University of Pennsylvania Working Papers in Linguistics

Volume 4
Issue 1 *A selection of papers from NWAVE 25*

Article 14

1-1-1997

# Frequency effects in Variable Lexical Phonology

James Meyers

Gregory R. Guy

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/pwpl/vol4/iss1/14
For more information, please contact libraryrepository@pobox.upenn.edu.

# Frequency effects in Variable Lexical Phonology

This working paper is available in University of Pennsylvania Working Papers in Linguistics: http://repository.upenn.edu/pwpl/vol4/iss1/14

# Frequency Effects in Variable Lexical Phonology

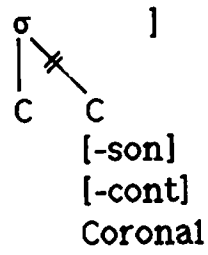
James Myers and Gregory R. Guy

## 1. Variable Lexical Phonology

The variable version of Lexical Phonology developed in Guy 1991, 1992 proposes that variable phonological processes like English Coronal Stop Deletion can apply both postlexically and lexically. Coronal Stop Deletion (CSD), which variably deletes clustered final /t/ or /d/ as represented in Figure 1, is well-known to have different rates of application in various morphological classes of English words. Variable Lexical Phonology explains these differences in terms of the contrasting derivational histories of the classes. Thus the high deletion rate in monomorphemic words like *lift*, where the final stop is underlying, is due to multiple exposures to the deletion rule, both within the lexicon and postlexically. Regular past tense forms like *laughed* only acquire the final cluster targeted by CSD through affixation at the end of the lexicon. Since they are therefore only subject to a postlexical application of the rule, they have low deletion rates.

Figure 1: English Coronal Stop Deletion

<variable, unmarked domain of application>




This model has significant implications for several areas of linguistic theory, and thus should be subject to stringent empirical tests. One of the most important consequences of this model





is that it predicts an exponential relation among the deletion rates in the various derivational classes; this prediction has been confirmed in several studies (e.g. Guy 1991, 1992, Santa Ana 1991), and some psycholinguistic implications have been tested in Myers (1996). In the present paper, we explore a further set of important predictions involving lexical frequency.

As will shortly become clear, Variable Lexical Phonology predicts that frequency should affect the rate of CSD in the class of Monomorphemic forms but not in the class of Regular past tense forms. Moreover, the model predicts that deletion rates should continue to be strongly affected by morphological class even when frequency is controlled. After we have discussed our data bearing on these predictions, we briefly consider an analysis of CSD in another dialect of English and show how the results found there complement the results of our own study. We will conclude that the results not only provide novel support for the Variable Lexical Phonology model, but also have interesting consequences for psycholinguistic models of morphological processing and for theoretical phonology.

## 2. Frequency Effects

We begin by explaining the basis of our predictions concerning frequency. There is substantial evidence from a number of sources that information about a word's rate of occurrence — its frequency — forms part of a speaker's knowledge of that word. So-called 'frequency effects' are in fact among the best-attested findings in the study of lexical access and retrieval. For example, frequency has been found to be a crucial factor affecting the speed with which words are produced or recognized (classic works include Forster and Chambers 1973, Whaley 1978). Because frequency information is unpredictable, it must be indicated in the lexicon. This means that frequency effects can be used as a diagnostic of lexicality: the existence of frequency effects in the behavior some class of linguistic constituents is an indication that those constituents themselves are stored in the lexicon.

One debate in which frequency effects have played an important role concerns the mental representation and processing of inflection. According to the view taken by Steven Pinker and

others, regularly inflected forms, including the regular past tense forms that will be discussed in this paper, are not stored as wholes in the lexicon, but rather are derived from the stems by a regular rule. Thus the regularly inflected word *laughed* is not found in the lexicon; only the stem *laugh* is (as assumed in Variable Lexical Phonology). By contrast, monomorphemic forms like *lift* and irregularly inflected forms like *found* are indeed stored as wholes.

On the other side of the debate are researchers such as Joan Bybee (e.g. Bybee 1995) and others who have maintained that even regularly inflected forms are stored as wholes in the lexicon. There is no past tense 'rule' as such; instead, novel inflected forms, as in *Clinton out-Republicaned the Republicans*, are formed by analogy to stored past tense forms.

If frequency effects can be used as a diagnostic of lexicality, these two views make distinct predictions. The claim that regular forms are not stored in the lexicon predicts that only monomorphemic and irregularly inflected forms will show frequency effects. The contrary claim, that regular forms are stored in the lexicon, predicts that they, too, will show frequency effects.

These predictions have been tested repeatedly in the psycholinguistic literature. In one typical experiment reported in Pinker (1991), subjects were shown verb stems on a computer screen and were asked to utter the past tense form as quickly as possible. With irregular verbs, subjects were faster to read high frequency past tense forms than low frequency past tense forms (stem frequencies were of course controlled). However, no frequency effect on the speed of response was found for regular past tense forms. Pinker and colleagues therefore concluded that subjects were deriving these forms on-line, and not retrieving them directly from the mental lexicon, where frequency effects reside.

Two properties of the variable rule of Coronal Stop Deletion suggest that it too can be exploited to address this debate about the processing of inflection. First of course, English happens to indicate regular past tense inflection with the segments, /t/ and /d/, that are subject to this rule. Second, it is known that phonetically-motivated processes, which CSD appears to be, are influenced by lexical frequency. For example, Phillips (1984) found that phonetically-motivated sound changes diffuse through the lexicon from more frequent to less frequent words. Similarly, Fidelholtz (1975)





found that the phonetically-motivated lexical rule of English vowel reduction applies more readily in higher frequency words like *mistake* than in lower frequency words like *mistook*.

Such frequency effects on variable phonology are essentially the variable analog of the 'lexical exceptions' familiar with invariant lexical rules (see for example Kiparsky 1982). 'Variable exceptionality,' as it might be called, leads to lexically-specific differences in rates of application. In particular, variable lexical rules affect higher frequency words at a higher rate than lower frequency words.

If the Variable Lexical Phonology model is correct, the frequency effect on CSD should therefore depend on the morphological status of the word-final /t/ and /d/. Specifically, we expect that Monomorphemic forms, being stored in the lexicon, will show a robust frequency effect, with higher frequency words like *past* showing a higher rate of deletion than lower frequency words like *priest*. By contrast, Regular past tense forms, being derived and not stored, should show no frequency effect at all: higher frequency words like *passed* and lower frequency words like *kissed* should show equal rates of deletion.

## 3. Methods

These predictions were tested on recordings of the conversational speech of two working-class informants in Philadelphia, one male and one female (approximately 75% of the tokens came from the female speaker). Tokens of words ending in /t/- or /d/-final clusters were coded as deleted if trained listeners could not hear any evidence of the stop; they were coded as retained if the stop had any audible reflex, including a glottal stop or an affricate derived from a stop-glide sequence. Tokens were also coded for phonetic environments: pre-consonantal, pre-vocalic, or pre-pausal. Finally, tokens were coded for morphological class: Regular past; Monomorphemic, which included strong past tense forms like *found*; and Semiweak past. The Semiweak class consisted of those irregular past tense forms that involve a suffix, such as *left* (past tense of *leave*; the adjective *left* was included in the Monomorphemic class).

As is standard in studies of CSD, certain words with very high frequencies that are known to have inordinately high deletion

rates were removed from the data set. These removed words were *and* and all words with the contraction *-n't* (following the practice of Guy 1991, 1992), as well as the words *just* and *went* (following Bybee 1996). In addition, all instances of the words *used* and *supposed* were removed, as these virtually always appeared in the lexicalized phrases *used to* and *supposed to*.

Because we were using the standard frequency counts of Kucera and Francis (1967), certain other tokens had to be removed as well. These included 12 instances of local proper names, such as *Lakehurst*, which have a frequency of 0 in the standard reference but were clearly of higher frequency in Philadelphia; nonlocal names, such as *Maryland*, were not removed. Also removed were all 17 compounds, such as *boyfriend*; the frequency of *boyfriend* is much lower than that of *friend*, and it was not clear which should be used in our analysis. The data set that remained after these adjustments contained a total of 1080 tokens. The class of Semi-weak forms was unfortunately too small to examine the effect of frequency (40 tokens of 5 types) and will not be discussed further.

Word frequency in Kucera and Francis (1967) is given as an integer representing the number of instances of that word in a corpus of one million words. Their original corpus was compiled from a variety of written material, including newspapers and novels, and although it may therefore not be ideal for the study of spoken language, it remains the largest and most widely used such corpus available. A computerized version of this corpus in the laboratory of Paul Luce at the State University of New York at Buffalo was used to determine lexical frequencies for all the words in our data set, ranging from 0 for *cheapest* and *bussed* to 1360 for *first* and 401 for *called*.

A cut-off point of 35 was used to classify tokens by frequency: tokens with a frequency equal to or below 35 were classified as low frequency and tokens with a frequency above 35 were classified as high frequency. This cut-off point was chosen to follow the procedure of Bybee (1996), who, as we will see, argues that regular forms are not derived on-line. Bybee motivates the choice herself by the fact that a frequency of 35 divides the set of past tense forms in the Kucera and Francis frequency list exactly in half. In Bybee's data set as well as ours, this criterion puts approx-





imately 20% of the tokens into the low frequency class and 80% into the high frequency class.

## 4. Results

The basic data are shown in Table 1.

Table 1: Variable Coronal Stop Deletion (Philadelphia)

| Monomorphemic* | Total | Deletions | Deletion % |
|---|---|---|---|
| Low frequency | 151 | 28 | 18.5 |
| High frequency | 573 | 194 | 33.9 |

| Regular** | Total | Deletions | Deletion % |
|---|---|---|---|
| Low frequency | 96 | 7 | 7.3 |
| High frequency | 220 | 18 | 8.2 |

*$\chi^2(1) = 13.182$, $p < .01$
**$\chi^2(1) = .073$, $p > .1$

A chi-square on the Monomorphemic class finds a significant effect of frequency on deletion rates, while a chi-square on the Regular class finds no such effect. An ANOVA finds significant effects for both morphology and frequency. The interaction between frequency and morphology is significant as well, which further supports the conclusion that frequency affects the Monomorphemic and Regular classes differently.

The fact that both morphology and frequency affect CSD independently is worth emphasizing. This is because an alternative explanation of the higher rates of deletion that have been found in Monomorphemic forms is that this is merely a frequency effect, since Monomorphemic forms tend to be of higher frequency than Regular past tense forms. For example, a chi-square on the above totals finds that the Monomorphemic class has a significantly higher proportion of high frequency tokens than the Regular class.

This frequency confound can be reduced by removing tokens in the Monomorphemic class that have frequencies above the highest frequency found in the Regular class. Doing this to our data set yields the results in the Table 2. The highest frequency in this frequency-capped Monomorphemic class is 399, very close to the highest frequency of 401 found in the Regular class.

Table 2: Frequency-balanced data sets

Monomorphemic* (max frequency = 399)

| | Total | Deletions | Deletion % |
|---|---|---|---|
| Low frequency | 151 | 28 | 18.5 |
| High frequency | 332 | 98 | 29.5 |

Regular (repeated from last table; max frequency = 401)

| | Total | Deletions | Deletion % |
|---|---|---|---|
| Low frequency | 96 | 7 | 7.3 |
| High frequency | 220 | 18 | 8.2 |

*$\chi^2(1) = 6.484$, $p < .025$

A chi-square test now finds no difference in low and high frequency ratios between the Regular class and the frequency-capped Monomorphemic class. An ANOVA still finds an overall effect of frequency on deletion, but only marginal significance (p=.0469). By contrast, the effect of morphology alone on deletion rates remains highly significant (p<.0001). Even more interesting, a chi-square on the frequency-capped Monomorphemic class still shows an effect of frequency, with CSD applying significantly more often in high frequency forms. In other words, even when the overall data set is controlled for frequency, frequency affects deletion rates within the Monomorphemic class but not within the Regular class.

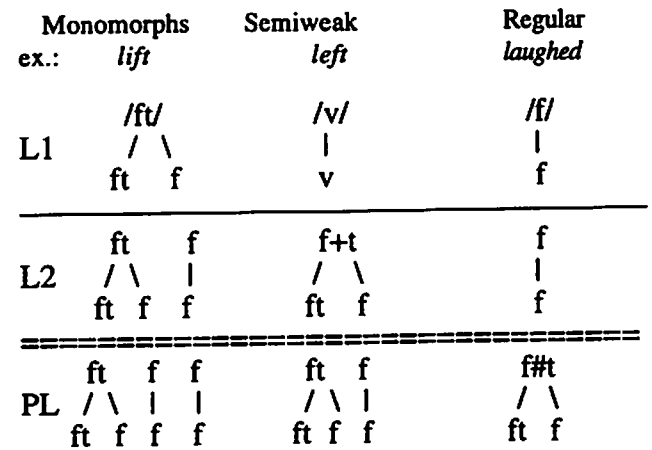
### 4.1. Exponential Effects

It is reasonable to ask an even more challenging question: Is an exponential relation still found in this frequency-controlled data set? Recall that Guy (1991) claimed that in the Variable Lexical Phonology model, Monomorphemic forms, which end in /t/ or /d/ underlyingly, have three chances to undergo variable deletion, twice lexically and once postlexically, while Regular forms have only one chance, namely postlexically. This is illustrated in Figure 2, where there are three pathways to surface deletion for the Monomorphs, two for Semiweaks, and just one for Regular pasts. If the probability that /t/ or /d/ will be retained is the same at each level -- call this p(r) -- and if the process operates independently at each level, we predict that the retention rate in Regular past forms will be p(r), while the retention rate in Monomorphemic forms will be the cube of p(r). This cubed retention rate in the Monomorphemic class will not merely be smaller than that found in the Regular class (because p(r) is less than 1), but smaller by a specific, statistically testable degree.





Figure 2: An exponential model of Coronal Stop Deletion (after Guy 1991, 1992)




In Table 3 we can see that the cube root of the observed retention rate for Monomorphemic forms is extremely close to the observed retention rate for Regular forms. This observation can be given statistical validity by comparing these observed rates with those expected given an estimated value for p(r). The simplest way

to estimate p(r) is to use the surface retention rate for the Regular class, 92.1%. A chi-square test finds no significant difference between the actual surface retention rates for the Monomorphemic and Regular classes and those that are predicted given this p(r) value. In other words, the exponential pattern is found even in the frequency-controlled data set, and therefore this pattern cannot be due to a frequency effect alone.

Table 3: Test of exponential hypothesis with frequency-balanced data sets

| | Total | Retentions | Ret.% | Est. pr |
|---|---|---|---|---|
| Mono | 438 | 357 | 81.5 | 93.4* |
| Reg | 316 | 291 | 92.1 | 92.1 |

*cube root of surface rate

## 4.2. CSD in Bybee 1996

The general observation we have reported here, that the Monomorphemic class shows a frequency effect in deletion rates while the Regular class does not, is precisely what is predicted by Variable Lexical Phonology, and supports the hypothesis that regularly inflected forms are NOT stored in the lexicon. However, Joan Bybee (1996), in an examination of Coronal Stop Deletion in the corpus of Los Angeles Chicano English collected by Otto Santa Ana (Santa Ana 1991), reports a frequency effect in Regular past tense forms. Bybee's data for Regular forms are presented in Table 4. A chi-square test does indicate a significant effect of frequency on the deletion rate.

Table 4: Coronal Stop Deletion in Los Angeles Chicano English (analysis by Bybee 1996)

Regular (non-prevocalic tokens only)

| | Total | Deletions | Deletion % |
|---|---|---|---|
| Low frequency | 58 | 11 | 18.9 |
| High frequency | 111 | 44 | 39.6 |





There are two major ways in which the data presented by Bybee differ from ours. First, the deletion rates in the dialect she examined are much higher than in the dialect we examined. Second, she restricted her examination to Regular tokens in non-prevocalic environments, that is, before consonants and pauses. This was done because these environments tend to favor deletion. We have no way to adjust the base deletion rate of the dialect we studied, but we too can boost deletion rates in our data set by following Bybee and including only tokens in non-prevocalic environments. These data are shown here. Again, however, there is no effect of frequency.

Table 5: Coronal Stop Deletion in Philadelphia in restricted phonological environments

Regular (non-prevocalic tokens only)

| | Total | Deletions | Deletion % |
|---|---|---|---|
| Low frequency | 73 | 7 | 9.6 |
| High frequency | 135 | 13 | 9.6 |

The fact that Bybee finds a frequency effect in Regular forms only in a dialect with an extremely high base deletion rate, and then apparently only in environments that boost deletion rates still higher, suggests that at the very least, the effect of frequency on Regular forms is not very strong. But does Bybee's finding threaten the claim made by us, Pinker and others that regularly inflected forms are not stored in the lexicon? It does, but only if one attempts to maintain the extreme position that Regular forms are *always* derived on-line. Such a position is untenable for independent reasons, however. Among other things, regularly inflected forms can come to take on unpredictable and therefore lexicalized aspects over time, which would be impossible if regular forms were never stored in memory. For example, speakers must remember that the regularly inflected plural form *glasses* describes a singular object. Similarly, the regular past tense forms in *used to* and *supposed to* now display irregular phonology. There is even evidence that an important factor in the lexicalization of regularly

inflected forms is lexical frequency. For instance, Stemberger and MacWhinney (1988) found that in both naturally occurring and experimentally-induced speech errors, inflections on regular forms are less likely to be shifted or exchanged if the forms are of high frequency. Regardless of their interest, however, such results do not negate our assumption that the on-line generation of regularly inflected forms is the default case.

## 5. Theoretical Implications

Thus far we have focussed primarily on the implications of our findings for the Variable Lexical Phonology model and for models of language production, but there are general implications for phonological theory that should be addressed as well. The theory of Lexical Phonology, upon which Variable Lexical Phonology is built, has lost considerable favor in the phonological climate of the mid-1990s, partly because its rule-driven formalism of level-ordering is incompatible with the currently fashionable paradigm of Optimality Theory (Prince and Smolensky 1993). As Kiparsky (1993) has shown, the exponential effect in CSD discovered by Guy (1991) can be modelled in Optimality Theory if one makes two fundamental assumptions. First, the presence or absence of /t/ and /d/ in different morphological classes is determined by independent well-formedness constraints, rather than by a single rule operating at different levels. Second, the ranking of these constraints is chosen randomly whenever a /t/-final or /d/-final form is uttered. It is easy to demonstrate, which we will not do here, that this scheme can be made to give rise to the exponential effect without the use of rules or level ordering.

However, one thing that this analysis cannot describe is the set of striking differences between the lexical and postlexical applications of Coronal Stop Deletion. Guy (1992) and Myers (1996) discuss some such differences, and the present paper reveals another: lexical applications are sensitive to frequency, while postlexical applications are not. While frequency effects on lexical rule application are easy to conceptualize within the framework of Lexical Phonology as a form of 'variable exceptionality,' as noted earlier, it is yet unclear how Optimality Theory can capture the lexical versus postlexical distinction without stipulation.





Turning back to our own research, a crucial question remains unanswered. While the present project has produced results that are quite consistent with work by Pinker and his colleagues, the exponential effect which inspired it is not. Pinker expects only two morphologically relevant classes: Monomorphemic forms, which are stored, and Regular forms, which are derived. However, Guy and Boyd (1990) and Guy (1991) were able to show that the Semiweak past tense forms behave as a distinct third class in their effect on Coronal Stop Deletion. Bybee (1996) suggests that the high rate of deletion in this class is due solely to high frequency, but this seems unlikely. The mean frequencies for the Monomorphemic and Semiweak classes in our data are virtually identical (360 versus 338), suggesting that if these classes behave distinctly, it is apparently not because of frequency. A much larger corpus of natural speech, one that includes a large number of Semiweak forms, both types and tokens, would be needed to determined how Semiweak forms are processed in speech production: by rule, analogy, or some combination of these.

## References

Bybee, J. (1995) "Regular morphology and the lexicon." *Language and Cognitive Processes* 10:425-455.

Bybee, J. (1996) "The phonology of the lexicon: Evidence from lexical diffusion." In M. Barlow and S. Kemmer (eds.) *Usage-Based Models of Language*. CSLI, Stanford University Press.

Fidelholtz, J. L. (1975) "Word frequency and vowel reduction in English," *CLS* 11:200-213.

Forster, K. I. and Chambers, S. M. (1973) "Lexical access and naming time,"*J. of Verbal Learning and Verbal Behavior* 12:627-35.

Guy, G. R. (1991) "Explanation in variable phonology:An exponential model of morphological constraints," *Language Variation and Change* 3:1-22.

Guy, G.R. (1992) "Contextual conditioning in variable lexical phonology," *Language Variation and Change* 3:223-239.

Guy, G. R. and Boyd, S. (1990) "The development of a morphological class,"*Language Variation and Change* 2:1-18.

Kiparsky, P. (1982) "Lexical phonology and morphology," in I. S. Yang (ed.) *Linguistics in the Morning Calm*, Linguistic Society of Korea, 135-160.

Kiparsky, P. (1993) "Variable rules." Paper presented at Rutgers Optimality Workshop, Rutgers University.

Kucera, H. and Francis, W. N. (1967) *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Myers, James (1996) "The categorical and gradient phonology of variable t-deletion in English." To appear in the *Proceedings of the International Workshop on Language Variation and Linguistic Theory*, University of Nijmegen, Netherlands.

Phillips, B. S. (1984) "Word frequency and the actuation of sound change." *Language* 60:320-342.

Pinker, S. (1991) "Rules of language," *Science* 253:530-5.

Prince, A. and Smolensky, P. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. Rutgers University and University of Colorado ms.

Santa Ana, O. (1991) *Phonetic Simplification Processes in English of the Barrio: A Cross-Generational Sociolinguistic Study of the Chicanos of Los Angeles*. University of Pennsylvania PhD thesis.

Stemberger, J. P. and MacWhinney, B. (1988) "Are inflected forms stored in the lexicon?" In M. Hammond and M. Noonan (eds.) *Theoretical Morphology*, San Diego: Academic Press, 101-116.

Whaley, C. P. (1978) "Word non-word classification time," *J. of Verbal Learning and Verbal Behavior* 17:143-154.

James Myers
Graduate Institute of Linguistics
National Chung Cheng University
Min-Hsiung, Chia-Yi 621
TAIWAN, ROC
*lngmeyers@ccunix.ccu.edu.tw*

Gregory R. Guy
York University
D. D. L. L., Ross S561
North York, Ontario M3J 1P3
CANADA
*guy@yorku.ca*



The *University of Pennsylvania Working Papers in Linguistics (PWPL)* is an occasional series produced by the Penn Linguistics Club, the graduate student organization of the Linguistics Department of the University of Pennsylvania.

Publication in this volume does not preclude submission of papers elsewhere; all copyright is retained by the authors of the individual papers.

Volumes of the Working Papers are available for $12, prepaid. Please see our web page for additional information.

**The PWPL Series Editors**

Alexis Dimitriadis
Laura Siegel
Clarissa Surek-Clark
Alexander Williams

**Editors for this Volume**

Charles Boberg
Miriam Meyerhoff
Stephanie Strassel
*and the PWPL series editors*

**How to reach the PWPL**

U. Penn Working Papers in Linguistics
Department of Linguistics
619 Williams Hall
University of Pennsylvania
Philadelphia, PA 19104–6305

*working-papers@ling.upenn.edu*
*http://www.ling.upenn.edu/papers/pwpl.html*