The Real Effect of Word Frequency on Phonetic Variation

Aaron J. Dinkin

1 Background

"Exemplar Theory" and "Usage-Based Phonology" are general names for a school of thought (see, e.g., Bybee 1999, 2000; Pierrehumbert 2002) that holds that the units of a speaker's phonological knowledge are memorized phonetic tokens of individual lexical items. Thus in producing a lexical item, the speaker's phonetic target is supposedly determined just by the average phonetic value of the stored exemplars of that item. This paper addresses a claim made in the Exemplar Theory literature about the relationship between lexical frequency and phonetic change in progress: It is frequently claimed that the Exemplar Theory literature implies that lexical items that are used more frequently should undergo regular sound changes more rapidly. This is because, each time a user of the language hears an innovative token of a word that is undergoing a change, then the average phonetic value of all the exemplars of that word heard so far will shift a little bit in the direction of the change. And so words that are heard more frequently will have had their phonetic averages shifted by that little bit in the direction of the change more frequently, and so they'll undergo the sound change more rapidly. Thus, to quote Pierrehumbert (2002), "high frequency words tend to lead Neogrammarian sound changes." Bybee (2000) cites several studies in which high-frequency words have been found to be undergoing sound change faster.

Labov (2003), on the other hand, examining an enormous amount of data on the fronting of the nuclei of the back upgliding diphthongs /uw/, /ow/, /aw/ in present-day American English, found that almost all variation could be accounted for purely by phonetic constraints. Word frequency played no role at all; high-frequency words were not in general any more or less advanced in the sound change in Labov's data than low-frequency words. This leads to a conundrum: It's clearly too strong to say that frequent words lead phonetic change as a general rule; there's no evidence for that at all in Labov's data. Therefore in the studies Bybee cites, there must be some other factor which is causing the more frequent words to be in the lead in those particular phonetic changes but not the changes studied by Labov. The results reported below will shed some light on what the actual relationship is between word frequency and sound change.

2 Methodology

This study in particular investigates the effect of word frequency on the frontness or backness of the short vowels /i e æ ʌ u/[1] of the English of the Northern United States, as defined by Labov et al. (2006): this region encompasses a large area on the southern side of the Great Lakes, including such cities as Buffalo, Cleveland, Detroit, Chicago, Milkwaukee, Minneapolis, and many others. In most of the North, most of the short vowels are involved in an ongoing chain shift known as the Northern Cities Shift. The relevant features of the Northern Cities Shift for the current study are its effects on the frontness and backness of the short vowels—in instrumental phonetic terms, its effects on the value of their second formants (F2). So what's relevant is that tokens of /æ/ that are leading the change should have higher F2, and leading tokens of /e/, /i/, and /ʌ/ should have lower F2. Like Labov (2003), for my data set I took advantage of the huge corpus of phonetic measurements collected for the Telsur survey of American English, reported in detail by Labov et al. (2006). This is a corpus of some 130,000 phonetic measurements of American English vowels, of which about 10% are short vowels from the Northern dialect region.

Tokens were coded for word frequency based on data from the Brown Corpus of Standard North American English.[2] All words that were among the five thousand most frequently-occurring words in the Brown Corpus were coded as "Top5000", and likewise for "Top500" and "Top200". Within the Top5000 group, each word was also coded for its exact frequency—that is, its exact number of occurrences within the Corpus. Finally, within the Top500 words, each word was also coded for its status as a function word or a lexical word; function words included prepositions, conjunctions, determiners, verbal auxiliaries, closed-class verbs like *have* and *be*, and the like.

For each short vowel phoneme, a multiple-regression analysis was run on all the F2 measurements of that phoneme in the Telsur data restricted to the Northern dialect region. The independent variables in the regression included both the word-frequency variables described above and all of the phonetic-environment variables that are included in the Telsur data.

---

[1] I use the notation of Labov et al. (2006) here: /i/ as in *pit*, /e/ as in *pet*, /æ/ as in *pat*, /ʌ/ as in *putt*, /u/ as in *put*. The vowel /o/ as in *pot* is excluded because it is phonologically a long vowel in the Northern United States (Labov & Baranowski 2006).

[2] My source of data on the frequency of words in the Brown corpus was *http://www.edict.com.hk/textanalyser/wordlists.htm*.

3 Results

Table 1 shows the results for /i/. The multiple regression found eleven phonetic variables plus the Top-5000 frequency variable as having statistically significant effects on backness of /i/: other things being equal, an /i/-word among the 5000 most frequent words of the Brown Corpus was on average about 60 Hz backer than a less frequent word. Since /i/ is being backed in the Northern Cities Shift, this is consistent with the Exemplar Theory claim that more frequent words will lead sound changes. Note, however, that word frequency has a smaller effect than any phonetic variable.

variable	coefficient	variable	coefficient
onset cluster	–489 Hz	labial onset	–119 Hz
liquid onset	–423 Hz	complex coda	–84 Hz
apical onset	–167 Hz	apical coda	–71 Hz
palatal onset	–151 Hz	/l/ coda	–69 Hz
nasal coda	+136 Hz	polysyllable	–66 Hz
labial coda	–122 Hz	**Top 5000**	**–57 Hz**

$p < .01\%$ $n = 2492$ constant = 2147 Hz $r^2 = 32\%$

Table 1: effects of frequency and phonetic variables on /i/ in the North.

Roughly the same thing holds for /e/, on Table 2: fifteen phonetic variables are statistically significant at the .01% level, and Top5000 is also significant but has the smallest effect. Here again the effect of word frequency is in the same direction as Exemplar Theory would predict—words in the top 5000 are 33 Hz backer, in the direction of the Northern Cities Shift.

variable	coefficient	variable	coefficient
apical coda	–353 Hz	stop coda	+127 Hz
labial coda	–324 Hz	liquid onset	–125 Hz
labdent. coda	–279 Hz	complex coda	–96 Hz
intdent. coda	–271 Hz	polysyllable	–83 Hz
nasal coda	+218 Hz	/l/ coda	–67 Hz
palatal coda	–216 Hz	voiced coda	+60 Hz
velar coda	–204 Hz	apical onset	–39 Hz
onset cluster	–162 Hz	**Top 5000**	**–33 Hz**

$p < .01\%$ $n = 2913$ constant = 2034 Hz $r^2 = 39\%$

Table 2: effects of frequency and phonetic variables on /e/ in the North.

However, when we move on to /æ/, the Exemplar Theory prediction breaks down. On Table 3, we see that tokens of /æ/ in the top 5000 words are backer than less frequent words, which is contrary to the Northern Cities Shift.

variable	coefficient	variable	coefficient
nasal coda	+275 Hz	stop coda	+94 Hz
velar coda	–207 Hz	labdent. coda	–79 Hz
apical coda	–152 Hz	voiced coda	+75 Hz
liquid onset	–134 Hz	apical onset	–63 Hz
onset cluser	–123 Hz	complex coda	+42 Hz
labial coda	–123 Hz	**Top 5000**	**–23 Hz**
polysyllable	–99 Hz		

$p \leq .01\%$ $n = 5091$ constant = 2058 Hz $r^2 = 30\%$

Table 3: effects of frequency and phonetic variables on /æ/ in the North.

Now, the tensing of /æ/ is basically a completed phase of the Northern Cities Shift, so this might not tell us very much about the relationship of frequency with sound change **in progress**. But the backing of /ʌ/ is a new and ongoing phase of the Northern Cities Shift, and on Table 4 we see that the most frequent tokens of wedge are **fronter**, again contrary to the shift. So, for /i/ and /e/, frequent words lead the Northern Cities Shift, but for /æ/ and /ʌ/, frequent words trail it. Therefore, frequent words leading sound change is clearly not the explanation for what's going on here.

variable	coefficient	variable	coefficient
/l/ coda	–287 Hz	palatal coda	+106 Hz
liquid onset	–147 Hz	polysyllable	+49 Hz
labial onset	–124 Hz	**Top 5000**	**+36 Hz**
onset cluster	–111 Hz	voiced coda	–32 Hz
apical coda	+110 Hz		

$p \leq .02\%$ $n = 1794$ constant = 1372 Hz $r^2 = 37\%$

Table 4: effects of frequency and phonetic variables on /ʌ/ in the North.

But if we disregard the particular directions of change in the Northern Cities Shift, the pattern of Tables 1–4 obvious. The front vowels, /i/, /e/, and /æ/, are backer in frequent words, regardless of the direction of sound change; /ʌ/, a back vowel, is fronter in more frequent words. Moreover, on Table 5 we find that the other short back vowel, /u/, is also fronter in the

most frequent words (although in this case the significant effect of frequency appears only for the Top200 variable; statistically significant effects do not emerge for Top5000 or even Top500). So the generalization is that short vowels are **more central** in frequent words: front vowels are backer, and back vowels are fronter.

variable	coefficient	variable	coefficient
apical onset	+253 Hz	**Top 200**	**+145 Hz**
palatal onset	+237 Hz	velar onset	+141 Hz
/l/ onset	–184 Hz	labial onset	–112 Hz

$p < .01\%$ $n = 731$ constant = 1267 Hz $r^2 = 68\%$

Table 5: effects of frequency and phonetic variables on /u/ in the North.

4 Beyond the North

Now, if such a tendency exists—that short vowels are more central in more frequent words—then we would that tendency to be structurally independent of the particular sound changes in progress in the North. In other words, we'd expect to be able to find short vowels to be more centralized in more frequent words in data from any region, or even in the aggregated data from all regions. And indeed we do: Table 6 summarizes the result of carrying out the same multiple-regression tests as in Tables 1–5 on the short-vowel measurements from the entire Telsur data set. Each vowel shows roughly the same frequency effects over the entire Telsur data set as it does when the data is restricted to the North.

vowel	/i/	/e/	/æ/	/ʌ/	/u/
effect of freq.	–61 Hz	–28 Hz	–18 Hz	+44 Hz	+80 Hz
n	10,182	11,466	17,147	6939	3197

$p < .01\%$ in all cases; freq. variable is Top200 for /u/, Top5000 otherwise.

Table 6: effects of frequency on short vowel F2 in the whole Telsur corpus.

So, we can conclude that the Northern Cities Shift, like the fronting of back upgliding vowels in Labov (2003), is not subject to frequency effects: short vowels show generally the same behavior with respect to word frequency in the area subject to the Northern Cities Shift as they do in North America overall. But the realization of short vowels across North American English as a whole does show a word-frequency effect: frequent words are more centralized. How do we interpret this?

5 Analysis

One possible explanation for the result that short vowels are more central in more frequent words is that the most frequent words tend to be function words, and function words are often unstressed, and their vowels get reduced to schwa. And so Exemplar Theory might predict that the speaker would be influenced by those unstressed tokens and end up centralizing the vowels in function words a bit even when those words stressed.[3] There are, in fact, some well-known cases of function words ending up phonemically less peripheral than lexical words with comparable phonological history: in dialects in which /æ/ is split into a tense and a lax phoneme, such as those of New York and Philadelphia, function words like *and* and *can* typically contain the lax phoneme even in phonological environments where the tense phoneme is usually found.

But function-word status was not found to have any statistically significant effect at all on F2 in most of the multiple-regression tests summarized above. For /e/ a marginally significant effect of function-word status appeared at $p = .2\%$ (compared to phonetic and word-frequency effects all with $p < .01\%$), with function words *fronter* by 68 Hz. So it seems as if the centralization tendency observed must actually be dependent on word frequency, not function-word status.

Phillips (1984), when discussing the relationship of word frequency with sound change, said "Changes affecting the most frequent words first typically involve either vowel reduction and eventual deletion or assimilation…. The thing to note about these sound changes is that they all have their basis in the physiology of speech." This is in sharp contrast to Pierrehumbert's blanket claim that "high-frequency words tend to lead Neogrammarian sound change". And in fact Phillips lists most of the examples cited by Bybee (2000) and shows that they all fit the description she gives—they consist for the most part of vowel weakening or deletion or assimilation, or in a few cases spirantization. With some abuse of terminology, we can put all of these changes in the broadly construed category of **lenition**—they all consist of reducing the articulatory effort to produce a word by reducing the number, duration, or intensity of articulatory gestures. So we can paraphrase Phillips as saying that high-frequency words tend to lead only sound changes of **lenition**.

On the other hand, the fronting of back upgliding diphthongs, which is

[3] The Telsur corpus of vowel-formant measurements includes only stressed tokens; so the actual reduction to schwa of unstressed tokens does not contribute directly to the statistical results described.

the subject of Labov (2003), is a **dissimilatory** sound change, which **increases** the number of articulatory gestures required to pronounce a word. So by what we may call "Phillips's principle", it's unsurprising that there's no word-frequency effect on this sound change. Likewise, the Northern Cities Shift is a complicated chain shift in which some vowels move one way and some another way, with no overall lenitory tendency; therefore we shouldn't be surprised that frequent words don't lead the change.

Furthermore, we can see the interaction of lenition and word frequency also in linguistic variation which isn't part of a change in progress. One of the standard examples in Exemplar Theory of frequent words leading a sound change is the finding from Bybee (2002) that frequent lexical items in English undergo deletion of final /t/ and /d/ more often. But as Abramowicz (2006) points out, *t*/*d*-deletion is generally regarded as a stable variable, not a change in progress, so frequency effects on *t*/*d*-deletion don't constitute evidence for claims about linguistic change. On the other hand, *t*/*d*-deletion **does** fall in the category being broadly referred to (in this paper) as lenition: synchronically, it's deletion of a segment, reducing the amount of articulatory effort it takes to pronounce a word.

Meanwhile, Abramowicz finds no frequency effect in his Philadelphia data set on the (ing) variable—that is, so-called "g-dropping", as in *walkin'*, *talkin'*, and so on. And the (ing) variable is **not** a case of lenition. It's just replacing a velar place of articulation with an apical one, without obviously reducing the amount of articulatory effort involved in pronouncing a word. So so far it seems as if Phillips's principle applies to stable variation as well as changes in progress: Frequent words are more subject to lenition than less frequent words.

A functional explanation of this phenomenon is attractive: Lenition has the effect of reducing the amount of articulatory effort required to produce a word, at the expense of rendering it phonetically less distinct—that is, closer, in phonetic terms, to other, similar words—and therefore more prone to misundertanding. Since less-frequent words are likely to be less familiar to the hearer, and therefore less expected and less easily remembered, they too may be more prone to misunderstanding than more frequent words. Under these assumptions, it seems reasonable that less frequent words should be less apt to undergo lenition, since they are more in need of the extra phonetic clarity afforded by distinct, non-lenited articulation than are more frequent, easily recognizable words.

Phillips's principle gives a rationale for the findings of this paper. All else being equal, centralizing short vowels can be construed as lenition in the broad sense: if the tongue moves a shorter distance from its default central position either to the front or back to produce a vowel, it's making less effort

to reach its target and taking less time to do so. So finding that short vowels in frequent words are on the whole more centralized than in less frequent words lines up again with the generalized version of Phillips's principle. If centralizing short vowels takes less articulatory effort, then we should expect short vowels to be more centralized in more frequent words, regardless of whether or not that centralization is part of a sound change in progress or even contrary to one.

So, we conclude that the real effect of word frequency on phonetic variation is not that more frequent words lead in regular sound change, as the Exemplar Theory literature says. The real effect of word frequency is that more frequent words are more subject, not to (diachronic) change per se, but to lenition—that is, variation in the direction of reduced articulatory effort, whether part of a sound change in progress or not.

6 Caveats and Conclusion

Some of the statistical results presented above show so me anomalous or ambiguous behavior that is worthy of mention. Prominent among these is that different measures of word frequency don't always yield the same result. For instance, on Table 5 above, Top200 is used as the frequency cutoff for /u/, while other cutoffs do not show statistically significant effects on F2; whereas Top5000 is used for all the other vowels considered. The selection of Top200 for /u/ and Top5000 for other phonemes can perhaps be ascribed just to the fact that there are relatively few lexical items that contain /u/.

But Top5000 and Top200 share the property of being categorical frequency variables, dividing up the lexicon into "more frequent words" and "less frequent words"; despite the fact that the location of the cutoff is different for /u/ than for other vowels, it is the same general approach that shows the F2 effect. However, the results are inconsistent if, instead of such a categorical cutoff, frequency is entered into the multiple regression as a gradient variable corresponding to each word's actual frequency in the Brown corpus. For some of the vowels this gradient frequency variable has no significant effect; for others, the effect is statistically significant but so small that it could account for only a few hertz' difference between the most and least significant words.

Also, some of the phonetic effects that do turn up as significant are bizarre. For example, Table 2 shows that having an apical coda has **six times** as strong a centralizing effect on /e/ as /l/ in the coda does. Although this comes out as statistically significant at the $p < .01\%$ level, it seems phonetically bizarre, since in individual cases /l/ is usually observed to have a strong backing effect on preceding vowels. The anomalousness of some of the pho-

netic effects found may cast some doubt on the validity of the frequency effects.

If word frequency really is the cause of the centralization effect that it seems to have from the above analysis, it's moderately surprising that robustly significant effects do not appear from more than one word-frequency variable. But on the other hand, it is encouraging that the results are consistent: the effects of word frequency presented above are in the direction of centralization for all five vowels, whether the data is restricted to the North or includes the whole Telsur corpus; and all the word-frequency results shown on Tables 1–6 are statistically significant at least to the level of $p = .01\%$. So even if it may not be word frequency directly that is having a centralizing effect on short vowels, at least it seems clear that some (perhaps subtler) factor related to word frequency is implicated. But, more importantly, it is certainly not sound change in progress *in general* that is led by more frequent words.

References

Abramowicz, Lukasz. 2007. Sociolinguistics meets exemplar theory: frequency and recency effects in (ing). *Penn Working Papers in Linguistics* 13.2.

Bybee, Joan. 1999. Usage-based phonology. In *Functionalism and formalism in linguistics, volume I: General papers,* ed. Michael Darnell, Edith Moravcsik, Frederick Newmeyer, Michael Noonan, & Kathleen Wheatley, 211–242. Amsterdam: John Benjamins.

Bybee, Joan. 2000. Lexicalization of sound change and alternating environments. In *Laboratory Phonology V: Acquisition and the Lexicon*, ed. Michael B. Broe and Janet M. Pierrehumbert, 250–268. 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.

Labov, William. 2003. Words floating on the surface of sound change. Talk presented at NWAVE 32, Philadelphia.

Labov, William, Sharon Ash, & Charles Boberg. 2006. *Atlas of North American English*. Berlin: Mouton/de Gruyter.

Labov, William & Maciej Baranowski. 2006. 50 msec. *Language Variation and Change* 18, 223–240.

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

Pierrehumbert, Janet. 2002. Word-specific phonetics. In *Laboratory Phonology VII*, ed. Carlos Gussenhoven & Natasha Warner, 101–140. Berlin: Mouton/de Gruyter.

Department of Linguistics
University of Pennsylvania
Philadelphia PA 19104
dinkin@ling.upenn.edu