The categorical and gradient phonology of variable t-deletion in English

James Myers
National Chung Cheng University

[Presented September 1995 at the International Workshop on Language Variation and Linguistic Theory, University of Nijmegen, Netherlands, and revised in early 1996. Repaginated and converted to PDF in August 2007, but otherwise unchanged. Read at your own risk.]

0. Introduction.

The purpose of this paper is to unpack some fundamental concepts that have been appearing more often in the literature recently, and show, through the use of a specific example, what sort of methodology may be appropriate to use in addressing these concepts empirically. The concepts are categoricality vs. gradience and variability vs. invariance; the example is -t,d deletion in American English (henceforth referred to as t-deletion); and the methodology advocated here involves the statistical analysis of the physical output of speech production. The central lesson of this paper, then, is that researchers interested in the above concepts, whether theoretical phonologists, sociolinguists, or psycholinguists, should prepare themselves to deal with the analysis of physical data.

1. Categorical and gradient phonology.

Categorical phonology is phonology that uses only categorical representations. Categorical representations are representations built solely out of discrete units, organized discretely. The discrete units may be segments, phonemes, phones, allophones (e.g. as described with IPA symbols), features, syllables, feet -- anything that comes from a countable set of basic elements. The discrete organization can be accomplished by arranging segments or feature bundles in a row, or by linking elements in a more complex fashion with autosegmental association lines; the key idea here is that all elements are discretely located. Disallowed here would be, say, a /p/ that overlaps in time "somewhat" with a following /a/. (For other discussions of categoricality and related concepts, see e.g. Labov 1994, Zsiga 1993, Sproat and Fujimura 1993; Pierrehumbert 1994 specifically notes that variability and categoricality are not mutually exclusive, a major theme of this paper as well.)

Categorical phonology contrasts with gradient phonology. This is phonology that involves gradient representations, which may be described only by reference to values along continuous physical scales. The gradience may result from the use of gradient organization, as in the partial overlapping example above. Zsiga (1993) provides a good overview of these concepts.

Note that categoricality and gradience, as defined here, properly refer to phonological representations, not to phonological processes. Moreover, because gradience makes reference to physical scales, the only way to determine if a given representation is gradient is to examine physical output. Intuition, and even perception experiments, do not allow for a certain determination of categoricality in production. There are many cases where native speakers and trained phoneticians were certain that a phonological fact involved categories, but where it turned out that the categories were misidentified, or even entirely illusory.

In the phenomenon of near merger, for example, native speakers of a dialect will respond in minimal pair tests as if two sounds were members of the same category and yet acoustic analysis shows that the sounds are consistently distinguished in production (New Yorkers declaring, for example, that "source" and "sauce" are pronounced identically; Labov 1994). Contra Sapir, even native speaker intuitions don't always help in establishing the psychological reality of phonological categories. A similar example is found in Read (1972), who discovered that six-year-old American beginning readers classify words like "track" and "check" as starting with the same sound, as indicated both by judgments and spelling; in adult speakers, and presumably children as well, the sounds, though similar due to the fricativization of /t/ before /r/ in American English, are not in fact identical in speech production. Allophonic categories are similarly problematic. Sproat and Fujimura (1993) demonstrate the gradient nature of allophonic variation in English /l/, showing that instead of the two discrete allophones of "dark /l/" and "light /l/" traditionally assumed in the phonological and phonetic literature, the realization of /l/ depends on complex prosodic factors that may affect the dorsal and apical gestures of /l/ in different ways.

If perception is not useful in testing a phonological phenomenon for gradience, how is gradience properly determined? The answer is a detailed physical analysis of speech production. As an example, consider vowel length in English. In most North American dialects, vowel length has been found to vary gradiently as an effect of the voicing and manner of the following consonant (i.e. vowel length variation is "phonetic"). Typical evidence is illustrated in the following figure, which shows that there appear to be no obvious categories into which the vowel lengths in varying contexts fall. That is, vowel length gradually increases from the _t environment to the _z environment. Describing this pattern with categorical notation would require a distinct length category for each environment, which would seem to miss the point.

(1) Vowel length distribution of long syllable nuclei in American English (Peterson and Lehiste 1960)

context	length (msec)
_t	210
_s	269
_r	296
_d	318
_z	390

Why do we need to concern ourselves with the distinction between categorical and gradient phonology? The primary reason here is the claim that the representations involved in lexical phenomena (i.e. lexical contrast, lexical phonology, "analogy", morphology and allomorphy) must be categorical, and cannot be gradient. This claim is made explicit in Kiparsky (1985) and Zsiga (1993), among other places, and may be termed the Lexical-Categorical Hypothesis. (This hypothesis has been challenged of late for reasons that will not be discussed here; see e.g. Flemming 1995, Kirchner 1995, Steriade 1996.)

Vowel length variation in Scottish English provides an example of the apparent validity of this hypothesis. The length of tense vowels and diphthongs in Scottish English is predictable by a generalization known as Aitken's Law (see e.g. McMahon 1991, Borowsky 1993), whereby tense vowels are long in open syllables and before voiceless continuants, and short otherwise. This state of affairs can be described informally with the following rule.

(2) Aitken's Law
[adapted from McMahon 1991]

$$\begin{matrix} \text{V} \\ [\text{+tense}] \end{matrix} \quad \rightarrow \quad \text{V: / \_} \begin{Bmatrix}]_{\sigma} \\ \begin{bmatrix} \text{+voice} \\ \text{+cont} \end{bmatrix} \end{Bmatrix}$$

Because vowel length does not make lexical distinctions in Scottish English, Aitken's Law would be considered allophonic were it not for the fact that it is sensitive to morphological structure. As the following examples show, syllables that are closed with regular suffixes are considered open for the purposes of the rule. Hence Aitken's Law must be considered a lexical rule (or a word-level rule, in Borowsky's formulation). Note that this is true even though it is not structure-preserving, in the sense of Kiparsky (1985) and elsewhere, in that it creates a structure (namely, a long vowel) that is not present underlyingly for lexical contrast.

(3)

V:	V
agree]d	greed]
brew]ed	brood]
stay]ed	staid]

The Lexical-Categorical Hypothesis predicts, therefore, that in contrast with American English, in Scottish English vowel length variation must be categorical, not gradient. The phonetic data available (McClure 1977, Agutter 1988a,b) are not ideal; McClure examined only his own productions, while Agutter's data contain gaps and display a lack of balance, making statistical analysis impossible. Nevertheless, by a reanalysis of Agutter's (1988a,b) data, McMahon (1991) has argued that Aitken's Law is indeed categorical, although its effects are overlaid with the gradient vowel length variation apparently ubiquitous in English dialects. Data for speakers of "Standard Scottish English" are given below.

(4) [after (14) in McMahon 1991:41]
Durations are given in msecs, with standard errors in parentheses.

	[-voice]	[+voice, -cont]	[+voice, +cont]
/ʌi, i/	118 (6.5)	166 (15.5)	230 (7.9)
/au/	149 (8.6)	195 (21.2)	212 (10.6)
/ɔ, ɪ/	104 (10.1)	142 (13.5)	166 (11.3)

McMahon's (1991) own argument for there being a categorical vowel length effect in Scottish English depends on a comparison with Received Pronunciation which does not have Aitken's Law, but a dialect-internal argument is also available. As can be seen in the following figure, the largest difference (64 msec) is found (a) with the tense vowels, which alone are subject to Aitken's Law, (b) at the boundary of the "Aitken's Law" category (i.e. the voiced continuant column). By contrast, note that the difference in vowel duration between the voiceless and voiced stop environments is about the same for vowels of all types, suggesting that this effect is due to general gradient vowel duration variation. The categoricality of Aitken's Law thus reveals itself in its respecting of categorical differences between (a) tense vs. lax vowels, and (b) voiced continuants vs. all other consonantal environments.

(5) Differences in duration in msecs [derived from previous figure]

	[-voice]	[+voice, -cont]	[+voice, +cont]
/ʌi, i/	48	**64**	
/au/	46	17	
/ɔ, ɪ/	38	24	

Notice that what we have done is address a question of theoretical phonology using physical measurements. In fact, it was not possible to answer the particular question we asked (Does Aitken's Law conform to the Lexical-Categorical Hypothesis?) without such measurements. The usual way of identifying categories in phonology, namely by reference to their function in making semantic distinctions, was not available to us because Aitken's Law is not structure-preserving. Since we could not define phonological categories by looking down into phonology from above (i.e. from semantics), we had to seek evidence for categories by looking up into phonology from below (i.e. from physics).

2. Variable phonology.

In contrast to the concepts of categoricality and gradience, which refer to representations, variable phonology refers to variability in input-output linking (via variability in rule application or constraint evaluation). This means that a given input will generate a given output with a probability of $0<p<1$. In invariant phonology, p is equal to 0 or 1.

Since the categorical/gradient dimension refers to representations while the variable/invariant dimension refers to input-output linkings, the concepts are to some extent independent of each other. Thus we expect there not only to be phonological processes that are neither gradient nor variable (e.g. a "classical" lexical rule), or both (e.g. many purely phonetic processes), but also processes that are both categorical and variable, or even, conceivably, both gradient and invariant (such as a rule which deterministically generates a distinct output for different inputs along a gradient scale). Since this paper is concerned with variable phonology, most useful for present purposes would be an established example of categorical variable phonology. Such is found in Finnish i-Spirantization (Kiparsky 1973:92-101), illustrated below.

(6) a. t s / _ +i

b.	/halut+i/	→ halusi	"wanted"
	/hakkat+i/	→ hakkasi	"hewed"
	/turpot+i/	→ turposi	"swelled"

The rule is structure-preserving and thus categorical. This is fortunate for the Lexical-Categorical Hypothesis, since the process is clearly lexical; i-Spirantization is sensitive to morphology (it only applies before the past tense suffix *-i*).

Of interest here is the fact that i-Spirantization is also variable. In words where Vowel Contraction deletes a prevocalic vowel, i-Spirantization does not always apply the same way with a given input. Namely, in some words it applies to the environment created by Vowel Contraction, in some words it does not apply at all, and in others it applies optionally. It is in these "optional" words that i-Spirantization behaves like a variable rule:

(7)	/piirt +i/	→	piirsi	"drew"
	/pit +i/	→	piti	"held"
	/kiit +i/	→	kiiti	"sped"
			kiisi	

One characteristic of this particular variable pattern is that the variability is restricted to specific lexical items. As it happens, however, not all categorical variable phonology is so restricted, not even in Finnish; see Anttila (this volume) for a particularly dramatic example of an apparently exceptionless categorical variable pattern.

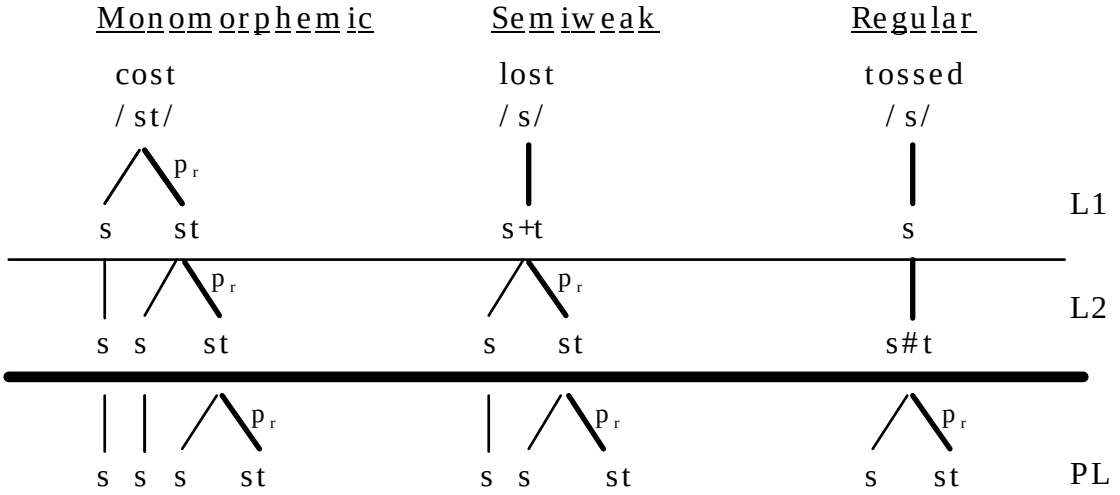
The fundamental characteristic of categorical variable phonology is upheld in this example in that a single input (e.g. /kiit +i/) is linked to discrete outputs (e.g. /kiiti/ and /kiisi/). Categorical variability thus contrasts with gradient variability, which would involve a single input being linked to a continuous physical scale. Vowel length variation that was both gradient and variable would generate a broad distribution of outputs along the scale defined by duration; some outputs may be more common than others, but examination of the distribution would show no evidence of discrete output categories (see Pierrehumbert 1994 for schematic histograms illustrating these ideas).

3. Variable t-deletion: categorical, gradient, or both?

In this setting, variable t-deletion in English becomes theoretically significant. This is because it is unambiguously a variable process, and yet has been argued by Guy (1991a,b) to be both lexical and postlexical. (See also Hinskens 1992 for a similar example of variable t-deletion in a group of Dutch dialects.) Given the Lexical-Categorical Hypothesis, this implies that t-deletion contains categorical components (the lexical applications of t-deletion) but may also include gradient components (the postlexical applications of t-deletion may be gradient, though they need not be). As we will see, however, t-deletion has usually been described as if it were entirely categorical. This will be seen to confront physical evidence on t-deletion from the literature, which in turn will force us to adopt somewhat unorthodox methodologies in an attempt to put the question on a firm empirical footing.

One aspect of the dilemma that will arise seems unshakable: Guy's evidence that in certain dialects of American English t-deletion is both lexical and postlexical. Specifically, Guy has shown that t-deletion interacts with morphology in such a way that it must be analyzed as applying both within the lexicon as well as in a domain larger than the word. Schematically this interaction can be illustrated as follows.

(8) [after discussion in Guy 1991a,b]



Guy, following Kiparsky (1982) and other analyses of lexical phonology in English, assumes that there are three relevant derivational levels involved, two lexical levels (L1 and L2) and a postlexical level (PL). The diagram visualizes the possible outcomes at each separate level for three morphologically distinct classes of word-final /t/. In monomorphemic (M) forms like "cost", the /t/ is available for deletion at all three levels. In semiweak (S) forms like "lost", the /t/ is suffixed before the L2 level, leaving t-deletion with only two chances to apply. Finally, in regular (R) forms like "tossed", the /t/ is not suffixed until just before the postlexical level, so that t-deletion can only apply once. The claim that t-deletion applies with the same probability at all three levels is indicated by the rate of retention for /t/, labeled p_r, appearing in appropriate locations in each level.

This analysis makes two distinct predictions. First, since the potential application of t-deletion is assumed to be independent at each level, the rate of retention p_r at each level can simply be multiplied to determine the rate of retention of /t/ at the surface. Thus the rate of retention should be $p_r{}^3$ in M forms, $p_r{}^2$ in S forms and p_r in R forms. This prediction is supported by statistical analyses of variable t-deletion in dialects of English (Guy 1991a). The second prediction concerns the differential effect of the preceding and following contexts on t-deletion. The environment before the /t/ is always word-initial, and thus its effect should vary across morphological class; word-internal information should have three chances to affect t-deletion in M forms, but only one chance in R forms. By contrast, the environment following the /t/ is always word-external information, and thus should affect all morphological classes equally, since the word-external environment only plays a role postlexically. This prediction is also confirmed (Guy 1991b).

Given the interests of this paper, we must ask what the representational nature of t-deletion is. Guy himself always describes it in categorical terms. He gives an informal characterization of the process as follows. This is clearly a categorical variable rule.

(9) [from Guy (1991a:8)]

-t, d Deletion <variable, probability of application = p_a>

$t \rightarrow <\varnothing> / C \_]$

Later he gives the rule in autosegmental notation. This version also constitutes a categorical rule, since both the input and output representations involve only categorical units organized with categorical association lines.

The categorical analysis is a natural consequence of the way Guy collected his data. Relevant tokens were taken from recorded natural speech samples and coded impressionistically as showing retention or deletion of the final stop; that is, trained listeners divided tokens into distinct categories. No claim was made that t-deletion was in fact categorical, however. The actual t-deletion process may well have been t-shortening, t-reduction or some other gradient process that was perceived as categorical, just as the gradient /l/ continuum is perceived as involving illusory "dark /l/" and "light /l/" categories. The method of coding the data prejudges the issue, and so perceptual data are of no use in addressing the questions asked in this paper.

Among other reasons, the issue has theoretical importance because at least one widely recognized analysis of Guy's data seems to require the categoricality assumption, namely the Optimality Theory analysis of Kiparsky (1993). In this analysis, t-deletion is described as the non-syllabification of /t/; ranked constraints select between two possible candidates for any given input, one where final /t/ is syllabified (and hence surfaces) and one where final /t/ is not syllabified (and hence does not surface). The analysis thus presupposes the existence of two distinct output categories for any given input. It is not immediately obvious how the analysis could be salvaged if it turned out that variable t-deletion were in fact gradient, and especially if it were gradient postlexically but categorical lexically (the analysis does not recognize this distinction).

As it happens, there is evidence that t-deletion is indeed gradient, at least in part. Browman and Goldstein (1990) analyzed relevant tokens from the AT&T Bell Laboratories archive of X-ray microbeam data on speech production (Fujimura, Kiritani, and Ishida 1973, Miller and Fujimura 1982) and found that in those cases where an underlying /t/ or /d/ is acoustically absent, the alveolar gesture is still clearly present, as indicated by the movement of the lead pellet attached to the tongue blade. What actually seems to happen when a phrase like "perfect memory" is produced as "perfec' memory" is that the labial articulation for the /m/ overlaps with the /t/ articulation, thus hiding its acoustic effect. Rather than a categorical deletion of /t/, then, we have a gradient change in the relative timing of the /t/ gesture and nearby gestures that may be reflected along the physical scale of duration from the onset of the /t/ to the onset of the following segment.

If t-deletion has a gradient component, and if Guy's Lexical Phonology analysis of t-deletion and the Lexical-Categorical Hypothesis are both correct, we must conclude that the postlexical application of t-deletion is gradient, while the two lexical applications are categorical.

This conclusion forces us to understand the lexical and postlexical variability of t-deletion as being due to two quite different phenomena. Lexically, variable t-deletion will involve the probabilistic generation of multiple categorical representations, where the allomorph with /t/ is chosen with a probability of p_r at each lexical level. Postlexically, however, variable t-deletion involves the generation of a distribution of tokens varying along a continuous physical scale, so that p_r now represents not the physically categorical retention of /t/, but rather something more like the probability of being perceived as ending in /t/.

Fundamental to Guy's analysis is the assumption that the value of p_r is the same at all levels. Within the present schema, however, the lexical and postlexical values for p_r can only come to match each other through diachronic or acquisitional forces; there can be no synchronic explanation for the existence of a single value for p_r, since the variability arises in different ways at the lexical and postlexical levels.

This sort of diachronic effect is probably quite common. As is standard in diachronic analyses of lexical phonology (see e.g. Kiparsky 1988), we assume that t-deletion was originally only postlexical. At this stage, the perceived value for p_r would have followed automatically from acoustic properties of the actual gradient distribution. If some children were then to reanalyze t-deletion as including a lexical component, they would only have to reinterpret the p_r value as reflecting the rate of generation of allomorphs with /t/; as Labov (1994) notes, probability matching is accomplished quite readily by a wide variety of organisms.

4. How should one investigate the nature of variable t-deletion?

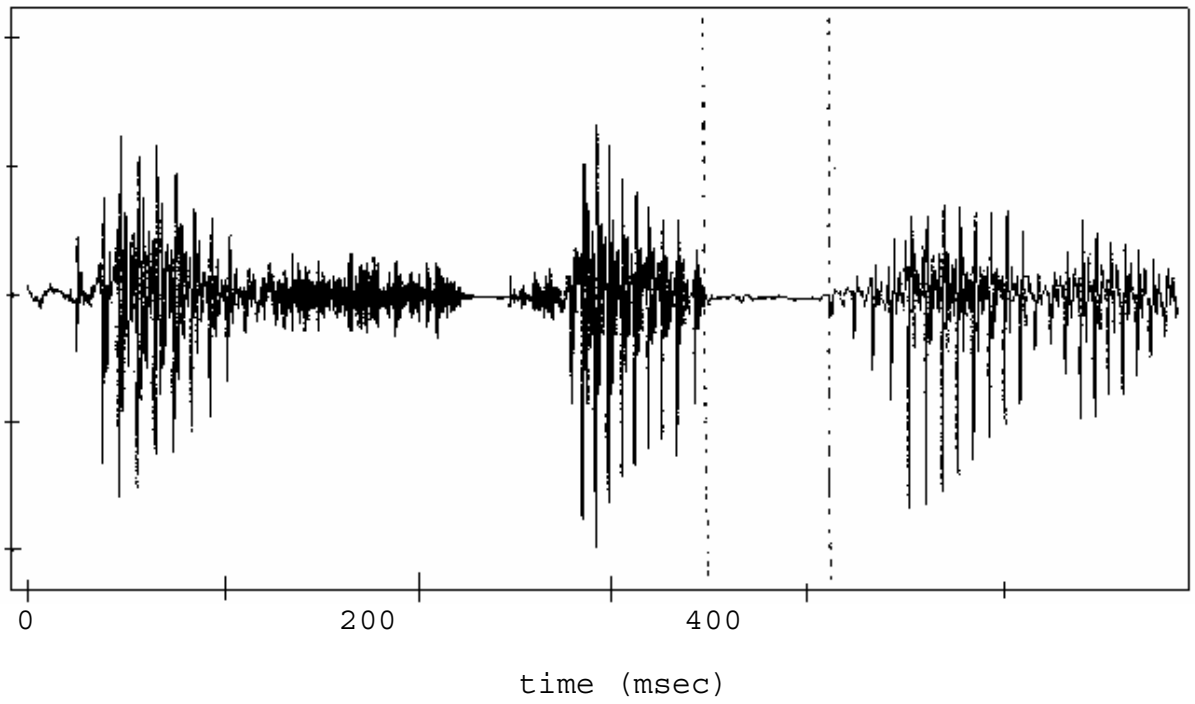
A major question raised by the analysis of t-deletion as both categorically and gradiently variable concerns the means of testing the claim empirically. As with non-structure-preserving lexical phonology, we cannot count on semantics to divide up the categories for us. Hence we must again turn to measures of values along a physical scale. A gradient process should produce a single broad distribution along such a scale, while categorical processes should produce evidence of discrete subdistributions.

There are numerous physical scales that could be used for t-deletion. Because we are interested in what the speaker intends articulatorily, and not what the perceiver hears, the ideal scale would be an articulatory one, perhaps involving the location of the tongue blade gesture for /t/ relative to the gestures of other articulators operating adjacent in time. As Browman and Goldstein (1990:363) note, however, "talking with lead pellets in your mouth and an X-ray gun pointed at your head hardly counts as a 'casual' situation"; because of such interference, there far fewer instances of t-deletion in the Bell Laboratories archive than occur in natural speech.

Acoustic methodologies, which are less likely to interfere with casual speech processes, have the obvious disadvantage of missing much interesting information. They cannot, for example, distinguish easily between unreleased /t/ from an absent /t/. More problems become clear when one considers precisely what it is one will be measuring. Stop closures can be seen most clearly in the acoustic waveform when surrounded by vowels; however, in this environment, the rate of t-deletion drops virtually to zero. The rate increases when /t/ follows a consonant of low sonority, like another stop, but in this environment the onset of /t/ is often impossible to make out.

The scale chosen for the present study was labeled <u>V-m duration</u>, defined as the duration from acoustic offset of the vowel in the target word to the onset of the following word (always "many"). Hence in "cost many", the V-m duration was measured from the offset of /ɔ/ to the onset of /m/. This necessarily includes the duration of the preceding consonant, which was unavoidable anyway in some cases because the preceding consonant was also a stop. It also includes any pause following the end of the word and the beginning of the next. The choice of /m/ to follow the target word provided an environment that was low enough in sonority to allow for the deletion of /t/ and yet whose onset was almost always clearly visible in the acoustic waveform. One example of the waveforms examined is shown below.

(10) "They stepped many..."
(V-m duration [between vertical lines] = 66 msec)



The observed values along the V-m duration continuum were used to generate distributions for R tokens like "tossed" and for M tokens like "cost." (S forms like "lost" were not examined because, as will be clear shortly, the predictions for distribution shape would not be distinct enough from those for M forms.) The predictions primarily concerned the shapes of these distributions. Thus R tokens should show a purely gradient pattern: a single broad (presumably normal) distribution. This is because in R tokens, t-deletion only occurs in the postlexical phonology, which according to the hypothesis being entertained here, is where t-deletion is gradient and not categorical.

By contrast, the distribution for M tokens should be more complex. Since t-deletion applies both lexically and postlexically in M tokens, its broad "gradient" distribution should be supplemented with evidence of categorical deletion of /t/. This will take the form of a second peak representing the V-m durations for tokens that have had /t/ deleted in the lexicon. We expect this second peak to be to the left (i.e. towards zero) relative to the broader peak, on the assumption that a token with a categorically deleted /t/ has a shorter V-m duration than one with a /t/ that is present, though gradiently shortened. This is presumably true whether or not /t/ is actually deleted, as in Guy's rule, or merely left unsyllabified, as in Kiparsky's (1993) analysis.

While these predictions are clear enough, we need to go beyond general descriptions of shapes of distributions and make the predictions statistically testable. Unfortunately, the specific difference we predict does not have a statistical test of its own. Statistical tests are designed to investigate the location and shape of single distributions, and what we are asking here is whether a distribution is made up of more than one distribution. Nevertheless, with careful use of existing tests it is possible to address our questions indirectly.

First, if the R and M distributions differ primarily in the presence of an extra peak on the left for the M tokens, we predict that the means for the R and M distributions will differ, with the

mean of the M distribution less than that of the R distribution; the second M peak will pull the overall mean downward.

Second, we might expect the variance to differ for the R and M distributions. Variance can be thought of as a measure of the "spreadedness" of a distribution, so if there is an extra peak in the M distribution, its variance may be higher. (This of course assumes that the "categorical" and "gradient" subdistributions do not entirely overlap.)

Third, we expect that the R distribution should be a normal distribution, while the M distribution, containing as it does an extra peak, should not be a normal distribution. (This prediction depends on the additional assumption that unimodal distributions tend to be normal.)

Finally, we predict differences in skewness. Skewness is a measure of a distribution's asymmetry; a distribution which spreads farther to the left of the mean than to the right is said to be negatively skewed. Again, we predict the R distribution to be normal, and hence not skewed. By contrast, we predict the M distribution, with its extra peak, to be negatively skewed relative to a normal distribution.

Skewness is one of the higher "moments about the mean." A still higher one is kurtosis, which is often taken to describe the "peakedness" of a normal distribution. Given this interpretation, we would predict the M distribution to have a lower kurtosis value, since with its extra peak it should be more spread out. However, kurtosis has been shown not to accurately predict the shape of a distribution, and so will not be discussed further.

It should be noted that we also make another prediction testable by the examination of distributions. This concerns the shape of the distribution associated with words, like "toss", that do not have a final /t/ at any level. If lexical t-deletion utterly removes all trace of /t/, we expect that the rimes of "cos(t)" (i.e. "cost" with the /t/ lexically removed) and "toss" should be physically identical. Thus the mean of the distribution of such "no-t" tokens along the V-m scale should be in the same location as the second peak hypothesized to appear at the left in the M distribution.

This is only true, though, if lexical t-deletion involves complete deletion of /t/. If /t/ is not in fact entirely deleted, the rimes of "cos(t)" and "toss" need not be identical. This state of affairs can arise in at least two ways without wreaking havoc on the Lexical-Categorical Hypothesis. First, lexical t-deletion may actually be a form of categorical reduction of /t/. That is, an underlying /t/ may be changed into another category, call it t*, that is different from a full [t] in some way (perhaps it is unreleased), but it is not entirely absent. In this scenario, lexical t-deletion would not be structure-preserving, but it would still be categorical. Another way would be to follow Kiparsky's (1993) analysis. If t-deletion actually involves the non-syllabification of /t/, rather than its deletion, then we would not necessarily require "cos(t)" to show the complete physical absence of an underlying /t/ (depending on one's views on the possible low-level effects of unsyllabified segments on articulation). In either scenario we don't make any prediction about the shape of the "no-t" distribution.

5. The experiment.

With this background, the experiment itself should be easy to understand. Twenty-one native speakers of American English were recorded (age range: 19-40, mean = 24; 18 males, 3 females). No attempt was made to control for dialect, but all speakers were born and raised in the Northeastern United States. The recordings were made at the State University of New York at Buffalo.

The target words examined were those listed below.

(11) Target words

M	R	no-t
accept	stepped	step
cost	tossed	toss
attract	tracked	track

The complete list of words recorded is given below. The S class words were excluded because it was determined that the predictions for the S distributions were not distinct enough from the predictions for M class tokens (both should contain a broad main distribution and a smaller, narrower peak to the left). The words ending in [d] were not used for acoustic phonetic reasons. In many tokens with [d], the V-m duration was not consistently measurable, partly because the offset of the vowel could not be distinguished from the offset of the following nasal or liquid. Moreover, for many speakers the vowels in "poll" and "fold" were not identical.

(12)

M	S	R	no-t
accept	kept	stepped	step
cost	lost	tossed	toss
attract		tracked	track
find		fined	fine
fold	told	polled	poll

Every attempt was made to balance the set of target words so that the only difference between the various classes was their morphological structure. Perfect balance was not possible, however. As the following figure shows, for example, target words vary in frequency, with the R class forms tending to be less frequent, as is typically the case.

(13) Frequencies of target words (Kucera and Francis 1967)

accept	72	stepped	41	step	131
cost	229	tossed	31	toss	9
attract	19	tracked	3	track	38

One factor that we can probably safely ignore is the number of syllables in the target words. Although "accept" and "attract" both have two syllables and "stepped" and "tracked" have one, this should not have any appreciable effect on durational measures made in the portion of the utterance that is of interest, since the duration of a stressed syllable in a word is determined by the number of syllables following, not by the total number of syllables (Nooteboom 1972, Gay 1981).

The target words always appeared in the same phonetic context, namely in the frame "They X many...". The sentence containing this frame was in turn embedded in natural-sounding passages, with the material after "many" consistent with the content of the passage. The passages were otherwise made as similar to each other as possible in style and length. In particular, each passage consisted of four sentences, and the sentence with the target word was always the third one. Each passage came in both present and past tense versions. M tokens were taken from the present tense passages with M target words; R tokens were taken from the past tense passages with R target words; and no-t tokens were taken from the present

tense passages with R target words. A sample set of M, R and no-t passages is shown below. Portions of the passages that were extracted for analysis are highlighted.

(14) M passage:

The violent storms rage along the coast. They wash out roads and bridges. They cost many people their lives. Hundreds of families have no water or electricity.

R passage:

The picnickers were getting ready to go home. They gathered up their dirty paper plates and other garbage. They tossed many cans into the trash. Then they put their picnic baskets into the car and drove away.

no-t passage:

The picnickers are getting ready to go home. They gather up their dirty paper plates and other garbage. They toss many cans into the trash. Then they put their picnic baskets into the car and drive away.

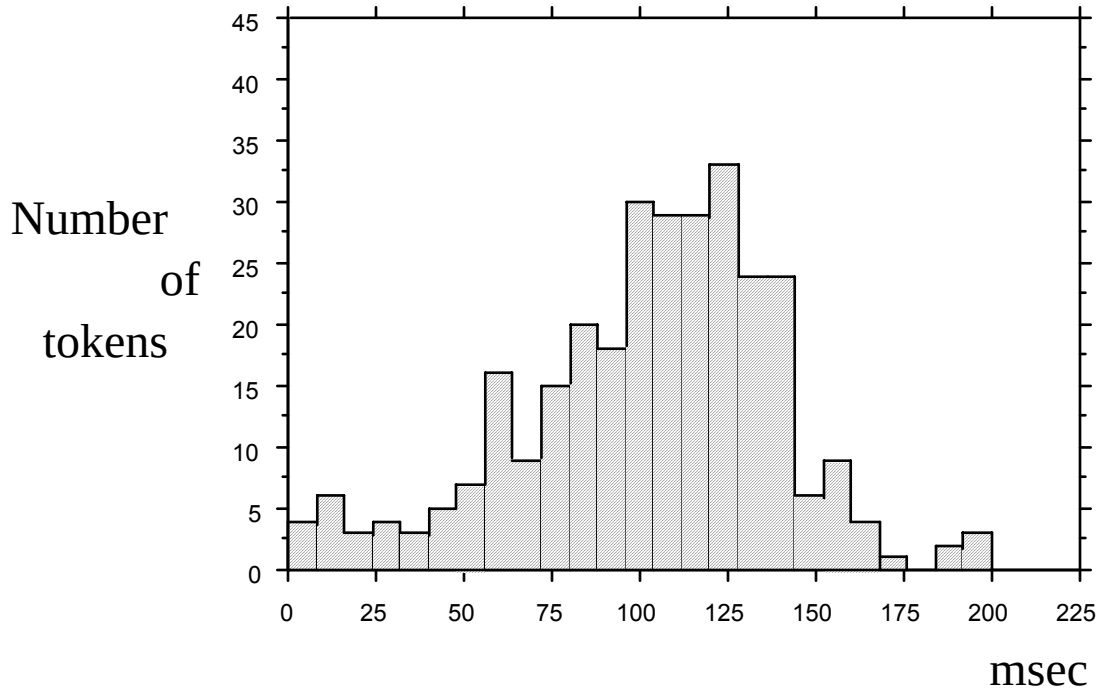
The general procedures for the study were as follows. Participants were given a packet with the 26 passages (one passage per each of 13 original target words, each in a present tense and past tense version). Packets were randomly ordered, differently for each speaker. Participants were instructed to read the whole set of passages aloud into a microphone in as natural a way as possible. Each participant was left alone in a recording room while reading aloud. After each read-through, the experimenter reentered the room and confirmed the completion of reading. This procedure was followed five times in total for each participant. Thus the variability in t-deletion in this study resulted both from variability across speakers and from variability within speakers due to the repetitions in reading.

After recording, the portions of the passages indicated above were digitized and examined with a waveform editor. V-m durations were measured for each token. With 21 participants, 5 repetitions and 3 each of M, R and no-t passages, each class should have contained 315 tokens. In fact, due to experimenter error, participants misspeaking and other problems (e.g. a few tokens where the interword pause was clearly due to hesitation), there were only 305 M tokens, 305 R tokens and 303 no-t tokens used in the analysis.

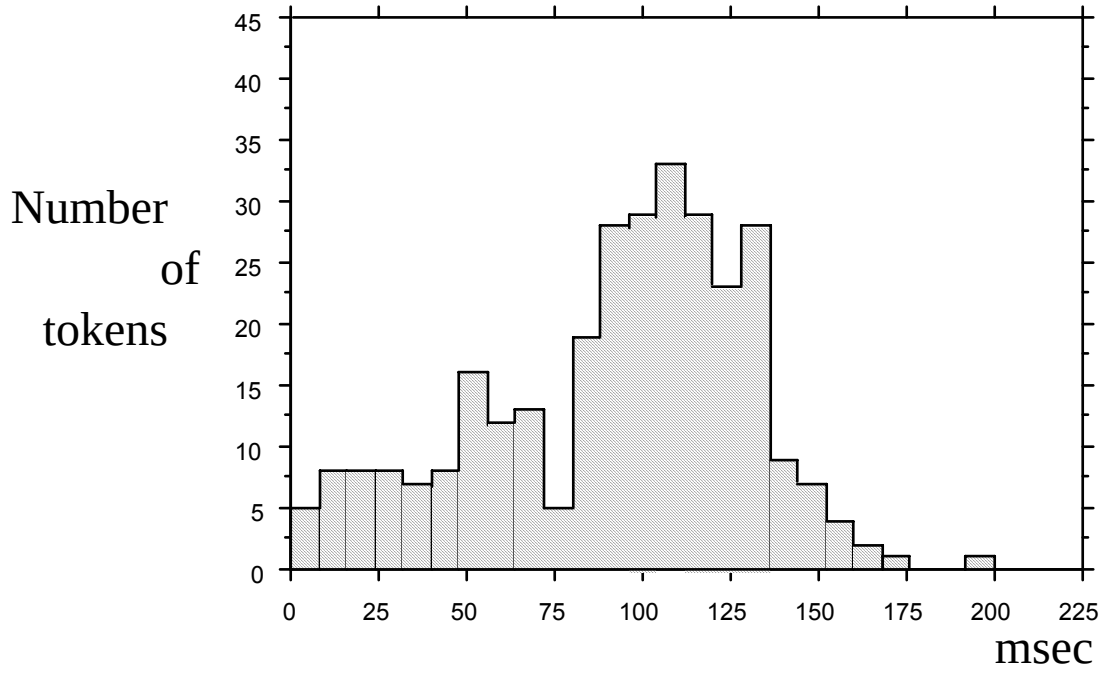
6. Results and discussion.

The distributions for the R, M and no-t tokens are shown in the following histograms.

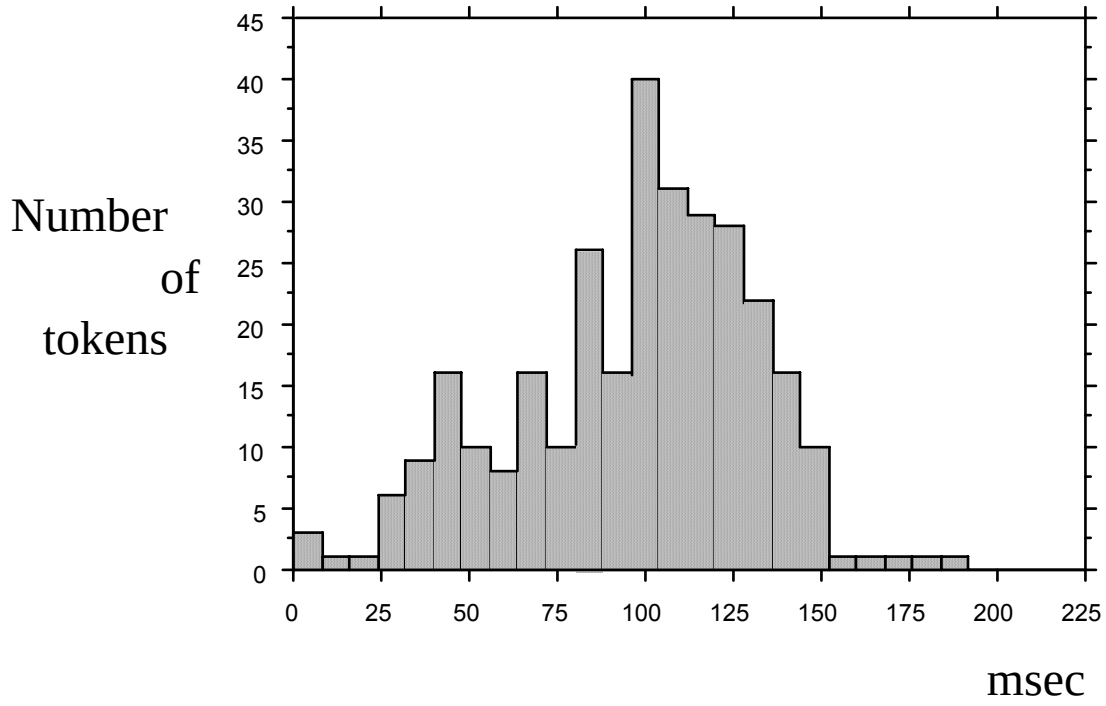
(15) R distribution



(16) M distribution



(17) no-t distribution



The first prediction tested concerns the difference in means between the M and R distributions. If there is a second peak to the left in the M distribution, we expect the mean for M to be lower. This is indeed the case (paired t-test, t(294)=-4.207, p<.0001).

(18)

Class	Mean V-m duration (msec)
R	102.12
M	93.14
R-M	9.54

These results, though consistent with the claim being tested, do not by themselves prove that some M tokens undergo categorical t-deletion, while R tokens never do. The difference in means may instead arise from the already established fact that M forms undergo t-deletion more often than R tokens. We need to show that M forms undergo t-deletion to a greater degree along the V-m duration scale, and not merely more often in the course of their derivation.

Although the crucial tests of this claim thus lie in the shape of the distributions, not in their means, it is instructive to examine the data on means a bit more closely. The table below provides the means (in msec) of the V-m duration participant scores (i.e. the mean of V-m durations taken from the five or fewer repetitions per participant), exploded to show differences across phonetic environment as well as morphological class. This is relevant, since V-m duration was necessarily defined to include the duration of the preceding consonant.

(19)

	R	M	no-t
p_	78.56	67.74	68.33
s_	123.80	119.60	119.90
k_	104.38	93.11	102.05

Under this regrouping, the V-m durations for R continue to be significantly longer than those for M (the no-t results will be discussed below), but it is clear that phonetic environment plays an important role as well. Moreover, the amount of variation is greater after /p/ than after /s/. The same generalizations follow from the table below giving the associated z-scores (which indicate distance above or below the mean in units of standard deviation). An analysis of variance shows that phonetic environment is an especially important factor, but morphological class continues to be significant.

(20)

	R	M	no-t
p_	-.60	-.93	-1.09
s_	.96	.75	.74
k_	.24	-.18	.16

These tables make two points. First, the mean of M remains significantly less than that of R, even if within-participant variability (i.e. due to multiple readings of the same word by the same speaker) is factored out. This is consistent with the difference being due to some sort of variable phonology, since the same sort of variability found within speakers is also found across speakers.

The second point also suggest a way in which the design of this sort of study could be improved. Because variation due to phonetic context is irrelevant to the present hypothesis, the finding that phonetic context causes a large amount of variation is to be regretted. In the original design, it was hoped that a sample of different words would increase the breadth of the generalizations. Unfortunately, because the English lexicon is not very generous with perfectly balanced stimuli, any attempt at cross-word generalizations will lead to the introduction of irrelevant variability.

The above discussion raises the issue of the sources of variability. The method as described assumes that in generating histograms whose shapes will be compared, it is irrelevant whether variability results from intra- or inter-speaker variation. This assumption will be confronted below, after the remainder of the results are presented.

We now to turn to variance. If the extra peak in the M distribution does not fully overlap with the main broad distribution, we expect the variance for the M distribution to be higher than that of the R distribution. As can be seen below, however, the statistical test chosen, which operates on the ratio of the M variance to the R variance, did not come out significant ($F(304)=1.114$, $p=.3483 > .05$). With hindsight this is not surprising, given the fact that the R and M distributions both extend down to 0 msec; there was thus no way for the hypothesized extra peak at the left of the M distribution to increase overall variance.

(21)

Class	Variance	
R	1400.373	
M	1559.624	
M/R	1.114	(not significant)

Other tests were concerned more directly with the shape of the distributions. As emphasized by Hopkins and Weeks (1990), from whom the statistical details of the following discussion are drawn, tests of distribution shape are well-developed but are not often used by social scientists. However, it should be kept in mind that these tests are designed to compare the shape of distributions to normal distributions, not to each other.

In the chi-square test for normality, a distribution is divided into a number of intervals. The total number of tokens in each interval may then be plotted against the expected number of tokens for the parallel interval in a normal distribution with the same mean and variance as the distribution under examination. The test then proceeds the same way as in a standard chi-square test for deviation from an expected distribution, except that two fewer degrees of freedom must be used, one for each estimated value (in this case, the estimates that the sample mean and the variance are the same as the population mean and variance, respectively). In order to maximize both the number of intervals and the number of tokens within each interval, the distributions were divided into ten intervals.

We predict that the R distribution will be normal, while the M distribution will not. These predictions are confirmed, as seen in the p-values, which show that the R distribution is not significantly different from normal while the M distribution is.

(22) Chi-square test for normality

Class	p-values	Compared to normal
R	$.1 > p > .05$	same
M	$.005 > p > .001$	different

The other measure of distribution shape used was skewness. There are two standard tests of skewness. The more intuitive one, unfortunately, cannot be used to draw valid inferences about significance, so both tests were used.

Since skewness represents asymmetry, a natural measure would involve the difference between the mean and the mode (location of peak); if the mean is less than the mode, this would indicate that the distribution spreads further to the left than a normal distribution, and so is negatively skewed. To standardize this measure, we would simply divide the difference by the standard deviation, which gives a value between -1 and 1. Unfortunately, the mode is not uniquely determinable in general, and so the median (the location of the central token in a list of tokens arranged by value along a scale) is used instead. Since the median tends to be about one-third as far from the mean as the mode, we can write the formula for skewness as follows:

(23) Skewness = 3(Mean-Median)/Standard Deviation

Using this formula, the skewness values for the R and M distributions go in the predicted direction, as indicated in the following figure. The minus signs indicate that both are skewed to the left, but the M distribution is skewed quite a bit more.

(24) Skewness (descriptive statistic)

Class	Skewness
R	-.107
M	-.531

The less intuitive, but inferentially superior, method of calculating skewness is represented in the following formula. If *n* is large (over 150), the z value calculated below can be used to draw inferences on the statistical difference in skewness relative to normal by using the familiar z-test.

(25) $z= (\Sigma z_i^3(n/((n-1)(n-2))))/\sqrt{(6/n)}$

Using this test, the skewness values and p-values are those shown below in (26). According to this test, both the M and R distributions are significantly skewed leftward from normal. The test does not address the question of relative skewness between the M and R distributions, however, which is what is really at issue. The difference between the chi-square and inferential skewness tests arises from the fact that the latter is much more powerful than the former; the chi-square test appears to be of just the right strength to distinguish the relative skewness of the two distributions.

(26) Skewness (inferential test)

Class	Skewness	p-value
R	-.411	.0016
M	-.291	.0162

In short, the predictions concerning the R and M distributions seem to be confirmed for the most part (acknowledging the methodological problems to be discussed more fully below). The R distribution does in fact have a higher mean and the M distribution differs from normal to a greater degree than does the R distribution. The statistical results are therefore consistent with the M distribution having an extra peak somewhere at or below 50 msec (as located by inspection of the histogram), though of course by no means do they demonstrate this conclusively.

Turning now to the no-t distribution, we find a more confusing picture. Recall that it was predicted that the no-t peak should coincide with the hypothesized peak in the M distribution representing tokens where /t/ was categorically deleted. This prediction is of course false, as should be obvious from the histograms. Instead, the no-t mean of 96.38 msec is not significantly different from either the R or M means. Moreover, the variance of the no-t distribution, while significantly less than the variance of both the R and M distributions, in no way indicates the narrow peak that was expected.

These results pose at least two challenges to the methodology attempted here. First, it is impossible to maintain the idea that the second hypothetical peak to the left in the M distribution is identical to the no-t distribution. In other words, if there is a second peak in the M distribution representing the output of categorical t-deletion, this output cannot be identical to the surface forms of no-t words: the rime of "cos(t)" is not identical to that of "toss." Fortunately, we were prepared for this eventuality. We only expect "cos(t)" and "toss" to show identical distributions for V-m duration if lexical t-deletion operates as Guy's rule implies, entirely removing final /t/. If t-deletion actually creates a category t* intermediate between /t/ and no-/t/, or allows /t/ to stay but leaves it unsyllabified, then we do not expect "cos(t)" and "toss" to behave identically.

On the other hand, we also do not expect the no-t mean to be identical with that of the R and M distributions, and the finding that it is poses the second challenge. Why shouldn't V-m durations for words like "tossed" and "toss" be statistically different if V-m duration is indeed a

measure of presence or absence of /t/? Informal listening to the tokens by the author seems to rule out the extreme possibility that word-final /t/ was always deleted (or that it was never deleted, for that matter). Moreover, as noted above, the variance of the no-t distribution (1155.018) was significantly less than that of the R and M distributions, which suggest that the R and M distribution involved more variability in t-deletion than did the no-t distribution, which of course involved none.

Apparently, however, by including the duration of the consonant preceding /t/ in the V-m duration measure, the power of this measure was considerably weakened. It is well known that a consonant in a syllable-final cluster is shorter than this same consonant appearing alone syllable-finally (e.g. Munhall, Fowler, Hawkins and Saltzman 1992). This compensatory shortening may have reduced the duration of the preceding consonant in words like "tossed" to such a degree that, when variability was added to the mix, any small difference in V-m duration between "toss" and "tossed" became insignificant. As the reader will recall, the necessity of measuring from the offset of the vowel and not of the preceding consonant followed from the fact that /t/ was preceded by different consonants in different words, a situation that in itself created a disturbing variability.

Finally, beyond the specific problems raised by the no-distribution and the V-m duration measure, we must address the issue regarding intra- and inter-speaker sources of variation. The danger in including both intra- and inter-speaker variability is that the data points are no longer independent; the five V-m durations for a speaker A will cluster around the same point, in contrast to the healthy scattering that would be seen in plotting one measure each from speakers A, B, C and so on. Nevertheless, in the present study, variability that arose across multiple readings by a participant was included together with variability due to differences across participants. The reason for this unorthodox mixture is easy to understand; the statistical tests used to analyze distribution shape require many more data points than the more familiar tests used to compare means. Given the stingy English lexicon, there are only two ways to generate enough data points per morphological class: vastly increase the number of speakers, or vastly increase the number of repetitions per speaker. The former is logistically difficult, while the latter violates the basic principle of psycholinguistic research whereby results should generalize from the sample to the real-life population at large, not to mention the sociolinguist's desire to understand speech communities in natural contexts and not merely the lone speaker-hearer jumping through articulatory hoops.

The conservative nature of statistics is also such to cast doubt on the heavy reliance on distribution shape in the present study. Experiments in the social sciences are virtually never designed to test hypotheses by means of distribution shape; if concepts like "kurtosis" come up at all, they are merely as a descriptive aside. Their importance in this study follows directly from the attempt to provide an unbiased analysis of t-deletion in speech production. Studies of the variability of t-deletion, independent of its categorical or gradient nature, are empirically straightforward, as they may rely on perceptual classification methods. Likewise, production studies of categoricality and gradience are empirically straightforward if one ignores variability. The difficulties arise when one attempts to address variability and gradience at the same time. The categoricality question precludes the use of perceptual judgments, while the variability question requires that variation in production be included as an item of study, and not an annoyance to be controlled. It is possible that a standard means-comparing method may be devised to answer the questions asked in this paper, though this seems unlikely. Readers intrigued by such quandaries are urged to ponder solutions themselves.

In spite of methodological misgivings, it should be remembered that the difference in R and M means does hold up even when intra-speaker variability is eliminated, and this fact,

combined with the findings regarding distribution shapes, suggest, at the very least, much promise for future research on the nature of variable t-deletion in produciton.

7. Conclusions.

Among the many points addressed in this paper, it is hoped that the reader will remember the following three. First, because of the partial independence of types of representations and types of input-output linkings, there should be a separate terminology associated with each. It has been proposed that the opposition categorical/gradient should be reserved for types of representations, rather than for nondeterminism in rule application. The nature of input-output linking should be indicated by separate terms, for which is proposed the opposition variable/invariant.

Second, using these clarified concepts, a syllogism was followed with implications for a number of disciplines interested in the sounds of language. Namely, if one accepts (a) Guy's (1991a,b) conclusion that t-deletion is both lexical and postlexical, (b) the Lexical-Categorical Hypothesis, and (c) Browman and Goldstein's (1990) evidence that t-deletion is gradient, then one is forced to accept the conclusion that t-deletion has both categorical and gradient components. Indeed, in all likelihood, t-deletion is not unique in this respect. Thus it behooves sociolinguists, theoretical phonologists and researchers interested in speech production (e.g. phoneticians and psycholinguists) to consider methods for determining the validity of this conclusion.

Finally, in the present attempt at addressing these questions using physical measurements, tentative evidence was provided that suggest that t-deletion may indeed have both categorical and gradient components. Of course, much remains to be done to clarify problematic aspects of the methodology, but it is hoped that the challenges and pitfalls have been identified clearly enough to guide future research on these issues of central importance to the study of variable phonology.

ACKNOWLEDGMENTS

The research reported in this paper was conducted in the laboratory of Jan Charles-Luce at the State University of New York. The author owes her and Paul Luce thanks for much assistance. Comments from the audience at the International Workshop on Language Variation and Linguistic Theory at the University of Nijmegen were also very useful, as were the careful critiques of Roeland van Hout, Frans Hinskens and an anonymous reviewer. Prof. van Hout also provided the tables in figures (19) and (20) and associated results. Above all, the author wishes to thank Greg Guy for providing the original inspiration. Naturally, the author takes full responsibility for all errors and misguided notions contained herein. The research was supported in part by NIDCD Grant No. 5-T32-DC00036-02.

REFERENCES

Agutter, A. (1988a) "The not-so-Scottish Vowel Length Rule." In J. M. Anderson and N. Macleod (eds.) Edinburgh studies in the English language. Edinburgh: John Donald, 120-132.

Agutter, A. (1988b) "The dangers of dialect parochialism: the Scottish Vowel Length Rule." In J. Fistak (ed.) Historical dialectology. Berlin: Mouton de Gruyter, 1-22.

Anttila, A. (1997) "How to derive variation from grammar." This volume.

Borowsky, T. (1993) "On the word level." In S. Hargus and E. M. Kaisse (eds.) Phonetics and phonology 4: studies in Lexical Phonology. Academic Press: Toronto, 199-234.

Browman, C. P. and Goldstein, L. (1990) "Tiers in articulatory phonology, with some implications for casual speech." In J. Kingston and M. E. Beckman (eds.) Papers in laboratory phonology 1; between the grammar and physics of speech. Cambridge University Press: Cambridge, 341-376.

Flemming, E. (1995) Perceptual features in phonology. Unpublished doctoral dissertation, UCLA.

Fujimura, O., Kiritani, S. and Ishida, H. (1973) "Computer controlled radiography for observation of movements of articulatory and other human organs," Computers in Biology and Medicine 3:371-384.

Gay, T. (1981) "Mechanisms in the control of speech rate," Phonetica 38: 148-158.

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

Guy, G. R. (1991b) "Contextual conditioning in variable lexical phonology," Language Variation and Change 3:223-239.

Hinskens, F. (1992) Dialect leveling in Limburg: structural and sociolinguistic aspects. Doctoral dissertation, University of Nijmegen.

Hopkins, K. D. and Weeks, D. L. (1990) "Tests for normality and measures of skewness and kurtosis: their place in research reporting," Educational and Psychological Measurement 50:717-729.

Kiparsky, P. (1973) "Phonological representations." In O. Fujimura (ed.) Three dimensions of linguistic theory. Tokyo Institute for Advanced Studies of Language: Tokyo, 1-136.

Kiparsky, Paul (1982) "Lexical phonology and morphology." In I. S. Yang (ed.) Linguistics in the morning calm. Seoul: Hanshin, 3-91.

Kiparsky, P. (1988) "Phonological change." In Frederick Newmeyer (ed.) Cambridge Survey of Linguistics, Vol. 1, 363-410. Cambridge University Press: Cambridge.

Kiparsky, P. (1993) "Variable rules." Paper presented at Rutgers Optimality Workshop 1, October 1993 and NWAVE 1994 (Stanford University).

Kirchner, R. (1995) "Contrastiveness is an epiphenomenon of constraint ranking," Proceedings of the Berkeley Linguistics Society 21.

Kucera, H. and Francis, W. N. (1967) Computational analysis of present-day American English. Providence: Brown University Press.

Labov, W. (1994) Principles of linguistic change: internal factors. Blackwell: Cambridge, MA.

McClure, J. D. (1977) "Vowel duration in a Scottish accent," Journal of the International Phonetic Association 7:10-15.

McMahon, A. M. S. (1991) "Lexical phonology and sound change: the case of the Scottish vowel length rule," Journal of Linguistics 27:29-53.

Miller, J. E. and Fujimura, O. (1982) "Graphic displays of combined presentations of acoustic and articulatory information," Bell System Technical Journal 61:799-810.

Munhall, K., Fowler, C., Hawkins, S., and Saltzman, E. (1992) "'Compensatory shortening' in monosyllables of spoken English," Journal of Phonetics 20:225-239.

Nooteboom, S. (1972) Production and perception of vowel duration." Unpublished doctoral dissertation, Utrecht University.

Peterson, G. E. and Lehiste, I. (1960) "Duration of syllable nuclei in English," Journal of the Acoustical Society of America 32:693-703.

Pierrehumbert, J. (1994) "Knowledge of variation." Papers from Parasession on Language Variation, Proceedings of Chicago Linguistics Society 30.

Read, C. (1972) "Pre-school children's knowledge of English phonology." In Language and learning: investigations and interpretations. Harvard Educational Review: Cambridge, MA.

Sproat, R., and Fujimura, O. (1993) "Allophonic variation in English /l/ and its implications for phonetic implementation," Journal of Phonetics 21:291-311.

Steriade, D. (1996) "Paradigm Uniformity and the Phonetics-Phonology Boundary," UCLA ms.

Zsiga, E. C. (1993) Features, gestures, and the temporal aspects of phonological organization. Unpublished doctoral dissertation, Yale University.

APPENDIX: Passages recorded for the study

Passages that were analyzed:

Monomorphemic:

Alice wants to become a doctor. She hopes she can get into the same medical school that her father went to. They accept many students each year. If she does get in, she knows her father will be very proud.

The violent storms rage along the coast. They wash out roads and bridges. They cost many people their lives. Hundreds of families have no water or electricity.

Everybody is talking about the new sculptures in the park. Even people who don't like the sculptures come to see them. They attract many visitors every day. People have very different opinions about them.

Regularly inflected:

Two little kids were playing in the rain. They found a big mud puddle. They stepped many times in the mud. They got themselves very messy.

The picnickers were getting ready to go home. They gathered up their dirty paper plates and other garbage. They tossed many cans into the trash. Then they put their picnic baskets into the car and drove away.

The hunters needed to get meat for their village. All day they searched through the forest. They tracked many kinds of animals. Only some animals, however, were good to eat.

no-t forms:

Two little kids are playing in the rain. They find a big mud puddle. They step many times in the mud. They get themselves very messy.

The picnickers are getting ready to go home. They gather up their dirty paper plates and other garbage. They toss many cans into the trash. Then they put their picnic baskets into the car and drive away.

The hunters need to get meat for their village. All day they search through the forest. They track many kinds of animals. Only some animals, however, are good to eat.

Passages that were not analyzed (not listed are the ten filler items, e.g. the past tense version of the "accept" passage)

Semiweak forms:

Mary's grandparents loved animals. Their house was filled with cats. They kept many dogs in their yard. When one of the dogs had puppies, Mary got to take one home.

The TV show tried out a new format. The producers were disappointed with the results. They lost many of their old viewers. Soon the TV show was back the way it was before.

Steve and Helen went camping. They set up their tents next to a river. They told many stories around the campfire. The next morning, they hiked to a beautiful lake.

Forms ending in [d]:

The kindergarten class is taking a field trip in the country. The children look around at all the living things. They find many insects in the grass. They see birds flying in the sky, too.

The traffic cops were out doing their job. They stopped cars that appeared to be going too fast. They fined many drivers for going over the speed limit. But they also took time to help out cars in trouble.

The people at the dry cleaners work hard. They iron and press dresses and skirts. They fold many pants and jackets. They do all they can to make their customers happy.

Patricia and Walt were running for class president. They had debates and press conferences. They polled many students before the election. Through it all, they still somehow managed to stay friends.