Sound change and hierarchical inference.

What is being inferred? Effects of words, phones and frequency.

Vsevolod Kapatsinski

University of Oregon

ABSTRACT: Articulatorily-motivated sound change is modeled as the outcome of two processes: 1) automatization of word production, which occurs in every instance of word use, and thus affects high-frequency words most, and 2) hierarchical inference during learning, in which words and phones (defining word groups) are assigned blame/credit for experienced pronunciations. I show that, unless we assume that learners also infer the effect of frequency on pronunciation, novel words should behave like run-of-the-mill known words, ones that have not become associated with either reduced or full pronunciations. The model predicts that, once the change has progressed far enough for the phone to become associated with the reduced pronunciation, novel words will be less reduced than existing words that, for any reason, have become associated with the unreduced variant. This prediction is tested using data from flapping in American English, a reductive sound change that has gone far enough for the reduced variant to be the default. Results are inconsistent with the model: novel words disfavor reduced pronunciations even more than real words that tend to occur in formal contexts and thus may be expected to be most associated with full pronunciations. Possible modifications to the model are discussed.

1. Introduction

1.1. The research question

Usage-based approaches to sound change suggest that the seeds of change, though carried along on the wind of social transmission, are planted by the biases inherent in the processes of perception, production, and learning performed by human beings (e.g. Bybee 2001: 196). In the present paper, we focus on a production-motivated sound change, hence it is production biases that are of particular importance. Speech production is continuously becoming more automated throughout one's lifetime (Bybee 2001: 11-12, Kapatsinski 2010a). This automatization is brought about in part by simplification of frequently performed articulatory routines (tempered perhaps by the need to be understood). Simplification manifests itself in the smoothing of transitions between adjacent and near-adjacent articulatory gestures, reduction in gesture magnitude and duration. For example, a full voiceless alveolar stop ([t]) changing into a voiced flap ([ɾ]) between vowels involves a reduction in closure duration, often accompanied by a reduction in closure magnitude as well as the absence of burst (see Figure 1). The resulting articulation provides a much smoother transition between vowels (de Jong 1998, Fujimura 1986), at the limit becoming difficult to notice on the spectrogram.

Usage-based linguists tend to follow Langacker's (1987) rejection of the Rule-List Fallacy. Langacker pointed out that a unit that has been stored in memory can be retrieved directly from memory on some occasions but can also be assembled out of its parts on others. More recently, this idea gave rise to the idea that there are many routes from form to meaning and from meaning to form, and that these routes operate in parallel, racing to produce the same output (Baayen et al. 1997, Beekhuizen et al. 2013, Kapatsinski 2010b). Given this perspective, the articulatory routines that are being automated are thought to come in many sizes. Plausibly, both words and individual phones have motor routines associated with them (e.g. Pierrehumbert 2001, 2002). Under the multiple-route perspective, the same word can then be produced by retrieving a word-level articulatory routine as well as by assembling a speech plan out of individual phones (or something in between).

Two mental representations are associated when activation of one leads to activation of the other. The formation of associations is thought to be an important part of learning. If the resulting associations are to reflect the structure of the input, i.e. if learning is to be successful, the processes of association formation should be functionally equivalent to statistical inference algorithms. The question we would like to ask in this paper is "what is being inferred?" given the data from sound change.

As suggested above, we assume that *both* words and phones can be assigned credit for how an instance of a phone in a certain word sounds when it is perceived, and both influence production (Pierrehumbert 2002). On this view, the phonological grammar and the lexicon exert parallel influences on pronunciation (without necessarily being separate cognitive modules). We are then faced with the task of determining what these influences are, or, in other words, assigning credit for how a given instance of a phone is pronounced to the identity of the phone and the identity of the word it is in (leaving aside for now the various contextual influences, see fn.1). Note that each observed instance of a phone must occur in some word, thus words can be thought of as forming non-overlapping groupings of the phone's instances, into which the full set of the phone's instances is completely divided. Because of this, the assignment of credit to words and phones can be modeled using *hierarchical* statistical inference.

An important unresolved question is whether the learner takes the words s/he has encountered as being representative of the population of words. Does s/he expect words s/he encounters in the future to be like the words she has already encountered? Or does s/he think that newly encountered words might differ systematically from words s/he already knows? In particular, if frequent words differ from rare words, as usage-based phonology suggests, does the learner catch onto this fact, expecting newly encountered (and therefore presumably rare) words to not be like the frequent words s/he already knows? While this hypothesis has not, to my knowledge, been proposed in the literature for pronunciation patterns, it is compatible with the notion of grammar as a default that is resorted to when lexical retrieval fails (e.g. Albright & Hayes 2003, Clahsen 1999, Pinker 1991, Kapatsinski 2010b). If the grammar is there to deal with novel inputs, then it would make sense for the learner to base their knowledge of how to deal with novel inputs on experience with rare/novel inputs. Alternatively, the learners may simply learn how known words and phones are pronounced without inferring anything about the relationship between word frequency and pronunciation? I take this to be the standard assumption in usage-based modeling of phonology, where the effect of frequency is due to automatic production pressure (e.g. Bybee 2001: 12). In the next section, I show that, as long as the existence of words and the existence of multiple pronunciations categorized into phonological categories (what we call *phones*) is assumed, the latter hypothesis predicts a U-shaped relationship between frequency and reduction in the later stages of a reductive sound change. We will then proceed to test this prediction using data from flapping in American English.

1.2. Frequency, reduction and inference

The most basic version of the model thus consists of the following parts: 1) there are two variants of a phone, reduced and unreduced; 2) every time a word is used, the likelihood of the reduced variant of the phone being used in that word is incremented; as a result, reduction advances further in frequent words than in rare ones; and 3) when a learner is exposed to the language, s/he learns the probabilities of producing each version of the phone as well as the way those probabilities are affected by lexical context. In other words, the child learns how often a certain phone is pronounced a certain way and that some words are pronounced exceptionally. This kind of word-specific phonetic learning appears to be necessary because lexical frequency does not account for all between-word variability in phone pronunciation; a residue of exceptionality remains after frequency is accounted for (Pierrehumbert 2002).

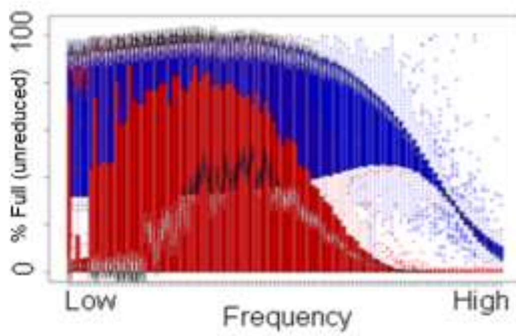
I assumed that the inference process if functionally equivalent to multilevel regression (as implemented in the lme4 package in R, Bates et al. 2013), where the learner was assumed, for any given phone, to estimate word-specific random intercepts as well as an overall intercept from which these deflect.[1] The word-specific intercepts are random because we often do not have enough data for an individual word to estimate the effect of that word on pronunciation: in any corpus, most words occur only once (Baayen

2001, Zipf 1935). By treating ‘word’ as a random effect, we avoid the need to estimate the effects of these individual rare words, by assuming that they will behave approximately like the average word, i.e. their reduction probabilities are drawn towards the mean reduction probability across all words. The extent to which a word is drawn to the mean is inversely proportional to its frequency. The less frequent a word, the less information we have about the effect of that word on pronunciation of the phone. Thus to know how an infrequent word behaves, we have to look at how words in general behave and estimate that this new word will behave like other words we know. Crucially, the learner is not assumed to estimate the effect of frequency, i.e. to learn the relationship between frequency and reduction. Rather, reduction is assumed to be due exclusively to production biases inherent to the human speech production apparatus (Bybee 2001). The simulation consisted of several generations of learners, where each generation learned based on the productions of the previous generation, and each generation incremented the learned reduction probabilities in proportion to word frequency.

Figure 1 shows the results of the simulation. The first generation, affected only by the reductive pressures of articulation, reduces frequent words more than the rarest words. However, as the sound change proceeds through the lexicon, the reduced variant of the phone becomes the default, used all the time with unknown words. Deviations from this default can be learned for individual words but the words need to be frequent enough for this learning to be successful. In other words, once the reduced variant overcomes the unreduced variant of the phone, individual words might become associated with the non-reduced variant but they have to be at least moderately frequent for the association to be formed. Thus, the effect of word frequency is predicted to be non-monotonic in the later stages of a reductive sound change. The most frequent words will be reduced because of two reasons: 1) the articulatory pressure towards reduction for the current generation, as well as 2) because they were reduced in the input to the current generation of learners and thus will be associated with the reduced variant of the phone by the current generation. The least frequent words will be reduced because they are not associated with any variant of the phone, and the reduced variant is more frequent. At intermediate frequency levels, some words, which happen to have been often used with the unreduced variant of the phone by previous generations, can become associated with the unreduced phone variant. This prediction of a U-shaped frequency effect in the later stages of an articulatorily-driven sound change is inevitable as long as 1) pronunciation variants are associated with probabilities, and 2) the relationship between word frequency and variant choice is due to inherent properties of the production system, rather than learning.

Figure 1. A boxplot showing the predictions from the simplest model for two generations of learners. Generation 2 in blue and Generation 10 in red. The figure shows reduction probability (y axis) as a

function of word frequency. One thousand re-runs of the model with different random samples of word frequencies (constant across generations) were used. The empty areas in the middles of boxes show 95% confidence intervals for the median reduction probability for each level of word frequency.



What might the exceptionally unreduced words be? Bybee (2002) and Raymond & Brown (2012) argue that words can become associated with reduction or non-reduction if they are often used in reduction-favoring or disfavoring contexts (see also Barth & Kapatsinski in press, Bybee 2003, Erker & Guy 2012 for other contextual features becoming associated with individual words). In the present approach, even if the choice of pronunciation variants is completely driven by non-lexical contextual factors to the first generation, it will become partially driven by lexical identity by the second generation. As long as the learner sometimes misattributes variant choice that is actually caused by a contextual influence to an effect of the individual word occurring in that context, a word that usually occurs in contexts favoring a pronunciation variant will acquire an association with that pronunciation variant. In particular, words used in formal contexts, which disfavor reduction, are more likely to become associated with the unreduced variant of the phone undergoing sound change (Raymond & Brown 2012).

While the prediction illustrated in Figure 1 is new to this paper, the model that generates it can be seen to implement well-accepted ideas in usage-based phonology. Thus, Bybee (2001: 8-9) argues that high token frequency leads to reduction as well as emancipation (freedom from analogical influences of the lexicon as a whole). The reason Bybee (2001) does not predict a U-shaped frequency effect for the later stages of an articulatorily-motivated ('phonetic') change can be found in the following quote:

> "The second effect of frequency [i.e., emancipation] seems to contradict the first, since it makes items more resistant to change, but it concerns change of a different kind. High frequency encourages phonetic change, but it renders items more conservative in the face of grammatical change or analogical change based on the analysis of other forms (Phillips 2001)." (Bybee 2001: 12)

Thus Bybee (2001), following Phillips (2001), thus assumes that analogical pressure affects only some changes. The theoretical motivation for this assumption is not entirely clear but may, perhaps, be traced to the questionable status of sublexical units and phonological categories in Bybee's model (e.g. Chapter 3 in Bybee 2001). If phonological categories are highly specific, then there are no alternative ways of pronouncing a phonological unit and therefore no analogical pressure to pronounce it in a certain way.

In contrast, the present paper assumes that analogical pressure is always 'on' and that it also affects 'phonetic' changes. Namely, I assume that a particular way of pronouncing a sublexical unit can spread from word to word, as suggested by Pierrehumbert (2002). This assumption is supported by the empirical results on new dialect acquisition in German et al. (2013), where speakers of American English were shown to rapidly learn new pronunciations for particular phones, e.g. a flapped pronunciation for an [ɹ], with no evidence that the learning was restricted to individual words experienced during training (see also Maye et al. 2008, McQueen et al. 2006, Peperkamp & Dupoux 2007). With this assumption, a U-shaped frequency effect is necessarily predicted for later stages of an articulatorily-motivated ('phonetic') change.

1.3. A test case: Flapping in American English

Flapping in American English presents a good case of an articulatorily-motivated sound change that has advanced far enough to test the prediction of a U-shaped frequency effect. Word-medial flapping examined here reduces an alveolar stop (/t/ or /d/) to a (shorter) flap or even an approximant when the stop is preceded by a vowel or a liquid and followed by an unstressed vowel (e.g. Kahn 1976, Warner & Tucker 2011). As one would expect of an articulatorily-motivated process (Bybee 2001, 2002), Patterson & Connine's (2001) found that flapping is more common in high-frequency words than in low-frequency words (operationalized using a somewhat arbitrary 60 words/million cutoff). The frequency effect was not replicated by Herd et al. (2010) for real words: words with frequencies greater than 25/million did not differ from those with frequencies lower than 25/million; and words with frequencies above 100/million did not differ from those not found in the Francis-Kučera (1982) corpus. However, non-words were reported to exhibit flapping significantly less often than real words (85% vs. 90%, p.511) and real words not found in the corpus exhibited a flapping rate as low as that of non-words (84%). Thus the data are consistent with low-frequency words and non-words (which are the lowest-frequency words possible) having a lower rate of flapping than words for which the speakers are likely to have stored whole-word production plans. The lack of flapping rate differences between the tested groups of real words in Herd et al. (2010) may be due to the stimuli not being designed to maximize frequency differences (and hence statistical power for finding a frequency effect). However, the word frequency effect for word-medial flaps was also not found by Warner & Tucker (2011), who used continuous frequencies from the much larger Corpus of Contemporary American English (Davies 2008) as a predictor, although the stimuli were again not designed to look for a frequency effect.

Bauer (2005) and Warner & Tucker (2011) also examined word-final flapping and found no word frequency effects. However, it is not clear that word frequency effects on reduction are expected for final sounds if the source of the effect is greater automaticity of production plans for frequent words (Bybee 2002, Kapatsinski 2010a). As Dell et al. (1997) argue, increasing automaticity increases anticipation of the future. In Dell et al. (1997), this leads to anticipatory speech errors. More generally, it might also lead to shortening of segments currently being produced as the upcoming segments become activated earlier. For the word-final segment, the upcoming segments are in the next word, hence the measure of how likely they are to be anticipated must refer to the frequency of a unit that includes the next word. Accordingly, Warner & Tucker (2011) did find phrase frequency effects on final stop reduction, with frequent phrases favoring more reduced realizations.

Flapping has advanced far enough in American English for the flap to be the dominant variant of /t/ and /d/ in the flapping environment. Previous empirical studies of flapping (Byrd 1993, Herd et al. 2010, Patterson & Connine 2001) all report that flaps account for 75%-99% of productions (depending on whether the flap is underlyingly /t/ or /d/, experimental task, speaker gender and dialect). The statistical dominance of the flap variant would lead one to expect that /t/ and /d/ would also be flapped in novel words, for which the speaker does not have a whole-word production plan stored in memory. In accordance with this expectation, Herd et al. (2010) report an 85% rate of flapping in novel words. Thus, American English flapping appears to be a good example of a process that the model described above should account for.

2. Methods

2.1. Stimuli

The crucial stimuli consisted of 15 triplets of closely matched disyllabic or trisyllabic words containing /t/ or /d/ in the flapping environment. The words were of three Types: colloquial words, formal words and non-words. Formal words were the words suspected to be most likely to favor the (rare for American English) full stop pronunciations. Colloquial words were suspected to favor reduced pronunciations.

According to the predictions illustrated in Figure 1, non-words should fall between the two classes of real words.

Colloquial and formal words were selected using the web interface of the BYU version of the British National Corpus at http://corpus.byu.edu/bnc/ (Davies 2004-). Colloquial words were significantly over-represented in what I considered informal genres: conversation, and dialogs from plays, movies and reality shows. Formal words were significantly over-represented in formal genres: parliamentary discourse, academic discourse, broadcast news, courtroom discourse, and public debates. Given the lower incidence of flapping in British English (e.g. German et al. 2013, Shockey 1984), Britishisms were then excluded from the “colloquial” set the words that are designed to be most likely to favor flapping. Britishisms were identified by comparing frequencies in SubtlexUS (Brysbaert & New 2009, available at http://expsy.ugent.be/subtlexus/) and SubtlexUK (van Heuven et al. 2014, available at http://crr.ugent.be/archives/1423), which are comparable corpora based on movie subtitles, and excluding the words that had higher frequencies in SubtlexUK.

Formal words and colloquial words were selected to be the most frequent words with a significant difference in probability of occurrence across genres. Nonetheless, the colloquial words in the sample are more frequent than the formal ones in all corpora I have tested: the British National Corpus, SubtlexUS, SubtlexUK and even the Michigan Corpus of Academic Spoken English, which was designed to over-represent ‘formal’ words (Simpson et al. 2003, available at http://quod.lib.umich.edu/m/micase/). Thus, colloquial words are expected to favor reduction for two reasons: their overall high frequency and the fact that they are mostly used in colloquial contexts favoring reduction.

The words were embedded in sentences (as shown in the Appendix). The sentences were presented orthographically on a computer screen in random order, with a different order for each participant. They were intermixed with 90 filler sentences that did not contain /t/ or /d/ in the flapping environment. All words were presented simultaneously in a single line of 14-point Times New Roman font, centered in the middle of the screen. Presentation was controlled using E-Prime 2.0 Professional software (Schneider et al. 2002).

2.2. Procedure

Each sentence was presented for a maximum of 5 seconds, with the whole experiment lasting a maximum of 11.25 minutes. The participants were instructed to read the sentences aloud as naturally as possible. Their productions were recorded directly onto the presentation computer using Sennheiser-HMD280 headsets. They were allowed to say the sentence more than once if they thought they made an error within the 5-second time limit. Participants were also allowed to continue on to the next sentence before the 5 second elapsed by pressing the left mouse button. On the rare occasions (<2%) that the participant repeated a sentence more than once, measurements were taken from the last pronunciation, as we are interested in what the speaker believes the correct pronunciation to be.

2.3. Participants

Thirty seven self-reported adult native speakers of American English with no history of speech, language, or hearing impairments participated in the experiment. They were recruited from the University of Oregon Psychology/Linguistics Human Subject Pool.

2.4. Exclusions

Triplets of trials were excluded from analysis if the speaker failed to produce the flapping environment or to produce either of the variants considered (an alveolar stop or a flap) in the environment for at least one of the stimuli in the triplet. For real-word cases in which the subject read a different word or a non-word were excluded. For non-word targets, cases in which the speaker read a real word were excluded. The total number of trials excluded on this basis was 144 (8.6%), largely consisting of 1) the triplets including the word ‘rating’ (read as ‘ratting’ by a third of the subjects, probably due to the presence of non-words and the preponderance of words with lax vowels and intervocalic ‘tt’), 2) triplets containing the word ‘emitter’ (often read with initial stress and therefore likely not recognized as a word by the participants who read it in that way), and 3) triplets containing the non-word ‘Abrimody’, often read as ‘Abrimbody’ or ‘Abribody’ (excluded because ‘body’ is a real word).

2.5. Measurements

For the remaining trials, the duration of the closure for /t/ or /d/ was measured using Praat (Boersma 2002). In the case of voiceless stops and true flaps, the duration of the closure was defined as lasting from the end of visible formants before closure to the beginning of formants after closure. In the case of approximant productions, the duration was defined as the period of decreased amplitude and waveform complexity, as illustrated in Figure 2. When there was no marked reduction in amplitude, as in Figure 3, the stop was coded as deleted with a duration of zero.

Figure 2. An approximant production of /t/ in the non-word ‘spating’. The duration was measured as 21 ms.

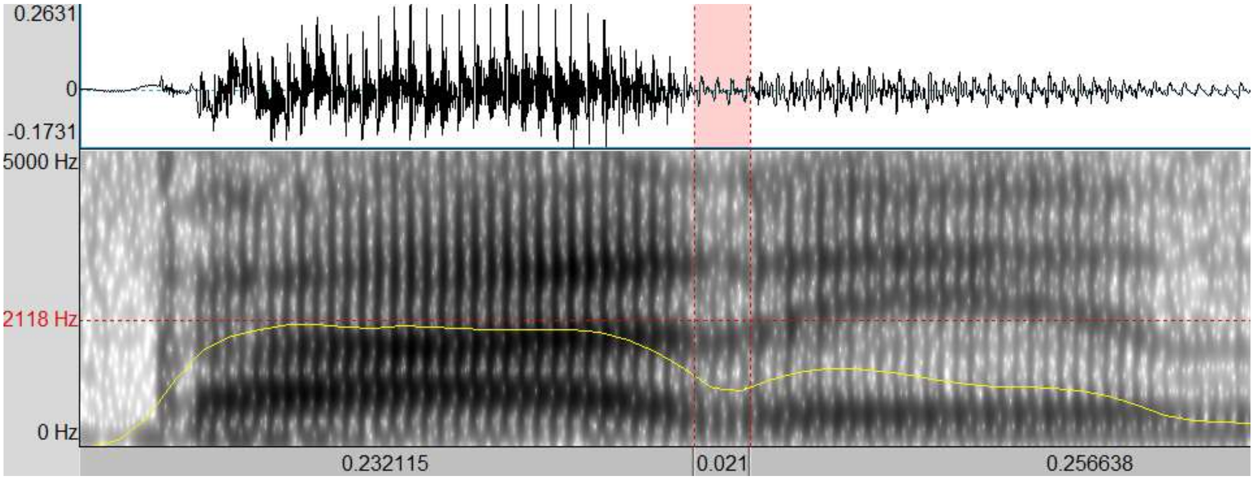
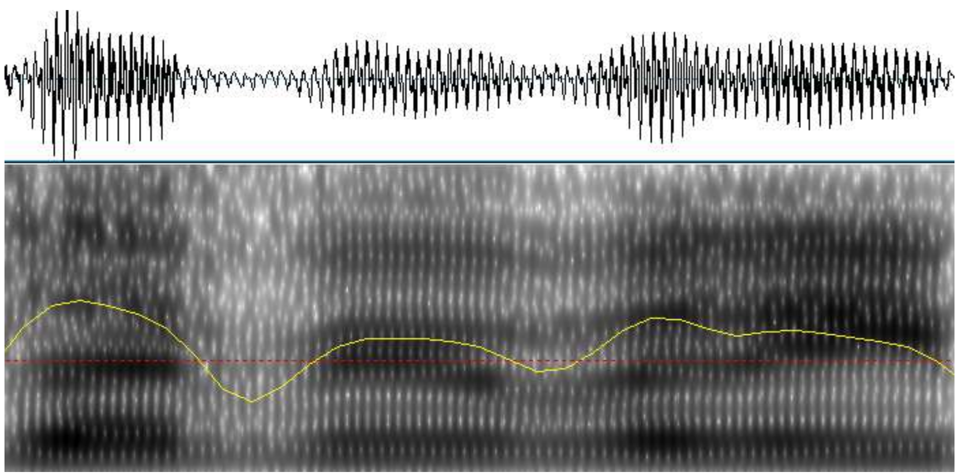
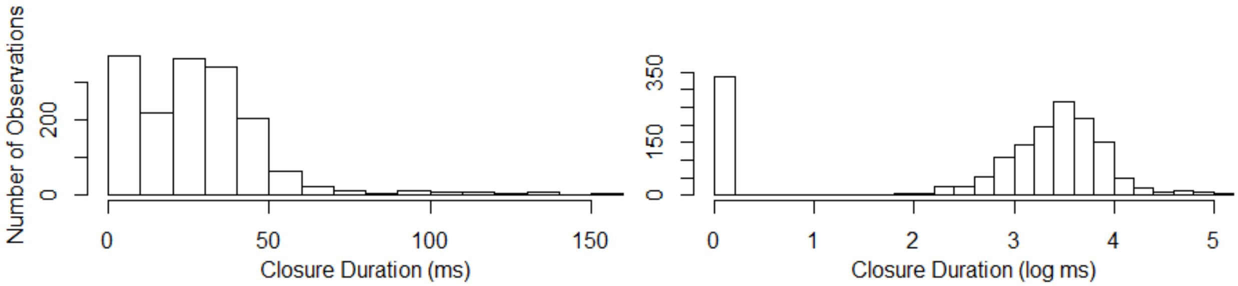


Figure 3. A token of ‘everybody’ in which closure duration was coded as zero.



As shown in Figure 4, the distribution of closure durations is skewed and multimodal (see also Herd et al. 2010). To approximate normality, we log transformed durations that were longer than 10 ms and shorter than 90 ms for analysis as a continuous dependent variable. These cut-offs align well with the duration cut-offs reported by Herd et al. (2010): 90 ms for the longest flap, and 9 ms for the shortest flap. We also defined a categorical dependent variable based on duration, with the categories 0 ms (deletion), 10-90 ms (flap) and >90ms (full stop).

Figure 4. The distribution of closure durations (left) and log durations, with 1 added to deal with zeroes.



The mean measured closure duration was 30 ms for /d/ and 35 ms for /t/; excluding real stops and deletions, the durations were 31 ms and 33 ms respectively. The durations of productions I considered flaps appear to be comparable to those considered flaps in previous studies: the means for /d/ and /t/ are 29.5 ms and 30.4 ms respectively in Herd et al. (2010), 26 ms and 30 ms in Sharf (1982), 37 ms and 34 ms in Lavoie (2000), 44 ms and 41 ms in Stathopoulos & Weismer (1983).

Duration of the preceding vowel was also measured but did not predict closure duration. Given that closure duration appears to be the primary attribute of a flap (see Fukaya & Byrd 2005, Port 1979, Stone & Hamlet 1982, and the review in Herd et al. 2010), preceding vowel duration was not included in the models.

I also considered two intensity-based dependent variables (measured in decibels), which are omitted from the discussion below: minimum intensity during closure, and the difference between minimum intensity during the flap and maximum intensity during the following vowel (e.g. Bauer 2005, Warner & Tucker 2011). (The preceding vowel is sometimes stressed, and sometimes not, which introduces an additional source of variance into its intensity, and led me not to include it in the calculations.) For both dependent variables, maximum amplitude of the following vowel served as a control predictor. The intended use of this variable was to to distinguish between true flaps and approximants, which can be of similar duration (Stone & Hamlet 1982). The locations of intensity minima and maxima was found using the intensity contour and the closure boundaries used for duration measurement. The intensities were approximately normally distributed. However, there were no significant effects of word type, single/double spelling, or voicing of the underlying (a.k.a. spelled) consonant on either of the intensity-based variables. The only significant predictor of intensity during the flap or intensity drop was intensity of the following vowel. Thus the intensity-based variables are omitted from the foregoing discussion.

2.6. Statistical Analysis

The data were analyzed using mixed-effects regression (linear for duration and intensity, logistic for the categorical version of the duration variable) using the lme4 package (Bates et al. 2013) in R (R Development Core Team 2013). The fixed-effects predictors were ‘Word Type’ (whether the word was colloquial, formal or a non-word), ‘Underlying Phoneme’ (/t/ or /d/) and ‘Spelling’ (single or double). The models included random intercepts for speaker and item triplet and random slopes for all fixed effects within speaker and for the crucial effect of Condition within item triplet, maximizing random-effects structure to minimize Type I errors (Barr et al. 2013). *p* values were derived using the normal distribution.

3. Results

There was an effect of word type on closure duration, along with marginal effects of spelling and underlying voicing. Formal words had longer closure durations than colloquial words. Unexpectedly, pseudowords patterned exactly like formal words, rather than being intermediate between formal and colloquial words.

Table 1. The effects of word type (formal words and pseudowords relative to colloquial words), spelling (single vs. double), and underlying phoneme (/t/ vs. /d/) on closure duration of the flap. Negative coefficients indicate favoring reduction (shorter flap duration), while positive ones indicate favoring full forms.

	b	*se(b)*	*t*	*p*
(Intercept)	3.30279	0.0422	78.26	<.001
Formal words	0.12992	0.04222	3.08	0.002
Non-words	0.13602	0.04495	3.03	0.002
Single spelling	-0.09273	0.04857	-1.91	0.06
Phoneme = /t/	0.06669	0.03892	1.71	0.09
Single /t/	0.01289	0.0609	0.21	0.83

Turning now to the categorical duration-based variable, with its three levels “0”, “flap” and “stop”, we find data inconsistent with the hypothesis. Table 2 examines the circumstances under which speakers produce closure. Here, Non-words significantly disfavor deletion (relative to colloquial words) while formal words do not. Rather than being between formal and colloquial words, non-words appear to disfavor reduction even more than formal words do (although the difference between Formal words and Non-words is not statistically significant). There is also a significant effect of spelling (or perhaps syllable structure): single spelling favors deletion.

Table 2. The effects of word type (formal words and pseudowords relative to colloquial words), spelling (single vs. double), and underlying phoneme (/t/ vs. /d/) on whether there is noticeable closure (flaps and full stops grouped together). Negative coefficients indicate favoring reduction (no closure), while positive ones indicate favoring full forms (flaps or stops).

	b	*se(b)*	*z*	*p*
(Intercept)	1.8623	0.3038	6.13	<.001
Formal words	0.3104	0.2409	1.288	0.2
Non-words	0.6615	0.2143	3.086	0.002
Single spelling	-0.7993	0.2734	-2.924	0.003
Phoneme = /t/	-0.1097	0.2507	-0.438	0.66
Single /t/	0.183	0.3318	0.551	0.58

Table 3 reports a comparison between stops and all reduced variants (with or without noticeable closure). Both Formal words and Non-words disfavor reduction, compared to Colloquial words. Once again, non-words do not fall between Formal and Colloquial words (although the difference between Formal words and Non-words is again not significant). There is also a significant effect of whether the phoneme is /t/ or /d/ (a.k.a. whether the spelling uses ‘t’ or ‘d’), with /t/ being more likely to be realized (or ‘t’ read) as a full stop.

Table 3. The effects of word type (formal words and pseudowords relative to colloquial words), spelling (single vs. double), and underlying phoneme (/t/ vs. /d/) on whether the speaker produced a full stop. Negative coefficients indicate favoring reduction (flap or no closure), while positive ones indicate favoring full forms (stops).

	b	*se(b)*	*z*	*p*
(Intercept)	-9.1085	1.5773	-5.775	<.001
Formal words	2.9266	1.0962	2.67	0.008
Non-words	3.3828	1.0887	3.107	0.002

Single spelling	-0.69	1.6283	-0.424	0.67
Phoneme = /t/	2.535	1.2346	2.053	0.04
Single /t/	-0.3005	1.7857	-0.168	0.87

4. Discussion

Words that usually occur in colloquial contexts, which should favor reduction and therefore flapping, were reduced more than words that usually occur in formal contexts. This finding supports the hypothesis that words can become associated with a certain pronunciation, even if that pronunciation is initially triggered by the context (Bybee 2002, Raymond & Brown 2012). However, given that colloquial words are more frequent than formal words in the present study, the result is also consistent with the hypothesis that reduction is associated with high frequency of use, regardless of the context in which the use occurs (Bybee 2001, Pierrehumbert 2001). For flaps, an effect of frequency for medial flaps in the same direction was previously found by Patterson & Connine (2001) with corpus data. However, no effect of word frequency was found by Bauer (2005), Herd et al. (2010), and Warner & Tucker (2011). While those studies used somewhat smaller frequency ranges than Patterson & Connine (2001), reducing statistical power for observing a frequency effect, our frequency range is even smaller. Hence no frequency effect may be expected with the present stimuli based on Bauer (2005), Herd et al. (2010), and Warner & Tucker (2011). The hypothesis that the actual predictor is not frequency itself but rather probability of occurring in a reduction-favoring context may thus be a more likely explanation (see also Raymond & Brown 2012).

The finding that Non-words behaved like the Formal words (or even more formal words) is consistent with the results of Herd et al. (2010) but unexpected under the simple model described in the introduction. According to that model, once the sound change has diffused through the lexicon far enough, the new, reduced pronunciation of the changing sound should become the default, used for low-frequency and, a fortiori, novel words. It is words of intermediate frequency that can remain exceptionally unreduced, provided there is some reason for them to be associated with non-reduction. For us, these are the Formal words, associated with non-reduction by virtue of frequently occurring in formal, reduction-disfavoring contexts. (If the change is articulatorily-motivated, high-frequency words, having always led the change, should remain the most reduced throughout the lifespan of the change.)

One may be tempted to claim that flapping has not yet diffused far enough in American English. However, this possibility is clearly ruled out by our data, as well as prior research. Only 3.4% of our tokens are full stops. Even for Non-words the rate is a low 5.8%. This figure is consistent with prior findings, from both corpus and experimental data (Byrd 1993, Herd et al. 2010, Patterson & Connine 2001). The flap is overwhelmingly more frequent. Yet, the full stop is favored by Non-words as much as it is favored by Formal words.

A more likely explanation lies in the fact that the present experiment employed a reading task, permitting the influence of orthography and the production of so-called “spelling pronunciations”. The influence of spelling on pronunciation should be strongest for Non-words, since they do not have existing articulatory routines pre-assembled in memory and hence have to be assembled piece-by-piece, a process that plausibly involves grapheme-phoneme conversion (e.g. Coltheart et al. 1983). The grapheme-phoneme conversion process might then ‘misfire’ resulting in production of stops in contexts appropriate for flaps. To test for this possibility, I conducted follow-up analyses testing the possible interactions between spelling (‘t’ vs. ‘d’, and single vs. double) and word/non-word status (i.e. collapsing Formal and Colloquial words into the Word category) for all the dependent variables. If it is the case that pronunciation of non-words is particularly affected by the orthography, *and* the orthography-to-phonology conversion process is such that ‘d’ is likely to map onto [d], ‘t’ onto [t], ‘dd’ onto [dd] and ‘tt’ onto [tt], resulting in more stop-like articulations in non-words, then we should expect significant interactions. Double letters should be more likely to be pronounced as geminate consonants in non-words, and ‘t’ should differ from ‘d’ more in non-words as well. However, there were no significant interactions. While this is a null result and hence inconclusive, it does not appear that the pronunciation of non-words is particularly driven by the context-independent grapheme-to-phoneme mappings required to explain the lack of reduction in non-words from an especially strong spelling influence.

I therefore consider it most plausible that the model described in the introduction needs some modification. A very radical revision would involve abandoning the idea of *sound* change. If the learner did not acquire probabilities of pronunciation variants of sublexical units, there would not be a default pronunciation variant for novel words. However, I do not consider this possibility very likely. First, it is unquestioningly true that not all parts of a high-frequency word reduce equally and that some sounds are more prone to reduction than others at any given time (Pierrehumbert 2002). Further, a sound’s proneness to reduction varies across languages, which means it must be learned for individual sounds as part of acquiring a language. Finally, findings of ready transfer of knowledge about sound pronunciations across words argue strongly in favor of sublexical units being probabilistically associated with pronunciations, or distributions thereof (German et al. 2013, McQueen et al. 2006).

One plausible explanation for the results is that an adult speaker may have beliefs about the likely characteristics of words that they encounter for the first time. Given their extensive experience with the language, the fact that the word is novel suggests that it occurs in contexts with which the speaker has had little experience. For our listeners, university students, most newly encountered come from formal, namely academic, contexts. They may therefore expect the novel words in our experiment to be of similar provenance and thus pronounce them in a more formal fashion. More work with other populations, and with oral presentation of novel words, is needed to provide strong support or refute this hypothesis.[2]

A related possibility is that the learner in fact does learn about the relationship between frequency and reduction as part of acquiring language. If each generation in our simulation learned about the monotonic relationship between frequency and reduction that comes from articulatory practice, the monotonic relationship would persist across generations, with each subsequent generation reducing low-frequency words least and high-frequency words most. Newly encountered words are low-frequency and therefore may be assumed to be less eligible for reduction.

5. Conclusion

I have implemented a simple model of articulatorily-motivated sound change, along the lines of Pierrehumbert (2002). In this model, 1) there are words and sounds, 2) a word's use causes reduction of the sounds in that word, and 3) words and sounds (modeled as groups of words) are associated with reduction probabilities. The model was shown to predict that an advanced articulatorily-motivated sound change will be more active in low-frequency (and especially novel) words than in medium-frequency words because some of the latter may become associated with non-reduction (e.g. due to frequent occurrence in a reduction-disfavoring context). We tested this prediction for American English flapping and found that, contrary to the prediction, novel words behave like words that tend to occur in reduction-disfavoring contexts. Several possible elaborations of the model were proposed. The one that appears most likely to the author is that language learners discover the relationship between reduction and frequency, thus coming to expect that unknown words are less likely to be reduced than more frequent words.

Appendix: Experimental Sentences

Colloquial	Formal	Non-word
Are you kidding me, Share?	Are you trading gas shares?	Are you cadding wug blares?
They are cheating on each other.	They are leading on each other.	They are meading on each other.
I would prefer a letter.	I would prefer the latter.	I would prefer a gatter.
She is looking for the butter.	She is looking at the jitter.	She is looking at The Witter.

I found the bullshitter.	I found the emitter.	I found the lenitter.
He bumped into his daddy.	He's read about that study.	He stood by the Gaddy.
How is everybody doing today?	How is the antibody doing that?	How is Don Abrimody doing now?
He found somebody online.	He came to embody this principle.	He came to Plembody today.
She is going to get even madder.	She is going to find the highest bidder.	She is going to get even gadder!
She is really into her knitting.	She is really into the setting.	She is really into her mitting.
He is getting fatter.	They just have to scatter.	They are getting snatter.
He is really into flirting with Jennifer.	He is really into rating the stimuli.	He is really into brating the blick.
That girl is so pretty!	The world needs this treaty!	The murl feeds the dretty.
The presenter is great at shutting up the audience.	The presenter is great at stating the obvious.	The presenter is great at spating the audience.
He always tells dirty jokes!	He always puts duty first!	He always bicks puty off.

Notes

[1] variant ~ b_0 + (1|word). A constant number was used as the single fixed-effects predictor because a fixed-effects predictor is required by the software. Additional influences on pronunciation can easily be incorporated as additional fixed or random effects in the equation (e.g. Labov 1969, Barth & Kapatsinski in press).

[2] This hypothesis is compatible with the idea that words are sorted into sublexica, to which different grammars apply (e.g. Becker & Gouskova 2012, Hayes 2014). The suggestion here is that the speaker can use frequency as one source of information about the sublexicon to which a word belongs.

References

Albright, Adam, and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational / experimental study. *Cognition* 90. 119-161.

Baayen, R. Harald. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer.

Baayen, R. Harald, Ton Dijkstra, & Robert Schreuder. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language* 37. 94-117.

Barr, Dale J., Roger Levy, Christoph Scheepers, & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68. 255-278.

Barth, Danielle, & Kapatsinski, Vsevolod. In press. 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 & Linguistic Theory*.

Bates, Douglas, Martin Maechler, Ben Bolker, & Steven Walker. 2013. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-4. http://CRAN.R-project.org/package=lme4

Bauer, Matt. 2005. Lenition of the flap in American English. *Penn Working Papers in Linguistics* 10.2.

Becker, Michael, & Maria Gouskova. 2012. Source-oriented generalizations as grammar inference in Russian vowel deletion. Ms., Indiana and NYU.

Beekhuizen, Barend, Rens Bod, and Willem Zuidema. 2013. Three design principles of language: The search for parsimony in redundancy. *Language and Speech* 56. 265-290.

Boersma, Paul. 2002. Praat, a system for doing phonetics by computer. *Glot international* 5. 341-345.

Brysbaert, Mark, & Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41. 977-990.

Bybee, Joan. 2001. *Phonology and Language Use*. Cambridge, UK: 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. 2003. Cognitive processes in grammaticalization. In Michael Tomasello (ed.), *The New Psychology of Language. Vol.2: Cognitive and Functional Approaches to Language Structure*, 145-167. Mahwah, NJ: Lawrence Erlbaum.

Byrd, Dani. 1993. 54,000 American stops. *UCLA Working Papers in Phonetics* 83. 97-116.

Clahsen, Harald. 1999. Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences* 22. 991-1013.

Coltheart, Max, Jacqueline Masterson, Sally Byng, Margot Prior, & Jane Riddoch. 1983. Surface dyslexia. *Quarterly Journal of Experimental Psychology* 35. 469-495.

Davies, Mark. 2004-. BYU-BNC. (Based on the British National Corpus from Oxford University Press). Available online at http://corpus.byu.edu/bnc/.

Davies, Mark. 2008. The corpus of contemporary American English: 425 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

de Jong, Kenneth. 1998. Stress-related variation in the articulation of coda alveolar stops: Flapping revisited. *Journal of Phonetics* 26. 283-310.

Dell, Gary S., Lisa K. Burger, & William R. Svec. 1997. Language production and serial order: A functional analysis and a model. *Psychological Review* 104. 123-147.

Erker, Daniel, & Gregory R. Guy. 2012. The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language* 88. 526-557.

Francis, W. Nelson, and Kučera, Henry. 1982. *Frequency Analysis of English Usage. Lexicon and Grammar*. Boston: Houghton Mifflin.

Fujimura, Osamu. 1986. Relative invariance of articulatory movements: an iceberg model. In Joseph S. Perkell and Dennis H. Klatt (eds.), *Invariance and Variability in Speech Processes*, 226-234. Hillsdale, NJ: Lawrence Erlbaum

Fukaya, Teruhiko, & Dani Byrd. 2005. An articulatory examination of word-final flapping at phrase edges and interiors. *Journal of the International Phonetic Association* 35. 45-58.

German, James S., Katy Carlson, & Janet B. Pierrehumbert. (2013). Reassignment of consonant allophones in rapid dialect acquisition. *Journal of Phonetics* 41. 228-248.

Hayes, Bruce. 2014. Comparative phonotactics. Paper presented at Chicago Linguistic Society 50.

Herd, Wendy, Allard Jongman, & Joan Sereno. 2010. An acoustic and perceptual analysis of /t/ and /d/ flaps in American English. *Journal of Phonetics* 38. 504-516.

Kahn, Daniel. 1976. *Syllable-based Generalizations in English Phonology*. Bloomington, IN: Indiana Linguistics Club.

Kapatsinski, Vsevolod. 2010a. Frequency of use leads to automaticity of production: Evidence from repair in conversation. *Language and Speech* 53. 71-105.

Kapatsinski, Vsevolod. 2010b. What is it I am writing? Lexical frequency effects in spelling Russian prefixes: Uncertainty and competition in an apparently regular system. *Corpus Linguistics and Linguistic Theory* 6. 157-215.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45. 715-762.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar. Vol. 1: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Lavoie, Lisa Marie. 2000. Phonological patterns and phonetic manifestations of consonant weakening. Ph.D. Dissertation, Department of Linguistics, Cornell University.

Maye, Jessica, Richard N. Aslin, & Michael K. Tanenhaus. 2008. The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science* 32. 543-562.

McQueen, James M., Anne Cutler, & Dennis Norris. 2006. Phonological abstraction in the mental lexicon. *Cognitive Science* 30. 1113-1126.

Patterson, David, & Cynthia M. Connine. 2001. Variant frequency in flap production. *Phonetica* 58. 254-275.

Peperkamp, Sharon, & Emmanuel Dupoux. 2007. Learning the mapping from surface to underlying representations in an artificial language. In Jennifer Cole & José Hualde (eds.), *Laboratory Phonology 9*, 315-338. Berlin/New York: Mouton de Gruyter.

Phillips, Betty S. 1984. Word frequency and the actuation of sound change. *Language* 60. 320-342.

Phillips, Betty S. 2001. Lexical diffusion, lexical frequency, and lexical analysis. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure*, 123-136. Amsterdam/Philadelphia: John Benjamins.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure*, 137-158. Amsterdam/Philadelphia: John Benjamins.

Pierrehumbert, Janet B. 2002. Word-specific phonetics. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory Phonology 7*, 101-139. Berlin/New York: Mouton de Gruyter.

Pinker, Steven. 1991. Rules of language. *Science* 253. 530-535.

Port, Robert F. 1979. The influence of tempo on stop closure duration as a cue for voicing and place. *Journal of Phonetics* 7. 45-56.

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/New York: Mouton de Gruyter.

Shockey, Linda. 1984. All in a flap: Long-term accommodation in phonology. *International Journal of the Sociology of Language* 1984. 87-96.

Schneider, Walter, Amy Eschman, & Anthony Zuccolotto. 2002. *E-Prime Reference Guide*. Psychology Software Tools, Inc.

Sharf, Donald J. 1962. Duration of post-stress intervocalic stops and preceding vowels. *Language and Speech* 5. 26–30.

Simpson, Rita C., David Y. Lee, & Sheryl Leicher. 2003. *MICASE Manual*. Ann Arbor: University of Michigan.

Stathopoulos, Elaine T., & Gary Weismer. 1983. Closure duration of stop consonants. *Journal of Phonetics* 11. 395–400.

Stone, Maureen, & Sandra Hamlet. 1982. Variations in jaw and tongue gestures observed during the production of unstressed /d/s and flaps. *Journal of Phonetics* 10. 401-415.

van Heuven, Walter J. B., P. Mandera, Emmanuel Keuleers, & Mark Brysbaert. 2014. SUBTLEX-UK: a new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology* 67. 1176-1190.

Warner, Natasha, & Benjamin V. Tucker. 2011. Phonetic variability of stops and flaps in spontaneous and careful speech. *Journal of the Acoustical Society of America* 130. 1606-1617.

Zipf, George Kingsley. 1935. *The Psycho-biology of Language*. Oxford: Mifflin Houghton.