or segments than words that are usually unpredictable. This correlation is independent of—and greater than—the known correlation between word lengths and frequency (Zipf, 1935). Frequent and predictable words are thus both shorter in length and also reduced in production. Several authors have proposed that these two phenomena are connected, and that probability-conditioned reduction leads to permanent representational change (Bybee, 2003; Lindblom et al., 1995; Mowrey and Pagliuca, 1995; Pierrehumbert, 2001).

1.5 The current study

This paper evaluates whether speakers have a bias favoring reduced productions of words that are typically encountered in reduced forms. If context-driven probabilistic reduction is represented at the word level, productions of usually-predictable and therefore usually-reduced words should be more reduced across contexts. This effect should be proportional to each word's informativity, since informativity measures how often that word is reduced online in predictable contexts. On the other hand, if contextual predictability only causes reduction online, usually-predictable words should only be reduced when they are actually predictable.

One way to look for the effect is the following: if a word that is usually predictable occurs in an unpredictable context, it should appear more reduced than

would be otherwise expected for that context. Equivalently, if a word that is usually unpredictable occurs in a predictable context, it should appear less reduced than would be expected for that context, since its representation is biased towards a clear form. This can be operationalized as the effect of word informativity on acoustic duration: high-informativity words—words that are usually unpredictable—should have longer durations than those words which are usually predictable, when all other factors are held equal.

This hypothesis is evaluated on word durations extracted from the Buckeye and Switchboard corpora using a series of linear mixed-effects regression models. Word informativity is included as a variable in the model, and the analysis should show that this variable is significantly associated with greater word duration if the hypothesis is true. The control variables include local probabilistic reduction—it is mathematically true that if words are shorter when they appear in predictable contexts, then the average token of a usually-predictable word will be shorter than the average token of a rarely-predictable word. However, the effects of local reduction are included as a separate parameter in the model. Therefore, if low-informativity words are only shorter on average because most or all of their tokens occur in predictable contexts, there will be no independent effect of the type-level informativity variable: token contextual probability will capture all of the possible variation arising from context.

Thus, once local probability is factored out of every word token's duration, a non-significant informativity effect would suggest that words have durations that can ultimately be derived from length, segmental content, syllable count, raw frequency, etc. On the other hand, if low-informativity words are shorter across all contexts, regardless of local token probability, then informativity should capture a significant amount of the remaining variance.

2 Materials and methods

2.1 Word duration data

Word durations were extracted from two sources of natural English speech, which were analyzed separately: the Buckeye Corpus of Conversational Speech (Pitt et al., 2007) and the NXT Switchboard Annotations (Calhoun et al., 2009) based on Switchboard-1 Release 2 (Godfrey and Holliman, 1997).

The Buckeye Corpus is a collection of interviews conducted around 1999–2000 in Columbus, Ohio. There are forty speakers in the corpus, each of whom was recorded for about one hour under the initial pretense that they were participating in a focus group on local issues. The speakers were balanced for age and sex, but all were white natives of Central Ohio belonging roughly to the middle and working class. The corpus itself contains the original audio recordings as well as several types of transcriptions and annotations. Word durations for this study were taken from the

timestamps provided for the word-level annotations. Additionally, each word token is annotated with two different segmental transcriptions. First, each token includes a dictionary-based transcription, which is the canonical or citation form for the word type, generated from automatic alignment software. Second, each individual token includes a close phonetic transcription, created by an annotator who hand-corrected the software-generated segments and timestamps for each token.

The Switchboard Corpus is a collection of telephone conversations conducted as a corporate research project in 1990–1991. The NXT-formatted subset used here included 642 annotated conversations between 358 speakers (Calhoun et al., 2010). Speakers from all areas of the United States were recruited through internal corporate and government listservs, through public electronic bulletin boards, and by peer-to-peer recruitment. No effort was made to balance speakers by age, sex, region, or socioeconomic class. Speakers were assigned to conversations with individuals they had not previously spoken with, and pairs were provided with one of 70 general-interest conversation topics selected by an automated operator. Word durations for this study were taken from the corpus annotation timestamps, which were created by hand-correction of automated word alignment. Below the word level, only dictionary-based segmental transcriptions were available for the words in Switchboard.

Only content words were included in the analysis, as it has been shown that function and content words respond differently to predictability effects (Bell et al., 2009), and function words are generally considered to be processed differently than content words (e.g., Levelt et al., 1999). Some content word tokens were excluded from the analysis for prosodic or other reasons, following standard practice (e.g., Bell et al., 2003, 2009; Gahl et al., 2012; Jurafsky et al., 2001, 2002). For the purposes of the following exclusions and for calculating speech rate, an utterance was defined as a stretch of speech by a single speaker that is delimited by pauses, disfluencies, or other interruptions greater than or equal to 500 milliseconds. Tokens were excluded if they were adjacent to a disfluency, a pause, or a filled pause; if they were utterance-initial or utterance-final; if the word was cliticized (e.g., *cousin's*); if the word type or bigram context was not found in the language model; if the utterance speech rate was more than 2.5 standard deviations away from the speaker's mean; if the word token duration was more than 2.5 standard deviations away from the word type's mean; or if there were no vowels or syllabic consonants in the word token's close phonetic transcription (in Buckeye). In the Buckeye Corpus, all data from speaker 35 were excluded due to a large number of transcription and alignment errors (Gahl et al., 2012).

2.2 Probabilistic language model

To test a hypothesis about word predictability, it is necessary to estimate inter-word probabilities from a source that belongs to the same genre of language to be analyzed. Research has shown that word probabilities estimated from corpora of

the same language register as the one to be modeled are much better associated with different word processing variables than probabilities estimated from larger but dissimilar corpora (Brysbaert et al., 2011; Brysbaert and New, 2009; Francom and Ussishkin, submitted). The Fisher English Training Part 2 Transcripts have a history of use by previous researchers looking to estimate probabilities in natural conversational speech (e.g., Arnon and Snider, 2010). Furthermore, they are a good genre-of-speech match with Buckeye and Switchboard, since all three corpora involve recorded informal conversations between individuals who are meeting for the first time.

The Fisher Part 2 corpus is a collection of English telephone conversations created at the Linguistic Data Consortium to aid speech-to-text dialogue systems (Cieri et al., 2005). There are 5,849 conversations of up to ten minutes each, totaling over 12 million words. Speakers were randomly assigned a conversation partner and one of 100 topics, with an effort to balance speakers by sex, age (although speakers older than 50 are under-represented), and geographical region (roughly one-fifth from the US North, Midland, South, and West dialect regions; with one-fifth from Canada or speaking non-US or non-native English varieties). Each speaker participated in usually 1–3 conversations in order to maximize inter-speaker variation within the corpus, and topics were selected for a range of vocabulary. The contents of Buckeye, Switchboard, and Fisher do not overlap.

Two bigram language models were calculated based on the Fisher transcripts. These models list the probabilities that each word will occur, given either the word before it (in one model) or the word after it (in the second model). Bigram models are standardly used in studies of predictability-based phonetic reduction, and focused research on predictability measures has shown negligible improvement in predicting reduction from trigram or more complicated models (Jurafsky et al., 2001). The probabilities were smoothed with the modified Kneser-Ney method described by Chen and Goodman (1998), using the SRILM Toolkit (Stolcke, 2002; Stolcke et al., 2011) with smoothing parameters optimized by the toolkit. The final estimates were used as measures of local contextual predictability. All other probabilistic measures were also estimated from the Fisher transcripts.

2.3 Variables

Each word token in the two corpora was annotated based on both the Fisher probability data as well as the type- and token-specific variables described here. First, word informativity for each word type w was calculated using equation 1, with probabilities taken from the smoothed Fisher language models. Two estimates of word informativity were calculated, with context c taken as either the preceding word or following word, respectively. Word informativity here is therefore the average of a word's bigram probability across the contexts that it occurs in, weighted by how frequently it occurs in each of those contexts. Informativity was calculated in bans, which uses log base 10.

Previous research has demonstrated that predictability given the following n-gram context is associated with greater and more reliable reduction effects than predictability given the preceding n-gram context. In fact, predictability given preceding context has often been reported as failing to reach significance in predicting English duration reduction (Jurafsky et al., 2001; Bell et al., 2009; Gahl, 2008), except for high-frequency function words. However, informativity in written corpora is usually calculated based on the preceding context (Piantadosi et al., 2011), and so preceding informativity is also included here.

Tokens were also annotated for a variety of control variables taken from previous literature on models of word duration. The collected data were analyzed with a series of linear mixed-effects models containing these variables as parameters, in order to evaluate the direction and significance of the association between informativity and word duration. The exact statistical procedure used to analyze the data is described in section 2.4.

Word durations, and all continuous control variables, were log-transformed (base 10) and centered around their respective means within each model. The distributions of these variables were found to be more normal in log space, and this practice follows previous research (e.g. Bell et al., 2009; see Kuperman and Bresnan, 2012 for discussion). Informativity results are qualitatively the same with and without log-transformation. In addition to the variables described in this section, each model listed below also includes per-word random intercepts, per-speaker random intercepts, and correlated per-speaker informativity slopes as controls for individual word-type and speaker idiosyncrasies.

Baseline duration In order to calculate a baseline expected duration for each word token, the Modular Architecture for Research on Speech Synthesis (MARY) text-to-speech system (Schröder and Trouvain, 2003) was used to analyze each utterance. This baseline was selected to control for the segmental length, content, and context of each word form. This method follows Demberg et al. (2012), who also use durations from the MARYTTS system as a baseline control when they evaluate the effects of syntactic predictability. Previous work on predictability effects has also used orthographic length, syllable count, simple segmental length, expected word durations estimated by summing average segment durations, and/or per-word random intercepts as statistical controls for word form length and content. Alternative analyses using different baselines measures are discussed in section 3.3.4.

The cmu-slt-hsmm voice package was used to calculate acoustic parameters for each segment and word token. This package has been trained on part of the CMU ARCTIC database (Komineck and Black, 2003) to estimate segment and word durations based on the phonological features of the current and adjacent segments, syllable structure and position, and word stress. To some extent, the package also models phrase accents and prosody, based on part-of-speech and utterance boundaries indicated with punctuation.

Utterances were sent to the MARYTTS system for analysis, and word durations were extracted from the realized acoustic parameters for each utterance. Demberg et al. (2012) show that this method of analyzing full utterances, which allows sentential context to be included, generates word duration estimates that are superior to single-word analysis. The final baseline estimates were log-transformed and centered.

Syllable count Number of syllabic segments in the word type's transcription; log-transformed and centered.

Speech rate Number of syllabic segments per second in each utterance; log-transformed and centered.

Bigram probability Two variables for the conditional probability of a word given the previous or following word, as estimated from the smoothed language models; log-transformed and centered.

Word frequency Raw token count of a word in the Fisher transcripts; log-transformed and centered.

Orthographic length Number of letters in the word's orthography; log-transformed and centered. Previous research has indicated that orthographic length may have an independent effect on word duration (Warner et al., 2004; Gahl, 2008), and it is important to include this variable to control for the previously-observed association between orthographic length and word informativity (Piantadosi et al., 2011).

Part of speech Coded as noun, verb, adjective, or adverb based on the annotations provided in the Buckeye and Switchboard corpora. Other parts of speech and proper nouns were excluded. In the models, this variable was treatment-coded with noun as the base level.

2.4 Model procedure

Linear mixed-effects models were fit to the Buckeye and Switchboard word duration data, using the full set of variables as predictors. Analysis was conducted using the lme4 package in R (Bates et al., 2013; R Core Team, 2013).

To help guard against model overfitting, backward model selection was done to remove predictors that did not significantly improve a model. Following this procedure, after each full model was fit, it was compared against a set of models that each had one fewer predictor. Each model in this set had a different predictor removed. If the full model was not significantly better than each of these models ($\alpha = 0.15$ based on log-likelihood fit), the predictor that contributed the least improvement to fit was removed from the full model. These steps were repeated until the final model was significantly better than all possible alternatives with one fewer predictor. The

final model was then compared against the original to confirm that it fit the data as well as the original. Final p-values for each effect were calculated by log-likelihood ratio tests that compared the fit of the final model with and without each variable.

For each corpus, a model was fit to all valid content word tokens from that corpus. However, word informativity is highly correlated with segment count (Piantadosi et al., 2011). Since duration is also predicted by segment count, this raises the concern that any effect of informativity on duration might be simply because informative words tend to have more segments. As a precaution against this confound, a baseline expected duration is included as a control variable, as described in 2.3. However, as a further precaution, additional models were fit over content words matched for a single length in each corpus. For example, one model was fit only to words with two segments, another model was fit only to words with three segments, etc. In this way, words could be compared exclusively with other words of the same segmental length. If informativity has an independent effect on word duration beyond simply the association with segment count, it should show up in every one of these models as well.

3 Results

3.1 Study 1: Buckeye

The data from Buckeye included 41,167 word tokens meeting the inclusion criteria, distributed among 3,429 types. Two predictors were found not to significantly improve fit, and were removed in order: (1) informativity given the previous word ($p > 0.5$), and (2) word frequency ($p > 0.5$). Random per-speaker informativity slopes provided a significant improvement in fit ($p < 0.0001$), and were retained the model. The final fixed and random effects estimates appear in tables 1 and 2, respectively, with durations in base-10 log seconds. Correlations between each pair of continuous variables appear in appendix table 1, and density plots showing the distribution of probabilistic variables appear in figure A.1.

Crucially, informativity given the following word was found to be significantly associated with word duration. Six additional models were fit over subsets of words grouped according to their dictionary transcription length, so that informativity effects on duration could be evaluated independently of segment count. A summary of the results appears in table 3. As before, informativity given the following word was found to be reliably associated with duration for words of up to seven segments long. Informativity given the previous word reached significance in the predicted direction for 2–segment words, but was non-significant for all other lengths. Predictors that did not improve fit and were removed from each additional model are listed in appendix table 2. For every model with a significant informativity effect in table 3, informativity was a reliable predictor (at least $p < 0.05$) both before and after the non-significant predictors were removed.

	β	SE	t	$p(\chi^2)$
Intercept	0.0257	0.0057	4.48	—
Baseline duration	0.5879	0.0150	39.32	< 0.0001
Syllable count	0.0592	0.0104	5.71	< 0.0001
Speech rate	−0.3406	0.0077	−43.97	< 0.0001
Bigram prob. given previous	−0.0102	0.0007	−15.00	< 0.0001
Bigram prob. given following	−0.0205	0.0007	−30.55	< 0.0001
Orthographic length	0.0437	0.0167	2.62	0.0089
Part of Speech = Adjective	0.0033	0.0032	1.04	(< 0.0001)
Part of Speech = Adverb	−0.0172	0.0042	−4.09	—
Part of Speech = Verb	−0.0275	0.0022	−12.54	—
Informativity given following	0.0244	0.0023	10.77	< 0.0001

Table 1: Fixed effects summary for model of Buckeye word durations.

Random effect	SD	Cor.
Word (intercept)	0.043	—
Speaker (intercept)	0.033	—
Speaker (inf., following)	0.007	0.061
Residual	0.098	—

Table 2: Random effects summary for model of Buckeye word durations.

			Inf. given previous				Inf. given following			
# Seg	Types	Tokens	β	SE	t	$p(\chi^2)$	β	SE	t	$p(\chi^2)$
2	51	2,734	0.0569	0.0217	2.62	0.0116	0.0805	0.0210	3.83	0.0002
3	536	13,282	—	—	—	—	0.0271	0.0050	5.47	0.0001
4	737	12,552	—	—	—	—	0.0198	0.0044	4.53	0.0001
5	660	4,660	—	—	—	—	0.0347	0.0046	7.50	0.0001
6	494	3,324	0.0143	0.0075	1.89	0.0616	0.0335	0.0080	4.22	0.0001
7	379	2,260	—	—	—	—	0.0252	0.0060	4.19	0.0001

Table 3: Informativity results for Buckeye models that compared words matched for segment count. # Seg = segment count; Types = content word types in Buckeye with that number of segments; Tokens = content word tokens with that number of segments.

3.2 Study 2: Switchboard

The data from Switchboard included 107,981 word tokens meeting the inclusion criteria, distributed among 4,997 types. Of the ten predictors, only word frequency did not significantly improve the model at $\alpha = 0.15$ and was removed ($p > 0.7$). The

fixed effects summary is given in table 4, and the random effects summary appears in table 5 (all durations in base-10 log seconds). Random per-speaker informativity slopes significantly improved the fit of the model ($p < 0.0001$). Correlations between each pair of continuous variables, and density plots for probabilistic variables, are given in the appendix. As in Buckeye, informativity given the following word captured a significant amount of duration variance ($p < 0.0001$). In this model, informativity given the previous word also reached statistical significance ($p < 0.05$).

A summary of the results of duration models over words matched for segment count appears in table 6. The results from Switchboard replicated those from Buckeye: informativity given the following word was significantly associated with word duration in the predicted duration for words of 2–7 segments. Informativity given the previous word reached significance for words of seven segments, but was eliminated or marginal for all other lengths. A complete list of non-significant predictors that were pruned from each model is given in appendix table 4. For every model in which informativity was statistically significant after non-significant predictors were removed, it was also significant when these predictors were retained in the model, except for the model of 2–segment words ($p < 0.11$ before non-significant predictors were removed).

	β	SE	t	$p(\chi^2)$
Intercept	0.0287	0.0023	12.62	—
Baseline duration	0.5363	0.0102	52.34	< 0.0001
Syllable count	0.0492	0.0070	7.01	< 0.0001
Speech rate	−0.3260	0.0044	−74.79	< 0.0001
Bigram prob. given previous	−0.0082	0.0005	−18.17	< 0.0001
Bigram prob. given following	−0.0227	0.0004	−53.91	< 0.0001
Orthographic length	0.1343	0.0115	11.69	< 0.0001
Part of Speech = Adjective	−0.0051	0.0021	−2.36	(< 0.0001)
Part of Speech = Adverb	−0.0186	0.0026	−7.18	—
Part of Speech = Verb	−0.0410	0.0017	−23.87	—
Informativity given previous	0.0040	0.0016	2.48	0.0131
Informativity given following	0.0142	0.0016	8.72	< 0.0001

Table 4: Fixed effects summary for model of Switchboard word durations.

3.3 Additional description and exploratory analysis

Figure 2 shows a sample of word types that appeared in the analysis. Log-frequency appears on the vertical axis and informativity given the following word is on the horizontal axis, with both measures calculated from the Fisher language model. *Nowadays* appears on the right side of the chart, with an informativity of 3.93 bans.

Random effect	SD	Cor.
Word (intercept)	0.039	—
Speaker (intercept)	0.028	—
Speaker (inf., previous)	0.006	−0.158
Speaker (inf., following)	0.007	0.089
Residual	0.100	—

Table 5: Random effects summary for model of Switchboard word durations.

			Inf. given previous				Inf. given following			
# Seg	Types	Tokens	β	SE	t	$p(\chi^2)$	β	SE	t	$p(\chi^2)$
2	81	6,136	—	—	—	—	0.0270	0.0119	2.27	0.0238
3	665	35,956	—	—	—	—	0.0158	0.0036	4.39	0.0001
4	983	29,892	0.0063	0.0034	1.84	0.0670	0.0143	0.0035	4.11	0.0001
5	940	14,586	0.0061	0.0035	1.75	0.0806	0.0173	0.0034	5.07	0.0001
6	758	8,372	—	—	—	—	0.0220	0.0037	5.98	0.0001
7	607	6,539	0.0097	0.0047	2.06	0.0400	0.0145	0.0041	3.53	0.0005

Table 6: Informativity results for Switchboard models that compared words matched for segment count.

Current has an informativity of 1.13 bans, which would put it to the left of the plot area.

3.3.1 Extension of previous work on correlations with informativity

For all content word types across the two corpora, the correlation between orthographic length and informativity given the previous word (Spearman's $\rho = 0.27$) is numerically similar to what was reported by Piantadosi et al. (2011) for all English word types (note that coefficients in A are calculated over tokens). This is marginally higher than the correlation between orthographic length and frequency ($\rho = 0.26$). Since function words were excluded from the data here, this finding helps address a possible concern that the correlation was influenced by the distribution of function words versus content words (see Mahowald et al., 2013). Furthermore, the correlation between orthographic length and informativity given the following word is slightly higher ($\rho = 0.29$).

3.3.2 Word frequency

Word frequency did not consistently improve the fit of the duration models, contra earlier research. However, this was found to be due to the inclusion of per-word

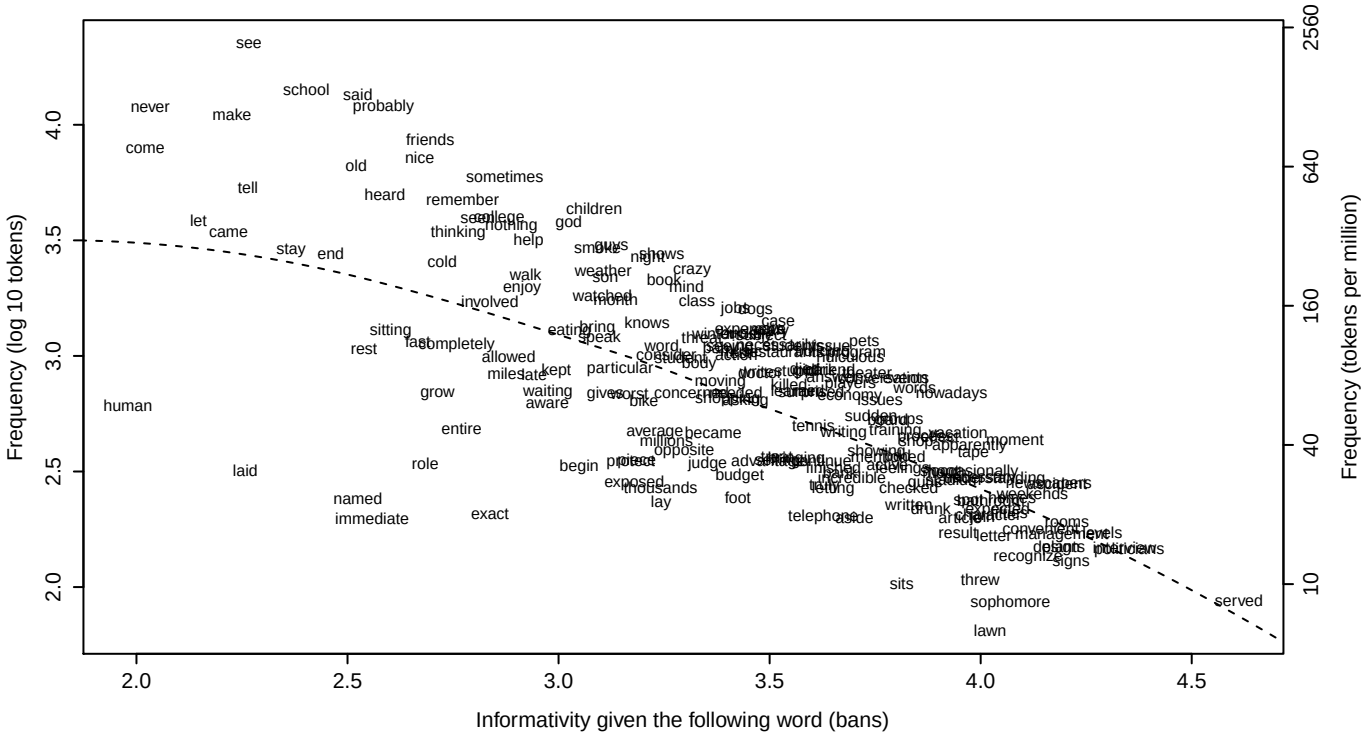

Figure 2: Sample of 200 word types that were observed in the combined data at least 10 times. Word types on the left are on average more predictable in context, and word types towards the top are more frequent. Dashed line shows local trend for all word types with at least 10 observations.

random intercepts in the current models, which captured most of the effect of frequency and some of the other per-type variables. Removing the per-word intercept parameter revealed an apparently reliable effect of word frequency in the *positive* direction (Buckeye: $\beta = 0.0092$, $t = 4.45$; Switchboard: $\beta = 0.0088$, $t = 7.04$). Furthermore, removing per-word intercepts made both informativity effects appear very reliable in the predicted direction (Buckeye: given previous: $\beta = 0.0087$, $t = 3.85$; given following: $\beta = 0.0402$, $t = 16.27$; Switchboard: given previous: $\beta = 0.0139$, $t = 11.65$; given following: $\beta = 0.0289$, $t = 24.75$).

The positive word frequency estimate does not suggest that more frequent words have longer durations. Instead, word frequency is acting as a suppressor variable for informativity (Friedman and Wall, 2005; Wurm and Fisicaro, 2014). Informativity is strongly correlated with word durations (Pearson's $r_{Y1} = 0.52$ for Buckeye) and frequency is less well correlated with duration ($r_{Y2} = -0.28$). However, informativity and frequency are well correlated with each other ($r_{12} = -0.58$). In the case where the inequality $r_{12} > {}^{r_{Y2}}/_{r_{Y1}}$ holds (here: $0.58 > {}^{0.28}/_{0.52}$) or is nearly satisfied[1],

[1]In a model with more than two predictors, the exact boundary also depends on the other relations among the predictors and with the outcome variable. Also, note that calculations here

the variable that is less well correlated with the outcome (frequency) may change sign.

Because frequency is much better correlated with informativity than with duration, in the model frequency is used to explain some of the error associated with informativity, rather than explaining a unique portion of duration variance. Wurm and Fisicaro (2014) show that, on simulated data with two correlated variables that fall into this region, the impact on the variable with the larger effect (here, informativity) is that this variable may have a slightly inflated estimate β, but that there is also a large loss of statistical power for detecting the variable's effect. Adelman et al. (2006) also report that frequency acts as a suppressor when it is entered into a model with a measure of contextual diversity (to predict reading times; see also McDonald and Shillcock, 2001), which is a variable that is conceptually similar to informativity.[2]

However, in the more conservative analysis—which includes the per-word intercept parameter—frequency simply fails to reach statistical significance at $\alpha = 0.15$ and is removed from the model. With this parameter, frequency was also found to be non-significant in each of the models by segment count for Switchboard, as reported in appendix table 4, and non-significant in four of the seven models by segment count for Buckeye, as reported in 2.

To evaluate whether a different corpus might produce different results, alternative bigram language models (forward and backward) were constructed from the 450-million-word Corpus of Contemporary American English (Davies, 2008). This corpus is larger, but it is mixed-genre and includes mostly written sources. This analysis included the original random effects structure and all of the parameters in the original model, except that the probabilistic measures (local bigram probability, frequency, and informativity) were replaced by estimates from the COCA language model. The COCA language model was smoothed using the Witten-Bell method (Witten and Bell, 1991), since counts for very low-frequency bigrams in COCA were not available.[3] In this analysis, frequency was found to be a reliable predictor in the expected direction (Buckeye: $\beta = -0.009$, $t = -2.38$, Switchboard: $\beta = -0.0111$, $t = -4.06$), while informativity given the following word retained its significance (Buckeye: $\beta = 0.0170$, $t = 4.65$, Switchboard: $\beta = 0.0081$, $t = 3.40$).

---

are done for frequency and duration prior to log-transformation, but the same condition exists for correlations after log-transformation.

[2]It is possible to remove this correlation and change the direction of the frequency effect by residualizing informativity on frequency. However, residualization does not change the estimates of the residualized variable (informativity), has a number of other undesirable effects on the model, and makes the interpretation of both residualized and residualizer variables problematic (Wurm and Fisicaro, 2014).

[3]The informativity effects were robust to the smoothing technique used on the original Fisher language model. For both datasets, model estimates for both bigram probability and informativity were similar when unsmoothed probabilities were used instead.

3.3.3 Predictability versus lexicalized two-word expressions

Given that the informativity measure is derived from bigram predictability, one question is whether predictability-driven reduction might be caused by lexicalized bigrams, rather than predictability-in-context per se. For example, one hypothesis might be that *human* is reduced in contexts like *human being*, *human rights*, *human nature*, *human life*, etc. not because *human* is predictable in those contexts, but because a bigram like *human rights* is stored as a single lexical entry. If such bigrams are processed as a single unit, it would be reasonable to expect them to be produced more quickly than two more independent words.

This question primarily has to do with the interpretation of bigram predictability, not informativity. If a word like *human* or *current* occurs in a common expression, it will be predictable in that context. Therefore, its reduction can be explained by its local bigram probability. If a word like *human* mainly occurs in such expressions, it will also have a low informativity value, since it is predictable on average. However, this could not be taken as evidence for an informativity effect. Wherever *human* occurs in a predictable expression, its reduction can be explained by local bigram probability. In order for there to be statistical evidence for an informativity effect, *human* must also be shortened in unpredictable contexts (or lengthened in predictable ones) beyond what the model would otherwise predict given the local relations between words.

Some previous work has addressed whether word reduction associated with local predictability should be interpreted as an effect of lexicalized bigrams. Bell et al. (2009) calculate pointwise mutual information for each bigram type in their data. This measure quantifies how dependent two words are on each other. They exclude tokens that occur in bigrams with a mutual information value in the top 15% of their data, but find that their bigram-predictability estimates are not qualitatively different with these data excluded. This suggests that predictability-driven reduction cannot be exclusively attributed to lexicalized bigrams. Jurafsky et al. (2001) study reduction of function word durations and vowel quality. They exclude tokens that have a bigram probability greater than the median bigram probability in their data. They find that local bigram predictability given the *following* word is robust to this exclusion, whereas local predictability given the *previous* word is still a significant predictor of duration reduction, but not of vowel reduction. Note that, for content word durations, Bell et al. (2009) did not originally find a significant effect of predictability given the previous word.

3.3.4 Alternative baseline durations

Recent work on probabilistic reduction has used corpus-derived estimates of word duration as a baseline for each word type. For example, Bell et al. (2009) and Gahl et al. (2012) calculate expected durations first by determining the mean duration of each segment type across the corpus. Then, for the segments in each word's

transcription, these mean segment durations are summed together to generate a baseline expectation for the word's total duration. For example, in the word *fish* [fɪʃ], the baseline duration would be the mean corpus duration of [f], plus the mean corpus duration of [ɪ], plus the mean corpus duration of [ʃ]. An alternative analysis was conducted using this measure as a baseline, rather than the MARYTTS duration model. For Buckeye, close phonetic transcriptions were available, and these were used in place of dictionary transcriptions. This means that expected durations in Buckeye were calculated based on each word token's transcription, rather than the word type's dictionary citation form. However, using dictionary transcriptions for Buckeye did not qualitatively change the results. With this baseline measure, informativity given the following word was significant (at least $p < 0.05$) in both of the full corpus models and all of the models stratified by segment count, for both corpora. Informativity given the previous word was also significant in some models using this baseline, but overall the effect was inconsistent.

A similar alternative baseline might also be adapted to take into account some of the effects of segmental context. Buz & Jaeger (p.c.) calculate baselines first by extracting mean durations of each segment type conditioned on the previous segment type. For example, the segment [t] has a different mean duration following [s] than following [n] or [ə], which reflects general articulatory constraints as well as language-specific phonetics and phonology. An alternative analysis was conducted using this biphone-sensitive measure as a baseline. For each word type, the mean durations of each segment type given the previous segment in the transcription were summed together. For example, in the phrase *a fish* [ə fɪʃ], the baseline duration of the token *fish* was taken to be the mean duration of [f] when it follows [ə] in the corpus, plus the mean duration of [ɪ] when it follows [f], plus the mean duration of [ʃ] when it follows [ɪ].

Using these baselines, in the Switchboard analysis, informativity given the following word was significant in the full corpus model and all of the models stratified by segment count. In Buckeye, for which close phonetic transcriptions were available, informativity given the following word was significant (at least $p < 0.05$) in the full model and all of the models stratified by segment count, except for the model of 6–segment words ($p < 0.10$). When dictionary transcriptions were used to generate the Buckeye baselines, the informativity effect in the model of 6–segment words was also significant. As before, informativity given the previous word was inconsistent across analyses and word lengths.

3.3.5 Segment deletion and duration

The effect of informativity on word durations might involve segment compression, and it might also involve categorical segment deletion. Figure 3 shows the mean percentage of segments that were transcribed as deleted from word tokens in the Buckeye data at each segment count. These numbers are very similar to those reported for content words by Johnson (2004). For that study, close transcriptions

were only available for about ⅓ of the full corpus. However, the data here also involve additional exclusions (such as exclusion of pause-adjacent tokens; see section 2.1). As observed by Johnson (2004) and replicated here, short content words with four or fewer segments in their dictionary transcriptions are rarely transcribed as having segments deleted ($< 5\%$ of segments deleted; although see also the segment deviation analysis in that study). Since there was an effect of informativity on word durations for these short words, it is more likely that the effect for these words involves compression rather than deletion.

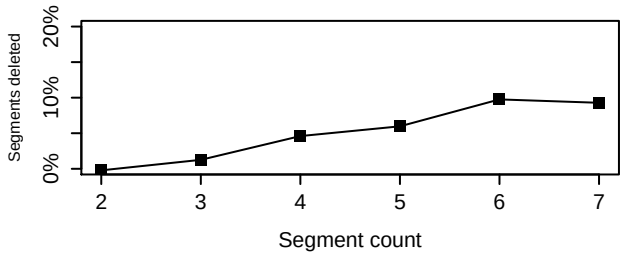

Figure 3: Mean percentage of segments deleted for word tokens of each segment count, based on comparisons of the dictionary and close phonetic transcriptions in the Study 1 Buckeye data.

Figure 4 shows the Spearman correlation coefficients between word informativity and the percentage of segments that were deleted in Buckeye tokens, divided by dictionary segment count. The figure also shows the coefficients between informativity and duration at each segment count. While there is a reliable correlation across different word lengths between informativity given the following word and duration, informativity is not well associated with deletion (as transcribed in the corpus) for shorter words of less than five segments. For longer words, full segment deletion probably also contributes to the duration effect.

To explore these two possible components of the duration effect, an alternative analysis was carried out using only those Buckeye word tokens which had a close phonetic transcription that was completely identical to the word type's dictionary transcription (20,296 of the original 41,167 tokens meeting the inclusion criteria). For these data, informativity given the following word had a statistically significant effect (at least $p < 0.05$) in the full model and for words of each segment count, except for 7–segment words, where the effect was non-significant ($p > 0.4$; with 408 tokens and 122 different types). The effect sizes for informativity were roughly the same as in the primary analysis; for 5 and 7–segment words the estimate was much smaller. Because informativity was found to influence duration for these words which had no segments deleted, compression is likely a key component of the informativity effect on word durations. However, deletion is not ruled out as a part of the effect, especially for longer word types for which full segment deletion is transcribed more often in general.

However, the analysis is potentially much more complex when considering seg-

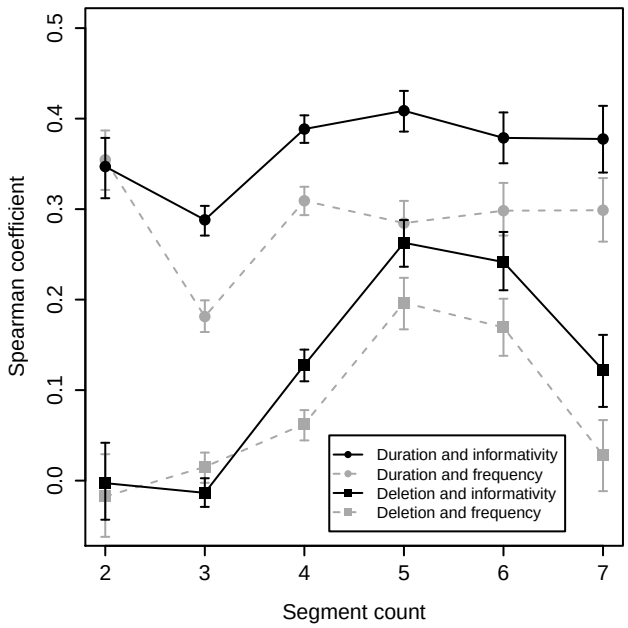

Figure 4: Spearman correlations between (i) informativity given the following word and (ii) duration (black circles) or percentage of segments deleted (black squares). Calculated over Study 1 Buckeye word tokens at each dictionary transcription length; error bars show bootstrapped 95% confidence intervals. Additional gray lines show correlations between frequency and duration or percentage of segments deleted. Lower informativity corresponds with shorter durations and higher deletion rates; deletion coefficients here are inverted for comparison with duration coefficients.

ment deletion in particular, instead of whole-word durations. For example, language-specific categorical deletion processes are typically restricted to particular contexts or classes of words, such as [t] reduction in English (Pitt et al., 2011) or schwa deletion in English or French (Connine et al., 2008; Racine and Grosjean, 2005). Many deletion processes are also sensitive to additional sub-lexical factors such as morphological structure (e.g., Labov et al., 1968; Guy, 1991; although see also Sugahara and Turk, 2009). A targeted study is likely better suited to analyze the interaction of word informativity with deletion processes (see Cohen Priva, 2012, especially Ch. 3).

4 General discussion

The main result of this paper is that the average contextual predictability of a word—word informativity—has a significant effect on that word's acoustic duration. A word that is usually unpredictable has a longer duration than a word that is usually predictable, independent of local contextual predictability, frequency, or segment count. The effect size is comparable to reduction associated with local

predictability. The effect was reliable for informativity given the following word, but not given the previous word. This difference is consistent with previous research on local probabilistic reduction in content words.

Since informativity captured a significant amount of the variance beyond local bigram probability, it is not the case that predictable words simply have shorter tokens on average because most of their tokens are reduced. Instead, even after the local probability of every token is taken into account, there is an additional effect of each word's average predictability in the hypothesized direction. If most tokens of a word are reduced in duration because they appear in high-probability contexts, the duration of that word was also found to be shorter overall, independent of token context.

4.1 Accounting for the informativity effect

Previous research has suggested that if a word is reduced often enough, speakers will encode that reduction in their representation of the word. This stored reduction will bias future productions of that word in all contexts. The current results support this prediction for probabilistic reduction. If a word is reduced very often because it often occurs in high-probability contexts, productions of that word are biased towards a more reduced form, even when it is not produced in a high-probability context. The informativity finding might be captured by several possible speech processing models, described below.

4.1.1 Exemplar-based and combined exemplar–abstract models

In an exemplar model, all phonetic detail of each incoming word token is stored as an exemplar of that word (Goldinger, 1996, 1998; Johnson, 1997, 2006, 2007). Representation is constructed from the distribution of previously-encountered exemplars. The forms that are most often heard and stored therefore have a greater influence on this distribution. Since low-informativity words are very commonly reduced, the distribution of exemplars of a low-informativity word will be biased towards reduced forms. To generate a production target, one or more exemplars are sampled from this distribution (Goldinger, 2000; Pierrehumbert, 2001). In some intermediate models, word exemplars that belong to a single category are compressed into a phonologically-abstract secondary representation that also influences the generation of a production target (Goldinger, 2007; Ernestus and Baayen, 2011; German et al., 2013; Ernestus, 2014). The production target is then passed to a system for phonetic implementation, during which motor-planning, articulatory, or other effects influence the final realization (Ernestus, 2014).

When speakers produce a low-informativity word, they will be more likely to sample reduced exemplars of that word when generating a target, even if the context would not otherwise trigger reduction (Goldinger, 2000; Pierrehumbert, 2002). During phonetic implementation, online probabilistic reduction may occur (Ernes-

tus, 2014), causing the word to be further reduced. Thus, low-informativity words should be reduced due to offline informativity effects—a higher likelihood of sampling reduced exemplars—as well as online contextual ones, which apply during phonetic implementation.

4.1.2 Abstract models with multiple variants

The results can also be described by a model with abstract phonological representations. In such a model, each word representation includes an unreduced citation form, and may also include several reduced variants that are sufficiently common in a language-user's experience. The importance that each variant form has for perception and production depends on their relative frequencies (Connine et al., 2008; Racine and Grosjean, 2005; Bürki et al., 2010; Pitt et al., 2011; Ranbom and Connine, 2007). It may also depend on factors such as orthography (Ranbom and Connine, 2007, 2011), a communicative pressure that favors unreduced forms (Pitt et al., 2011), and probabilistic knowledge about articulatory contexts (Mitterer and McQueen, 2009).

If a reduced variant of a word type has a higher relative frequency, it will be more accessible in production (Bürki et al., 2010), and more often accessed in spontaneous speech. In this case, the informativity effect would represent the proportion with which the unreduced variant is selected in favor of the reduced variant. For example, consider a scenario in which the word *current* has two abstract variant forms stored in lexical representation. One form is the unreduced citation variant [kɻ̩.ɪnt], and the other is a reduced variant [kɻ̃ʔ̩]. The analysis presented here models word duration, not the selection of variant forms. Duration is not represented in these abstract forms, but it is empirically true that the unreduced variant has an average duration of 350ms, while the reduced variant has an average duration of 200ms. If speakers usually select the unreduced variant, on average the word type will have a longer duration that is closer to 350ms. If speakers usually select the reduced variant, on average the word type will have a shorter duration that is closer to 200ms.

Given that unreduced variants can reasonably be assumed to always be longer than reduced variants (or at least not shorter), the informativity parameter then describes the tendency for speakers to select unreduced variants in general. If informativity has a positive coefficient, that means that the model predicts that high-informativity word types will have longer durations on average, because speakers prefer to select the unreduced variant more often than the reduced variant.

4.1.3 Rational speech production

There are other representational models that might result in an informativity effect. For example, it may be the case that each word type has a default phonetic target for production. If a speaker chooses to deviate from this default, such as to reduce or hyper-articulate a word, it is costly to do so in motor planning (even though

reduction might save on articulatory effort) and the speaker may not always reach their deviant target. Therefore, in order to minimize both planning and articulatory costs and maximize effectiveness, a rational speaker will select a default target for each type that is most similar to the tokens that usually need to be produced. If reduction is usually called for, then the rational speaker will choose a reduced default form, so that it is only rarely necessary to deviate from that form and incur costs (thanks to Roger Levy for this suggestion). In this model, informativity is a goal-oriented effect chosen directly by the speaker, rather than an indirect consequence of representation.

4.1.4 Efficient articulation

A related model might also refer to word-specific articulatory timing specifications. In this model, words are specified for tighter or looser alignments of each necessary articulatory gesture. This results in a range of possible reductions during fast speech that is unique to each word (Lavoie, 2002). Low-informativity words usually occur in predictable contexts. In such contexts, a speaker is more likely to be understood, and consequently is more likely to accept an imprecise production of a target word as an adequate acoustic realization of that word. A speaker will gradually learn through experience that some or all gestures in such a word can have looser timings, yet still produce an acceptable acoustic form (following Jaeger and Ferreira, 2013; Jaeger, 2013). Over time, the word will acquire less strict gestural alignment specifications. Because of this, gestural overlap (and, potentially, a greater degree of acoustic reduction) will become more likely for this word in all contexts.

For example, Jaeger and Ferreira (2013) suggest that it is primarily reductions of low-confusability words or word forms that become acceptable variants in common usage. By definition, low-informativity words tend to be more predictable and are unlikely to be very confusable in context.

4.1.5 Direct knowledge of average word probabilities

It may instead be the case that an informativity metric is represented directly at the word level. In other words, language-users would not store reduction or reduced variant forms, but instead would directly track how predictable a word type is on average. This would be stored in addition to (implicit or explicit) knowledge about specific inter-word relationships. In production, speakers would then use their knowledge of both factors, plus frequency, to determine online how much articulatory effort to give to a word. The results of Lee and Goldrick (2008), who argued that speakers perceptually track both average predictability as well as local predictability of segments within syllables, suggest that language-users may in some way use informativity-like knowledge in perception.

Extending this account to production would require that informativity-driven reduction be a consequence of audience-design considerations. An online processing

model of probabilistic reduction is more difficult to reconcile with an informativity effect. In these models, locally-predictable words gain a boost in activation from nearby words or related constructions that are usually associated with them, which speeds production. The current results show that low-informativity words are shorter even in unpredictable contexts where no nearby syntactic or semantic associates would cause pre-activation. An online processing model of informativity would thus require an additional mechanism to explain how pre-activation occurs even in contexts where there are no associated words to activate the target.

An online audience-design account of informativity is possible. In this account, speakers are aware of the fact that their own model of word probabilities may not be exactly the same as their listeners' models. Since informativity is an averaged probability, it would be prudent for the speaker to adjust an extreme estimate of local probability by also weighing how probable the target word usually is, on average. Note that the primary informativity effect involved predictability given the following word, not the previous word. If speakers reduce words when they are predictable for the listener, this must be accommodated by research showing that listeners interpret words based on the following context as well as the preceding context (Szostak and Pitt, 2013).

In this account, speakers would store probabilistic knowledge of how predictable each word tends to be, rather than storing the effects of reduction. This would still likely entail lexical representation of informativity, since this knowledge is word-specific. This account would also require that speakers then balance at least three sources of probability (context-specific, average context, and frequency) in choosing how much articulation a word requires for efficient communication. By contrast, an offline representational account of informativity gives rise to the observed effect and requires neither active goal-oriented behavior on the part of the speaker nor fine-grained negotiation of multiple word-specific probabilities during each articulation.

4.2 Summary

Previous research has shown that words are reduced when they occur in predictable contexts. The results of this paper show that words which are typically predictable in context are reduced even when they occur in unpredictable contexts. This phenomenon is predicted by models in which probabilistic reduction is stored in representation, and this stored reduction has a stronger effect on processing when it is relatively more frequent.

The effect was found to be robust across English word types of different lengths and segmental content, for two large corpora and various implementations of a probabilistic language model. None of a large set of control variables could fully account for the relationship between informativity and duration. Future work might evaluate different representational accounts with laboratory production studies, in not only English but also other languages. In particular, such work might help explain the correlation between word lengths and informativity across languages.

Acknowledgments

Thanks to Roger Levy for helpful discussion, instruction, and advice, and for comments on earlier drafts of this paper; also thanks to Eric Bakovic, Esteban Buz, Anne Canter, Uriel Cohen Priva, Gabriel Doyle, Marc Garellek, Gwendolyn Gillingham, Florian Jaeger, Andrew Kehler, Toben Mintz, Emily Morgan, Mark Myslin, three anonymous reviewers, the New Zealand Institute of Language, Brain, and Behaviour, the UC San Diego Computational Psycholinguistics Lab, and the audience at Architectures and Mechanisms for Language Processing 2013. This project was supported by a National Science Foundation Graduate Research Fellowship. Any errors are mine.

References

Adelman, J. S., Brown, G. D., and Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823.

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1):67–82.

Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

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

Baese-Berk, M. and Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and Cognitive Processes*, 24(4):527–554.

Baker, R. E. and Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech*, 52(4):391–413.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., and Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, 42(1):1–22.

Bates, D., Maechler, M., and Bolker, B. (2013). lme4.0: Linear mixed-effects models using s4 classes. R package version 0.999999-4.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, 60(1):92–111.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, 113(2):1001.

Browman, C. P. and Goldstein, L. (1990). Tiers in articulatory phonology, with

some implications for casual speech. *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, pages 341–376.

Brysbaert, M., Keuleers, E., and New, B. (2011). Assessing the usefulness of google books word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, 2.

Brysbaert, M. and New, B. (2009). Moving beyond kucera and francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, 41(4):977–990.

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

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

Bybee, J. (2003). Mechanisms of change in grammaticization: the role of frequency. In Joseph, B. D. and Janda, R. D., editors, *The Handbook of Historical Linguistics*, pages 602–623. Blackwell, Oxford.

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

Byrd, D. (1996). A phase window model framework for articulatory timing. *Phonology*, 13:139–169.

Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (2010). The NXT-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419.

Calhoun, S., Carletta, J., Jurafsky, D., Nissim, M., Ostendorf, M., and Zaenen, A. (2009). NXT switchboard annotations. Technical report, Linguistic Data Consortium, Philadelphia.

Chen, S. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University.

Cieri, C., Graff, D., Kimball, O., Miller, D., and Walker, K. (2005). *Fisher English Training Part 2*. Linguistic Data Consortium, Philadelphia.

Clopper, C. G. and Pierrehumbert, J. B. (2008). Effects of semantic predictability and regional dialect on vowel space reduction. *The Journal of the Acoustical Society of America*, 124(3):1682.

Cohen Priva, U. (2008). Using information content to predict phone deletion. In Abner, N. and Bishop, J., editors, *Proceedings of the 27th West Coast Conference on Formal Linguistics*, pages 90–98.

Cohen Priva, U. (2012). *Sign and signal: deriving linguistic generalizations from information utility*. PhD thesis, Stanford University.

Connine, C. M. and Pinnow, E. (2006). Phonological variation in spoken word recognition: Episodes and abstractions. *The Linguistic Review*, 23(3).

Connine, C. M., Ranbom, L. J., and Patterson, D. J. (2008). Processing variant forms in spoken word recognition: The role of variant frequency. *Perception & Psychophysics*, 70(3):403–411.

Davies, M. (2008). The corpus of contemporary american english: 450 million words, 1990-present.

Demberg, V., Sayeed, A. B., Gorinski, P. J., and Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 356–367.

Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*, 142:27–41.

Ernestus, M., Baayen, H., and Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language*, 81(1-3):162–173.

Ernestus, M. and Baayen, R. H. (2011). Corpora and exemplars in phonology. In Goldsmith, J. A., Riggle, J., and Yu, A. C., editors, *The Handbook of Phonological Theory*, pages 374–400. Blackwell.

Everett, C., Miller, Z., Nelson, K., Soare, V., and Vinson, J. (2011). Reduction of brazilian portuguese vowels in semantically predictable contexts. In *Proceedings of the 17th International Congress of Phonetic Sciences*, pages 651–654, Hong Kong.

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

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

Gahl, S. and Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 80:748–775.

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4):789–806.

German, J. S., Carlson, K., and Pierrehumbert, J. B. (2013). Reassignment of consonant allophones in rapid dialect acquisition. *Journal of Phonetics*, 41(3-4):228–248.

Godfrey, J. J. and Holliman, E. (1997). Switchboard-1 release-2. Technical report, Linguistic Data Consortium, Philadelphia.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 22(5):1166–1182.

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, 195(2):251–279.

Goldinger, S. D. (2000). The role of perceptual episodes in lexical processing. In *Proceedings of the Workshop on Spoken Word Access Processes*, pages 155–158, Nijmegen, The Netherlands. Max-Planck Institute for Psycholinguistics.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In *Proceedings of the 16th International Congress of Phonetic Sciences*, pages 49–54.

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

Hooper, J. (1976). Word frequency in lexical diffusion and the source of morphophonological change. In Christie, W. J., editor, *Current Progress in Historical Linguistics*, pages 95–105. North-Holland, Amsterdam.

Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in Psychology*, 4.

Jaeger, T. F. and Ferreira, V. (2013). Seeking predictions from a predictive framework. *Behavioral and Brain Sciences*, 36(4):31–32.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson, K. and Mullenix, J., editors, *Talker Variability in Speech Processing*, pages 145–165. Academic Press, San Diego.

Johnson, K. (2004). Massive reduction in conversational american english. In Yoneyama, K. and Maekawa, K., editors, *Proceedings of the 1st Session of the 10th International Symposium on Spontaneous Speech: Data and Analysis*, pages 29–54, Tokyo, Japan. The National International Institute for Japanese Language.

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

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. In Sole, M., Beddor, P., and Ohala, M., editors, *Experimental Approaches to Phonology. In Honor of John Ohala.*, pages 25–40. Oxford University Press.

Jurafsky, D., Bell, A., and Girand, C. (2002). The role of the lemma in form variation. *Laboratory Phonology*, 7:3–34.

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

Kahn, J. M. and Arnold, J. E. (2012). A processing-centered look at the contribution of givenness to durational reduction. *Journal of Memory and Language*, 67(3):311–325.

Kemps, R., Ernestus, M., Schreuder, R., and Baayen, H. (2004). Processing reduced word forms: The suffix restoration effect. *Brain and Language*, 90(1-3):117–127.

Komineck, J. and Black, A. W. (2003). CMU ARCTIC databases for speech synthesis. Technical Report CMU-LTI-03-177, Carnegie Mellon University.

Kuperman, V. and Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of*

Memory and Language, 66(4):588–611.

Labov, W., Cohen, P., Robins, C., and Lewis, J. (1968). A study of the non-standard english of negro and puerto rican speakers in new york city. Technical Report Coorperative Research Project 3288, Vol. I, Columbia University.

Lavoie, L. (2002). Some influences on the realization of for and four in american english. *Journal of the International Phonetic Association*, 32(2):175–202.

Lee, Y. and Goldrick, M. (2008). The emergence of sub-syllabic representations. *Journal of Memory and Language*, 59(2):155–168.

Levelt, W. J., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22:1–75.

Levy, R. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*.

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

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the h & h theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modeling*, pages 403–439. Kluwer, Dordrecht.

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

Mahowald, K., Fedorenko, E., Piantadosi, S. T., and Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

McDonald, S. A. and Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44(3):295–322.

Mitterer, H. and McQueen, J. M. (2009). Processing reduced word-forms in speech perception using probabilistic knowledge about speech production. *Journal of Experimental Psychology: Human Perception and Performance*, 35(1):244–263.

Moore-Cantwell, C. (2013). Syntactic predictability influences duration. In *Proceedings of Meetings on Acoustics*, volume 19.

Mowrey, R. and Pagliuca, W. (1995). The reductive character of articulatory evolution. *Rivista di Linguistica*, 7(1):37–124.

Munson, B. (2007). Lexical access, lexical representation, and vowel production. *Laboratory Phonology*, 9:201–228.

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

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In Bybee, J. and Hopper, P., editors, *Frequency effects and the emergence of lexical structure: studies in language*, pages 137–157. John Benjamins, Amsterdam.

Pierrehumbert, J. (2002). Word-specific phonetics. *Laboratory Phonology*, 7:101–139.

Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). Buckeye corpus of conversational speech (2nd release). Technical report, Columbus, OH.

Pitt, M. A. (2009). How are pronunciation variants of spoken words recognized? a test of generalization to newly learned words. *Journal of Memory and Language*, 61(1):19–36.

Pitt, M. A., Dilley, L., and Tat, M. (2011). Exploring the role of exposure frequency in recognizing pronunciation variants. *Journal of Phonetics*, 39(3):304–311.

Pluymaekers, M., Ernestus, M., and Baayen, R. H. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, 62(2-4):146–159.

R Core Team (2013). R: A language and environment for statistical computing.

Racine, I. and Grosjean, F. (2005). Le coût de l'effacement du schwa lors de la reconnaissance des mots en français. *Canadian Journal of Experimental Psychology*, 59:240–254.

Ranbom, L. and Connine, C. (2007). Lexical representation of phonological variation in spoken word recognition. *Journal of Memory and Language*, 57(2):273–298.

Ranbom, L. J. and Connine, C. M. (2011). Silent letters are activated in spoken word recognition. *Language and Cognitive Processes*, 26(2):236–261.

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

Stolcke, A. (2002). SRILM — an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO.

Stolcke, A., Zheng, J., Wang, W., and Abrash, V. (2011). SRILM at sixteen: Update and outlook. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, HI.

Sugahara, M. and Turk, A. (2009). Durational correlates of english sublexical constituent structure. *Phonology*, 26(3):477.

Szostak, C. M. and Pitt, M. A. (2013). The prolonged influence of subsequent context on spoken word recognition. *Attention, Perception, & Psychophysics*, 75(7):1533–1546.

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

van Son, R. J. J. H. and Pols, L. C. W. (2003). How efficient is speech? In *Proceedings of the 25th Institute of Phonetic Sciences*, volume 25, pages 171–184.

Warner, N., Jongman, A., Sereno, J., and Kemps, R. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception:

evidence from dutch. *Journal of Phonetics*, 32(2):251–276.

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

Witten, I. H. and Bell, T. C. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094.

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

Yao, Y. (2009). Understanding VOT variation in spontaneous speech. In *Proc. 18th International Congress of Linguists (CIL XVIII)*, page 2943.

Zipf, G. K. (1935). *The Psychobiology of Language*. Houghton-Mifflin, Boston.

A Predictor summaries

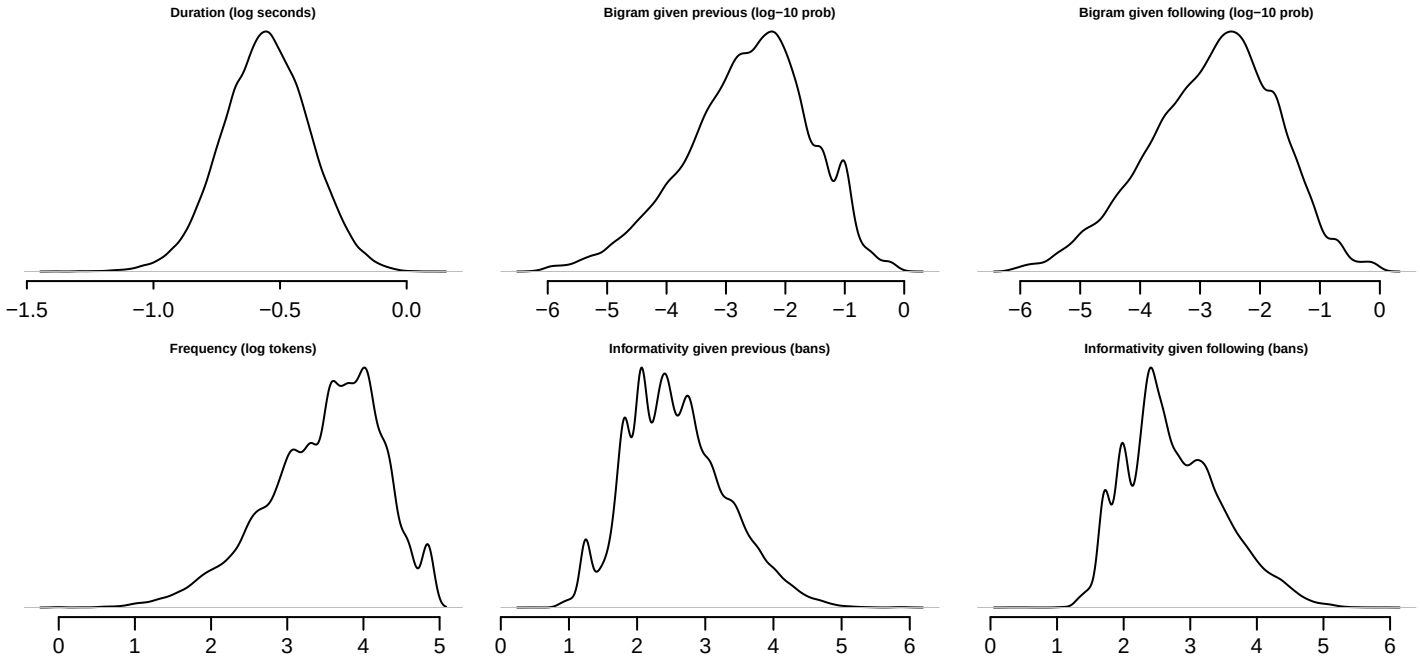

Figure A.1: Density plots showing the distribution of the probabilistic variables in the Buckeye Corpus. Variables were centered before being entered into the model.

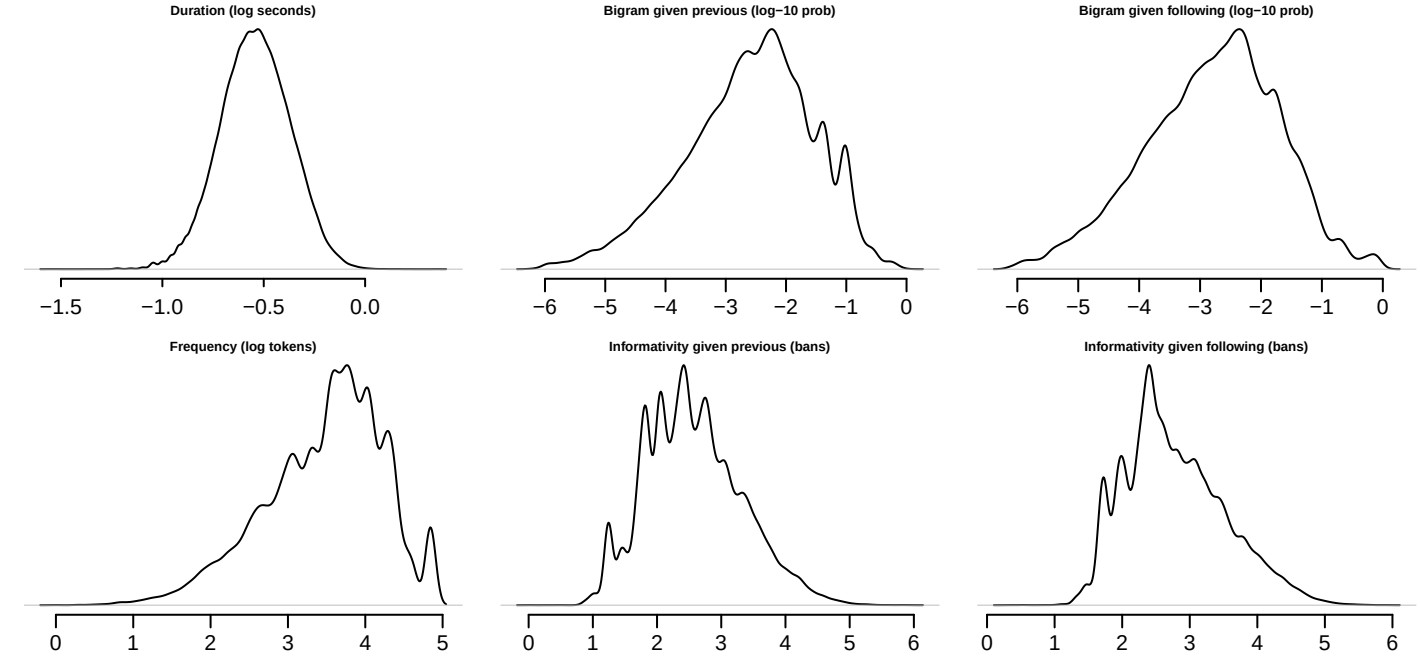

Figure A.2: Density plots showing the distribution of the probabilistic variables in the Switchboard Corpus. Variables were centered before being entered into the model.

	Seg	Dur	Mary	Syl	Rate	Bg-P	Bg-F	Freq	Orth	Inf-P	Inf-F
Seg	1.00	0.57	0.86	0.76	0.02	−0.26	−0.25	−0.37	0.87	0.35	0.36
Dur	0.57	1.00	0.63	0.48	−0.21	−0.32	−0.43	−0.44	0.55	0.37	0.50
Mary	0.86	0.63	1.00	0.71	0.01	−0.21	−0.25	−0.31	0.83	0.28	0.35
Syl	0.76	0.48	0.71	1.00	0.06	−0.23	−0.19	−0.29	0.76	0.30	0.29
Rate	0.02	−0.21	0.01	0.06	1.00	0.03	0.02	0.02	0.02	−0.01	−0.02
Bg-P	−0.26	−0.32	−0.21	−0.23	0.03	1.00	0.38	0.64	−0.22	−0.73	−0.54
Bg-F	−0.25	−0.43	−0.25	−0.19	0.02	0.38	1.00	0.58	−0.23	−0.49	−0.70
Freq	−0.37	−0.44	−0.31	−0.29	0.02	0.64	0.58	1.00	−0.31	−0.86	−0.84
Orth	0.87	0.55	0.83	0.76	0.02	−0.22	−0.23	−0.31	1.00	0.28	0.33
Inf-P	0.35	0.37	0.28	0.30	−0.01	−0.73	−0.49	−0.86	0.28	1.00	0.71
Inf-F	0.36	0.50	0.35	0.29	−0.02	−0.54	−0.70	−0.84	0.33	0.71	1.00

Table 1: Spearman correlations between variables in Buckeye data. Seg = segment count, Dur = empirical duration, Mary = baseline duration, Syl = syllable count, Rate = speech rate, Bg-P = bigram probability given previous word, Bg-F = bigram probability given the following word, Freq = word frequency, Orth = orthographic length, Inf-P = informativity given the previous word, Inf-F = informativity given the following word.

# Seg	Non-significant predictors removed
2	Syl ($p > 0.8$), Orth ($p > 0.6$), Bg-P ($p > 0.15$)
3	Orth ($p > 0.5$), Syl ($p > 0.2$), Freq ($p > 0.2$), Inf-P ($p > 0.4$)
4	Orth ($p > 0.4$), Freq ($p > 0.4$), Inf-P ($p > 0.2$), Syl ($p > 0.15$)
5	Inf-P ($p > 0.8$), Orth ($p > 0.7$), Freq ($p > 0.4$)
6	—
7	Freq ($p > 0.7$), Inf-P ($p > 0.5$)

Table 2: Predictors removed at $\alpha = 0.15$ in each Buckeye model by segment count. Abbreviations given in caption to figure 1.

	Seg	Dur	Mary	Syl	Rate	Bg-P	Bg-F	Freq	Orth	Inf-P	Inf-F
Seg	1.00	0.60	0.86	0.67	0.00	−0.27	−0.24	−0.38	0.87	0.35	0.36
Dur	0.60	1.00	0.65	0.46	−0.20	−0.34	−0.44	−0.46	0.56	0.40	0.50
Mary	0.86	0.65	1.00	0.64	0.01	−0.24	−0.26	−0.35	0.82	0.31	0.37
Syl	0.67	0.46	0.64	1.00	0.07	−0.26	−0.22	−0.32	0.67	0.34	0.33
Rate	0.00	−0.20	0.01	0.07	1.00	0.02	0.01	0.03	0.01	−0.02	−0.02
Bg-P	−0.27	−0.34	−0.24	−0.26	0.02	1.00	0.39	0.65	−0.23	−0.74	−0.55
Bg-F	−0.24	−0.44	−0.26	−0.22	0.01	0.39	1.00	0.60	−0.23	−0.50	−0.71
Freq	−0.38	−0.46	−0.35	−0.32	0.03	0.65	0.60	1.00	−0.32	−0.86	−0.84
Orth	0.87	0.56	0.82	0.67	0.01	−0.23	−0.23	−0.32	1.00	0.29	0.35
Inf-P	0.35	0.40	0.31	0.34	−0.02	−0.74	−0.50	−0.86	0.29	1.00	0.71
Inf-F	0.36	0.50	0.37	0.33	−0.02	−0.55	−0.71	−0.84	0.35	0.71	1.00

Table 3: Spearman correlations between variables in Switchboard data. Abbreviations given in caption to figure 1.

# Seg	Non-significant predictors removed
2	Inf-P ($p > 0.7$), Freq ($p > 0.8$), Syl ($p > 0.3$)
3	Orth ($p > 0.9$), Freq ($p > 0.8$), Inf-P ($p > 0.6$)
4	Freq ($p > 0.7$), Syl ($p > 0.3$)
5	Freq ($p > 0.3$)
6	Freq ($p > 0.9$), Orth ($p > 0.3$), Syl ($p > 0.2$), Inf-P ($p > 0.15$)
7	Freq ($p > 0.9$), Orth ($p > 0.6$)

Table 4: Predictors removed at $\alpha = 0.15$ in each Switchboard model by segment count.