To appear in Bybee, Joan and Paul Hopper (eds.). 2000. *Frequency and the emergence of linguistic structure.* Amsterdam: John Benjamins.

Probabilistic Relations between Words: Evidence from Reduction in Lexical Production

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D. Raymond
University of Colorado, Boulder

1 Introduction

The ideas of frequency and predictability have played a fundamental role in models of human language processing for well over a hundred years (Schuchardt, 1885; Jespersen, 1922; Zipf, 1929; Martinet, 1960; Oldfield & Wingfield, 1965; Fidelholz, 1975; Jescheniak & Levelt, 1994; Bybee, 1996). While most psycholinguistic models have thus long included word frequency as a component, recent models have proposed more generally that probabilistic information about words, phrases, and other linguistic structure is represented in the minds of language users and plays a role in language comprehension (Jurafsky, 1996; MacDonald, 1993; McRae, Spivey-Knowlton, & Tanenhaus, 1998; Narayanan & Jurafsky, 1998; Trueswell & Tanenhaus, 1994) production (Gregory, Raymond, Bell, Fosler-Lussier, & Jurafsky, 1999; Roland & Jurafsky, 2000) and learning (Brent & Cartwright, 1996; Landauer & Dumais, 1997; Saffran, Aslin, & Newport, 1996; Seidenberg & MacDonald, 1999).

In recent papers (Bell, Jurafsky, Fosler-Lussier, Girand, & Gildea, 1999; Gregory *et al.*, 1999; Jurafsky, Bell, Fosler-Lussier, Girand, & Raymond, 1998), we have been studying the role of predictability and frequency in lexical production. Our goal is to understand the many factors that affect production variability as reflected in reduction processes such as vowel reduction, durational shortening, or final segmental deletion of words in spontaneous speech. One proposal that has resulted from this work is the *Probabilistic Reduction Hypothesis*: word forms are reduced when they have a higher probability. The probability of a word is conditioned on many aspects of its context, including neighboring words, syntactic and lexical structure, semantic expectations, and discourse factors. This proposal thus generalizes over earlier models which refer only to word frequency (Zipf, 1929; Fidelholz, 1975; Rhodes, 1992, 1996) or predictability (Fowler & Housum, 1987).

In this paper we focus on a particular domain of probabilistic linguistic knowledge in lexical production: the role of local probabilistic relations between words.

Our previous research as well as research by others (Bush, 1999; Bybee & Scheibman, 1999; Krug, 1998) suggests that words which are strongly related to or predictable from neighboring words, such as collocations (sequences of commonly cooccurring words), are more likely to be phonologically reduced.

This paper extends our earlier studies of reduction, arguing that these probabilistic relations between words should be interpreted as evidence for emergent linguistic structure, and more specifically as evidence that probabilistic relations between words are represented in the mind of the speaker. Testing the claim requires showing that probabilistic relations are represented very generally across words. We therefore examine probabilistic relations with function words as well as with content words, with frequent words as well as with infrequent words. It is also crucial to understand the exact nature of these probabilistic effects. We thus study various probabilistic measures of a word's predictability from neighboring words, and test the effects of each on various types of reduction. Our conclusions support the probabilistic reduction hypothesis; more probable words are more likely to be reduced. The results suggest that probabilistic relations between words must play a role in the mental representation of language.

Our experiments are based on two distinct datasets, each drawn from 38,000 words that were phonetically hand-transcribed from American English telephone conversations (Greenberg, Ellis, & Hollenback, 1996). The first dataset consists of 5618 of the 9000 tokens of the 10 most frequent function words: *I*, *and*, *the*, *that*, *a*, *you*, *to*, *of*, *it*, and *in*. The second focuses on 2042 of the 3000 content word tokens whose lexical form ends in a t or d. Each observation is coded with its duration and pronunciation as well as contextual factors such as the local rate of speech, surrounding segmental context and nearby disfluencies. We use linear and logistic regression to control for contextual factors and study the extent to which various probabilistic measures of lexical predictability account for reduction of word forms, as indicated by vowel reduction, deletion of final t or d, and durational shortening. Throughout this paper we will use the term 'reduced' to refer to forms that have undergone any of these processes.

2 Measures of Probabilistic Relations between Words

The Probabilistic Reduction Hypothesis claims that words are more reduced when they are more predictable or probable. There are many ways to measure the probability of a word. This section discusses a number of local measures that we have studied, although we will mainly report on two measures: the conditional probability of the target word given the preceding word and the conditional probability of the target word given the following word.

The simplest measure of word probability is called the *prior probability*. The prior probability of a word is the probability without considering any contextual factors ('prior' to seeing any other information). The prior probability is usually estimated by using the *relative frequency* of the word in a sufficiently large corpus. The relative frequency is the frequency of the word divided by the total number of word tokens in the corpus:

$$P(w_i) = \frac{C(w_i)}{\sum_j C(w_j)} = \frac{C(w_i)}{N} \tag{1}$$

The relative frequency is thus a normalized version of word frequency similar to information in frequency dictionaries such as Francis and Kučera (1982). Throughout the paper we use the term *relative frequency* rather than prior probability, although the reader should keep in mind that frequencies are estimates of the probability of a word's occurrence independent of context. We also consider the relative frequencies of the preceding and following words.

Probability can also be measured with respect to neighboring words. We use two measures (the *joint probability* and the *conditional probability*) of the predictability of a word given the previous word. The *joint probability* of two words $P(w_{i-1}w_i)$ may be thought of as the prior probability of the two words taken together, and is estimated by just looking at the relative frequency of the two words together in a corpus:

$$P(w_{i-1}w_i) = \frac{C(w_{i-1}w_i)}{N} \tag{2}$$

This is a variant of what Krug (1998) called the *string frequency* of the two words.

The *conditional probability of a word given the previous word* is also sometimes called the *transitional probability* (Bush, 1999; Saffran *et al.*, 1996). The conditional probability of a particular target word w_i given a previous word w_{i-1} is estimated from a sufficiently large corpus, by counting the number of times the two words occur together $C(w_{i-1}w_i)$, and dividing by $C(w_{i-1})$, the number of times that the first word occurs:

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} \tag{3}$$

The difference between the conditional and joint probability is that the conditional probability controls for the frequency of the conditioning word. For example, pairs of words can have a high joint probability merely because the individual words are of high frequency (e.g., *of the*). The conditional probability would be high only if the second word was particularly likely to follow the first. Most measures of word cohesion, such as conditional probability and mutual information, are based on such metrics which control for the frequencies of one or both of the words (Manning & Schütze, 1999).

In addition to considering the preceding word, the effect of the following word may be measured by the two corresponding probabilities. The *joint probability of a word with the next word* $P(w_i w_{i+1})$ is estimated by the relative frequency of the two words together:

$$P(w_i w_{i+1}) = \frac{C(w_i w_{i+1})}{N} \tag{4}$$

Similarly, the *conditional probability of the target word given the next word* $P(w_i|w_{i+1})$ is the probability of the target word w_i given the next word w_{i+1}. This may be viewed as the predictability of a word given the word the speaker is about to say, and is estimated as follows:

$$P(w_i|w_{i+1}) = \frac{C(w_i w_{i+1})}{C(w_{i+1})} \tag{5}$$

Finally, we mention briefly four other measures that played a smaller role in our analyses. We considered a number of *trigram probability* measures. Two of these were the conditional probability of the target given the *two previous* words $P(w_i|w_{i-2}w_{i-1})$, and the conditional probability of the target given the *two following* words $P(w_i|w_{i+1}w_{i+2})$. Neither of these turned out to predict reduction after we controlled for the (bigram) conditional probabilities of the previous and following words. The conditional probability of the target given the *two surrounding* words is the probability of the target given one word preceding and one word following the target $P(w_i|w_{i-1} \cdots w_{i+1})$. This trigram was a significant predictor in some analyses. It is estimated as follows:

$$P(w_i|w_{i-1} \cdots w_{i+1}) = \frac{C(w_{i-1} w_i w_{i+1})}{C(w_{i-1} \cdots w_{i+1})} \tag{6}$$

Table 1: Summary of probabilistic measures and high probability examples.

Measure		**Examples**
Relative Frequency	$P(w_i)$	**just, right**
Joint of Target with Next Word	$P(w_i w_{i+1})$	**kind** *of*
Joint of Target with Previous	$P(w_{i-1} w_i)$	*a* **lot**
Conditional of Target given Previous	$P(w_i \mid w_{i-1})$	*Supreme* **Court**
Conditional of Target given Next	$P(w_i \mid w_{i+1})$	**United** *States*
Conditional of Target given Surrounding	$P(w_i \mid w_{i-1} \cdots w_{i+1})$	*little* **bit** *more*

Table 1 contains a summary of the probabilistic measures and some examples of high probability items from the dataset for each measure. The reader can obtain some idea of the ways that these different measures of local predictability rank word combinations in Tables 6–8 in Appendix 2.

The actual computation we used for estimating these probabilities was somewhat more complex than the simple explanations above. Since our 38,000 word corpus was far too small to estimate word probabilities, we used the entire 2.4 million word Switchboard corpus (from which our corpus was drawn) instead. See Jurafsky *et al.* (1998) for details about the backoff and discounting methods that we used to smooth the estimates of very low frequency items. We then took the log of these probabilities for use in our regression analyses.

In this paper we report mainly the effects of conditional probabilities. In general, however, we find that most of the measures (conditional probability, joint probability, various relative frequencies of words) show some effect on reduction. Given their definitional interdependence, this is not surprising. If one wishes to pick a single measure of probability for convenience in reporting, it makes sense to pick one which combines several independent measures, such as mutual information (which combines the joint, the relative frequency of the target, and the relative frequency of the neighboring word) or conditional probability (which combines joint probability and the relative frequency of the neighboring word). We chose conditional probability because for this particular data set it was a better single measure than joint probability.

In Gregory *et al.* (1999) we considered the *mutual information* (Fano, 1961) of the target word and the following word. There we showed that mutual information produces very similar results to the conditional probability of the target word given the following word. For this reason, and because mutual information turns out to be an inappropriate metric for our analyses of function words,[1] we report on conditional probability rather than mutual information in this paper.

In general, the most predictive model of any data is obtained by using a combination of (independent) measures rather than one single measure. Thus, for example, in some cases we found that a combination of conditional probability, joint probability, and relative frequency all play a role in reduction. See Appendix 1 for further discussion of the relationships between conditional probability, joint probability, and relative frequency of the previous word.

3 Effects of Predictability on Function Words

Our first experiment studied the 10 most frequent English function words in the Switchboard corpus. (These are also the ten most frequent words in the corpus.)

[1] This is because mutual information includes the relative frequency of the target word. Since the function word analysis was based on only ten types of function words, this relative frequency component will merely act to distinguish the ten items, rather than to represent their frequencies, as it would with a larger sample.

3.1 The Function Word Dataset

The function word dataset was drawn from the Switchboard corpus of telephone conversations between strangers, collected in the early 1990s (Godfrey, Holliman, & McDaniel, 1992). The corpus contains 2430 conversations averaging 6 minutes each, totaling 240 hours of speech and about 3 million words spoken by over 500 speakers. The corpus was collected at Texas Instruments, mostly by soliciting paid volunteers who were connected to other volunteers via a robot telephone operator. Conversations were then transcribed by court reporters into a word-by-word text.

Approximately four hours of speech from these conversations were phonetically hand-transcribed by students at UC Berkeley (Greenberg *et al.*, 1996) as follows. The speech files were automatically segmented into pseudo-utterances at turn boundaries or at silences of 500 ms or more, and a rough automatic phonetic transcription was generated. The transcribers were given these utterances along with the text and rough phonetic transcriptions. They then corrected the phonetic transcription, using an augmented version of the ARPAbet, and marked syllable boundaries, from which durations of each syllable were computed.

The phonetically-transcribed corpus contains roughly 38,000 transcribed words (tokens). The function word dataset consists of all instances of the 10 most frequent English function words: *I*, *and*, *the*, *that*, *a*, *you*, *to*, *of*, *it* , and *in*. This subcorpus contained about 9,000 word tokens. Our analyses are based on the 5,618 tokens remaining after excluding various non-comparable items (see §3.3).

Each observation was coded for two dependent factors reflecting reduction:

vowel reduction: We coded the vowel of each function word as *full* or *reduced*. The full vowels included basic citation or clarification pronunciations, e.g. [ði] for *the*, as well as other non-reduced vowels. The reduced vowels that occurred in the function words were [ə] and [ɨ].[2] Table 2 shows full and reduced-vowel pronunciations of the function-words, while Figure 1 shows the relative proportions of each vowel type by function word.

duration in milliseconds: the duration of the word in milliseconds.

3.2 The Regression Analysis

We used multiple regression to evaluate the effects of our predictability factors on reduction. A regression analysis is a statistical model that predicts a *response variable* (in this case, the word duration, or the frequency of vowel reduction) based

[2]In general we relied on Berkeley transcriptions for our coding, although we did do some data cleanup, including eliminating some observations we judged likely to be in error; see Jurafsky *et al.* (1998) for details.

Table 2: Common pronunciations of the 10 function words by vowel type.

	Full	**Reduced**
a	[eɪ] [ʌ],[ɪ]	[ə],[ɨ]
the	[ði],[i],[di] [ðʌ],[ðɪ],[ʌ]	[ðə],[ðɨ], [ə]
in	[ɪn],[ɪ],[ɪɾ̃], [ɛn],[ʌn],[æn]	[ɨn],[n̩],[ən]
of	[ʌv],[ʌ],[ʌvv] [ɪ],[i],[ɑ]	[ə],[əv],[əf]
to	[tu],[tʉ],[ɾu] [tʊ],[tɪ],[tʌ]	[tə],[tɨ],[ə]
and	[æn],[ænd],[æɾ̃] [ɛn],[ɪn],[ʌn]	[ɨn],[n̩],[ən]
that	[ðæ],[ðæt], [æ] [ðɛ],[ðɛt],[ðɛɾ]	[ðɨt], [ðɨ], [ðɨɾ]
I	[aɪ] [ɑ],[ʌ],[æ]	[ə]
it	[ɪ],[ɪt],[ɪɾ] [ʊt],[ʊ],[ʌ]	[ɨ],[ə],[ət]
you	[yu],[u],[yʉ] [yɪ],[ɪ],[i]	[yɨ],[y],[ɨ]

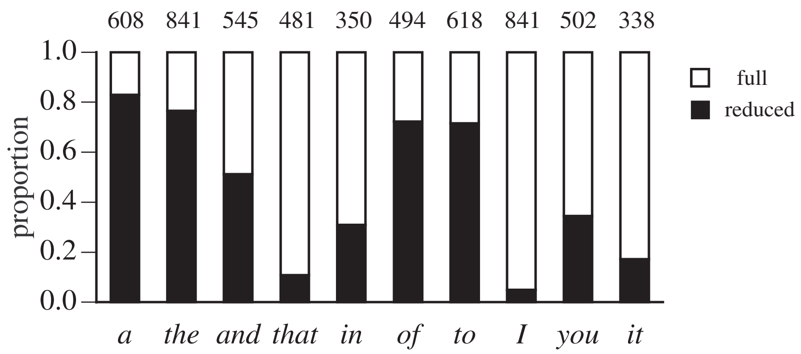

Figure 1: Proportion of full and reduced forms for the 10 function words. Total occurrences appear above.

on contributions from a number of other *explanatory factors* (Agresti, 1996). Thus when we report that an effect was significant, it is meant to be understood that it is a significant parameter in a model that also includes the other significant variables. In other words, after accounting for the effects of the other explanatory variables, adding the explanatory variable in question produced a significantly better account of the variation in the response variable. For duration, which is a continuous variable, we used ordinary linear regression to model the log duration of the word. For vowel quality, which is a categorical variable, we used logistic regression.

3.3 Control Factors

The reduction processes are each influenced by multiple structural and performance factors that must be controlled to assess the contribution of the probability measures to reduction. We briefly review these factors here and our method of controlling for them. First, we excluded tokens of function words based on the following three factors:

Planning problems: We removed function words which are immediately followed by disfluencies indicative of 'planning problems' (pauses, filled pauses *uh* and *um*, or repetitions), since they tend to have less-reduced pronunciations (Fox Tree & Clark, 1997; Jurafsky *et al.*, 1998; Bell *et al.*, 1999; Shriberg, 1999). We also removed words that were preceded by filled pauses since preceding pauses might affect durational patterns.

Phrase boundary position: We removed words which are initial or final in our pseudo-utterances. The pseudo-utterances of our datasets are bounded by turns or long pauses, although they do include multiple intonational phrases in some cases. Thus words which were initial or final in our pseudo-utterances included most words which are turn- or utterance-initial or final. Such words are known to have different durational patterns.

Special forms We removed cliticized function words (e.g., *you've*, *I've*, *it's*) and the variant *an* of the indefinite article *a*.

We then controlled other variables known or suspected to affect reduction by entering them first in the regression model. Thus the base model for an analysis was a regression on the following set of control factors:

Rate of Speech: Speech researchers have long noted the association between faster speech, informal styles, and more reduced forms. For a recent quantitative account of rate effects in Switchboard, see Fosler-Lussier and Morgan (1998). We measured rate of speech at a given function word by taking the number of syllables per second in the smallest pause-bounded region containing the word. Our regression models included both log rate and log squared rate.

Segmental Context: A general fact about reduction processes is that the form of a word is influenced by the segmental context—for example, consonant deletion is favored when a segment is preceded by or followed by a consonant. We controlled for the class (consonant or vowel) of the following segment.

Syllable type of target: We coded the target word for syllable type (open or closed) (e.g., *it* vs. *a*). This variable interacts closely with segmental context.

Reduction of following vowel: The prosodic pattern of the utterance plays a crucial role in reduction. Since our current dataset does not mark stress or accent, the only prosodic control was whether the vowel in the syllable following the target word was reduced or full. (This partially controls for stress since the reduction of the following vowel should correlate with its stress level, and hence the stress level of the target word.)

We also included a number of terms for the interactions between these variables.

Several factors that have been reported to influence reduction were not controlled in this study. First, our definition of words was quite simplified; we assume that anything bounded by spaces in the text transcriptions was a word. Thus *Supreme Court* and *most of* were each considered two words, although we controlled for this simplification in the experiments described in §4. Other factors not controlled included additional aspects of the preceding segment environment (e.g., vowel identity and coda identity), prosodic structure (including position and metrical prominence) and social variables (register, age, gender, race, social class, etc.). We did control for some of these social variables in our earlier work (Bell *et al.*, 1999) and still found robust effects of the predictability measures. Control of reduction of the following vowel and of pseudo-utterance position in our analyses partially controls effects of prosodic structure, stress, and accent.

The fact that the 10 words in this dataset were all very frequent limited our ability to study relative frequency. (The most common word, *I*, is about 3 times more frequent than the least common word *in*, compared to an overall ratio of probability of about 100,000 to 1 for the highest and lowest frequency words in the entire corpus.) What variation there is, moreover, is inextricably confounded with the effects of form and patterns of combination of the individual items. Since it is consequently not possible to obtain useful inferences about the effects of relative frequency with the function words dataset, this variable is omitted from the analyses.

3.4 Results

3.4.1 Vowel Reduction in Function Words

We first tested the relationship between the target word and the previous word, by adding the conditional probability of the target word given the previous word $P(w_i|w_{i-1})$ to the regression equation after a base model that included the control variables. Predictability from the previous word was a significant predictor of reduction ($p < .0001$). The higher the conditional probability of the target given the previous word, the greater the expected likelihood of vowel reduction in the function word target.

The predicted likelihood of a reduced vowel in words which were highly predictable from the preceding word (at the 95th percentile of conditional probability) was 48 percent, whereas the likelihood of a reduced vowel in low predictability words (at the 5th percentile) was 24 percent.

Reduction of the target word is also affected by its probabilistic relations with the following word. Higher conditional probabilities of the target word given the following word $P(w_i|w_{i+1})$ were again a predictor of a greater likelihood of reduction ($p = .002$).

The predicted likelihood of a reduced vowel in words which were highly predictable from the following word (at the 95th percentile of conditional probability) was 42 percent, whereas the likelihood of a reduced vowel in low predictability words (at the 5th percentile) was 35 percent. Note that the magnitude of the effect was a good deal weaker than that with the previous word.

Even after accounting for the individual effects of the conditional probability of the preceding and following words, there is a small additional significant effect of the preceding and following words together, as measured by the conditional trigram probability given the two surrounding words ($P(w_i|w_{i-1} \cdots w_{i+1})$ ($p < .02$).

3.5 Function Word Duration

We found similar effects of predictability on function word duration. The conditional probability of the target word given the previous word $P(w_i|w_{i-1})$ was a significant predictor of durational shortening ($p < .0001$). The higher the conditional probability of the target given the previous word, the shorter the target word. High conditional probability tokens (at the 95th percentile of the conditional probability) have a predicted duration of 92 ms; low conditional probability tokens (at the 5th percentile) have a predicted duration of 118 ms.

A similar effect on shortening was found for the relationship of the target word with the following word. The conditional probability of the target word given the following word $P(w_i|w_{i+1})$ was again a strong predictor of shortening; the higher the probability of the target word given the following word, the shorter the target was ($p < .0001$). Tokens which were highly probable given the following word (at the 95th percentile of the conditional probability) have a predicted duration of 99 ms; tokens with low probability given the following (at the 5th percentile) have a predicted duration of 123 ms.

As with vowel reduction, there is a small additional significant effect of the preceding and following words together, as measured by the conditional probability given the two surrounding words ($p < .0001$).

3.6 Independence of Duration and Vowel Reduction

The fact that the vowels in function words are reduced when the words are more predictable could be modeled as a categorical, non-gradient effect. That is, based on predictability, speakers could be making some sort of categorical choice in lexical production between two possible vowels, one full and one reduced. But the results on durational shortening cannot be modeled categorically. The effect of predictability on shortening is a gradient, non-categorical one.

It is possible, however, that the shortening effects that we observe for function words might be solely a consequence of the vowel reduction effects, since reduced vowels are indeed durationally shorter than full vowels. If shortening was only a consequence of vowel selection, there might be no evidence for a gradient effect of probability on reduction. In order to test whether the effects of probability on shortening were completely due to vowel reduction, we added a variable to the base model for duration that coded whether the function word's vowel was reduced or full.

We found that all the probabilistic variables remain robustly significant predictors of duration, even after controlling for vowel reduction. That is, predictability not only affects vowel reduction, but has an additional independent non-categorical effect on word duration.

As further confirmation, we looked at the full and reduced vowels separately to see whether the shortening effects occurred in words with full vowels as well as words with reduced vowels. Indeed, higher probability predicted durational shortening both in the words with full vowels and words with reduced vowels. For words with full vowels and words with reduced vowels, those that had higher conditional probabilities (given either the previous or following word) were significantly shorter than those with lower conditional probabilities ($p = .0001$).

These results confirm that there is an effect of predictability on reduction that is continuous and not purely categorical, suggesting that the domain of applicability of the Probabilistic Reduction Hypothesis includes linguistic levels that allow continuous phenomena.[3]

3.7 The Function Word Dataset: Discussion

The results for the function word dataset show that function words that are more predictable are shorter and more likely to have reduced vowels, supporting the

[3]In order to ensure that the durational effects have some continuous component, we would also need to control for presence or absence of consonants. While we couldn't do a full analysis here, we did examine the durations of a subset of 2878 items in which all consonants were present. Even after controlling for these categorical factors (vowel quality and consonant presence), target words were still shorter when they had a high conditional probability given the following word, or a high joint probability with the previous word.

Probabilistic Reduction Hypothesis. The conditional probability of the target word given the preceding word and given the following one both play a role, on both duration and deletion. The magnitudes of the duration effects are fairly substantial, in the order of 20 ms or more, or about 20 percent, over the range of the conditional probabilities (excluding the highest and lowest five percent of the items).

The fact that there are effects of predictability on duration in addition to the effects on vowel reduction, and that they affect both full and reduced vowels, suggests that some of the effects of predictability on reduction are continuous and non-categorical. Under one possible model of these effects, the categorical vowel reduction effects could be the result of lexicalization or grammaticalization leading to segmental changes in the lexicon or grammar, while the continuous duration effects are on-line effects, perhaps mediated in part by prosodic structure, but not represented in lexicalized differences. Our results do not allow us to make any conclusions about such a possible model. Indeed, while our results, like many results on variation phenomena, could arise from two qualitatively different processes, one applying more generally across items and processes and one the result of lexicalizations and grammaticalizations, these need not map cleanly into categorical and non-categorical reductions. At least some vowel reduction may be gradient, and it is conceivable that some of the duration effects demonstrated above could arise from lexicalization. Thus the actual delineation of a model of the effects of predictability on reduction remains to be done.

4 Lexical versus Collocation Effects

So far we have shown that the conditional probability of a function word given the surrounding words is a significant predictor of reduced vowels and shorter durations. Shortening effects seems to provide strong evidence that probabilistic links between words are represented in the mind of the speaker.

But an examination of the high probability word pairs in Tables 6–8 (Appendix 2) raises a potential problem. Many of these pairs (like *sort of* or *kind of*) might be single lexical items rather than word pairs (*sorta*, *kinda*). This classification as high-probability word pairs would then stem from the fact that we rely on a purely orthographic definition of a word (i.e., words are separated by white space). Perhaps our results concerning the effect of predictability on reduction are merely facts about such recently emergent words like *sorta*, and not facts about probabilistic relations between words that are accessed separately. That is, perhaps our results are purely lexical rather than syntactic (e.g., word-order) facts about reduction.

In order to test this hypothesis, it is necessary to show that higher predictability is associated with increased reduction even in word combinations that are not lexicalized. Based on the intuitions that many pairs of words with high conditional probability may be lexicalized (see the top half of Tables 7 or 8) and word pairs with low conditional probabilities are likely not (see the bottom half of Tables 7 or 8), we split the function word observations into two groups of high and low conditional probabilities. Table 3 shows the 10 sequences with the highest conditional probabilities from the **lower** half of the range. Looking at these tables, these words are less likely to be lexically combined with their neighbors, and yet their duration is still affected by both the conditional probability given the preceding and the conditional probability given the following word. The higher the probability of the word given its neighbor, the shorter the word.

Table 3: The 10 most probable function word sequences in context from the lower half of the probability range, according to two probability measures. Function words in this lower range did show effects of durational shortening due to higher probability.

Conditional Probability Given Previous Word $P(w_i \vert w_{i-1})$	**Conditional Probability Given Next Word** $P(w_i \vert w_{i+1})$
Top 10 of lower half	Top 10 of lower half
them **and**	**a** chocolate
sometime **in**	**a** law
differences **of**	**a** crime
bet **that**	**the** old
homes **that**	**the** gun
does **that**	**you** must
where **the**	**the** Mastercard
been **a**	(oil) **and** filter
with **a**	**the** north
fine **and**	**I** do

For each of these groups, we tested the effects of conditional probability given the previous word on both vowel reduction and durational shortening. Each test was then repeated for the conditional probability given the following word. Since lexicalized sequences of words should have high conditional probabilities, if the effects we find are limited to lexicalizations, we should find that our effects only hold for the upper halves of the conditional probabilities.

Considering first the effects of the preceding word, we found that there was no significant effect of conditional probability on vowel reduction in the low group, but there was a significant effect of conditional probability in the high group. These results lend some support for the influence of lexicalization. For duration, however, conditional probability of the preceding word had a significant effect for both groups, although it did appear to be somewhat stronger for the high group.

The results for following word effects did not support the lexicalization hypothesis. Conditional probability of the following word was just as good a predictor of vowel reduction in the low probability group as in the high probability group.

We were surprised to find that the duration of tokens in the high group was *not* affected by conditional probability given the following word, even though durations in the low group were shorter for higher conditional probabilities. This suggests that there may be a ceiling that limits its effect on duration.

While these results are preliminary, and invite further analysis, they suggest two conclusions. First, more predictable words are more reduced even if they are in a low probability group and unlikely to be lexically combined with a neighboring word. Thus we find clear evidence for probabilistic relations between words. Second, particularly for the predictability from the previous word, the high group shows a stronger effect of predictability on reduction. This suggests that there is some reduction in duration may be due to the lexicalization of word pairs.

5 Effects of Predictability on Final-t/d Content Words

Our previous results show that function words which are very predictable from neighboring words (i.e., have high conditional probability given the previous or following word) are more reduced. Even though these results show that probabilistic relations hold over the full range of predictabilities for function words, it is possible that they would not hold for content words. This might be true, for example, if function words are more likely to cliticize, lexicalize, or collocate with neighboring words than content words, or if probabilistic relations between words were to only apply at the higher ranges of predictability that are more typical of function words. Because content words have a much wider range of frequencies than function words, they also allow us to investigate the role of target word frequency. We therefore turn to content words to see if they are also reduced when they are more probable.

5.1 The Final-t/d Content Word Dataset

The Final-t/d Content Word dataset is again drawn from the 38,000 word phonetically-transcribed Switchboard database. (See §3.1 for details.) The database contained about 3000 content words ending in t or d. Eliminating observations to control for factors discussed below left 2042 word tokens in our analyses. Table 4 shows some common examples, together with frequencies per million words from the entire 2.4 million word Switchboard corpus.

Each observation was coded for two dependent reduction factors:[4]

[4]Our earlier work also considered other reduction factors; see Jurafsky *et al.* (1998) for our results on deletion of coda obstruents in function words (*it*, *that*, *and*, *of*) and Gregory *et al.* (1999) on tapping in final-t/d words.

Table 4: The 30 most frequent words in the Final-t/d dataset, with counts from the 2.4 million word Switchboard corpus, but renormalized (divided by 2.4) to be counts-per-million.

Word	**Frequency**	**Word**	**Frequency**	**Word**	**Frequency**
want	12,836	last	887	read	604
just	8,781	bit	863	part	585
lot	3,685	first	834	fact	585
good	3,225	thought	826	heard	523
kind	3,103	need	826	made	521
put	1,226	sort	823	start	484
said	1,190	old	818	least	461
went	1,153	great	793	point	460
used	941	bad	669	state	452
most	899	quite	628	let	442

Deletion of final consonant: Final t-d deletion is defined as the absence of a pronounced oral stop segment corresponding to a final t or d in words. A final t or d was coded as deleted if in the Greenberg *et al.* (1996) transcription the t or d was not transcribed as phonetically realized. For example, the phrase 'but the' was often pronounced [bəðə] in the dataset, with no segment corresponding to the t in *but*. Table 5 shows examples of full and t/d-deleted forms.

Duration in milliseconds: The hand-coded duration of the word in milliseconds.

Table 5: Examples of full (including tapped) and reduced (i.e., deletion of final t or d) forms from the final-t/d dataset.

Word	**Full and Tapped Forms**	**Forms with Deleted t or d**
mind	[maɪnd]	[maɪn], [maɪ]
about	[əbʌd],[baʊt]	[bæ]
made	[maɪd], [meɪɾ]	[meɪ]
most	[moʊst], [moʊt]	[moʊs] [m]
lot	[lɑt], [lɑɾ]	[lɑ]

5.2 Control Factors

As with the function word analyses, we excluded tokens of words which occurred in disfluent contexts, or initially or finally in pseudo-utterances. We also excluded polysyllabic words from the duration analyses to make the items more comparable.

Other factors were controlled by including them in the regression model before considering the predictability factors. They included variables already discussed—rate of speech, rate of speech squared, whether the next vowel was reduced or not, following segment type (consonant or vowel), and whether the word coda included a consonant cluster. The base model also included the following additional factors:

Inflectional status: Fasold (1972), Labov (1972), Bybee (1996) and others noted that a final t or d which functions as a past tense morpheme (e.g., *missed* or *kept*) is less likely to be deleted than a t or d which is not (e.g. *mist*).

Identity of the underlying segment: We coded the identity of the underlying final segment (t or d).

Number of syllables: The number of syllables in the word is of course correlated with both word frequency and word duration (for the deletion analysis only, since the duration analysis was limited to monosyllabic words).

5.3 Results

Using multiple regression, the predictability measures were tested on the two shortening variables of deletion and duration by adding them to each of the regression models after the base model. Recall that in the function word experiment we did not include the relative frequency of the target word as a factor. For the content words, however, this factor was included. Note that while targets are content words, preceding and following words may be function words.

5.4 Duration

The duration analysis was performed on 1412 tokens of the final-t/d content words.

We found a strong effect of the relative frequency of the target word ($p < .0001$). Overall, high frequency words (at the 95th percentile of frequency) were 18% shorter than low frequency words (at the 5th percentile).

The conditional probability of the target given the next word significantly affected duration: more predictable words were shorter ($p < .0001$). Words with high conditional probability (at the 95th percentile of the conditional probability given the next word) were 12% shorter than low conditional probability words (at the 5th percentile).

Both the conditional probability of the target given the previous word ($p = .0009$) and the joint probability of the target with the previous word ($p = .046$) significantly affected duration. This instance is complicated in that no one factor adequately represents the effects on duration.

5.5 Deletion

The deletion analysis was performed on 2042 tokens of t/d-final content words.

Again, we found a strong effect of relative frequency ($p < .0001$). High frequency words (at the 95th percentile) were 2.0 times more likely to have deleted final t or d than the lowest frequency words (at the 5th percentile).

The conditional probability of the target given the previous word did not significantly affect deletion. The only previous word variable that affected deletion in target words was the relative frequency of the previous word. More frequent previous words lead to less deletion in the target word ($p = .007$).

We had found in earlier work (Gregory *et al.*, 1999) that deletion was not sensitive to predictability effects from the following word. This result was confirmed in our current results. Neither the conditional probability of the target word given the next word nor the relative frequency of the next word predicted deletion of final t or d.

5.6 Final-t/d Content Word Dataset: Discussion

Content words with higher relative frequencies (prior probabilities) are shorter and are more likely to have deleted final t or d than content words with lower relative frequencies. As is the case with all of our results, this is true even after controlling for rate of speech, number of syllables, and other factors. The effect of target word frequency was the strongest overall factor affecting reduction of content words, and provides support for the Probabilistic Reduction Hypothesis.

In addition to the effect of relative frequency, we also found an effect of conditional probability. Content words which have a higher conditional probability given the following word are shorter, although not more likely to undergo final segment deletion.

Overall, however, the effects of conditional probability on reduction are much weaker in content words than we saw in function words. Conditional probabilities of the targets given either the following or the previous word had no effect on deletion. We also found no effect of the conditional probability of the target word given the previous word on duration. Failure to find effects may be due to the smaller number of observations in the content word dataset or the general lower frequencies of content words.

The only effect of the previous word was an effect of previous-word relative frequency. High-frequency previous words led to *longer* target forms and *less* final-t/d deletion. Unlike the effects of joint and conditional probabilities which plausibly represent the predictability of the target word, the effect of previous (or following) word frequency has no immediate interpretation. We are currently investigating two possible explanations for the role of previous-word frequency. One possibility is based on the fact that the previous-word frequency is in the denominator of the equation defining the conditional probability of the word given the previous word (Equation 3; see also Appendix 1). Perhaps the effect of previous word frequency is really a consequence of conditional probability, but the size of our content-word dataset is too small to see the effects of the numerator of Equation 3. This could be due to the fact that the counts for any two-word combinations are lower than the counts for single words.

Another possibility is that the lengthening of content words after frequent previous words is a prosodic effect. For example, if the previous word is frequent, it is less likely to be stressed or accented, which might raise the probability that the current word is stressed or accented, and hence that it is less likely to be reduced.

Prosodic effects might also explain the asymmetric effect of surrounding words (the preceding word played little role in final deletion). This likely illustrates that not all reduction processes are affected in the same way by probabilistic variables. (Gregory *et al.* (1999), for example, found a different pattern for tapping of final t and d.) The asymmetry of this particular case is perhaps understandable from the fact that final deletion is a word edge effect, in the terminology of the phonological of prosodic domains. It would be worth investigating whether such edge processes are systematically less sensitive to the probability conditioning effects of material across the prosodic boundary they mark.

6 Conclusion

The fundamental result of these analyses is that we find evidence for the Probabilistic Reduction Hypothesis. In general, more probable words are reduced, whether they are content or function words. Predictability from neighboring words played a strong role in the high-frequency function words. The content words exhibited weaker effects of surrounding context, but strong effects of relative frequency. Thus all of our measures of local predictability play a role in at least some reduction processes, and all the reduction processes are influenced by some predictability measures. By showing that probabilistic factors influence lexical production, our results also provide general support for probabilistic models of human language processing (Jurafsky, 1996; Seidenberg & MacDonald, 1999)

Our analyses also show that predictability links between words are a key factor in such probabilistic models. We showed, using several kinds of evidence, that the effect of the neighboring word on reduction was not necessarily due to lexicalization. This includes evidence that the effect of predictability on reduction applies both to content and function words, and that the effect of predictability applies both to the higher and to the lower ranges of predictability. The fact that the shortening effects are independent of vowel reduction also tends to support this hypothesis, since such gradient processes are more likely at production processing levels after lexical items have been merged into a prosodic frame.

This is an ongoing research effort, and we are currently extending these results in a number of directions, including further examination of the slightly different effects on context versus function words, use of larger and more general datasets, and effect of other measures of collocation and predictability.

Acknowledgements

This project was partially supported by NSF IIS-9733067. Thanks to Eric Fosler-Lussier for supplying us with the N-gram probability distributions and collaborating on many earlier related projects, Cynthia Girand for helping with the construction of the original function word database and collaboration on related projects, Joan Bybee for inspiring this line of research and for useful feedback and comments, Matthew Dryer and David Perlmutter for (independently) suggesting that we check more carefully if the effects of predictability were confined to the lexicon, and to the audiences of various talks, including the CMU workshop and the linguistics department at UC San Diego.

Appendix 1: Joint versus Conditional Probability

In the body of this paper we reported on the conditional probability as a measure of word predictability. This appendix summarizes a slightly different way of looking at conditional probability.

Recall that the conditional probability of the target word given the previous word is estimated from two counts:

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} \tag{7}$$

An alternative computation substitutes probabilities for the constituent counts, since the probabilities are just the counts divided by a normalizing constant, and the normalizing constants cancel:

$$P(w_i|w_{i-1}) = \frac{P(w_{i-1}w_i)}{P(w_{i-1})} \tag{8}$$

Thus the conditional probability is made up of two probabilities: the *joint probability with the previous word* $P(w_{i-1}w_i)$ (which may be thought of as the 'relative frequency of the two words occurring together') and the *relative frequency of the previous word* $P(w_{i-1})$. This means that instead of using the conditional probability in the regression equation to predict reduction, we can add in the two relative frequencies instead as independent factors.

Adding in these two factors to the regression (directly after the base model, i.e., without the *conditional probability given previous* variable) showed that both play a role in vowel reduction ($p < .0001$).

Probability	Equation	Regression Coefficient
Joint Probability of Target with Previous Word	$P(w_{i-1}w_i)$	$-.503$
Relative Frequency of Previous Word	$P(w_{i-1})$	$+.724$

The *regression coefficient* gives the weight that the regression assigned to each factor. The negative coefficient for the joint probability means that the higher the joint, the more likely the word's vowel is reduced. By contrast, the coefficient is positive for previous word probability. This means that a higher previous word probability predicts **less reduction**. This is what we would expect from the probabilistic model, since the prior probability of the previous word is in the denominator in Equation 8.

The difference between the analyses is that the conditional probability essentially holds the relative weights of the joint and preceding word probabilities equal, whereas in the second analysis they are free to vary. The regression is essentially telling us, for this set of data, that the joint probability should be weighted somewhat less heavily than the previous word's relative frequency. We can see the relationship a different way by combining the conditional probability with the joint.

Probability	Equation	Regression Coefficient
Conditional Probability of Target Given Previous	$P(w_{i-1}\|w_i)$	$-.724$
Joint Probability of Target with Previous	$P(w_{i-1}w_i)$	$+.221$

It is not a coincidence that the coefficient of the conditional probability ($-.724$) is the same magnitude as the coefficient of the previous word's relative frequency in the first analysis. The first analysis gives the relative weights of the two basic (log) probabilities. Since the weight of the relative frequency must be $.724$, and in the second analysis its only expression is through the denominator of the conditional probability, the conditional probability must have a weight of $-.724$. Thus the coefficient of the joint probability ($+.221$) in this regression exactly compensates for the difference between the joint and the prior probabilities in the second analysis ($-.503 + .724$).

These results (and similar ones for the conditional probability of the target given the following word) suggest that the components of conditional probability may be playing slightly different roles in reduction, and reflect different causes. This is clearly an area that calls for further study.

Appendix 2: Examples of Conditional Probabilities

Table 6: The function word contexts with the highest conditional probabilities, according to three probability measures. Target function words are in boldface. Note that *of* and *to* are most likely to collocate with the previous word, while *I*, *the* and *a* tend to collocate with the following word. *To* is most predictable from the surrounding two words.

Highest Probability Given Previous Word $P(w_i \vert w_{i-1})$	**Highest Probability Given Next Word** $P(w_i \vert w_{i+1})$	**Highest Probability Given Surrounding Word** $P(w_i \vert w_{i-1} \cdots w_{i+1})$
rid **of**	**I** guess	going **to** be
supposed **to**	**I** mean	well **I** guess
tends **to**	**the** midwest	know **I** mean
ought **to**	**a** lot	have **a** lot
kind **of**	**a** shame	do **a** lot
able **to**	**the** Kurds	supposed **to** be
sort **of**	**the** wintertime	used **to** be
compared **to**	**in** terms	matter **of** fact
kinds **of**	**the** same	quite **a** bit
tend **to**	**the** United	kind **of** thing

Table 7: Effects of the previous word. The final-t/d content words with the highest and lowest conditional probabilities given the previous word, and the highest and lowest joint probabilities with the previous word. The target word is in boldface.

Highest Probability Given Previous Word	**Highest Joint Probability with Previous Word**
supreme **court**	a **lot**
Amsterdam **Holland**	i **get**
doctoral **student**	i **just**
sesame **street**	a **good**
capital **punishment**	it's **just**
Harrison **Ford**	little **bit**
German **shepherd**	i **thought**
awful **lot**	was **just**
backyard's **great**	it **just**
raters **loved**	my **husband**

Lowest Probability Given Previous Word	**Lowest Joint Probability with Previous Word**
and **punished**	non **colored**
and **proceed**	blind **sided**
and **shred**	tongue **pressed**
and **disinterested**	Arizona **used**
and **sauerkraut**	girls **kind**
and **closed**	Lehrer **report**
and **gold**	student **discount**
and **touched**	tomatoes **next**
and **ironside**	soccer **filed**
and **bloomed**	families **end**

Table 8: Effects of the following word. The final-t/d content words with the highest and lowest conditional probabilities given the next word, and the highest and lowest joint probabilities with the next word. The target word is in boldface.

Highest Probability Given Next Word	**Highest Joint Probability with Next Word**
United States	**kind** of
good heavens	**lot** of
last resort	**want** to
east coast	**sort** of
need trimming	**used** to
Burt Reynolds	**need** to
called crier	**just** a
government entities	**most** of
good fellas	**part** of
grapefruit citron	**went** to

Lowest Probability Given Next Word	**Lowest Joint Probability with Next Word**
threatened i	**eight** engines
hold i	**installed** the
ragged i	**harmed** you
indoctrinated i	**engaged** to
England i	**unemployment** insurance
liberated i	**determined** and
road the	**filmed** in
draft the	**blind** sided
misclassed the	**dependent** you
installed the	**homemade** pasta

References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. John Wiley & Sons, New York.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., & Gildea, D. (1999). Forms of English function words – Effects of disfluencies, turn position, age and sex, and predictability. In *Proceedings of ICPhS-99*.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.

Bush, N. (1999). The predictive value of transitional probability for word-boundary palatalization in English. Master's thesis, University of New Mexico, Albuquerque, NM.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of *don't* in English. *Linguistics*, *37*(4), 575–596.

Bybee, J. L. (1996). The phonology of the lexicon: evidence from lexical diffusion. In Barlow, M., & Kemmer, S. (Eds.), *Usage-based Models of Language*.

Fano, R. M. (1961). *Transmission of information; a statistical theory of communications*. MIT Press.

Fasold, R. W. (1972). *Tense marking in Black English*. Center for Applied Linguistics, Washington, D.C.

Fidelholz, J. (1975). Word frequency and vowel reduction in English. In *CLS-75*, pp. 200–213. University of Chicago.

Fosler-Lussier, E., & Morgan, N. (1998). Effects of speaking rate and word frequency on conversational pronunciations. In *ESCA Tutorial and Research Workshop on Modeling pronunciation variation for automatic speech recognition*.

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, *26*, 489–504.

Fox Tree, J. E., & Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, *62*, 151–167.

Francis, W. N., & Kučera, H. (1982). *Frequency Analysis of English Usage*. Houghton Mifflin, Boston.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *IEEE ICASSP-92*, pp. 517–520. IEEE.

Greenberg, S., Ellis, D., & Hollenback, J. (1996). Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In *ICSLP-96*, pp. S24–27 Philadelphia, PA.

Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. In *CLS-99*. University of Chicago.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 824–843.

Jespersen, O. (1922). *Language*. Henry Holt, New York.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*, 137–194.

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., & Raymond, W. D. (1998). Reduction of English function words in Switchboard. In *ICSLP-98*, Vol. 7, pp. 3111–3114 Sydney.

Krug, M. (1998). String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change. *Journal of English Linguistics*, *26*, 286–320.

Labov, W. (1972). The internal evolution of linguistic rules. In Stockwell, R. P., & Macaulay, R. K. S. (Eds.), *Linguistic Change and Generative Theory*, pp. 101–171. Indiana University Press, Bloomington.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, *32*, 692–715.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Martinet, A. (Ed.). (1960). *Elements of General Linguistics*. University of Chicago Press, Chicago.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*, 283–312.

Narayanan, S., & Jurafsky, D. (1998). Bayesian models of human sentence processing. In *COGSCI-98*, pp. 752–757 Madison, WI. LEA.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*, 273–281.

Rhodes, R. A. (1992). Flapping in American English. In Dressler, W. U., Prinzhorn, M., & Rennison, J. (Eds.), *Proceedings of the 7th International Phonology Meeting*, pp. 217–232. Rosenberg and Sellier.

Rhodes, R. A. (1996). English reduced vowels and the nature of natural processes. In Hurch, B., & Rhodes, R. A. (Eds.), *Natural Phonology: The State of the Art*, pp. 239–259. Mouton de Gruyter.

Roland, D., & Jurafsky, D. (2000). Verb sense and verb subcategorization probabilities. In Merlo, P., & Stevenson, S. (Eds.), *Volume from CUNY-98*. Benjamins, Amsterdam.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical cues in language acquisition: Word segmentation by infants. In *COGSCI-96*, pp. 376–380.

Schuchardt, H. (1885). *Über die Lautgesetze*. Cited in Jespersen (1923).

Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, *23*, 569–588.

Shriberg, E. (1999). Phonetic consequences of speech disfluency. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS-99)*, Vol. I, pp. 619–622 San Francisco.

Trueswell, J. C., & Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In Clifton, Jr., C., Frazier, L., & Rayner, K. (Eds.), *Perspectives on Sentence Processing*, pp. 155–179. Lawrence Erlbaum, Hillsdale, NJ.

Zipf, G. K. (1929). Relative frequency as a determinant of phoentic change. *Harvard Studies in Classical Philology*, *15*, 1–95.