EGG School, Banja Luka
July/August 2018

Does word frequency affect phonology? Reasons to be cautious...1

Patrick Honeybone
University of Edinburgh
patrick.honeybone@ed.ac.uk

The contents of this session

1. Frequency effects – what's it all about...?
2. What kinds of frequency effects are there?
3. High frequency effects and low frequency effects
4. 'Tiny-word-based effects' (word-reduction) and segmental-category-type effects
5. What's really at issue?

Frequency effects – what's it all about...?

Here's a possible definition of 'frequency effect' for our purposes:

- a phenomenon which is relevant to phonology in some way, the patterning of which is constrained by lexical token frequency

In such things,

- the patterning of a phonological phenomenon is claimed to be affected by the differential frequency of use of words in which the phonological environment required by the phonological phenomenon is found

One thing to be clear about:

- we're talking about token frequency – not type frequency
  - token frequency is sometimes called text frequency
  - that is, it's referring to the frequency of occurrence in texts

Type frequency refers to the number of distinct entries in the lexicon that feature a particular structure, whereas token frequency refers to language use.

To exemplify...

- one famous count for English was done by Fry (1947), [here from Taylor (2012)]

Consonant	Token frequency	Type frequency
n	7.58%	6.48%
t	6.42%	6.95%
d	5.14%	4.32%
s	4.81%	6.88%
l	3.66%	5.56%
ð	3.56%	0.12%
r	3.51%	4.68%
m	3.22%	3.01%
k	3.09%	4.56%
w	2.81%	0.93%
z	2.46%	4.05%
v	2.00%	1.22%
b	1.97%	2.21%
f	1.79%	1.79%
p	1.78%	3.16%
h	1.46%	0.75%
ŋ	1.15%	1.86%
g	1.05%	1.27%
ʃ	0.96%	1.24%
j	0.88%	0.72%
ʤ	0.60%	0.79%
ʧ	0.41%	0.54%
θ	0.37%	0.33%
ʒ	0.10%	0.07%

Token frequency effects, driven by the frequency of use of items have been claimed to exist in both synchronic and diachronic phonology.

In a sense, people have 'always' known ('obviously') about such things...

goodbye < *god by with you*
hiya < *how are you*

Stampe (1979) points out that that this kind of thing can be live in variation:

- *I don't know* can reduce to [ãn̪õn̪ũ]
- *I dent noses* cannot reduce like this

This kind of lexicalisation of reduced forms (Kiparsky 2016) only occurs to highly frequent strings

- it's sporadic (unpredictable?) and can be accounted for in any model

Other claims have been made that with more far-reaching potential importance.

As soon as the neogrammarians' exceptionlessness hypothesis was proposed, it was argued to be mistaken

- Schuchardt (1885) wrote: "The greater or lesser frequency in the use of individual words that plays such a prominent role in analogical formation is also of great importance for their phonetic transformation, not within rather small differences, but within significant ones. Rarely-used words drag behind; very frequently used ones hurry ahead. Exceptions to the sound laws are formed in both groups."
  - this expresses the basic frequency argument: words behave differently in phonological changes according to how frequently speakers use them
  - this is an inherently lexically-specific factor – frequency of use is not driven by phonological factors

Recent work, with roots in the 1970s, but starting really in the 2000s, has picked this up and run with it.

Phillips (2006) uses **Coronal Stop Deletion** as a basic example of a frequency effect

- in Dutch and (American and some other varieties of) English, there is variation between realisations of words like those below, in which forms with a final coronal stop following another consonant occur alongside forms without the coronal stop:

English:			Dutch:		
told	[toʊld]	[toʊl]	*kiest*	[kiːst]	[kiːs]
held	[hɛld]	[hɛl]	*danst*	[dɑnst]	[dɑns]
felt	[fɛlt]	[fɛl]	*wast*	[wɑst]	[wɑs]
built	[bɪlt]	[bɪl]	*wist*	[wɪst]	[wɪs]
sent	[sɛnt]	[sɛn]	*moest*	[muːst]	[muːs]
meant	[mɛnt]	[mɛn]	*buigt*	[bœyxt]	[bœyx]
lent	[lɛnt]	[lɛn]	*lacht*	[lɑxt]	[lɑx]
kept	[kɛpt]	[kɛp]	*bracht*	[brɑxt]	[brɑx]
slept	[slɛpt]	[slɛp]	*krijgt*	[krɛixt]	[krɛix]
left	[lɛft]	[lɛf]	*vliegt*	[fliːxt]	[fliːx]
lost	[lɑst]	[lɑs]	*mocht*	[mɔxt]	[mɔx]
			zegt	[zɛxt]	[zɛx]

- for English, the Coronal Stop Deletion (CSD) rule can be seen as: t,d → Ø / C__#
- in Dutch, the rule can be seen as: t → Ø / s,x__#
  - what's so interesting about that...?

The interest lies in the correlation of the commonness of deletion in particular words and the frequency with which those words are used, as in the following data (from Phillips, 2006)

- the zeros imply that some words may not undergo CSD at all
- *CELEX* = a frequency database from the Centre for Lexical Information, based on a corpus of 17.9 million words (16.6 million from written texts; 1.3 million from dialogue)
  - the figures for frequency given here are 'raw word form frequencies' = the number of times each words occurs in the CELEX corpus

Phonetic environment	Verb	% Deletion	*CELEX* – raw word form frequency	
			More susceptible to deletion	Less susceptible to deletion
-ld	told	68	1763	
	held	0		765
-lt	felt	55	1449	
	built	0		456
-nt	sent	25	551	
	meant	0		515
	lent	0		25
-pt	kept	66	750	
	slept	50		120
-ft/st	left	25	1503	
	lost	0		759

The claim is that:

- once phonological environment is considered – there is a frequency effect

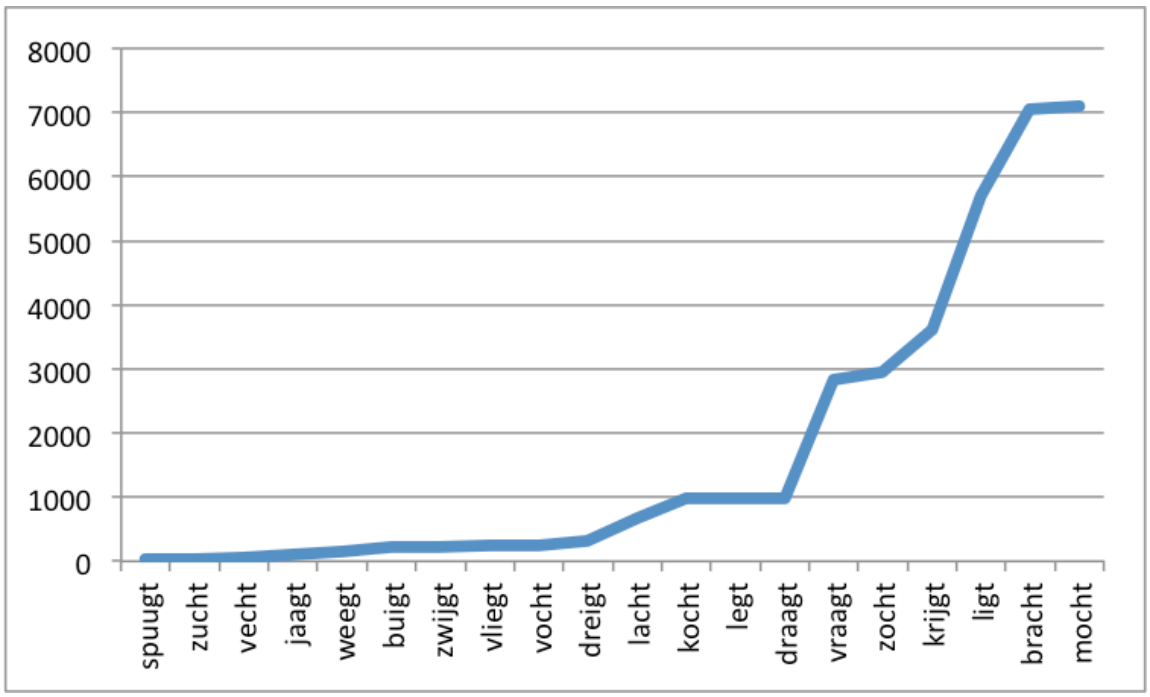
More detailed data for Dutch CSD (also from Phillips, 2006) shows a gradient correlation, at least for the environment /x __

- we would expect, if frequency really is driving this:
  - the frequency with which different words are used increases gradually
  - the proportion of deletion fundamentally should follow this gradual increase

Word	CELEX frequency	% Deletion	Average % for 0, 1–100, 101–1000, over 1000
Phonetic environment [st]			
dorst	0	10	10.00
vriest	22	15	
barst	66	3	
wast	71	14	10.67
blaast	104	16	
danst	105	9	
kiest	400	14	
leest	555	18	14.25
wist	19986	34	
moest	31941	42	38.00
Phonetic environment [xt]			
spuugt	24	8	
zucht	27	11	
vecht	63	11	10.00
jaagt	101	15	
weegt	144	16	
buigt	214	17	
zwijgt	235	12	
vliegt	243	16	
vocht	250	13	
dreigt	330	12	
lacht	678	13	
kocht	981	19	
legt	987	19	
draagt	991	11	14.82
vraagt	2840	16	
zocht	2955	24	
krijgt	3614	30	
ligt	5693	18	
bracht	7061	32	
mocht	7089	56	
zegt	9502	27	29.00
dacht	19358	29	29.00

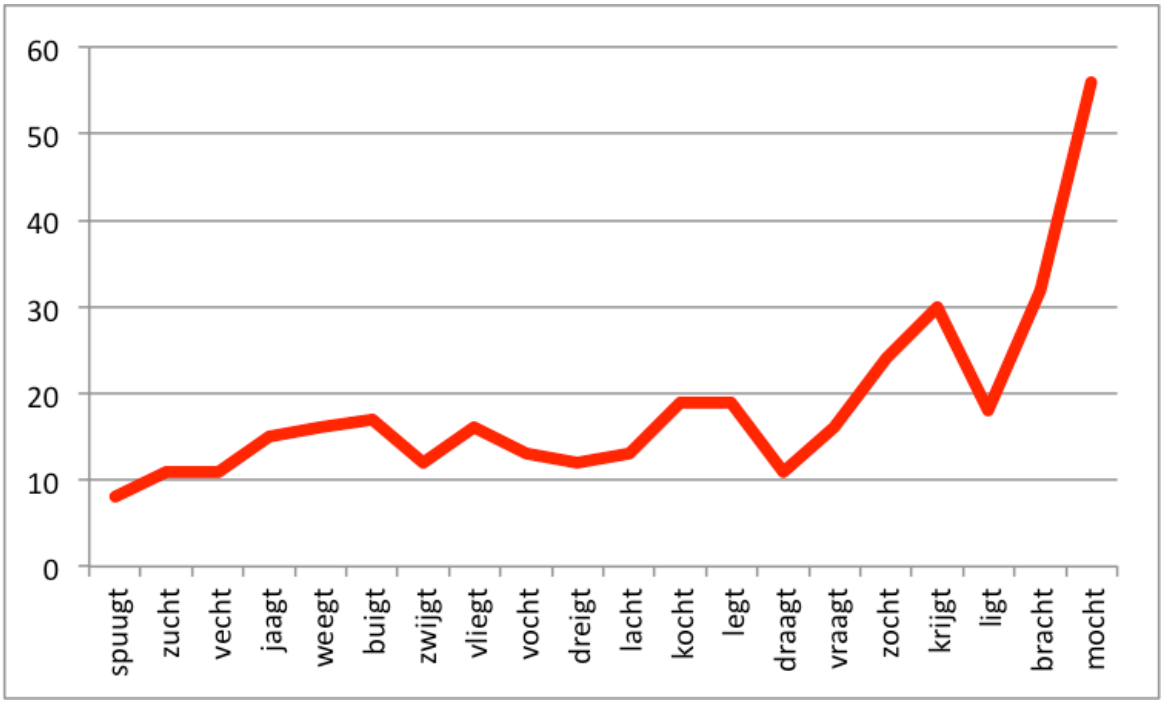
CELEX frequency counts for first 20 of the words that in Phillips' (2006) list are as follows:

- frequency increases gradually



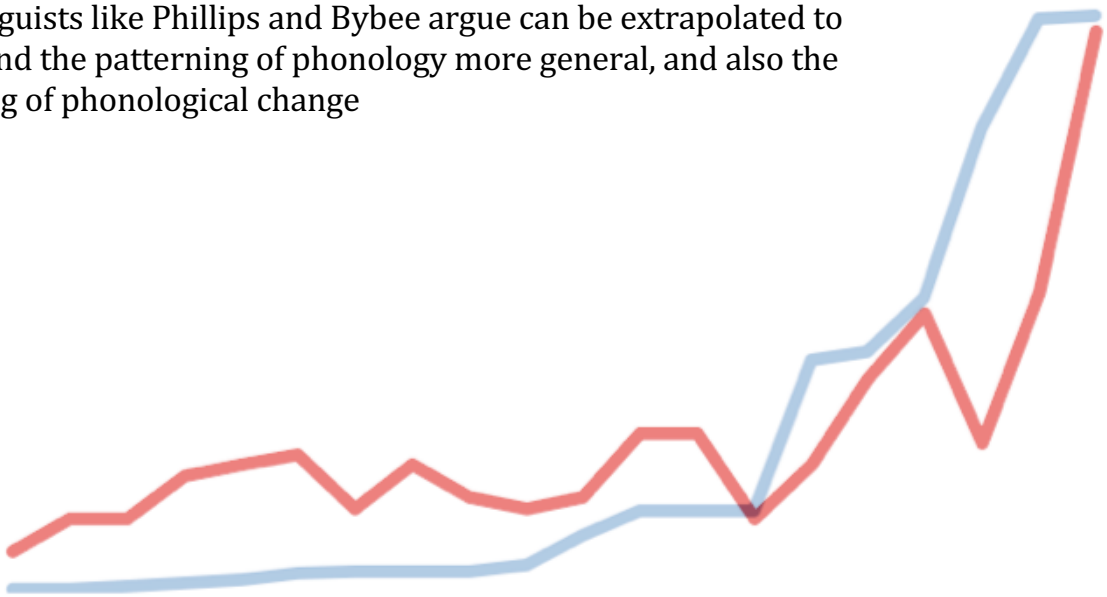
Percentage deletion of /t/ in those same words:

- deletion increases gradually



It seems that there is a fair correlation between the frequency with which words are used by speakers and how susceptible coronal stops are to deletion

- it seems that something which is specific to individual lexical items – their frequency of occurrence – influences the extent to which (or perhaps even *whether*) they are involved in a change
- this can be seen as evidence for a frequency effect in contemporary variation, which linguists like Phillips and Bybee argue can be extrapolated to understand the patterning of phonology more general, and also the patterning of phonological change



Syncope in English

Bybee/Hooper has argued many times that the behaviour of syncope in English is also constrained by frequency

- in this syncope, [ə] is deleted in certain prosodic and melodic environments

Hooper (1978) says:

> The processes to be discussed are in a variable state. A few words seem to have lost their schwas entirely, e.g. *every, camera, family, general, chocolate* (Zwicky 1972:283); some words can be pronounced with or without schwas, e.g. *elaborate, happening, leveling*; while still others seem to resist schwa-deletion, e.g. *infirmary, mockery, perjury*. There is a great deal of variation among individual speakers

As Kiparsky (2016) summarises, the claim is that frequency influences the extent to which a word undergoes this process, which is "more advanced in words of higher frequency (such as those just named) than in words of lower frequency" (Bybee 2001, 11)

High frequency word:	*every*	[∅]
Mid frequency word:	*memory*	[∅ ~ ə]
Low frequency word:	*mammary*	[ə]

Bybee (2000) sets out some precise figures:

TABLE 9.1. Words Undergoing Reduction at Differential Rates due to Word Frequency

No Schwa	*Syllabic* [*r*]	*Schwa* + [*r*]
every (492)	memory (91)	mammary (0)
	salary (51)	artillery (11)
	summary (21)	summery (0)
	nursery.(14)	cursory (4)
evening (149) (noun)		evening (0) (verb + *ing*)

Frequency figures from Francis and Kučera 1982.

"*time* and *thyme* are not homophones"

Another example of a relevant phenomenon has been claimed by Gahl (2008)

- this does not focus on segmental phenomena, but on the pronunciation of whole words

The measurements involved consider the duration of chunks of speech

- such durations are massively variable
- Maslowski (2015) shows some of this in terms of variation in the pronunciation of the phrase 'I see' in an elicitation task:

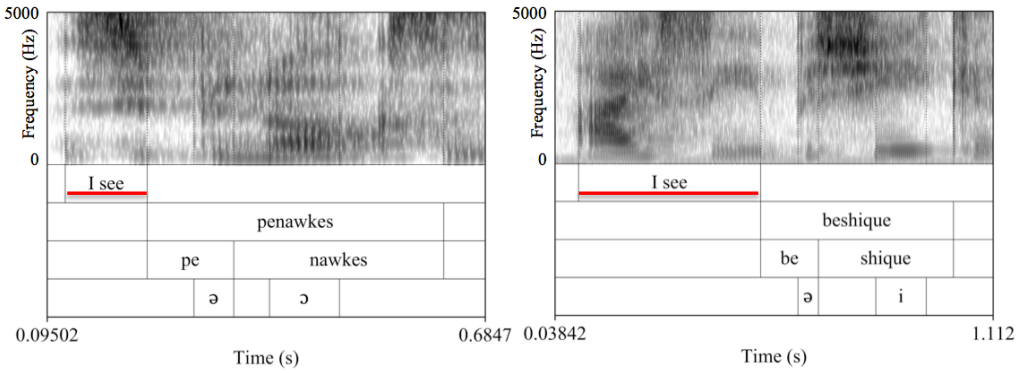


Figure 6: Spectrograms of the two productions of I see *with the most extreme durations in test sentences spoken by two different participants. On the left, the shortest production of* I see *(110 msec.) is visible; on the right, the longest production of* I see *is shown (446 msec.).*

This strand of work relevant here argues that there are principles that explain parts of this variation

- more frequent words are reduced more than less frequent words
  - the shortening of frequent words is typically described as reduction

As Gahl (2008) points out, this should mean that words which are typically transcribed as 'the same' will be pronounced differently

- *time* [tʰaɪm] – high frequency = more likely to reduce
- *thyme* [tʰaɪm] – low frequency = less likely to reduce
- *for* [fɔː] – high frequency = more likely to reduce
- *four* [fɔː] – low frequency = less likely to reduce

Gahl (2008) controls for a range of factors in a corpus-based study and argues that this is, indeed, the case:

VARIABLE	B	β	SE	*t*	VIF
intercept	−0.5247		0.103497	−5.07	
low-fq duration[b]	0.2141	0.2823	0.039524	5.416	1.1004
m-score[c]	−0.2213	−0.1565	0.073207	−3.023	1.0847
noun proportion	0.1034	0.2178	0.024098	4.292	1.0427
speaking rate[f]	−0.0492	−0.1386	0.020312	−2.422	1.3258
bigram probability[h]	−0.0171	−0.1826	0.005315	−3.21	1.3104
pauses[g]	0.2813	0.1187	0.136587	2.06	1.3447
log frequency[h]	−0.0297	−0.2471	0.00669	−4.433	1.2581

TABLE 3. Summary of regression model of durations of high-frequency homophones (N = 220); B = raw unstandardized coefficient, β = standardized coefficient, SE = standard error, t = t value, VIF = variance inflation factor.

Crucially for the current study, the log frequency of a word was a significant predictor of word duration when all other factors were controlled for: as frequency increases, word duration decreases, when other factors are held constant.

English preterites

Practically all verbs in English form their past tense in a phonologically simple way

I pay	*I paid*	*I rub*	*I rubbed*	*I pick*	*I picked*
[peɪ]	[peɪd]	[rʌb]	[rʌbd]	[pɪk]	[pɪkt]
I fill	*I filled*	*I ease*	*I eased*	*I heap*	*I heaped*
[fɪl]	[fɪld]	[iːz]	[iːzd]	[hiːp]	[hiːpt]
I slam	*I slammed*	*I heave*	*I heaved*	*I miss*	*I missed*
[slam]	[slamd]	[hiːv]	[hiːvd]	[mɪs]	[mɪst]

As is well-known, however, some forms show an extra vowel:

- the precise nature of the vowel varies from accent to accent:

I heat	*I heated*	*I heed*	*I heeded*
[hiːt]	[hiːtɪd]	[hiːd]	[hiːdɪd]

On this basis, the UR of the past-tense morpheme is typically assumed to end in /d/.

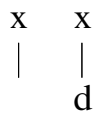
Regular preterite formation can be understood as the interaction of two rules

$$\begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \\ +\text{consonantal} \\ +\text{anterior} \\ +\text{coronal} \\ +\text{voice} \end{bmatrix} \rightarrow [\alpha\text{voice}] \;/\; [\alpha\text{voice}]\_\#$$

$$\emptyset \rightarrow \text{ɪ} \;/\; \begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \\ +\text{consonantal} \\ \alpha\text{anterior} \\ \beta\text{coronal} \end{bmatrix} \_\_ \begin{bmatrix} -\text{sonorant} \\ -\text{continuant} \\ +\text{consonantal} \\ \alpha\text{anterior} \\ \beta\text{coronal} \end{bmatrix} \#$$

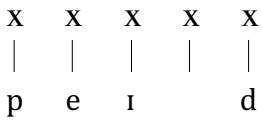
	heaved	*heaped*	*heated*
UR	/hiːv+d/	/hiːp+d/	/hiːt+d/
Epenthesis	—	—	hitɪd
Assimilation	—	hipt	—
SR	[hiːvd]	[hiːpt]	[hiːtɪd]

A less derivational, representational solution along fundamentally the same lines is given in Gussmann (2002), which assumes that the past tense morpheme is:

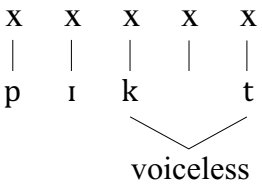


And implies a derivation along these lines:

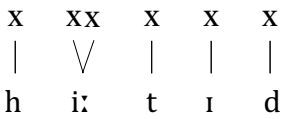
a.



b.



c.



However, several verbs form their preterite in an 'irregular' way:

I drive	*I drove*	aɪ	oː
I write	*I wrote*		
I shoot	*I shot*	uː	ɒ
I choose	*I chose*	uː	oː
I know	*I knew*	oː	ɪʊ
I grow	*I grew*		

Such forms have been derived by rules, but are typically now seen to involve more than one UR for the morpheme ('suppletion')

'use letters to record language'

VERB

/raɪt/
/roːt/$_{past}$

'Elsewhere' ordering and blocking will account for all this:

	heaved /hiːv+PAST/	*heaped* /hiːp+PAST/	*heated* /hiːt+PAST/	*wrote* /raɪt+PAST/
specific PAST	—	—	—	/roːt/
regular PAST	/hiːv+d/	/hiːp+d/	/hiːt+d/	—
UR	/hiːv+d/	/hiːp+d/	/hiːt+d/	/roːt/
Epenthesis	—	—	hitɪd	—
Assimilation	—	hipt	—	—
SR	[hiːvd]	[hiːpt]	[hiːtɪd]	[roːt]

If we go back to Old English, the situation is different.

- there were several 'classes' of strong verbs, which followed the same ablaut patterns

'Class I'

I drive	*I drove*	ModE
ic drīfe	*ic drāf*	OE
I write	*I wrote*	ModE
ic wrīte	*ic wrāt*	OE
I bide	*I bided*	ModE
ic bīde	*ic bād*	OE
I sneak	*I sneaked*	ModE
ic snīce	*ic snāc*	OE

‘Class II’

I shoot	*I shot*	ModE
ic scēote	*ic scēat*	OE
I choose	*I chose*	ModE
ic cēose	*ic cēas*	OE
I shove	*I shoved*	ModE
ic scūfe	*ic scēaf*	OE
I float	*I floated*	ModE
ic flēote	*ic flēat*	OE

‘class VII’

I know	*I knew*	ModE
ic cnāwe	*ic cnēow*	OE
I grow	*I grew*	ModE
ic grōwe	*ic grēow*	OE
I sow	*I sowed*	ModE
ic sāwe	*ic sēow*	OE
I flow	*I flowed*	ModE
ic flōwe	*ic flēow*	OE

Hooper/Bybee (1976, 2001) has often explained that the regularisation of strong preterite forms affects infrequent verbs before frequent verbs – the numbers are frequency counts

Strong Verbs			*Strong Verbs That Have Become Weak*	
Class I				
*drive	208		bide	1
*rise	280		reap	5
*ride	150		*slit	8
write	599		*sneak	11
*bite	128	Partially leveled		
		*shine 35		
Average frequency	273.00		Average frequency	6.25
Class II				
choose	177		rue	6
*fly	119		seethe	0
*shoot	187		*smoke	59
lose	274		*float	23
flee	40		shove	16
Average frequency	159.40		Average frequency	32.50
Class VII				
*fall	338		*wax	19
*hold	498		weep	31
know	1227		*beat	96
grow	257		hew	1
blow	81		*leap	42
			mow	1
Average frequency	473.80		sow	3
			*flow	95
			*row	53
			Average frequency	37.89

Hooper (1976) continues...

A problem with the results displayed in table 2.3 is that the frequency count used was based on Modern English, but the analogical leveling took place sometime during the last ten centuries. However, since the results show such a striking difference in frequency between leveled and nonleveled forms, I do not think a more accurate frequency count would alter the general picture. A way to avoid this problem would be to study modern leveling. One case I have investigated involves the six verbs *creep*, *keep*, *leap*, *leave*, *sleep*, and *weep*, all of which have a past form with a lax vowel (due to the Middle English laxing mentioned earlier). Of these verbs, three, *creep*, *leap*, and *weep*, all may have, at least marginally, a past forms with a tense vowel, *creeped*, *leaped*, and *weeped*. The other three verbs are in no way threatened by leveling; past forms **keeped*, **leaved*, **sleeped* are clearly out of the question. Now consider the frequency differences among these verbs, in table 2.4. Again the hypothesis that less frequent forms are leveled first is supported.

This table is adapted from Coetzee (2007), including some of the figures Bybee is referring to:

Less likely to regularize		*More likely to regularize*	
Present	*Raw frequency*	*Present*	*Raw frequency*
keep	348	*creep*	19
leave	345	*leap*	20
sleep	106	*weep*	22
drive	174	*dive*	32

Diatonic Stress Shift

Chen & Wang (1975) and Phillips (2006) consider a phonological change that they describe as the emergence of 'diatonic pairs' in English

- this is also known as Diatonic Stress Shift
- 'diatones' are noun-verb pairs which contrast in their stress pattern, such as:
  cónvict$_N$ ~ *convíct*$_V$
  récord$_N$ ~ *recórd*$_V$
  éxport$_N$ ~ *expórt*$_V$
- 'monotones' are noun-verb pairs which don't vary in their stress pattern, such as:
  contról$_N$ ~ *contról*$_V$

The number of diatonic pairs has gradually increased over several centuries

- the change involves in the creation of diatones from monotones (Diatonic Stress Shift)
  - in monotonic pairs, both have σσ́
  - in DSS, σσ́$_V$ *stays as* σσ́$_V$, but σσ́$_N$ > σ́σ$_N$
  - previously both forms of the following had final stress: *prefix*, *discount*, *export*, *contract*
  - they are now diatonic, but many similar forms are not: *assault*, *dislike*, *exchange*, *control*

Based on Sherman (1973), Chen & Wang (1975) plot the course of Diatonic Stress Shift in the history of English

- in 1570, there were only three diatonic pairs – all other N~V pairs were monotones
  - *récord*$_{N}$ ~ *recórd*$_{V}$
  - *rébel*$_{N}$ ~ *rebél*$_{V}$

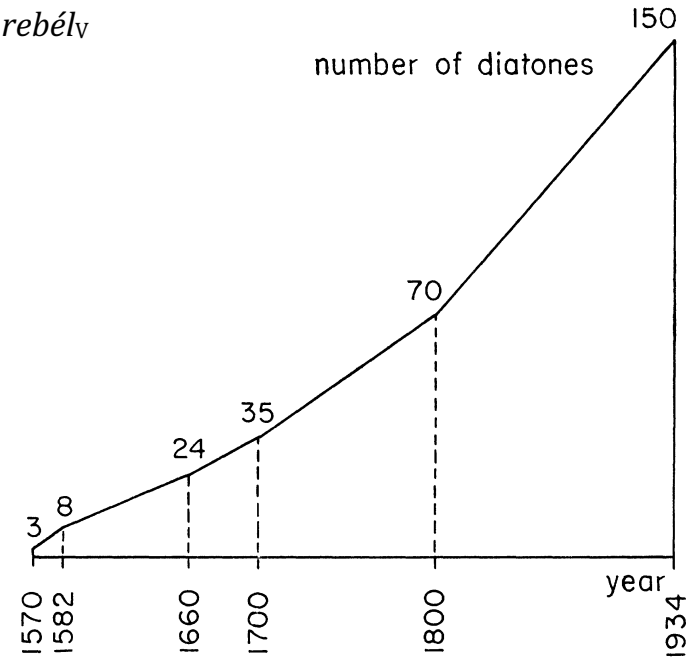


FIGURE 2. Increase in number of diatonic N–V homographs as a function of time (based on Sherman 1973). Only disyllabic pairs are counted.

The assumption is that DSS is a change affecting English over a long period, but not all eligible words are affected by a change at the same rate

- the spreading through the lexicon takes time
  - and, crucially for our purposes, the "words which have undergone the Diatonic Stress Shift have lower frequency than those which have not" (Sonderegger 2010)

The observant among you will have noticed that...

- there are **different kinds of frequency effects**

Two types of frequency effect are often recognised (Bybee 2001, Phillips 2006)

frequency effects

'high frequency effects'

- the *most* frequent words engage most in some phonological behaviour
- a *promoting* effect of frequency

'low frequency effects'

- the *least* frequent words engage most in some phonological behaviour
- a *conserving* effect of frequency

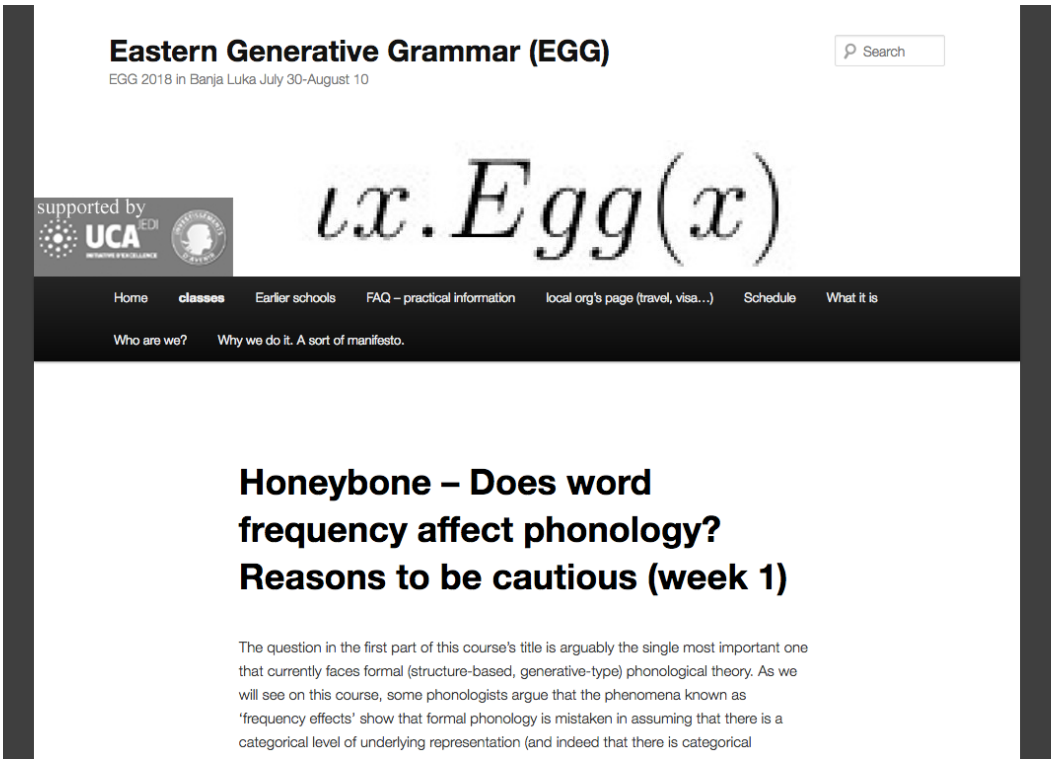
On the basis of the phenomena that we have seen, we can also differentiate between:

- whole word effects = 'tiny-word-based effects'
- segmental-category type effects

Kiparsky (2016) distinguishes between "an imperceptible phonetic effect of a few milliseconds, or neutralization to a categorically distinct pronunciation"

Why should we care about all that?

Someone does...



Is it just me?

Bybee (2007, 5)

> A newcomer to the field of linguistics might be surprised to learn that for most of the twentieth century facts about the frequency of use of particular words, phrases, or constructions were considered irrelevant to the study of linguistic structure.

Gahl (2008, 491)

> I agree with the observation that 'parsimony cannot be assumed to be a property of the language system; it is only something to which accounts of its underlying principles aspire' (O'Seaghdha 1999:51). The underlying principle of recognizing that frequency may shape every aspect of language and speech is simple.