Danielle Barth*

Effects of average and specific context probability on reduction of function words *BE* and *HAVE*

https://doi.org/10.1515/lingvan-2018-0055
Received April 28, 2017; accepted September 27, 2018

Abstract: In a study of word shortening of *HAVE* and contraction of *BE*, it is found that both high transitional probability and high average context probability (low informativity) result in reduction. Previous studies have found this effect for content words and this study extend the findings to function words. Average context probability is by construction type, showing that words are shorter in constructions with high average predictability, namely in perfect constructions for *HAVE* and in future and progressive constructions for *BE*. These findings show that in cases of grammaticalization, it is not an increase in frequency that results in reduction, but a decrease in informativity.

Keywords: predictability; informativity; function words; reduction; contraction.

1 Reduction and grammaticalization

1.1 Overview

Word frequency, probability and informativity all affect word realization. The present study examines the effect of construction informativity as a predictor of reduction for grammaticalized *BE* and *HAVE*. Grammaticalization is the process by which a word develops a new, grammatical meaning. For instance, the copula *BE* grammaticalized to have passive and progressive auxiliary meanings and later the progressive auxiliary grammaticalized to have a future meaning (*she is a woman* > *she is seen, she is reading* > *she is going to/gonna read*). The copula *BE* is then the source for the grammaticalized words progressive *BE* and passive *BE*. One of the consequences of grammaticalization can be phonetic erosion: *reduction* of the new word in comparison to the source word (Gabelentz 1891; Bybee and Pagliuca 1985; Givón 1985; Heine 1993, Heine 2003; Hopper and Traugott 1993; Bybee et al. 1994; Lehmann 1995). Reduction may take place because of a speaker's desire to differentiate the grammaticalized words from their source words, or because the grammaticalized word increases in frequency in comparison to the source word, and more frequent words are shorter than less frequent words (Bybee 2001, Bybee 2007; Traugott 2011). However, grammaticalization research also shows that words do not grammaticalize on their own, but in specific contexts (see, for example Bybee 2002; Diewald 2002; Heine 1993; Hopper and Traugott 1993; Traugott 2003). So, although frequency effects have been well studied, something else seems to be needed to explain the effects of context.

Construction informativity – an information theoretic measure calculated as the average probability across contexts of a construction – offers a different approach to understanding why reduction occurs in new contexts. When a word grammaticalizes, it is used in a different construction from its source construction. The new construction has a narrower meaning and more limited context of use. The word *BE* is a prime example of this tendency for the grammaticalized usages to have restricted scope – the passive and progressive auxiliaries have a much more limited context than the copula, as they can only co-occur with past and present participles respectively. A future auxiliary has an even more limited context as it can only occur with a present

***Corresponding author: Danielle Barth,** Australian National University, Centre of Excellence for the Dynamics of Language, H. C. Coombs Building, Fellows Road, Acton, Australian Capital Territory 2601, Australia, E-mail: danielle.barth@anu.edu.au

participle of the verb *go*. As the contexts in which a given word may occur become restricted, the occurrence of the word is more predictable in those contexts, and therefore less informative. Hence, a decrease in informativity (and an increase in probability) may be a better predictor of reduction during grammaticalization than frequency.

The probabilistic reduction hypothesis (Gregory et al. 1999; Jurafsky et al. 2001; Bell et al. 2003; Jaeger and Buz 2016) states that word forms are reduced when they are probable given the current context, and are enhanced when they are less probable. Context includes local context (neighboring words), syntactic or lexical structure, semantic or style expectations, and discourse factors (Bell et al. 2003). Seyfarth (2014) found that content words that occur in highly probable contexts often tend to be reduced, even when in an unpredictable context. This shows that speakers are sensitive to tendencies associated with words because they produce reduced speech even when incongruent with the current context. The average informativity of the words accounts for the same effect size as the probability of specific context of the words.

After discussing the probabilistic reduction and the motivation for expanding this line of research to function words in the remainder of Section 1, I present the data, dependent and independent variables for a multimodel analysis of word shortening and contraction in Section 2. In Section 3, I present results showing that informativity, among other predictors, has a significant influence on reduction. I end the paper with a discussion of how function word reduction impacts lexical representation and grammaticalization theory.

1.2 Why does reduction occur in predictable contexts?

The literature on probabilistic reduction has focused on content words, and has found that the main motivations for reduction are frequency and probability. There are several types of theories accounting for why these effects exist. Proponents of speaker-internal models argue that words are reduced when they are highly frequent or highly probable. It takes less time for a speaker to retrieve the word from memory in the course of speech production, because that particular word is available due its high frequency or its probable context (Ferreira 2008; Gahl 2008; Bell et al. 2009; Gahl et al. 2012; inter alia). As a consequence, frequent or predictable strings have a higher resting activation, because they are retrieved from the memory more often (Fedorenko et al. 2006; Jaeger and Tily 2010).

Proponents of listener-oriented models argue that words reduce when the speaker assesses them as easy for a listener to predict from context. Less signal (in terms of milliseconds of duration) is necessary for a listener to decode the meaning of the word and therefore to understand the speaker's intent, so a speaker can get away with a shortened, more coarticulated, reduced pronunciation, saving effort (Lindblom 1990; Lindblom et al. 1995; Galati and Brennan 2010; Wedel 2012; inter alia). Communicative models argue that speakers want to maintain a more constant rate of information transfer, so more time is spent on informative words, less on predictable ones (cf. van Son and Pols 2003; Pluymaekers et al. 2005; Aylett and Turk 2006).

As an extension of listener-oriented accounts, in an exemplar account, a word's phonological representation is a distribution of previously encountered productions that is sensitive to context (Pierrehumbert 2001; Wedel 2006). Frequent words may reduce over time, on average, because they can be interpreted correctly more often, even with a reduced acoustic signal. An increasingly higher proportion of the representations are reduced ones, shifting the average realizations towards more reduced forms (Pierrehumbert 2002). Additional evidence suggests that speakers store information about context probability (Arnon and Cohen Priva 2013; Seyfarth 2014; Jaeger and Buz 2016) and that informativity accounts for reduction better than frequency (Cohen Priva and Jaeger 2018; Hall et al. 2018).

The present study aims to place the focus on function words instead of content words, in particular, *BE* and *HAVE*, because they not only have different meanings but different meaning (construction) frequencies. Using these words, construction informativity is evaluated as a predictor of function word reduction. Data from the Buckeye corpus of American English (Pitt et al. 2007) is used to test whether the reduction which accompanies grammaticalization is better predicted by informativity rather than frequency.

1.3 Content words vs function words

Most previous studies of probabilistic reduction have investigated content words, excluding function words due to the assumption that content and function words are processed differently (à la Levelt et al. 1999; Ullman 2001). The few studies examining function words show that they are less affected by repeated mention (Bell et al. 2009) and more affected by predictability given the preceding context (Jurafsky et al. 2001), but both content and function words reduce in cases of high following context probability (Bell et al. 2009). Studies of contraction of the function words *HAVE* and *BE* indicate that contraction is more likely when *BE* or *HAVE* is highly probable given the context (Krug 1998; Frank and Jaeger 2008; Bresnan and Spencer 2016; Barth and Kapatsinski 2017).

In all its meanings *BE* is a function word, allowing a controlled look at expression differences among grammatical meanings of the same word. Some meanings of *HAVE* are grammatical, but the possessive meaning is lexical, allowing an examination of grammatical vs. lexical meanings of the same word form.

2 Methodology

2.1 Data

The Buckeye Corpus (Pitt et al. 2007) contains 40 interviews with residents of Central Ohio. All speakers are white middle- or working-class individuals. Speakers are balanced for age and gender in the corpus. Sound files were force aligned and then hand-corrected by the corpus creators. Time stamps from the corpus data were used to calculate the durations used in the present study. Tokens were limited to words occurring utterance medially and not occurring next to pauses or disfluencies, hesitations or word re-starts.

2.2 Dependent variables: duration and contraction

In the current study, reduction is operationalized as being short in milliseconds or contracted. Other studies of probabilistic reduction have also included consonant lenition, flapping, stop release, vowel centralization, increased coarticulation, word omission, and lack of intonational prominence (Fowler and Housum 1987; Byrd 1994; van Bergem 1995; Bybee 2001; Gahl and Garnsey 2004; Aylett and Turk 2006; Gahl 2008; Bell et al. 2009; Jaeger and Buz 2016). The dependent variables to be investigated are contraction for *BE* words and duration for *HAVE* words. Duration is the dependent variable for *HAVE* because outside of perfect constructions it does not contract. Earlier investigations (Barth 2015) have shown that construction informativity affects *BE* contraction, not *BE* duration (which is affected by specific probability context among other influences), and so the focus here is on *BE* contraction. All words in the study are monosyllabic: *am, are, is, had, has, have*.

The duration of *HAVE* is normalized. For each target word, duration is measured in milliseconds, then normalized by the speech rate of the utterance, defined as word duration/(sentence duration/syllables in utterance). An utterance is defined as a string of words with no pauses. A long normalized duration will be over 1 (longer than the average syllable in the utterance), and a short one less than 1 (shorter than average syllable in the utterance). Function words often undergo segment reduction, and a reduced form may actually be the most frequent form, so a baseline duration (partially) based on segment durations à la Bell et al. 2009 or Jaeger and Buz 2016 is not ideal. The measure used here takes into account speech rate at the context level, in order to bring out the shorter time spent on function words in comparison to other words in the utterance. Per-word random intercepts (see Section 2.3.6) are also included in the model. Table 1 shows the number of tokens and ranges for normalized duration and raw duration in milliseconds for *HAVE (had, has, have)*.[1] Duration is scaled and centered.

1 For *HAVE* duration, only non-contracted forms are included as tokens for the analysis, however both contracted and not contracted uses of *HAVE* are included in calculations of construction informativity (see Section 2.3.1) ($n = 3{,}790$).

Table 1: Word distribution and length ranges.

Word	*n*	Range in milliseconds	Normalized range	Mean of normalized values
had	747	32.000–469.582	0.207–2.510	0.976
has	273	55.837–893.854	0.397–3.136	1.195
have	1674	24.623–696.048	0.190–2.768	0.996

Table 2: Contraction by lemma in the Buckeye Corpus.

Word	Contracted	Non-contracted	Total
am	1008	32	1040
are	1478	502	1980
is	4887	921	5808

The dependent variable contraction is binary: in a given context a word has either been contracted or it has not. Contraction is only examined here in contexts where it is possible. So only in present tense inflections, and, for example, tokens are excluded that have an *is* following a word ending in a sibilant such as *Texas, this, which, etc*. ($n = 8828$). Table 2 shows the number of contracted and contractible non-contracted tokens of *BE (am, are, is)*.[2]

2.3 Independent variables

2.3.1 Construction

BE tokens come from three different constructions: copula, future, or progressive. *HAVE* tokens come from either modal, possessive, or perfect constructions. Are there any construction-specific effects of reduction beyond the specific following joint or transitional probability of the target instances? Construction type was annotated by the author using context for disambiguation.

Table 3 shows several characteristics of the constructions: frequency, construction informativity, informativity given following word and informativity given preceding word. These measures are presented as properties of the constructions that may be responsible for, or contribute to, reduction. High frequency is a

Table 3: Informativity and frequency by construction.

Construction	Following informativity	Preceding informativity	Construction frequency
BE			
Copular	1.359	0.822	9611
Progressive	0.999	0.860	1665
Future	0.668	0.805	511
Passive	1.052	0.954	286
HAVE			
Possessive	1.939	1.326	1841
Modal	1.543	1.237	470
Perfect	0.855	1.169	2694

Note: Informativity and frequency values combined from both contraction and duration data. Frequency is calculated using construction frequencies from the Buckeye Corpus, and informativity is calculated using frequencies from the COCA (Davies 2008).

2 For *BE* contraction, only contract licensed forms are included as tokens for the analysis, however, contractible and non-contractible tokens ($n = 12{,}661$) are included in calculations of average transitional probability, including the inflections *was* and *were* to capture overall construction probability (see Section 2.3.1).

potential motivation for reduction (Bybee 2001, Bybee 2007; Traugott 2011). Informativity (Cohen Priva 2008, Cohen Priva 2012; Piantadosi et al. 2011) is a kind of context probability, important because words (or segments) may be infrequent, but always predictable through context, or frequent but sometimes surprising given the context. Informativity is calculated as $-\sum$ [P(context|word) * log(P(word|context))]. For Construction Informativity, I calculate the informativity of a word in a particular construction. Context is *n* of the following or preceding word for a particular construction. This is summed across all inflections of a word that occur in a particular construction: $-\sum$ [P(context|construction) * log(P(construction|context))]. For example, following construction informativity for the progressive construction is calculated using the sum of probabilities of progressive auxiliaries *was, were, am, are* and *is* given their following context. In this manner, informativity is calculated for the constructions generally, not just the specific contexts. Informativity given the following word (Seyfarth 2014) varies by construction as the construction type limits the word following the target. Preceding informativity (cf. Piantadosi et al. 2011) is similar across construction types. These measures cannot be compared directly in one regression model due to multicollinearity. Therefore, after showing that construction is a significant predictor, models with either preceding informativity, following informativity or log construction frequency are directly compared.

2.3.2 Transitional and joint probability

Tokens were coded for preceding and following transitional context probability using word and bigram frequencies from the COCA (Davies 2008). Probability measures are logarithmically transformed. Joint probability is the probability that two words occur together. Transitional probability is the likelihood the target will occur after/before a particular word: -log((w-1, w)/w) or -log((w, w+1)/w). Preceding transitional probability is a forward transitional probability, and following transitional probability is a backward transitional probability. In the models, only preceding joint probability and following transitional probability were included, as these performed better than their counterparts. For the function words under investigation, the preceding element was usually a pronoun. For the possessive and copula constructions, the following element was often a determiner; and determiners were only possible in these construction contexts. For modal constructions, the following element was almost always *to* as in *have to X*. Because the construction limits the following elements, which are often function words as well, following transitional probability was calculated using the frequency of the following word, even if it was a function word, rather than looking to the next content word context, which would have been less restricted due to construction. Table 4 below shows bigram examples of *BE* with particularly high and low joint and transitional probabilities. In the models, these variables are scaled and centered.

2.3.3 Speaker speech rate

Average speech rate is calculated per speaker as the average time in milliseconds that it takes a speaker to produce one syllable. Each utterance that contained a target word is measured for its length and divided by the number of syllables in that utterance, which is then averaged for each speaker. This measure is for a speaker's overall speech rate, not just speed of the particular phrase the token is in, which is accounted for by normalizing the dependent variable (see Section 2.2.1). This variable is scaled and centered.

Table 4: Bigram probability examples from the Buckeye Corpus.

Probability measure	High probability bigram	Value	Low probability bigram	Value
Preceding joint	*it is*	2.60673	*Krushchev is*	8.65321
Following joint	*is a*	2.87876	*are crummy*	8.65321
Preceding transitional	*there is*	0.23407	*whether was*	4.82295
Following transitional	*was born*	0.06226	*am during*	4.98528

2.3.4 Word position in utterance

The position of a word within an utterance is calculated based on the starting position of the token given the overall length of an utterance in characters. Position in the utterance is controlled for as Bell et al. 2003 found lengthening effects due to position, particularly utterance initial and final positions. Tokens in utterance initial or final positions are excluded, but there may still be contributions by the position in an intonational phrase. This variable is scaled and centered.

2.3.5 Speaker

Speaker ($n = 40$) is included as a random intercept in the mixed effects regression models in the multimodel inferencing procedure.

2.3.6 Word

Word is included as a random intercept in the models below. Only monosyllabic inflections of words are included.

2.4 Multimodel inferencing

A mixed-effects linear regression model is built to examine the shortening of *HAVE* and a generalized linear (logistic) regression model is built to examine the contraction of *BE* using *R* (R Core Team 2017) and packages {lme4} (Bates et al. 2015), {lmerTest} (Kuznetsova et al. 2016) and {MuMIn} (Bartoń 2018). The kind of regression analysis done here is known as multimodel inferencing – an alternative to selecting a single “best” model, which is often only marginally more predictive than the “next best” model, particularly for language data where there is high model selection uncertainty due to the natural redundancy of language (Burnham and Anderson 2002; Kuperman and Bresnan 2012; Barth and Kapatsinski 2017). Instead, multimodel inferencing makes predictions based on a set of plausible models. The plausible models are decided by building all possible models out of a given set of predictors, up to a specified maximal model, and adding a null model with no predictors, and the set is then ranked by their corrected Akaike Information Criterion or *AICc* (Akaike 1973). Only the models with possible predictive value are reported (those that are within Δ 2 in *AICc* of the best performing model) as only these have substantial empirical support based on the data (Burnham and Anderson 2002), along with the null model, as a way of comparing how much support the best performing models have. All models reported below substantially outperform the corresponding null model. Cumulative probability (CP) of a predictor is the summed AICc scores of models that include that predictor. Predictors that perform well in many models, and especially many highly predictive models, have a high CP score, indicating that they are highly probable of being predictive. Akaike weight (w) is the probability that the model is the most predictive one. Models that differ in their predictors, but have similar weights, are likely to be just as predictive, demonstrating model uncertainty and why a single model is not the best way to understand the predictors’ effect on variation.

The models include random intercepts. It is not possible to include predictor per speaker slopes in this multimodel inferencing because models are built with the possible varying combinations of predictors, and some models in the set do not converge with predictor per speaker slopes. Single maximal models with all variables, and construction per speaker slopes shows that variance is quite low.[3]

The independent variables described in Section 2.3 are tested for their contribution to both word shortening (*HAVE*) and contraction (*BE*), including whether the construction type is predictive. This study aims

3 For *HAVE* duration, variance of modal construction per speaker is 0.003, $SD = 0.059$, variance of perfect construction per speaker is 0.002, SD 0.051, for *BE* contraction, variance of future construction per speaker is 0.065, $SD = 0.255$, variance of passive construction per speaker is 0.016, $SD = 0.125$, variance of progressive construction per speaker is 0.034, $SD = 0.183$.

to investigate the contribution of frequency versus informativity to reduction, but a direct test of these two predictors results in high multicollinearity (cf. Tables 11 and 12 in Appendix). Where construction has a high Cumulative Probability, I make a comparison of models in which the Preceding Informativity, Following Informativity and construction Frequency (the *construction* properties – see Table 3) are directly tested against one another (Table 7 for *HAVE*, Table 10 for *BE*).

3 Results

3.1 *HAVE* word shortening

Table 5 shows, for *HAVE* in the Buckeye Corpus, the contribution of each predictor to word shortening. Following Transitional Probability is significant, as less probable following contexts are associated with longer targets ($\overline{\beta} = 0.051$; $z = 5.964$; $p < 0.001$). Construction type also matters, as perfect constructions are also associated with shorter targets than possessive constructions ($\overline{\beta} = -0.32$; $z = 13.11$; $p < 0.001$). A later position of the target in the utterance is associated with longer durations ($\overline{\beta} = 0.034$; $z = 5.113$; $p < 0.001$). This may be due to utterance-final phonological effects, although targets that occurred at the end of utterances or before pauses were not included. All three of these factors had a CP (cumulative probability) of one, meaning that they were very likely to contribute positively to model fit. Preceding Joint Probability ($\overline{\beta} = 0.011$; $z = 1.511$; $p = 0.13$) and speech rate ($\overline{\beta} = 0.011$; $z = 0.944$; $p = 0.345$) had no probable effect on normalized duration. These factors had much lower CP, meaning they less often contributed to a good model fit.

Table 6 shows that there are four models that predict the duration variation of *HAVE* reasonably well. The best four models out of the full set of possible models have the full Akaike weight of all models combined. These models include Construction, Following Transitional Probability (Following TP) and Position in the utterance, meaning that these factors are beneficial to include in the model. Other models that include Preceding Joint Probability or Speaker Speech Rate do not improve the fit, and so a more parsimonious single model should leave them out.

Table 5: Multimodel inferencing results for *HAVE* duration from the Buckeye corpus.

Predictor	$\overline{\beta}$	$\overline{\sigma}$	LoCI	HiCI	z	p	CP
(Intercept)	**1.102**	**0.064**	**0.977**	**1.227**	**17.282**	**0.000**	**NA**
Construction: modal	−0.021	0.019	−0.057	0.016	1.108	0.268	1
Construction: perfect	**−0.320**	**0.024**	**−0.368**	**−0.272**	**13.110**	**0.000**	
Following transitional probability	**0.051**	**0.008**	**0.034**	**0.067**	**5.964**	**0.000**	**1**
Word position in utterance	**0.034**	**0.007**	**0.021**	**0.048**	**5.113**	**0.000**	**1**
Speaker speech rate	0.011	0.012	−0.012	0.035	0.944	0.345	0.36
Preceding joint probability	0.011	0.007	−0.003	0.025	1.511	0.131	0.53

Note: Reported values are conditional averages with adjusted standard error, $n = 2694$. $\overline{\beta}$ and $\overline{\sigma}$ indicate average coefficient and error terms. Significant predictors are in bold.

Table 6: Models of *HAVE* duration with a Δ below 2.

Model factors	*k*	*df*	log likelihood	AICc	Δ	*w*
2345	4	9	−972.17	1962.42	0	0.34
235	3	8	−973.31	1962.68	0.26	0.30
12345	5	10	−971.74	1963.55	1.14	0.19
1235	4	9	−972.87	1963.80	1.39	0.17
(Null)	0	4	−1189.44	2386.90	424.48	0

Note: Cutoff: $\Delta < 2$, 1 = Speaker speech rate, 2 = Construction, 3 = Following TP, 4 = Preceding JP, 5 = Word position in utterance.

At this point we have established that Construction *is* predictive of *HAVE* duration, our critical parameter for *HAVE* reduction. But which construction property is (cf. Table 3) most predictive? Models were built that contained all the predictors already present, but with either Construction Frequency (logged, scaled and centered), Following Informativity (scaled and centered) or Preceding Informativity (scaled and centered) replacing construction. Each model has random intercepts for word and speaker, and random slopes for the construction variable of interest, by speaker. The outcomes of the models are ranked in Table 7 by their AIC scores. Following Informativity results in the lowest AIC score, meaning it is more predictive of *HAVE* duration than frequency.

3.2 *BE* contraction

Table 8 presents the contribution of each predictor from the multimodel output for contraction of *BE*. Positive values of coefficients are associated with more contraction; negative values are associated with less contraction. Several factors have a CP of one, meaning they are extremely likely to influence contraction. Construction is predictive of contraction, with progressive ($\overline{\beta} = 0.56$; $z = 4.17$; $p < 0.001$) and future ($\overline{\beta} = 0.87$; $z = 3.65$; $p < 0.001$) constructions associated with more contraction than copula constructions, but not passives ($\overline{\beta} = 0.03$; $z = 0.10$; $p = 0.92$). High following transitional probability ($\overline{\beta} = -0.25$; $z = 5.63$; $p < 0.001$) and high Preceding Joint Probability ($\overline{\beta} = -1.66$; $z = 39.65$; $p < 0.001$) are associated with lower levels of contraction, meaning the less probable the particular context, the less likely contraction is to occur. Preceding Joint Probability has a strong effect because contraction is strongly associated with pronouns that are in themselves frequent. The highest rates of contraction are likely to occur earlier in the utterance ($\overline{\beta} = -0.49$; $z = 12.53$; $p < 0.00$), because often the auxiliary contracts with the subject, which is often the first word in an utterance. Faster speech rates are marginally associated ($\overline{\beta} = -0.13$; $z = 0.97$; $p = 0.07$) with lower levels of contraction, but this does not reach significance.

Model rankings for *BE* contraction are presented in Table 9. There are two reasonable models given this set of predictors: either with all predictors (*Akaike weight* = 0.63), or excluding speech rate (*Akaike weight* = 0.37). Both of these models have a much better fit than the null model, which was the least good model out of all possible models. Unlike *HAVE* duration, almost of the predictors have a strong effect on *BE* contraction and there are fewer acceptable models for predicting contraction.

Again, we see that construction as a factor is predictive in the model for *BE* contraction. And, as for *HAVE* duration (3.1), models were built that contained all the predictors present in the models above, but

Table 7: Model Comparison with varied construction predictors, *HAVE*.

Predictor varied	AIC	BIC	Rank
Following informativity	1990.1	2055	1
Preceding informativity	2027.8	2092.7	2
Log construction frequency	2114.5	2179.3	3

Table 8: Multimodel inferencing results of *BE* contraction from the Buckeye corpus.

Predictor	$\overline{\beta}$	$\overline{\sigma}$	LoCI	HiCI	z	p	CP
(Intercept)	**2.398**	**0.346**	**1.721**	**3.076**	6.939	**0.000**	**NA**
Construction: future	**0.875**	**0.240**	**0.406**	**1.345**	3.653	**0.000**	**1**
Construction: passive	0.032	0.308	−0.573	0.636	0.102	0.918	
Construction: progressive	**0.560**	**0.134**	**0.297**	**0.823**	4.167	**0.000**	
Following transitional probability	**−0.248**	**0.044**	**−0.334**	**−0.161**	5.631	**0.000**	**1**
Preceding joint probability	**−1.659**	**0.042**	**−1.741**	**−1.577**	39.651	**0.000**	**1**
Word position in utterance	**−0.492**	**0.039**	**−0.569**	**−0.415**	12.533	**0.000**	**1**
Speaker Speech Rate	−0.131	0.073	−0.274	0.011	0.968	0.071	0.63

Note: Reported values are coefficients with shrinkage and adjusted standard error, $n = 8828$. $\overline{\beta}$ and $\overline{\sigma}$ indicate average coefficient and error terms. Significant predictors are in bold.

Table 9: Models of *BE* contraction with a Δ below 2.

Model factors	*k*	*df*	log likelihood	AICc	Δ	*w*
12345	5	10	−2120.37	4260.76	0	0.63
2345	4	9	−2121.91	4261.84	1.08	0.37
(Null)	0	3	−3774.77	7555.55	3294.79	0

Note: Cutoff: $\Delta < 2$, 1 = Speaker Speech Rate, 2 = Construction, 3 = Following TP, 4 = Preceding JP, 5 = Word Position in Utterance.

Table 10: Model Comparison with varied construction predictors, *BE*.

Predictor varied	AIC	BIC	Rank
Following informativity	4262.8	4333.6	1
Log construction frequency	4269	4339.9	2
Preceding informativity	4285.9	4356.8	3

either Construction Frequency (logged, scaled and centered), Following Informativity (scaled and centered) or Preceding Informativity (scaled and centered) used in place of Construction to rank the contribution of the construction properties (cf. Table 3). Each model has random intercepts for word and speaker, and random slopes for the construction variable of interest by speaker. The outcomes of the models are ranked in Table 10 by their AIC scores. Following Informativity results in the lowest AIC score, meaning that, as with *HAVE* duration, it predicts reduction better than frequency does.

4 Discussion

4.1 Probabilistic reduction of function words

Studies of probabilistic reduction show that speaker behavior is affected by the frequency and probability of words. Studies such as Gahl (2008) show that the more frequent word of a homophone pair (*time* vs. *thyme*) is the shorter one. However, Gahl (2008) examines only content word homophone pairs with different spellings, expressly excluding function words. Seyfarth (2014) shows that low average informativity of words (not controlling for homophones) results in shorter durations, controlling for segment count, syllable count and frequency among other factors. He finds that word informativity and local predictability have similar effect sizes. However, he too excludes function words from his analyses. Previous studies examining function words, such as Bell et al. (2003), find that once factors such as speaking rate, segmental context, pitch accent and contextual predictability are accounted for, there is no frequency effect for different meanings of words such as *to, that, of, you, I, and, the, a* and *it*. Previous studies of contraction of *BE* and *HAVE* (Krug 1998; Frank and Jaeger 2008; Bresnan and Spencer 2016) examine multiple factors including probability, but not meaning differences or meaning frequency.

The present study shows that, contra Bell et al. 2003, there are production differences even for a given function word, due to different meanings. That is, the predictability of the token in its particular context is very important, suggesting that there is an online component to reduction (Gregory et al. 1999; Jurafsky et al. 2001). However, average probability of a word's meaning across contexts is also important, suggesting that phonetic detail is stored with forms linked to meanings (Torres Cacoullos 1999; Bybee 2002; Pierrehumbert 2002; Raymond and Brown 2012). In order for average statistical properties associated with a word's meaning to affect its production – as this study has shown– there must be storage for (at least certain) function words.

Future studies should focus on other function words to determine to what extent the findings reported here are restricted to these words.[4] Corpus studies that examine polysemous meanings of words, or words

4 Though there is no reason *a priori* that the findings will not be generalizable.

that are homonymous, are time-consuming to perform as they require at least some hand-coding. However the detailed analyses presented here has paid dividends in proving, contra Levelt et al. 1999, that probabilistic reduction and average probability effects are in operation for function words, adding a further similarity in content and function word production.

4.2 Grammaticalization

Bybee (2002), Bybee (2007) posits that it is the jump in frequency of new grammatical constructions that leads to reduction of grammaticalized elements. However, the construction meanings for *BE* and *HAVE* that have highest frequency in the Buckeye Corpus are actually the oldest constructions (copula and possessive respectively: see Table 3). The model comparison in Section 3 shows that construction informativity performs better than construction frequency in predicting reduction. This is contra the frequentist account, but captures the meaning narrowing that comes about through grammaticalization. Lower informativity is an inevitable consequence of grammaticalization and better predictor for phonetic erosion than frequency increase, which may not actually be a consequence of all cases of grammaticalization. For auxiliaries in particular, informativity given the following word matters. Take for example the three *HAVE* constructions: possessive, modal and perfect. When *HAVE* is in the possessive construction, many words may follow it, including determiners, adverbs, etc. and many nouns (a large open class). This means that *HAVE* as a possessive is highly unlikely to be predictable given any of those words. When *HAVE* is used as a semi-modal auxiliary, it must occur before *to*. Although that is a very limited context, *to* is so frequent that the transitional probability is still not as high as when *HAVE* is a perfect auxiliary. When *HAVE* is used as a perfect auxiliary, it occurs before past participles, and sometimes adverbs or negator words (*he has not been to the park recently*). In the cases of adverbs or words like *not*, *HAVE* is not predictable. However, in the case of the past participles, *HAVE* is highly predictable because one of the few words that can occur, and occurs with any regularity, before past participles is *HAVE* and these past participles are relatively infrequent, meaning that, together, the auxiliary and following word have a high transitional probability. *BE* may occur before some past participles (i.e. in passives) but the past participles that occur in passive constructions only minimally overlap with those that occur in perfect constructions. Past participles are sometimes used as adjectives, but these kinds of past participles also only partially overlap with perfect construction participles. Perfect *HAVE* is highly likely to occur before the kinds of past participles that occur in perfect constructions, resulting in high average following transitional probability.

The results of the present study show that an increase in frequency cannot be the reason for the higher levels of reduction. However, the increase in predictability due to the narrower contexts that come about through grammaticalization (Heine 1993; Hopper and Traugott 1993; Bybee 2002; Diewald 2002; Traugott 2003) is a good reason for the higher level of reduction. Other aspects associated with the construction, such as context of use (Torres Cacoullos 1999; Raymond and Brown 2012) and usage frequency (Bybee 2002, Bybee 2007; Alba 2008; Torres Cacoullos and Walker 2011) also matter. However, the measure of average following transitional probability quantitatively captures the intuition that grammaticalization researchers have had for over a century that more grammatical (and therefore less informative and more predictable, a.k.a. 'bleached') information is more subject to reduction (phonetic erosion) than source lexical items. This account also receives support from being congruent with psycholinguistic research showing the same tendencies are found throughout language production as speakers spend less time and effort on producing items that are predictable.

5 Summary

The study of *HAVE* and *BE* presented here shows that:

a) construction probability as well as context specific probability is predictive of reduction for *BE* and *HAVE;* and
b) context probability (already known to affect content words) also affects function words.

In other words, English speakers make use of context- and construction-specific reduction strategies in addition to more general word reduction strategies. This result was obtained by focusing closely on the behavior in context of two function words, rather than attempting to generalize over all words. Further work on other corpora and other function words would prove equally fruitful.

Acknowledgements: Thanks go to the attendees of the 2016 Predictability Symposium at Western Sydney University and to Seamus Donnelly, Susan Ford and three anonymous reviewers for their comments on drafts of this paper.

References

Akaike, Hirotogu. 1973. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Cáski (eds.), *Second International Symposium on Information Theory* 267–281. Akademiai Kaidó: Budapest.
Alba, Matthew, C. 2008. Ratio frequency: Insights into usage effects on phonological structure from hiatus resolution in New Mexican Spanish. *Studies in Hispanic and Lusophone Linguistics* 1(2). 247–286.
Arnon, Inbal & Uriel Cohen Priva. 2013. Time and again: The changing effect of word and multiword frequency and constituency on phonetic duration. *Language and Speech* 56(3). 349–371.
Aylett, Matthew & Alice Turk. 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America* 119. 3048–3058.
Barth, Danielle. 2015. *To BE and to HAVE: Function word production in child speech, child directed speech and inter-adult speech*. Eugene, OR: University of Oregon dissertation.
Barth, Danielle & Vsevolod Kapatsinski. 2017. A multimodel inference approach to categorical variant choice: Construction, priming and frequency effects on the choice between full and contracted forms of *am*, *are* and *is*. *Corpus Linguistics and Linguistic Theory* 13(2). 1–58.
Bartoń, Kamil. 2018. MuMIN: Multi-Model inference. R package version 1.40.4. CRAN.r-project.org/package=MuMIn.
Bates, Douglas, Martin Maechler, Ben Bolker, & Steven Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.
Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle L. Gregory & Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation.*The Journal of the Acoustical Society of America* 113(2). 1001–1024.
Bell, Alan, Jason M. Brenier, Michelle L. Gregory, Cynthia Girand & Daniel Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1). 92–111.
Bresnan, Joan & Jessica Spencer. 2016. Frequency effects in spoken syntax: 'Have' and 'be' contraction. In Stefan Grondelaers & Roeland van Hout (eds.), *New ways of analysing syntactic variation*, Berlin: De Gruyter Mouton.
Burnham, Kenneth P. & David R. Anderson. 2002. *Model selection and multimodel inference: A practical Information-Theoretic approach*, 2nd edition. New York: Springer.
Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.
Bybee, Joan. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14. 261–290.
Bybee, Joan. 2007. *Frequency of use and the organization of language*. Oxford & New York: Oxford University Press.
Bybee, Joan, & William Pagliuca. 1985. Cross-linguistic comparison and the development of grammatical meaning. In Jacek Fisiak (ed.), *Historical semantics, historical word formation*, 59–83. Berlin: Mouton de Gruyter.
Bybee, Joan, Revere Dale Perkins & William Pagliuca. 1994. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. Chicago: University of Chicago Press.
Byrd, Dani. 1994. Relations of sex and dialect to reduction. *Speech Communication* 15(1). 39–54.
Cohen Priva, Uriel. 2008. Using information content to predict phone deletion. In *Proceedings of the 27th West Coast Conference on Formal Linguistics*, 90–98.
Cohen Priva, Uriel. 2012. *Sign and signal: Deriving linguistic generalizations from information utility*. Palo Alto, CA: Stanford University dissertation.
Cohen Priva, Uriel & T. Floriean Jaeger. 2018. The interdependence of frequency, predictability, and informativity in the segmental domain. *Linguistics Vanguard* 4(s2). doi:10.1515/lingvan-2017-0028 (accessed 24 Sep 2018).
Davies, Mark. 2008. The Corpus of Contemporary American English (COCA): 400+ million words, 1990-present. Available online at http://www.americancorpus.org.
Diewald, Gabriele. 2002. A model for relevant types of contexts in grammaticalization. *Typological Studies in Language* 49. 103–120.

Fedorenko, Evelina, Edward Gibson & Douglas Rhode. 2006. The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language* 55. 541–553.

Ferreira, Victor S. 2008. Ambiguity, accessibility, and a division of labor for communicative success. In Brian H. Ross (ed.), *The psychology of learning and motivation* 210–246. Amsterdam: Elsevier.

Fowler, Carol A. & Jonathan Housum. 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.

Frank, Austin F. & T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In Brad C. Love, Ken McRae & Vladimir M. Sloutsky (eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, 939–944. Austin, TX: Cognitive Science Society.

Gabelentz, Georg von der. 1891. *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig: T. O. Weigel.

Galati, Alexia & Susan E. Brennan. 2010. Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language* 62(1). 35–51.

Gahl, Susanne. 2008. "Time" and "thyme" are not homophones: Word durations in spontaneous speech. *Language* 84. 474–496.

Gahl, Susanne & Susan Marie Garnsey. 2004. Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language* 80(4). 748–775.

Gahl, Susanne, Yao Yao & Keith Johnson. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66. 789–806.

Givón, Talmy. 1985. Iconicity, isomorphism and non-arbitrary coding in syntax. In John Haiman (ed.), *Iconicity in syntax*, 187–219. Amsterdam: John Benjamins.

Gregory, Michelle L., William D. Raymond, Alan Bell, Eric Fosler-Lussier & Daniel Jurafsky. 1999. The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society* 35. 151–166.

Hall, Kathleen Currie, Elizabeth Hume, T. Florian Jaeger & Andrew Wedel. 2018. The role of predictability in shaping phonological patterns. *Linguistics Vanguard* 4(s2). doi:10.1515/lingvan-2017-0027 (accessed 24 Sep 2018).

Heine, Bernd. 1993. *Auxiliaries*. Oxford: Oxford University Press.

Heine, Bernd. 2003. Grammaticalization. In Brian D. Joseph & Richard D. Janda (eds.), *The Handbook of Historical Linguistics*, 575–601. Oxford: Blackwell.

Hopper, Paul J., & Elizabeth Closs Traugott. 1993. *Grammaticalization*. Cambridge: Cambridge University Press.

Jaeger, T. Florian & Esteban Buz. 2016. Signal reduction and linguistic encoding. *Handbook of psycholinguistics*. Wiley-Blackwell.

Jaeger, T. Florian & Harry Tily. 2010. On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(3). 323–335.

Jurafsky, Daniel, Alan Bell, Michelle L. Gregory & William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan Bybee, and Paul J. Hopper. (eds.), *Frequency and the emergence of linguistic structure*, 229–254. Amsterdam: Benjamins.

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

Kuperman, Victor & Joan Bresnan. 2012. The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language* 66. 588–611.

Kuznetsova, Alexandra, Per Bruun Brockhoff & Rune Hauo Bojesen Christensen. 2016. lmerTest: Tests in linear mixed effects models. R package version 2.0-33. CRAN.r-project.org/package=lmerTest.

Lehmann, Christian. 1995. *Thoughts on grammaticalization*. München: Lincom Europa.

Levelt, Willem J. M., Ardi Roelofs & Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22. 1–75.

Lindblom, Björn. 1990. Explaining phonetic variation: A sketch of the H&H theory. In William J. Hardcastle & Alain Marchal (eds.), *Speech production and speech modelling*, 403–439. Dordrecht: Kluwer.

Lindblom, Björn, Susan Guion, Susan Hura, Seung-Jae Moon, & Raquel Willerman. 1995. Is sound change adaptive? *Rivista di linguistica* 7. 5–36.

Piantadosi, Steven T., Harry Tily, & Edward Gibson. 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, Janet. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In Joan Bybee & Paul J. Hopper. (eds.), *Frequency and the emergence of linguistic structure*, 137–157. Amsterdam: Benjamins.

Pierrehumbert, Janet. 2002. Word-specific phonetics. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory phonology VII (phonology and phonetics)*, 101–140. Berlin: Mouton de Gruyter.

Pitt, Mark A., Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume & Eric Fosler-Lussier. 2007. Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).
Pluymaekers, Mark, Mirjam Ernestus & Harald R. Baayen. 2005. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62(2–4). 146–159.
R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org.
Raymond, William D. & Esther L. Brown. 2012. Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In Stefan Th. Gries & Dagmar Divjak (eds.), *Frequency effects in language learning and processing*, 35–52. Berlin: Mouton De Gruyter.
Seyfarth, Scott. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133. 140–155.
Torres Cacoullos, Rena. 1999. Construction frequency and reductive change: Diachronic and register variation in Spanish clitic climbing. *Language Variation and Change* 11(2). 143–170.
Torres Cacoullos, Rena & James A. Walker. 2011. Collocations in grammaticalization and variation. In Bernd Heine & Heiko Narrog (eds.) *The Oxford handbook of grammaticalization*, 225–238. Oxford: Oxford University Press.
Traugott, Elizabeth Closs. 2003. Constructions in grammaticalization. In Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 624–647. Oxford: Blackwell.
Traugott, Elizabeth Closs. 2011. Grammaticalization and mechanisms of change. In Bernd Heine & Heiko Narrog, (eds.), *The Oxford handbook of grammaticalization*, 19–31. Oxford: Oxford University Press.
Ullman, Michael T. 2001. A neurocognitive perspective on language: The declarative/procedural model. *Neuroscience* 2. 717–727.
van Bergem, Dick R. 1995. *Acoustic and lexical vowel reduction*. Amsterdam: IFOTT.
van Son, Rob. J. J. H. and Louis C. W. Pols. 2003. Information structure and efficiency in speech production. Paper presented at the 8th European Conference on Speech Communication and Technology (Eurospeech 2003) Geneva, Switzerland, 1–4 September.
Wedel, Andrew B. 2006. Exemplar models, evolution and language change. *The Linguistic Review* 23. 247–274.
Wedel, Andrew. 2012. Lexical contrast maintenance and the development of sublexical contrast systems. *Language and Cognition* 4. 319–355.

Appendix

This appendix contains correlation tables for the numeric predictors in each model.

Table 11: Correlations of numeric predictors in *HAVE* duration model.

	Preceding JP	Following TP	Speaker speech rate	Word position	Following informativity	Preceding informativity	Construction log frequency
Preceding JP	1						
Following TP	−0.024	1					
Speaker speech Rate	0.040	−0.075	1				
Word position	0.140	−0.003	0.019	1			
Following informativity	−0.159	0.572	−0.056	−0.010	1		
Preceding informativity	−0.140	0.559	−0.049	−0.006	0.980	1	
Construction log frequency	0.084	−0.022	0.03	0.021	−0.026	0.175	1

Table 12: Correlations of numeric predictors in *BE* contraction model.

	Preceding JP	Following TP	Speaker speech rate	Word position	Following informativity	Preceding informativity	Construction log frequency
Preceding JP	1						
Following TP	0.103	1					
Speaker speech rate	0.015	0.009	1				
Word position	0.199	0.073	−0.015	1			
Following informativity	−0.023	0.268	0.000	0.011	1		
Preceding informativity	0.033	−0.066	−0.003	0.018	−0.35	1	
Construction log frequency	−0.029	0.248	0.002	0.005	0.957	−0.575	1