Senior Thesis
12/12/00
Christine R. Lattin

The potential of data sharing in quantitative sociolinguistics: /t,d/ deletion as a case study

Introduction:

Quantitative sociolinguistics is the study of linguistic variation in an empirical, statistically relevant way, in an attempt to find patterns relating variation to both internal (linguistic) and external (social) factors.[1] An important part of this process is the collection and annotation of databases, or "corpora," of speech. Traditionally, corpora have been collected using tape recorders, and then annotated by playing and replaying the tapes. The current state of computing technology, however, now allows the collection, annotation, analysis, and even summarization and presentation of linguistic data to rest entirely within the digital domain of computers and the world-wide web. These data are easily shared, which encourages a range of positive practices within the field of sociolinguistics, such as the reuse of existing data for new purposes; the comparison of results across studies; the use of stable data as a benchmark with which to compare new models and methodologies; and the measurement of interannotator consistency. However, the use of shared speech corpora in sociolinguistics also raises a number of theoretical and methodological concerns on issues such as speaker selection and the elicitation of data.

Since June 2000, I have been the head annotator for a project addressing some of these issues at the Linguistic Data Consortium (LDC) at the University of Pennsylvania in Philadelphia. This project, designed by sociolinguists Chris Cieri and Stephanie Strassel, explores potential benefits and problems associated with data sharing via a case study analyzing an already well-documented linguistic variable, /t,d/ deletion in English word-final consonant clusters.[2] We will eventually examine /t,d/ deletion as it appears in four large speech corpora, all collected to support speech engineering technology development but

capable of being reannotated to fit sociolinguistic purposes. This paper will look at /t,d/ deletion results in the first corpus, TIMIT (which stands for Texas Instruments and MIT, where the data was collected); this database consists of speakers reading groups of sentences chosen for phonetic richness. I will compare our results with results from past studies of /t,d/ deletion that used more traditional sociolinguistic methodology, and examine ways in which we may wish to alter our approach with the next three corpora. I will discuss both the positive and negative aspects of using this particular corpus of data and reannotated corpora in general, and suggest ways in which corpora could be made more broadly useful to the sociolinguistic community.

1.0 The background of this study

1. 1 The development of quantitative sociolinguistics

By examining the development of quantitative sociolinguistics, we will be able to trace the development of sociolinguistic methodology into what it is today, and better understand the ways in which sharing corpora of data differs from what is usually done.

Quantitative sociolinguistics has its origins in the study of dialect geography, which focused on finding the geographical boundaries of the distribution of linguistic features (Wardhaugh, 130). Dialect geography employed many assumptions drawn from historical linguistics, such as the belief that language changes internally as speakers become separated across time and space, and any variation among dialects can be traced to the influence of these two factors. Variation within a dialect can then be attributed to either dialect mixing, the existence in one locality of two or more dialects which allow a speaker to draw now on one dialect and then on the other, or to free variation, random linguistic fluctuation of no significance (Wardhaugh, 136). This view of variation is ultimately unsatisfactory. No suitable theory has ever been proposed to explain either dialect mixing or free variation, and it fails to account for the fact that so-called free variation, upon closer examination, is not random at all, but related to complex linguistic and social factors.

The methodology of dialect geography also left a lot to be desired. Because it had its origins in historical linguistics, it was mostly interested in finding the most historical form of a language in a given area. It focused almost exclusively on rural areas, which were thought to be more linguistically conservative and generally easier to survey. By extension, the most conservative speakers in these rural areas were seen as the oldest, most socially isolated people, who had had little formal education. These informants were prized above all others, and dialect geographers based their studies mostly upon the speech of these kind of people. Furthermore, the labels "socially isolated" and "little formal education" (among others) were based solely on the judgements of the person collecting the data, who often gave no clue in their work as to how or why these labels were applied (Wardhaugh, 134 Wolfram 1969).

There was very little scientific rigor in the sampling of these populations; the areas and informants chosen for study and the classification of these informants were all based upon the biases of the dialect geographer, rather than upon any objective, replicable set of methods. Dialect geography also relied on the notion of a community of speakers (or "speech community") that was isolated and homogeneous. As population shifted to heterogeneous urban centers, the need for more complex analyses became stronger. Also, dialect geography does not address the fact that speech is affected by more than just one's geographic origin. Something more was needed.

Enter quantitative sociolinguistics, which attempts to find patterns relating social and linguistic variation rather than dismissing it as random. It focuses on the study of linguistic variables, and attempts to look at language variation in an objective, quantifiable way. There are two main types of variables: ones that have discrete variants, and ones that must be measured on some kind of scale or continuum. For example, the / E / variable, in words like *swimming* and *fishing*, can either be realized as /n/ [sw=m3n] or as /E/ [sw=m3E]. Other variables, such as amount of nasalization, or fronting of vowels, might be measured using some kind of weighting system. Being able to quantify linguistic variation in this way

allows linguists to measure and compare the rates at which different variants occur, with the ultimate goal of relating these different variants to a range of linguistic and social factors.

Once a linguistic variable has been identified, the next issue becomes how to collect data correlating these linguistic variables with social variables. The focus of sociolinguistics is expansive far beyond that of dialect geography, examining both the rural and the urban, as well as upper and lower classes, male and female, and young and old. These speech communities are based not just on geography, but also on group membership and social ties. Sociolinguists have looked to the methodology of sociology to help identify and quantify these variables.

Certain variables are fairly easy to quantify—age for example, and sex (although, it should be noted, gender encompasses far more than biological sex, a matter which sociolinguistics has just begun to address). Race and ethnicity are harder, because they are much more subjective, and social class is even more complicated. Linguists use a number of different scales for classifying people when they try to place individuals within a social system, similar to the ways in which sociologists have done this. They use a variety of factors including but not limited to: occupation, education, income level, and type and location of housing. They also assign different weights to these factors, depending on the relative importance they decide each has to the others (Wardhaugh, 143, see Labov 1966 or Trudgill 1974 for examples).

This means that the resulting social-class designation given to an individual may vary from study to study. These designations end up being somewhat subjective, not completely unlike the judgements of dialect geographers described above. Also, these class categories are outwardly imposed, not necessarily a part of how the subjects would identify themselves. Despite these problems, sociolinguists have found many interesting patterns relating language to social class, so they continue to use this variable along with more straightforward variables like age, region and sex. We can see, however, that this is one variable which may complicate data sharing, especially among linguists who use

dramatically different scales to measure social class. This arises as a complication in TIMIT for another reason; social demographic data is not given at all. The closest factor we have to social class in TIMIT is education, which is only one factor of several that linguists typically use to determine class.

Once a sociolinguist has decided which social variables should be taken into account and how they should be classified, and formed a hypothesis about possible relationships between social and linguistic variation, the next task is collecting data to confirm or refute this hypothesis. He or she must make a plan to elicit relevant data and then collect such data from a representative sample of speakers. And in doing this, a sociolinguist should try to obtain data that is as objective as possible. None of this is easy.

One of the first problems that arises is what is known as the "observer's paradox." How can sociolinguists be sure that the data they collect is not contaminated by the investigation process itself? How do they know that the way in which a person speaks to an outsider is going to be similar to the way in which that person speaks to their friends, their relatives, their co-workers?

Researchers deal with this problem by trying to elicit a wide variety of speech from subjects, using different kinds of interview circumstances. There are four main distinctions that sociolinguists typically use to distinguish interview situations: 1) a casual situation, such as speech outside the formal interview; conversation with someone not doing the interviewing; the recall of childhood rhymes; or the narration of a story about feeling like one's life was in danger 2) a formal interview situation 3) the reading aloud of a story and 4) the reading aloud of lists and pairs of words. (Wardhaugh, 149). An interview situation which elicits all of these behaviors will cover very casual speech to very formal speech.

Not only do linguists hope to catch subjects off-guard in order to obtain speech that shows the least observer effects (category one, the "vernacular"), but they also get a chance to see the effects of these different speech styles by comparing data from across the four. Style-switching has been shown to have pronounced effects on linguistic variation in a

number of studies, and is a common phenomenon in everyone's daily life, as people talk differently to their boss than they do to their best friend.

Although all of these situations give us useful data that represent some aspect of a subject's speech, category one speech has long been prized as the most desirable data, the "truest" kind of speech. This brings up what may be another problem with data sharing; linguists may have very different ways of eliciting speech in category one (and perhaps the other three), and might not trust other linguists' methods. The TIMIT data is category three or four, speakers reading sentences aloud, and is not rife with the methodological problems involved with collecting the "pure" data of category one. However, linguists might still have some questions about how it was collected. Were speakers allowed to read the sentences over beforehand? Could they practice them? We are unable to answer these questions, which is certainly less than ideal. A more ideal situation would be a corpus with detailed notes about how data was collected, so that even if a linguist did not agree with the methods, he or she would know exactly how he or she disagreed, and could take that into consideration when analyzing the data.

Another part of a linguist's task is choosing a sample: finding a representative group of speakers. The best kind of sample is a random one, with which the judgements of the investigator do not interfere. However, most sociolinguists use a judgement sample instead, in which the investigator chooses subjects based on a set of criteria such as age, sex, or class. This is a result of quantitative studies striving for representativeness, which is important if a study is to be statistically sound. A study that randomly selects two men and eighteen women as subjects, and then wishes to talk about language variation and sex, is less statistically sound than one that judgmentally selects ten men and ten women. Also, if a researcher is looking for speech representative of Detroit natives, he or she probably wants to make sure the subjects selected have lived in Detroit longer than a year.

This problem decreases with larger subject pools—a study that randomly selected 200 men and 1800 women would be fairly statistically sound, and if most people in Detroit

have lived there a long time, a few one-year residents in a large sample would not make much of a statistical impact. Unfortunately, subject pools of this size are not often an option for sociolinguists, so they continue to use judgement samples. One advantage of sharing data is that a linguist may not be getting a random sample, but at least it is probably free of his or her own biases. Also, some corpora (like TIMIT) are much larger than the ones linguists typically handle, and give the statistical advantages offered by large subject pools. Another advantage to large subject pools is that certain speech constructions are relatively rare, making it necessary to have a huge data set to obtain a significant number of them.

It is important to consider, however, that TIMIT may have a large subject pool, but it only contains a small sample of speech from each speaker (about one minute). This is the complete opposite of more standard sociolinguistic studies, which have continued to focus on the idea of speech communities created by shared group membership and social ties. In order to accurately represent these speech communities, sociolinguists typically collect a lot of data from what they decide are a few representative speakers. Thus, the more traditional approach gives us a very complete picture of the speech of a few people who all share a speech community, whereas with TIMIT we get a brief sketch of the speech of a lot of people from a range of speech communities. TIMIT would not be very useful for someone looking to do a small community-based study, but it would be useful for linguists interested in finding broad patterns that emerge from a lot of people's speech.

Quantitative sociolinguists need to try to employ proper statistical procedures not just in sampling but also in what comes next, analyzing data and testing hypotheses. Summary statistics and more complicated statistical tests[3] are used to support or reject initial hypotheses. But although linguists use these tests and show their results, very little of the raw data makes its way into published papers. Different sociolinguists may choose to handle even similar sets of data very differently; for example, one linguist might group African Americans and whites of the same social class together, while another might look at them separately. Or, as with /t,d/ deletion, different kinds of data might be thrown out. In

our study, we chose not to look at contractions (like *couldn't)* as potential environments for deletion.[4] Differences in data handling may lead to very different results and conclusions. This is another way in which data sharing contributes to good scholarship in sociolinguistics; someone else's results have much more credibility if we can manipulate the data and obtain similar results for ourselves.

This also brings up the question of how objective and consistent measurements of linguistic data can be. Even with the same exact set of data, two observers may make different observations—one may hear an aspirated [t], where another thinks it sounds unaspirated. Good studies strive for consistency in data annotation, and may have some kind of test of interannotator agreement, involving reannotation of a certain percentage (say 5%) of data to see how much data coding varies from observer to observer.

But even with these kinds of tests, questions arise about how objective the second annotator truly is, and even about which samples are picked for reannotation. How do we know that they are not the least ambiguous samples? Other ways of solving the problem of annotator disagreement, such as acoustic analyses, are extremely time consuming and always have cases of ambiguity. Again we see that data sharing gives more credibility to sociolinguistic studies, this time by allowing measurements of annotator agreement by objective outsiders. With this project, not only will outside linguists have access to our data, they will also have access to our annotation. This way they can both listen to the data and decide if they agree or disagree with the choices we made in annotation.

It is important to keep in mind that even at this point, many of the practices that are designed to give more objectivity and scientific soundness to sociolinguistic studies (such as the use of statistical techniques or tests of interannotator agreement) are not uniform among sociolinguists. Linguists who are wary about these practices are likely to be even more wary of the prospect of sharing their data. But if a linguist does not trust an outsider to look at his or her data and come up with similar results, what does that say about the validity of his or her conclusions?

Some linguists, like Bill Labov, are against sharing data because of personal commitments to a speech community with which they have fostered a relationship. Labov tells his subjects what their data is going to be used for, and who, exactly, will have access to it. While no-one is going to lambaste Labov for not breaking trust with his subjects, it seems like there are also sensitive, committed ways of sharing data. Linguists could ask subjects if the data they give could be shared with other linguists, and explain the kinds of studies it could be used for. If a speaker does not want his or her data shared, or used for particular kinds of studies, then keeping that promise is important; but if a speaker doesn't care, neither should a linguist. Also, certain sets of data are simply not as sensitive as others. A two year study of a Harlem neighborhood or a Navajo village, where a linguist might have to fight to gain speakers' trust, is just much more sensitive than a database consisting of casual ten-minute telephone interviews with subjects around the country.

Other linguists have a somewhat antagonistic view of data sharing, based on the difficulty involved in collecting databases. This "collect your own damn data" attitude, while understandable, is not particularly productive. It is hard for the field of sociolinguistics to advance if linguists regard one another with suspicion and hostility instead of as seeing each other as valued colleagues. Full credit should certainly be given to the sources of corpora whenever they are used; also, no one is asking linguists to share data they are still in the midst of using. These considerations might help assuage some linguists' fears.

Although data sharing has not really been a part of sociolinguistic methodology thus far, it has too many compelling benefits for anyone to dismiss it lightly. As we have seen in this section, other than providing useful additional data for linguists, data sharing is also an important step in increasing the academic rigor of this relatively young field.

1.2. The variable of /t,d/ deletion

Coronal stop (or /t,d/) deletion in word-final consonant clusters such as *just* [.Zst], *hand* [h‹nd], or *packed* [p‹kt] occurs with varying frequency depending on a number of internal and external constraints. It is part of a larger pattern of final stop deletion, (others include -*sp* and -*sk* clusters) but word-final /t,d/ is particularly interesting because 1) it is so common that it is relatively easy to obtain a large amount of relevant data 2) most past tense verbs are formed in English using an -*ed* suffix, which provides an interesting grammatical context that other clusters do not (Labov et al. 1968, Neu 1980, Guy 1990). This deletion process is universally variable (every speaker deletes some of the time) and universally constrained (every speaker will have more deletion before consonants than vowels, etc.) (Guy 1994).

/T,d/ deletion is an ideal candidate for the purposes of our project because it has been examined in any number of quantitative sociolinguistic studies, and the findings have been very consistent with each other. These past studies provide a good baseline of comparison, and yet there are still interesting questions to explore with /t,d/ deletion, both with this variable and in linguistic theory.

The many studies of /t,d/ deletion have suggested that deletion rate is influenced by both internal constraints, such as phonological, grammatical and prosodic factors, and external constraints, such as speaking style, social class, and race. We will review these studies and their findings briefly here.

1.2.1 Internal constraints

<u>Grammatical conditioning</u>

Many of the words containing a final consonant cluster ending in a [t] or [d] are past tense verbs, formed by suffixing an -*ed* ending to a root verb (*laughed, missed, packed*) (note that the orthographic *e* is not usually pronounced). These bimorphemes can be directly compared to monomorphemes such as *loft, mist* and *pact.* All studies that have examined this variable have found the deletion rate of /t,d/ in bimorphemes to be

considerably less than that of monomorphemes (see Labov and Cohen 1967, Wolfram 1969, Fasold 1972, Guy 1980, Neu 1980, and Boberg 1993 for examples). It has been argued that the lower deletion rate in past tense verbs may be a result of the increased functional load of the [t] or [d] in conveying "past tense" meaning, and avoiding the potential ambiguity that could result from its deletion.

However, there is also a third category examined by most studies of /t,d/ deletion, composed of verbs that form their past tense by suffixation of /t,d/ as well as some kind of stem change (*sold, heard*, and *kept* are examples of this; in the literature they are called "semiweak" or "ambiguous" verbs). This class of verbs is small but important, containing many frequently-used verbs. Most studies that have examined this ambiguous class of verbs have found that their deletion rate is lower than that of monomorphemes, but higher than that of regular past tense verbs (see Labov et al. 1968, Labov 1975, Guy 1980). Neu did not find a significant difference between the deletion rate of regular verbs and ambiguous verbs, and Boberg found no significant difference between the deletion rates of ambiguous verbs and monomorphemes, but both admitted in their studies that this was probably the result of a small data pool.

Greg Guy (1994, 2000) has found an exponential relationship in several sets of /t,d/ deletion data, with monomorphemes retaining /t,d/ at approximately the cube of the rate of regular past tense verbs, and irregular verbs retaining /t,d/ at about the square of the rate of regular verbs. He has used these data to support his theory of variable lexical phonology: monomorphemes are exposed to deletion at all derivational levels, semiweak verbs only after level one of the lexicon, and regular past tense verbs only exposed post-lexically (Guy 2000). The more the exposures, he concludes, the higher the rate of deletion, forming this structured, exponential pattern. This is an interesting theory, but not one that has been explored by many other studies of /t,d/ deletion.

Another question that has been asked is whether deletion is higher in preterit forms of regular past tense verbs than it is in adjectival or participial forms. The loss of the /t,d/ in

these adjectival or participial forms (such as *the divorced men*, or *they have divorced*) could be less crucial than preterits (such as *they divorced*) because it does not have as heavy a functional load, and its deletion would not necessarily lead to ambiguity. Fasold examined this and found that the deletion rate of adjectival and participial forms was slightly higher (55.4%) as compared to preterit forms (49.5%) but he did not find these numbers to be significant in a Chi-square test. Wolfram (1969) found that the deletion rate of regular past tense forms was consistent regardless of its grammatical function, but did not give any data to support this finding.

Phonological factors

There are two phonological factors which have a strong influence on /t,d/ deletion, the segment following the [t] or [d], and the segment preceding it. Various studies have looked at these two factors in very different ways.

- Following environment

The most important distinction in following environment is that of a consonant versus a vowel; a following consonant encourages deletion (*best friend*), a following vowel discourages it (*worst enemy*). Early studies, such as Labov et al. (1967) and Wolfram (1969) focused on this distinction alone. Later, the effects of a following pause were also examined, and found to be dependent on geographical region (the only variable that appears to be so linked). Following pause seems to inhibit deletion for some speakers, such as Detroit natives and Philadelphians, but to encourage it for others, such as New Yorkers (Fasold 1972, Guy 1980).

Other studies began to look at following environment with more detail, and found that the deletion rates of following glides and liquids were between those of consonants and vowels, which can perhaps be explained by their manner of articulation, which is more vowel-like than consonants but more consonant-like than vowels (Guy 1980, Neu 1980).

It has been suggested that the effects of following segment may be (at least partially) a result of resyllabification (Guy 1991, Reynolds 1994). Thus, a following vowel would inhibit deletion but encourage resyllabification, which would get rid of a complex coda and form instead a more favorable CV- onset (for example, *shocked Ann* would become *shock dAnn*, but *shocked Steve* cannot become *shock dSteve*). Based on this analysis, one would expect that certain liquids and glides would pattern differently, which can be summarized as follows:

Possible resyllabification	No possible resyllabification
rhotics (r) (*train*)	laterals (l) *(tlain)
w + unrounded vowels (*twist*)	w + rounded vowels *(twube)
y + rounded vowels (*tube*)**	y + unrounded vowels *(tyist)

**some pronunciations

In an investigation of resyllabification, Labov (1995) found that following rhotics did have a very low deletion rate, as did following *y*. However, following *w* had a very high deletion rate, much higher, in fact, than the deletion rate of following laterals. This evidence against resyllabification may be partially due to the fact that he did not look at the glide-vowel combinations, but in this paper, Labov states a lot of other convincing evidence against resyllabification. Resyllabification, he says, is usually considered to apply to a single consonant between two vowels; using it to explain /t,d/ deletion in final consonant clusters is trying to expand it in a direction never before predicted (9). He also listened closely to instances of speech where resyllabification might have occurred, listening for the aspiration that occurs with word-initial [t] in English. He found very weak evidence for it, except in the case of following *y*, where it sometimes occurs (but also, it should be noted, as a minority event).

- Preceding environment

This factor has been looked at in a variety of ways; most studies have examined the manner of articulation of the preceding segment (Wolfram 1969, Fasold 1972, Guy 1980). This has led to somewhat messy results, with lowest deletion occurring after stops, and then

after nasals for some speakers, but after fricatives for others, which has been linked to factors as diverse as geography and gender (Guy 1980, Neu 1980). Recently, place of articulation has also been considered, and seems to have a significant effect. Deletion after alveolar fricatives and nasals ([s], [z], [n]) is higher than after non-alveolars (Neu 1980, Boberg 1993). This may partially explain the hodgepodge of results from just looking at the manner of articulation—if one group of speakers has more alveolar nasals (or fricatives) than the other, nasals (or fricatives) may appear to favor deletion more.

It has been proposed by Guy and Boberg (1997) that preceding alveolars may favor deletion due to the Obligatory Contour Principle (OCP), where back-to-back segments that share too many features are undesirable. [t] or [d] share more features with the alveolar nasal [n], the alveolar fricatives [s] and [z] (they group the postalveolars [R] and [Å] with these as well) and the stops [p] [b] [k] and [g] than they share with any other segments. This can be seen below:

Segment	Features
[t,d]	[+coronal, -sonorant, -continuant]
[s,z,R,Å]	[+cor, -son, +cont]
[n]	[+cor, +son, -cont]
[p,b,k,g]	[-cor, -son, -cont]

Thus, these three preceding environments might favor deletion more than any others, which Guy and Boberg indeed found to be the case.

The OCP may also be responsible for the slightly higher deletion rate of [d] over [t] observed by Guy and Boberg—most of the time, a voiceless consonant is followed by [t] and a voiced consonant by [d], but in clusters with [l], this is not the case. Thus, we can have both the word *cold* and *colt,* and because [l] and [d] also share a [+voiced] feature, [d] is more likely to be deleted than [t] in these clusters, which might lead to a slightly higher deletion rate overall.

Another, similar factor that has been observed is that the deletion rate of /t,d/ is higher when the preceding and the following environment are the same (Wolfram 1969, Neu 1980). This may be due in part to the OCP, or to some other factor.

- Other phonological factors

There are several other phonological factors which have been shown to play a role in /t,d/ deletion, such as stress, cluster length and cluster complexity. As these are all relatively low-level constraints, they are not going to be directly examined by this study, but are worth mentioning. Depending on the results we obtain from this study, we may wish to expand our focus to include some or all of them in the next corpus we examine.

Fasold (1972) looked at the effects of stress and found that unstressed syllables encourage deletion (*fo.cused*), while stressed syllables discourage it (*tossed*).
In his1980 study, Guy briefly mentions that words with triple clusters (*next, edged, lapsed*) show a higher probability of deletion than double clusters (*mind, shopped, lift*).

He also mentions that deletion appears to be higher in clusters that show more articulatory complexity, which he measures in the number of changes in place of articulation required to make the cluster. Thus, an [st] cluster, with no changes in place of articulation, is easier to produce than an [ft] or [kt] cluster, which each require one. More difficult to produce would be a [skt] cluster (as in *asked*), with two changes. If this is extended to include the place of articulation of the following word, this factor can be measured from 0 (as in *missed out*) to 4 (as in *asked Brian*). His preliminary results confirmed that higher articulatory complexity encouraged deletion. However, like stress and complexity of cluster, this is a fairly low-level constraint and not examined in most /t,d/ deletion studies.

Prosodic constraints

· Speaking style

Style seems to have a fairly important effect on /t,d/ deletion, with higher deletion rates in casual styles like conversation between groups of friends, and lower deletion rates in formal styles like reading aloud (see Labov et al. 1967, Wolfram 1969, Fasold 1972, Guy 1980). This is expected, given what we know about style shifting in other cases of stable variation (Trudgill, 94).

Deletion also seems to increase in proportion to increased speech rate, but there is no set of standards for measuring and coding rate of speech, so it is difficult to look at this factor quantitatively (Guy 1980).

1.2.2. External constraints

· Social class

As is the case with most sociolinguistic studies, social class in /t,d/ deletion studies has been determined in different ways. Also, many different groups have been compared, for example, "middle class" to "working class" in Wolfram (1969), and "upper class" to "working class" in Fasold (1972). Despite these differences, however, most studies have observed higher deletion rates in lower classes than in higher classes (Labov et al. 1967, Wolfram 1969, Fasold 1972). Again, this is to be expected, based on what we know about the correlation between social class and the use of stable variants (Trudgill, 34).

· Race

Many of the early studies of /t,d/ deletion focused on African American communities in urban centers—New York (Labov et al. 1967), Detroit (Wolfram 1969) and Washington DC (Fasold 1972). It is important to keep in mind that while African American Vernacular English (AAVE) is by no means homogenous, and although not all African Americans use the vernacular, African American speakers have been shown to share

certain linguistic features (Trudgill, 49-50). With /t,d/ deletion, black speakers seem to have a higher overall deletion rate and a different ranking of /t,d/ deletion constraints from speakers of standard white English. Labov et al. (1968) found that in spontaneous speech, the phonological constraint (following consonant vs. following vowel) is dominant over the grammatical constraint (monomorpheme vs. bimorpheme) for most black speakers. This is the exact opposite of what is found in the spontaneous speech of white speakers.

Wolfram also looked to see if racial isolation would have any effects on /t,d/ deletion, and found no strong correlations.

· Geographical background

As stated above, it appears that the only feature that is linked to geography is the effect of following pause, with higher deletion in New York than in Philadelphia or Detroit.

· Sex

Most studies that have examined this factor have either found that women show a slightly lower rate of deletion (Wolfram 1969, Neu 1980) or that there is no significant difference between women and men (Fasold 1972). This seems to be a fairly low-level constraint.

· Age

This appears to be a somewhat complex factor, and another low-level constraint. Wolfram (1969) could not find any clear correlation between age and /t,d/ deletion, at least not without mixing in the effects of social class. Fasold (1972) found that adults deleted less than adolescents, who deleted slightly less than children.

Age seems to be related to the grammatical conditioning of /t,d/ deletion. Guy and Boyd (1990) found that Philadelphia children deleted the /t,d/ of ambiguous verbs at almost as high a rate as monomorphemes, until sometime in adolescence, where they begin to delete

at a much lower rate (due, they say, to a reanalysis of these words as different from monomorphemes). Some adults go through another change, much later in life (around 45 years and older) where the deletion of ambiguous verbs lowers dramatically again. Guy and Boyd say this is because they have reanalyzed ambiguous verbs again, as more similar to regular verbs.[5]

1.3 The importance of this study[6]

This project, known as DASL (Data and Annotations for Sociolinguistics), will eventually investigate the process of /t,d/ deletion in four large digital speech corpora, for the purpose of assessing some of the benefits and drawbacks involved in using shared databases.

There are many compelling reasons to share data, some of which were discussed in section 1.1. Data sharing allows linguists to work with corpora that may not be collected completely objectively, but at least are probably free of their own biases. It also allows linguists to observe each others' data classification, annotation and manipulation, which gives their results more credibility.

There are many additional reasons to share data. Sociolinguistic corpora are both expensive and time-consuming to collect, which is a good reason to reuse them rather than let them sit on a shelf to gather dust. Shared data particularly benefit young researchers who rarely have the time or the financial resources to collect their own data, yet still want the valuable skills gained from working with real corpora. Also, shared data are very beneficial to a linguist who may just want a low-cost way to test a theory.

Old databases that other linguists have put aside might be invaluable to one's own research, especially as it is sometimes impossible to obtain similar data sets. For example, if a researcher wishes to know how the speech in a community has been changing over the past forty years, tapes of data from the 1960s or 70s would be incredibly helpful, but physically impossible to collect anew.

There is also a lot of value in having stable data sets to which the entire sociolinguistic community has access. When each linguist is using his or her own corpora to test his or her own theories, it can be hard to make cross-comparisons and assess the weaknesses and strengths of particular theories. Stable data sets that are widely available could be used to test competing theories by doing direct comparisons: How descriptively adequate is Theory A in describing patterns found in the data set, versus Theories B and C? Also, linguists could multiply annotate the same data set for more variables, which would be a resource for testing whether different variables display similar patterns.

While using corpora from other sociolinguistic studies seems like it would be the most helpful for linguists, there are also benefits involved in using corpora like the ones examined here.

These corpora were collected using grants and commercial funding, in order to support speech engineering technology development. Some of them can be used free of charge, while others come at a cost far less expensive than the cost of collection. Essentially, with these corpora, the sociolinguistic community gets access to data intended and paid for by others. Above and beyond the cost and time saving is the unique nature of these databases. These corpora are larger than linguists would typically have access to, and have much larger, geographically diffuse subject pools.

However, while the benefits of shared data are many, it is important to note that sharing data does not diminish the value of new data collection. Both the researcher and the research community benefit from new contributions. The researcher gains skills and a unique appreciation of the subject pool that can only be developed through spending time with the speech community he or she is studying. The research community gains not only a new set of data but also, hopefully, new perspectives and methodological approaches.

2.0 Study overview and methodology

2.1 The focus of this study

DASL will eventually examine four published, publicly-available corpora, each created for the development and testing of speech technology tools. A team of annotators, headed by myself, will code the corpora for /t,d/ deletion, and then analyze the data using summary statistics, as well as statistical tools such as Chi-square and VARBRUL, a multi-variate analysis package used in sociolinguistic research. Since the corpora and annotations used in this study are accessible via the worldwide web, interannotator agreement can be measured not just by the annotation team, but by the larger sociolinguistic community as well.

In addition to the empirical study of /t,d/ deletion and the methodological questions about the use of shared speech corpora in sociolinguistics, this project will address several additional questions. How do the corpora used in this study relate to the data most commonly used in quantitative sociolinguistics (i.e., recordings of sociolinguistic interviews)? Do the insights gained from the large-scale study of a geographically diffuse subject pool differ qualitatively from speech community studies? What is the rate of interannotator consistency for the task of coding /t,d/ deletion? And can studies of similar variables be organized on a large scale with teams of non-specialist annotators?

The corpora to be examined are:

- The TIMIT Acoustic Continuous Speech Corpus—groups of phonetically rich sentences (722 distinct sentences), read by 630 speakers from across the United States
- Switchboard—2400 five-minute telephone conversations among 543 distinct speakers who don't know each other and who speak about an assigned topic
- CallHome American English—120 thirty-minute telephone calls among pairs of family members or close friends
- 1996 American English Broadcast News Speech—104 hours of television and radio broadcast from 11 different American news programs

The data have already been transcribed (in standard orthographic English for all four, and also phonetically for TIMIT) and segmented so that speech can be retrieved in separate blocks. TIMIT and Switchboard data are accessible word by word; in the other corpora, data are accessible by speaker turns, defined by how long the speaker speaks before another speaker interrupts, or takes his or her turn. Within long speaker turns, individual pause groups are segmented.

Table 1: Comparison of the four corpora in amount and type of data.

Corpus	**Minutes**	**# of Speakers**	**Data Type**
TIMIT	630	630	Phonetically rich sentences
Switchboard-1	12000	543	Short conversations among strangers on constrained topics
CallHome American English	1200	~240	Long conversations among intimates on free topics
American English Broadcast News	6240	Currently unknown; ~500+	Broadcast news

At this point in the study, only the TIMIT corpus has been fully annotated and analyzed for /t,d/ deletion. Eventually, this project will present a more comprehensive picture of the use of corpora than only one corpus can give us, but that does not mean that we can not benefit from examining this database alone. On the contrary, by examining /t,d/ deletion in just TIMIT, we can both focus in on more detail and see how to change our approach as we handle the next three corpora.

Also, although we will probably train additional annotators to help me work through the next three corpora, TIMIT was small enough (2059 tokens) that I was able to annotate it by myself. The results presented in this paper are based on my own annotation, with independent annotators currently in the process of checking between 5-10% of tokens for agreement.

This particular corpus consists of employees of Texas Instruments and MIT (hence, TIMIT) reading groups of phonetically-rich sentences aloud. The data from this corpus is probably similar to some of the most formal, text-reading data obtained by sociolinguists in studies such as Labov et al. (1967), Wolfram (1969) and Fasold (1972). These are highly self-conscious speakers, reading sentences they probably wouldn't say in the course of a normal conversation (for example, *Irish youngsters enjoy fresh kippers for breakfast*). We can expect that the /t,d/ deletion rate might be rather low overall, since these speakers are probably going to be focused on their enunciation and pronunciation. /t,d/ deletion, however, is not a variable that speakers have shown much awareness of (unlike, say, the rhoticization of the speech of Boston natives, in sentences like *Pahk the cah at Hahvahd Yahd*). Thus, we cannot really expect that speakers will be intentionally not deleting /t,d/ in any kind of a conscious manner, although we can expect some style-shifting to take place, depending on the comfort level of the speakers.

2.2 The annotation procedure

The annotation tool used in this project is a multi-purpose, interactive interface: it selects tokens of potential interest, displays them, and allows users to code them, save their results and export these results to a spreadsheet or statistical analysis package.

First, we used the interface to isolate sentences of potential interest from the corpus. Since TIMIT is segmented at the word level, if we can find an utterance relevant to /t,d/ deletion, locating the corresponding speech is simple. We searched the orthographic transcripts of the speech files of TIMIT, choosing not to use its phonetic transcript in order to obtain consistency with the other three corpora (which do not have phonetic transcripts).

Because we used orthographic transcripts, our queries are somewhat complicated. The annotation tool accepts a query in the form of a regular expression, which sorts through a file and picks out examples that fit a flexible pattern. For example, the regular expression [Oh s.*t.] will match strings like "**Oh s**ay can you see by the dawn's early ligh**t**" and "**Oh**

shoo**t**!" (Wall et al, 1996). Regular expressions must be formed in accordance with a particular syntax, which varies slightly depending on which programming language one uses.[7]

In the future we would like to use a pronouncing lexicon as an intermediary to the search. A pronouncing lexicon is a dictionary of all the words used in a corpus, matched with their phonetic transcriptions (and sometimes morphological analyses as well). This would provide the list of English words susceptible to /t,d/ deletion in each corpus. The interface would then search for those words in the transcripts, making the search easier and more efficient. Although orthographic transcripts work mostly satisfactorily in providing tokens of /t,d/ deletion, it would be extremely difficult, if not impossible, to formulate a regular expression that would search orthographic transcripts for certain phonetic variables, especially variation in vowels.

We formulated a regular expression query[8] to give us all examples of /t,d/ at the end of a word that follow a consonant or the letter *e* (because we want past tense forms like *packed* [pʻkt]) which are not followed by a *t* or *d* (which makes it nearly impossible to tell if the /t,d/ has been deleted). We also formulated the query so it would exclude contracted forms like *couldn't* and *haven't,* which are a strange case—they have a very high deletion rate (Fasold 1972) but are not quite monomorphemes, because they are composed of a verb plus a contracted negative particle.

Orthography and pronunciation are not a perfect match; this query gave us all applicable examples, but also many that were not applicable. Some of the most glaring "false hits" were removed by a series of filters; during the process of annotation I just ignored those that remained. Three filters were created. One removed all cases of *and*,[9] which has been shown to have unusually high rates of both deletion and occurrence. In Neu's 1980 study, *and* accounted for 41.0% of the tokens, and had a deletion rate of 90.0%. This does not occur with similar words, like *sand*. Almost all studies of /t,d/ deletion eliminate *and* from analysis (see Neu 1980, Boberg 1993).

Our second filter removed words with *-et* endings.[10] Although we want words ending in orthographic *-ed* so we get all tokens of past tense verbs, we do not want words ending in *-et*, because they do not contain word-final consonant clusters but [3t] (like the word *pocket* [p"k3t]).

The third filter removed words with *-ted* or *-ded* endings, like the verbs *heated* [hi~3d] and *needed* [ni~3d].[11] Again, these do not contain word-final consonant clusters but [3d] or [3t].

At this point in the process, the corpus has been sorted, filtered, and prepared for annotation. The final set of tokens selected for coding is displayed as a list in the annotation interface. Each token is shown with surrounding words and a list of factors to be coded (Figure 1). These factors can be altered easily depending on which linguistic variable is being studied and what factors examined. Each factor appears as a radio button, and to code a token we simply click on the button corresponding to the relevant factor. A comments field also appears after each token so we can record notes.

Figure 1. Example token in the annotation interface.

To hear a token, we simply click on the word containing the [t] or [d], which is played with the following word. We can also play the entire sentence in which the token occurs, or zoom in on a fragment of speech smaller than the word level. All computer platforms (PCs, Macintoshes, and UNIX systems) are capable of supporting this interface, as long as they have access to the world-wide web.

2. 3 Coding guidelines

Past studies of /t,d/ deletion have examined a wide range of factors, and coded these factors in slightly different ways. This is why it is particularly important to describe in detail the factors and coding guidelines used in our study.

We chose to examine four factor groups: status of the dependent variable; morphological category; preceding segment and following segment. Factor groups were annotated in the following ways:

The status of the dependent variable

· Deleted:

The /t,d/ segment has been completely deleted.

· Retained:

The /t,d/ segment is retained. Although a basic deleted/retained distinction was usually sufficient, in some cases the final [t] or [d] was phonetically altered. The segment was often unreleased, glottalized, or (less commonly) flapped. I coded all of these variant realizations of /t,d/ as retained, and noted the variation in the comments field.

· N/A:

Because the regular expression query used to generate the list of potential /t,d/ tokens used orthographic transcripts, even with several filters in place some words which were not tokens of /t,d/ appeared in final list of words to be reviewed. (For example, based on its orthography, the word *would* [wId] appears to be a possible /t,d/ token, but it is not.) I coded these tokens as N/A, excluding them from the final analysis.

Morphological category

· Monomorpheme:

/t,d/ appears in a word of just a single morpheme, i.e., *old.*

· Irregular/ambiguous past tense:

/t,d/ appears in an irregular past tense verb, i.e., *told.*

· Regular past tense:

/t,d/ appears in a regular past tense verb, i.e., *rolled.*

What, precisely, was considered to be part of the irregular/ambiguous past tense category changed over the course of this project and is discussed in detail in section 3.1.

We were also interested in seeing how far the "functional load" argument discussed above could be extended. With this in mind, I also coded regular past tense tokens as "preterit" (in tokens like *they married*) or "participial" (*they have been married*), making these notes in the comments field.

Preceding segment

Although many previous studies adopted a five-way distinction to categorize the preceding segment, our study adopted a seven-way distinction. We chose this finer-grained coding scheme because of the studies that indicate that preceding [s] and other alveolar segments favor deletion much more strongly than their non-alveolar counterparts (see Neu 1980, Boberg 1993, Guy and Boberg 1997, also discussed above). Thus, we chose to code preceding alveolars separately. Our seven factors are:

· Lateral [l]
· Rhotic [N]
· Alveolar nasal [n]
· Non-alveolar nasal [m] or [E]
· Stop [p,b,k,g] also affricates [.] and [V]
· Alveolar fricative [s,z]
· Non-alveolar fricative [f,v,U,R,Å]

It is important to note that we made a slightly different choice than certain other studies (specifically Guy and Boberg1997) by coding the fricatives [R] and [Å] with the non-alveolar fricatives. There were so few tokens with a preceding [R] or [Å], however, that it should not strongly influence our results.

In coding this factor group, I was sometimes faced with a situation where the preceding segment had been reduced. This typically occurred with the liquids [N] and [l], but occasionally with other phonemes as well. If the preceding segment was phonetically reduced in this way, I noted this in the comments field, but still indicated the preceding environment as [N] or [l], because the segment was phonemically present although phonetically altered.

From time to time, complete deletion of the preceding phoneme took place (for example, *government* realized as [gZvNm3t]). Here, the token was annotated according to the actual preceding sound—in this example, the preceding sound was a vowel, which made the token non-applicable. The only case where I indicated an environment that was deleted instead of what I heard was with a preceding lateral [l], because even completely vocalized laterals have been shown to effect surrounding segments. (For examples of this, see Sharon Ash's 1982 dissertation on the vocalization of [l] in Philadelphia.)

Following segment:

We broke this factor down into a seven-way distinction, specifically to see how the theory of resyllabification discussed above holds up in our corpora (see Guy 1991, Labov 1995). We distinguished possible resyllabifying environments (rhotics, clustering glides and vowels) from environments where resyllabification is not possible (obstruents, laterals, and non-clustering glides). Pause is a separate element, which is particularly interesting in its connections with geographical region. Our seven factors are:

- Obstruents (stops, fricatives and nasals)
- Lateral [l]
- Rhotic [N]
- Clustering glides ([w] + unrounded V, [ju])
- Non-clustering glides ([w] + rounded V, [j] elsewhere)
- Vowel
- Pause (silence follows /t,d/ segment)

Pause environments were distinguished in two ways. Some occurred at the end of a sentence, when the subject finished reading. These were pretty clear-cut, with no possible annotator variation. Others occurred mid-sentence, and were left up to my judgement to determine. These mid-sentence pauses often occurred after a comma or during a list of some kind, but also sometimes occurred in other places, such as when the speaker needed to take a breath. This is where interannotator agreement is especially important, since these choices could conceivably vary quite a bit.

In coding following segments, I occasionally encountered cases where the segment had been entirely deleted. This happened most frequently with [h], particularly when the following word is an unstressed pronoun (*her*, *his*). For example, the token *encouraged her* was sometimes realized as [=nk2N3gd 3N]. In such cases, I coded the following segment as a vowel, since the segment had been completely deleted, and made a note in the comments field.

2.4 Background information

When the TIMIT data were collected, some social information on the speakers were collected as well. The background information given for each speaker consists of sex, birthdate, geographical region, education level, race and height. The information on geographical region is perhaps the most problematic. Region in TIMIT is broken down into a few very large groups—Army Brat, New England, New York City, Northeast, South, North Midland, South Midland and West. Although these groupings are somewhat well-motivated (they are based, presumably, on some of the broad dialect boundaries used by dialectologists) we do not know exactly which states and regions are part of these groups because this information is hard to come by for TIMIT.[12]

Since the people collecting these data were not sociolinguists, they did not quantify background information in the same ways that linguists have. Few quantitative sociolinguists would group the dialect(s) of New Orleans with those of Atlanta and Virginia

under the broad category of "Southern." This category, they might say, would be too large to be descriptively useful. Also, we do not have information on what prompted this response—was the question “where do you live”? Or “where did you grow up”? These two questions elicit very different answers in a number of people.

Similarly, most sociolinguistic studies have some kind of factor quantifying social class. Here, the closest category we have to social class is educational level, which, as we have already discussed in detail, is in itself is no unequivocal indicator of class.

Another problem that arises looking at the background information is the low diversity of the speaker group. After all, this is no speaker group carefully hand-picked in order to obtain equal or near-equal numbers of different factors. The speakers are all employees of Texas Instruments and MIT in 1985. There are two times as many men as there are women. Everyone is fairly well-educated, with the majority of speakers having earned their Bachelor's Degree or above. There are few African American speakers and only a handful of Asian, Latino, and Native American speakers.

Instead of dismissing this somewhat problematic background information, we can instead use it with a grain of salt. For example, we clearly cannot make any generalizations about Native American /t,d/ deletion based on four tokens from two speakers, although we might be able to say something about African American deletion based on 67 tokens from 26 speakers.

As has already been discussed, it may even be good to get away from the biases of sociolinguists. This allows us to reaffirm that the categories linguists create are both genuinely useful and adequately descriptive by noting the problems (or lack thereof) that crop up with these categories. We can also question the sociolinguistic biases that go into speaker selection. One sees these sociolinguistic biases in every study one reads—for example, in Wolfram's 1969 study of /t,d/ deletion among African American speakers in Detroit, he decides to only use informants who have lived in Detroit for at least ten years; adults should have lived there for half their lives, while children and teenagers should be

native Detroiters. Also, he decides, there should be a "sufficient" amount of discourse to expect a "reasonable" variety of syntactic structures (15-16). He says he knows it restricts his study's randomness, but it is necessary. Is it really? Using a speaker sample like TIMIT, chosen and grouped according to an agenda different than ours, we can begin to ask that kind of question.

3.0 Results

Out of 1577 tokens of word-final consonant clusters ending in a [t] or [d], 518 were deleted, giving us a total deletion rate of 32.8%. Some sentences (and, by extension, tokens) occurred more than once, as can be seen in Table 2:

Table 2 : Percentage of tokens that occurred multiple times in TIMIT.

Number of occurrences	**Percentage of tokens**
1	35.4% (n=558)
2	3.6% (n=58)
6	4.2% (n=66)
7	50.6% (n=798)
14	1.7% (n=28)
All other numbers	4.4% (n=69)

This repetition of tokens (and in particular, highly unusual ones like *Encyclopedias seldom present anecdotal evidence*) may have certain effects on the results, which will be discussed in later sections. Note, however, that over one third of the data are original tokens.

A VARBRUL analysis was done on the data, the full results of which can be seen in Appendix A. I will not discuss these results in depth, except where they pertain to particular factors.

3.1 Internal constraints

Grammatical conditioning

Originally, we coded the data as monomorpheme, regular verb or irregular verb. This irregular verb category included all irregular verbs—the modal verb *must*, the lexical anomaly *went* (present tense *go*), and strong verbs like *built* (present tense *build*). This gave us a high deletion rate for irregular verbs, 41.1% (n=107), which was higher than the rate for both regular verbs (22.6%, n=513) and monomorphemes (37.4%, n=958). Since these numbers do not correlate with most of the literature, we searched for factors that could be skewing our data, giving us such a high rate of irregular verb deletion.

We noticed that the deletion rate of the modal verb *must* was much higher the total of irregular verbs (76.5%) and also that the number of *must* tokens was higher than of any other irregular verb (n=34), so we tried excluding these tokens from the total. This gave us a deletion rate of 24.7% (n=73), just above the rate of regular verbs. This correlated nicely with the results of previous studies, which usually finds the deletion rate of irregular verbs to be between that of regular verbs and monomorphemes.

Excluding *must* makes sense, because it has only one form and does not affix an *-ed* ending for a "past tense" meaning. Thus, its final [t] does not have a *+past tense* significance and deleting it would not weaken its status as a *+past* verb. We then reclassified *must* as a monomorpheme, where it did not affect the overall deletion rate. This certainly brings up the question of how past studies treated *must*, of which we found no mention in any of the literature. It seems somewhat likely that the problems we had with *must* are partially due to the unique nature of TIMIT; there was quite a bit of repetition of tokens containing *must*, which might lead to a higher occurrence rate than in normal speech.

Examining this category of irregular verbs again, we also realized that we had included strong verbs, which other studies did not—the verbs *built* (present tense *build*), *found* (*find*), *hold* (*held*), and *wind* (*wound*). Studies such as Guy (1980) have removed these verbs from the "ambiguous-irregular" category because the /t,d/ is found in the present as well as the past tense, not affixed as a *+past* suffix. Therefore, the final [t] or [d] does

not necessarily carry any *+past* functional load, and deleting it would not be the same as deleting the final coronal stop in the past tense of a semi-weak verb like *felt* (present tense *feel*). With these, the literature indicates that they were excluded from the ambiguous verb category but not what was done with them instead. We do not know if they were eliminated from the analysis or treated as monomorphemes. In our study they were reclassified as monomorphemes, which, again, did not affect the overall monomorpheme deletion rate.

Many studies have also excluded the verb *went* (present tense *go*) from the ambiguous verb category, since there is no clear vowel change and /t,d/ suffix relationship between its present and past forms (for examples of this treatment see Guy 1980, Guy and Boyd 1990, and Boberg 1993). This is a unique lexical case whose forms are perhaps memorized separately. Most studies classify *went* as a monomorpheme, which we also did.

So after redefining our irregular category to exclude *must*, *went*, and strong verbs like *built*, we obtained a deletion rate of 15.4%—now much *lower* than that of regular verbs. This does not correlate with most of the literature, but this is probably due to the greatly reduced n (only 39 tokens). A Chi-square test found that the difference between the deletion rates of regular and irregular verbs was not significant at the $p < 0.05$ level.

These findings can be summarized in Table 3.

Table 3. The deletion rate of /t,d/ in monomorphemes, regular verbs, and the three different classifications of irregular verbs.

Monomorphemes	Irregular verbs (excluding nothing)	Irregular verbs (excluding *must*)	Irregular verbs (excluding *must*, *went* and strong verbs)	Regular verbs
36.6% (n=1081)	41.1% (n=107)	24.7% (n=73)	15.4% (n=39)	22.6% (n=513)

We can see here one of the disadvantages of using this particular set of data, which only gives us 39 tokens of /t,d/ in ambiguous verbs. Because it consists of a set of sentences created and selected for phonetic richness, it contains far fewer ambiguous verbs than would probably occur in the course of a normal conversation.

Since the data do not fit any of the usual patterns, they also do not fit the exponential pattern that Greg Guy (1994, 2000) proposed as evidence for variable lexical phonology. We do not have enough data in the ambiguous verb category to be able to support or discredit his hypothesis.

We were also interested in applying the "functional load" argument to see if /t,d/ deletion would be lower in the participial forms of regular verbs (in sentences such as *They would have planned*) as opposed to the preterit forms of regular verbs (in sentences like *They planned*). (I classed adjectival forms, like *the planned meeting*, as monomorphemes, based on the difficulty involved in determining which adjectives were derived from verbs and which ones were just adjectives.)

In sentences like the participial example given above, deletion of the [d] would not make the sentence ambiguous—a participial always has a [t] or [d] ending, there can be no other interpretation of the sentence whether it is pronounced or not. However, in the preterit example, deleting the [d] would give us the present tense form of the verb and might lead to an ambiguity (although, it should be noted, there are other cues that give a sentence a *+past* reading, such as lexical items like *yesterday*). Overall, however, one might expect a higher deletion rate in participial forms than in preterit forms.

This, in fact, seems to be the case. Deletion in participial forms is 35.8% (n=137), quite a bit higher than the 17.5% deletion found in preterit forms (n=342). A Chi-square test of these data showed the difference to be significant at the $p < 0.001$ level.

Phonological factors

· Following environment

The correlation of following environment and /t,d/ deletion can be seen in Figure 3.

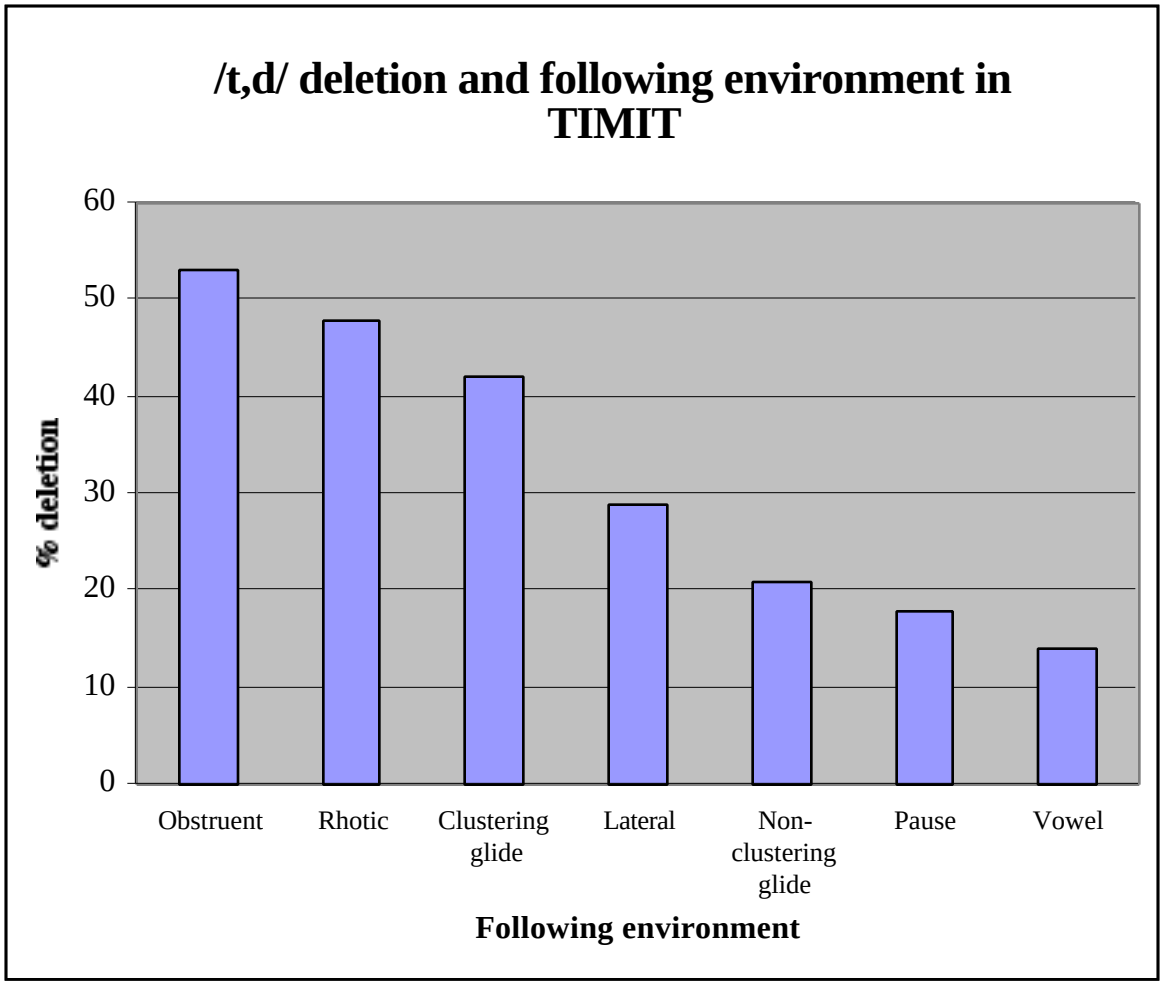

Figure 3. The correlation of following environment and /t,d/ deletion.

The deletion rate of obstruents (52.9%, n=607) was much higher than that of vowels (13.7%, n=527), with the deletion of liquids, glides, and pause in between. The broad patterns of this data conform nicely with previous studies.

Our data do not support Guy's resyllabification hypothesis; in fact, they are the exact opposite of what one would expect if resyllabification had a strong effect, with higher deletion rates for following rhotics (48.2%, n=56) than for laterals (29.4%, n=17) and higher deletion rates for so-called clustering glides (41.9%, n=105) than for non-clustering glides (21.4%, n=14). Two things should be considered about these findings, however: one, that they are based on very little data for laterals and non-clustering glides, and two, that they do not really correlate with Labov's 1995 paper either, where he found that following rhotics had a very low deletion rate.

Interestingly, as in Labov's study, following [w] and [j] seem to have very different effects on deletion, regardless of whether they are part of "clustering" or "non-clustering" glides. The deletion rate of [w] (48.2%, n=85) was much higher than that of [j] (17.6%, n=34), just as he found. A Chi-square test found this difference to be significant at the $p < 0.01$ level.

The effects of following pause will be discussed more fully in the section on geographical region.

- Preceding environment

The correlation of preceding environment and /t,d/ deletion can be seen in Figure 4.

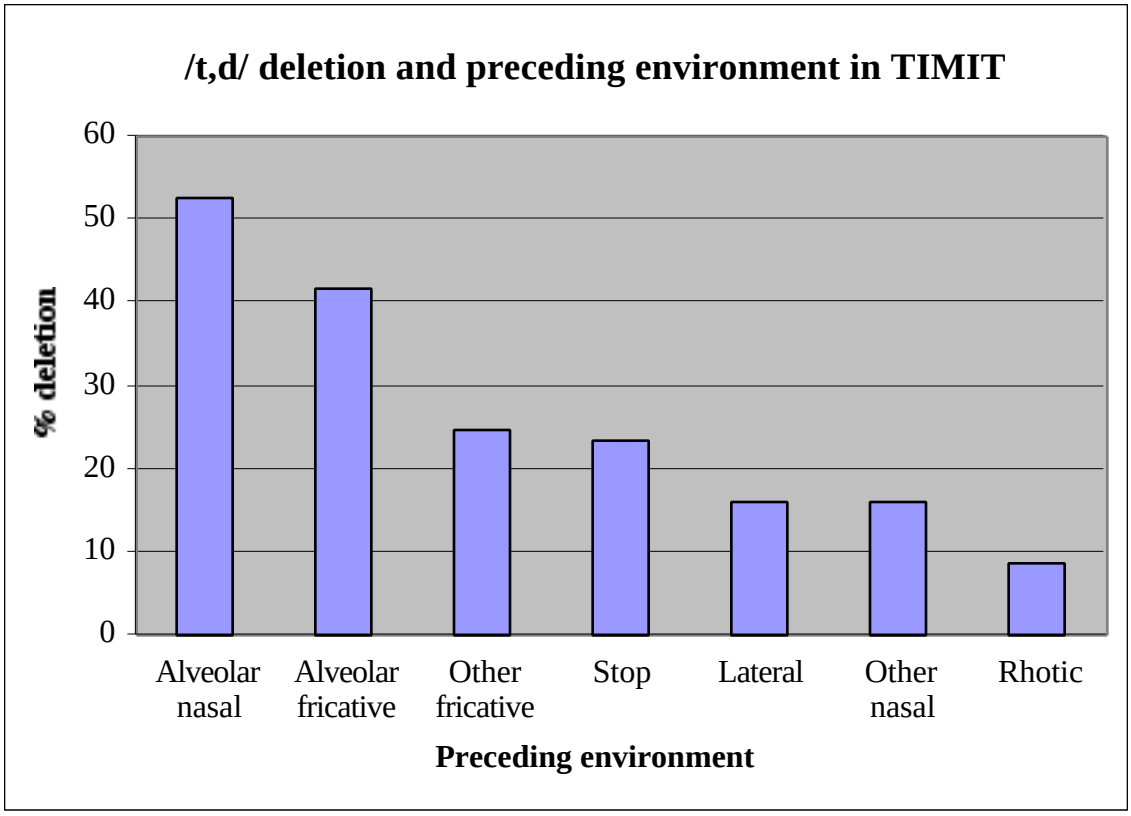

Figure 4. The correlation of preceding environment and /t,d/ deletion.

These data seem to fall into four groups—the alveolar nasals and fricatives have the highest deletion (52.7-41.7%), followed by other fricatives and stops (24.7-23.4%), followed by laterals and other nasals (16.1-16.0%), followed finally by rhotics (8.7%). Our data pattern nicely with the studies that have shown alveolar nasals and fricatives tend to encourage deletion. The deletion rate after these segments was 52.7% (n=432) and 41.7% (n=391), respectively. These data give support to the Obligatory Contour Principle, and also to Guy's conjecture that [N] is not, in fact, [+coronal], since we would expect a higher deletion rate based on the OCP if it were.

The OCP also predicts, however, a similar deletion rate for stops, alveolar nasals and fricatives, which we see is not the case here. Guy and Boberg (1997) found similar rates of deletion for all three of these preceding environments, which makes sense based on the

OCP, since stops have as many features in common with [t] and [d] as alveolar fricatives and nasals. These data suggest that maybe place of articulation is more important than manner of articulation in determining which environments are more prone to deletion. The VARBRUL data in Appendix A even suggest a three-way distinction, as the factor weight for alveolar nasals (0.756) is quite different from that of both alveolar fricatives (0.635) and stops (0.426).

· Same preceding and following segment

During the process of annotation, I noticed an unusually high rate of /t,d/ deletion where the preceding and following segments were both voiceless alveolar fricatives [s]. There were several examples of this, in phrases such as *processed soybeans* and *nearest synagogue*. We calculated deletion in this environment and found it confirmed my initial impressions, with a high deletion rate of 88.9% (n=36). Following previous studies, we looked at all the environments where the preceding and following segments were the same (*stopped passing, zoologist saw*) and found an overall deletion rate of 71.4% (n=56).

However, when the liquids [N] (*guard rail*) and [l] (*old lady*) were calculated separately, we obtained a deletion rate of 87.0% (n=46) for obstruents and 0% (n=10) for liquids. It is hard to say anything conclusive about this as it is based on only 10 tokens of liquids, but it seems that having the same preceding and following segment might strongly encourage deletion in the case of obstruents, and discourage it in the case of liquids. A Chi-square test found this difference between obstruents and liquids to be significant at the $p < 0.001$ level.

3.2 External constraints

· Education

The correlation of education and /t,d/ deletion can be seen in Figure 5.

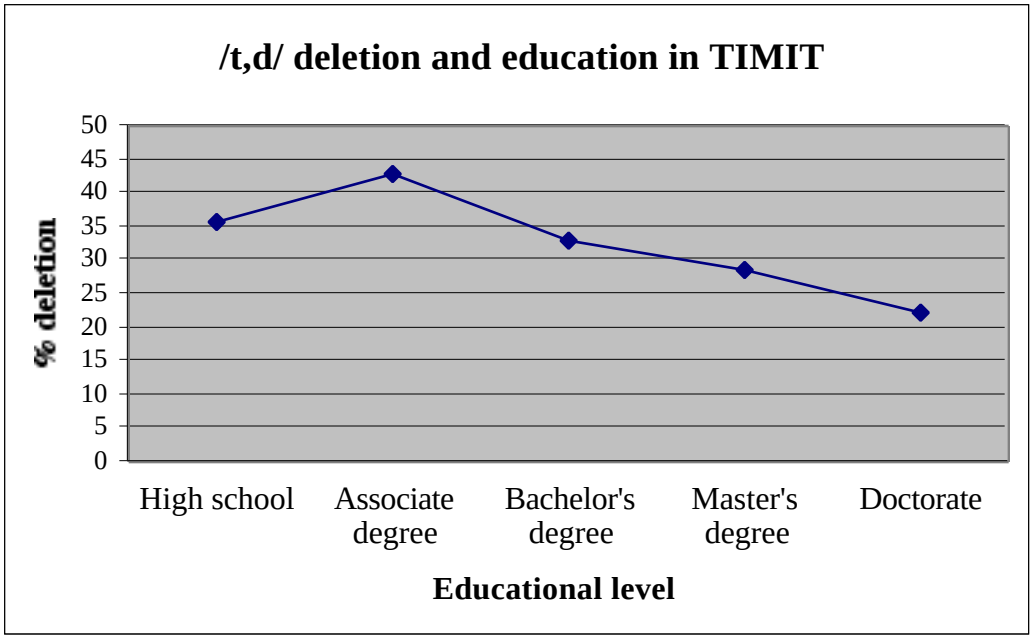

Figure 5: The correlation of education and /t,d/ deletion.

We can see from this graph that educational level and deletion seem to be related, with a decrease in deletion as educational level increases, except for a sharp increase for speakers with Associate degrees.

We may be able to explain this trend somewhat by linking education with social class. As stated earlier, education is not the only factor that determines social class, but it is an important one. /t,d/ deletion studies have consistently shown speakers of lower classes deleting more than higher classes, which follows larger patterns relating social class and variation. This is also shown in this graph, except for the somewhat perplexing numbers of deletion in individuals with Associate degrees. We cannot help but notice that this seems to be the exact opposite of the hypercorrection pattern that is often found among the lower middle class in the most formal environments (Trudgill, 95). Hypercorrection is a result of speakers in the lower middle class trying to use prestige forms of pronunciation, so much so that they overdo it. Here, we seem to have a pattern that suggests that speakers with

Associate degrees are trying not to use prestige forms of pronunciation and overdoing it. These results certainly merit a closer look and further investigation.

- Race

The correlation of race and /t,d/ deletion can be seen in Figure 6.

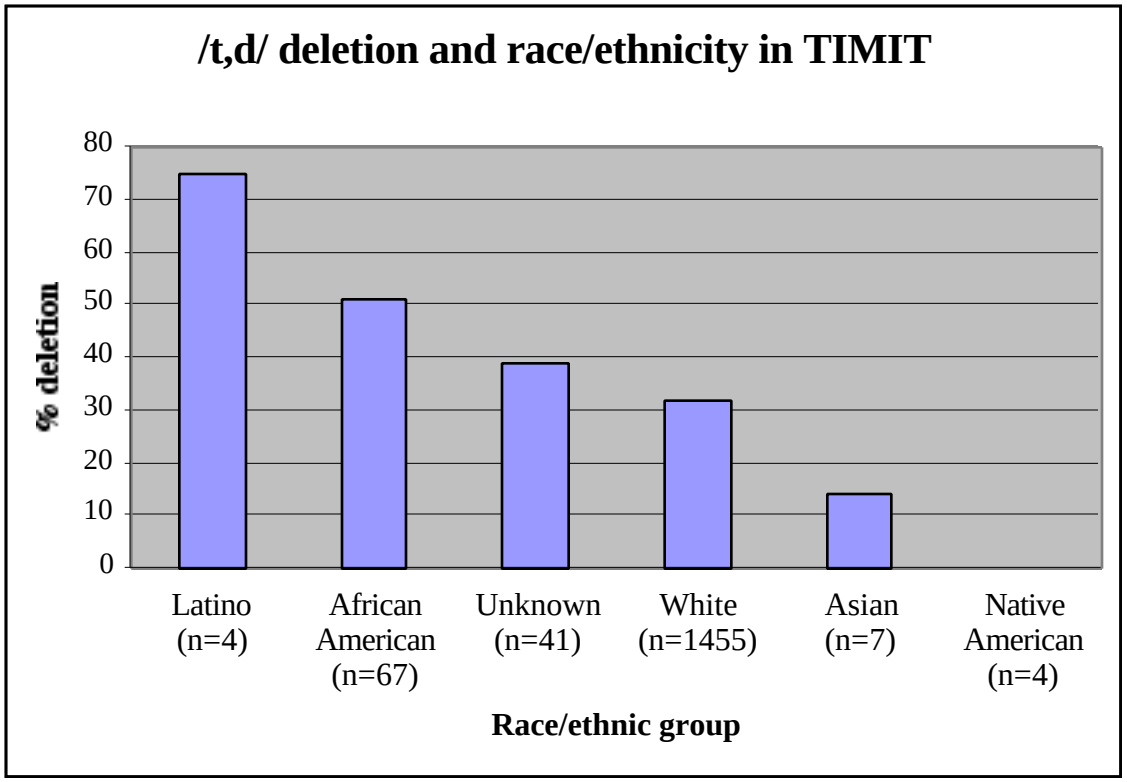

Figure 6. The correlation of race and /t,d/ deletion.

This is one of those areas where we really see the shortcomings of the lack of diversity of the subject pool. We can't really say anything about Latino, Asian, or Native American speakers, because there are just not enough tokens (four, seven and four, respectively). And the "Unknown" group, subjects who chose not to classify themselves based on race, doesn't allow us to say much of anything either. We don't know whether

these subjects were persons of color that didn't fit into the above categories, people of mixed ancestry, or just people who didn't feel like responding to this question.

The two groups that we might be able to say something about are the African American and white groups (even though the number of tokens in each group are vastly unequal, 67 from black speakers and 1455 from white speakers). African American speakers here seem to delete at a higher rate (50.7%) than white speakers (31.9%), which may be due to differences between African American Vernacular English and Standard English, or may be connected to issues of social class or region.

As we can see in Table 4, when we cross-tabulate race with our factor closest to class (education), African American speakers delete at a higher rate than white speakers in every educational level, so this pattern cannot immediately be dismissed as one of class alone. However, there are so few tokens in most of these categories that we can't rule out social class, either.

Table 4. The correlation of education with race in /t,d/ deletion.

	Tokens from white speakers	**Tokens from African American speakers**
High school	32.5% (n=169)	50.0% (n=24)
Associate degree	40.4% (n=52)	100% (n=3)
Bachelors degree	32.8% (n=822)	39.3% (n=28)
Masters degree	27.7% (n=329)	66.7% (n=6)
Doctorate	22.4% (n=58)	N/A
Unknown	56.0% (n=25)	66.7% (n=6)
Total	31.9% (n=1455)	50.7% (n=67)

Another factor that might be responsible for this higher incidence of deletion in African American speakers may be not race, but region. Most of the African American speakers used in this study were from Southern states (46 out of 67 tokens, or 68.7%, came from Southern African American speakers). As we will see in the next section, the Southern region had a higher deletion rate than other regions, so we cross-tabulated race with region to see what effect it might have.

Unfortunately, most of the regions did not have enough tokens of African American speakers to say anything significant about them (Army Brat, New England, New York, North Midland, South Midland, Northern and Western regions all had nine or fewer tokens from African American speakers). But Southern African Americans deleted /t,d/ at a rate of 58.7% (n=46), compared to 33.7% (n=208) in Southern whites. This difference was shown to be significant in a Chi-square test, with $p < 0.01$.

These data seem to indicate that there is more going on to these race figures than just class or region can explain. Higher deletion rates in African American speakers than in white speakers follows the patterns of several past studies of /t,d/ deletion (Wolfram 1969, Fasold 1972), and of variation patterns in general (Trudgill, 45-46).

- Geographical region

The correlation of region with /t,d/ deletion can be seen in Figure 7.

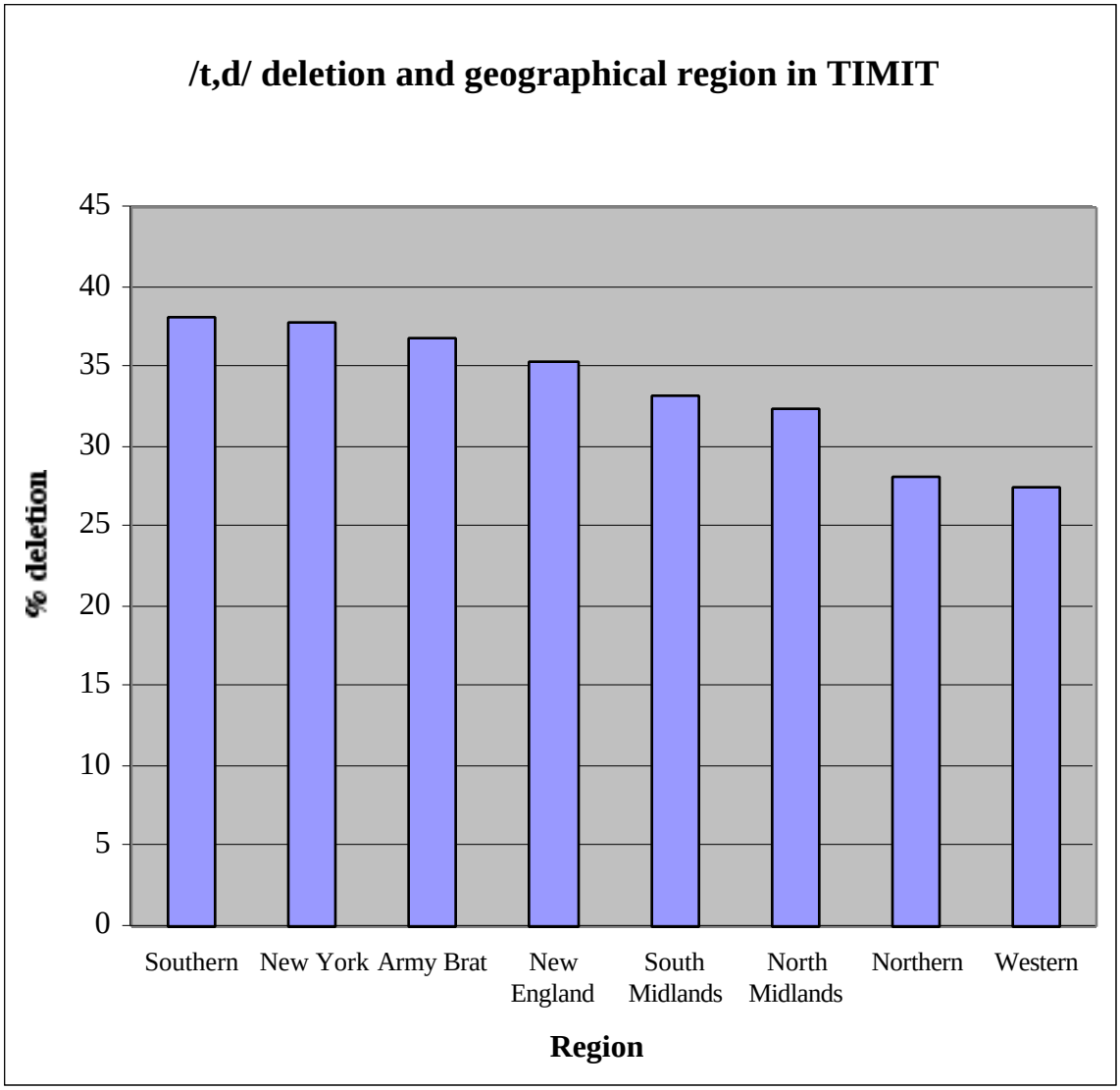

Figure 7. The correlation of region and /t,d/ deletion.

There is a slight but noticeable effect that geographical region seems to have upon /t,d/ deletion. Deletion is highest in speakers from the South (38.2%, n=254), but as we saw above, this may be due at least in part to the large number of African American speakers who compose this group (they make up roughly 20% of the whole). Deletion is also fairly high in New York City (37.8%, n=119), for so-called "Army Brats" (36.8%, n=76), and in New England (35.4%, n=113). The North and South Midland clump fairly close together,

with 33.2% (n=229) for the South Midland and 32.4% (n=278) for the North. Lowest deletion is found in the Northern states (28.2%, n=262) and the West (27.5%, n=247).

No previous studies have looked at /t,d/ deletion in quite this fashion, breaking up the states into broad geographical regions and looking for patterns. They have instead looked at a particular city, or other smallish, discrete regions. It is interesting that even in breaking up the states this way we see some distinct groupings. However, when we did our multi-variate analysis on these data using VARBRUL, region was not selected as a significant factor in /t,d/ deletion.

One of the few cross-geographical comparisons that has been frequently made is the different effects of following pause on /t,d/ deletion. The correlation of geographical region and /t,d/ deletion before a pause can be seen in Figure 8.

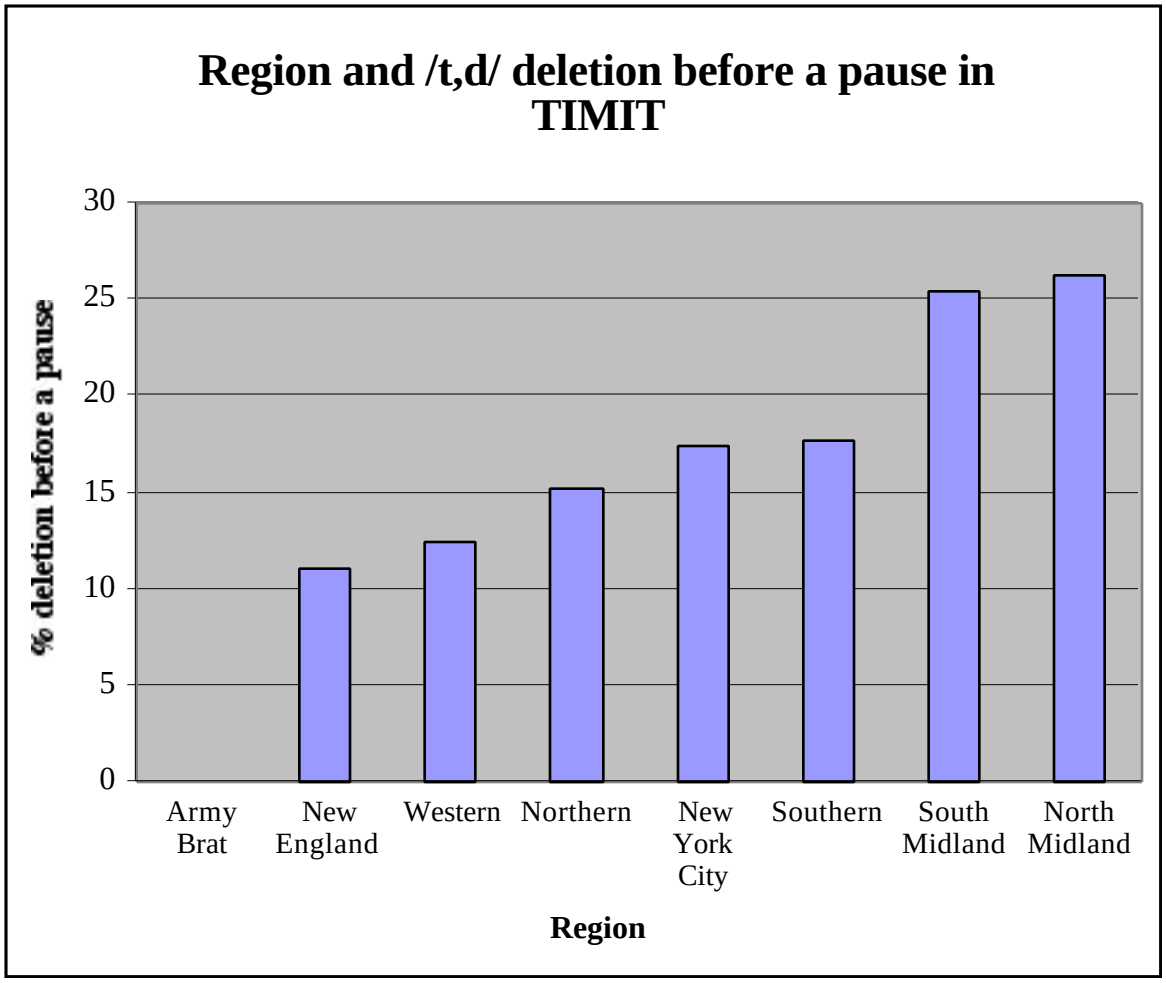

Figure 8. The correlation of geographical region and /t,d/ deletion before a pause.

As has been found in previous studies, deletion before a pause seems to pattern differently in different regions. There is an especially large gap between the North and South Midland regions and all other regions. When grouped in this way, the deletion rates before a pause in the Midlands are significantly higher than they are in all the other regions ($p < 0.025$ in a Chi-square test).

However, the ways in which these regions pattern does not seem to match previous studies very closely. New York City was shown in Guy's 1980 study to have a high deletion rate before a pause, similar to the effect of a following consonant. Here, however,

we see that the deletion rate is 17.4%, more like the effect of a following vowel (13.7%) than an obstruent (52.9%). This is only based on 23 tokens, however, which may not give us an accurate picture.

This nonconformity with previous studies could be due to several things. In some cases, there might not be enough tokens of /t,d/ deletion before a pause, as in the case of New York. Also, these regions encompass too many states and areas, which may negate the effects of some of the smaller regions within them (Philadelphia might have very different patterning that the rest of the North Midland region, for example). A third possibility might be due to the formality of these data. Speakers might be more conscious of articulating final clusters before a pause because they are concentrating on reading aloud and doing it well. This environment, more than many others, seems like it might be more sensitive to style-shifting.

Here we can see the disadvantages of using data that are not grouped geographically in ways that we would choose. If we were doing a study focusing on the effects of following pause on /t,d/ deletion, we would also be sure to elicit enough tokens, and in enough different styles, to be able to draw stronger conclusions.

- Sex

Overall, men had a /t,d/ deletion rate of 31.6% (n=1077) and women had a slightly higher deletion rate of 35.5% (n=501). A Chi-square test of these data did not find them to be significant at the $p < 0.10$ level, nor did our VARBRUL analysis select sex as a significant factor. These results correlate with studies such as Fasold (1972), which did not find a significant difference between the deletion of men and women.

- Age

The correlation of age and /t,d/ deletion can be seen in Figure 9.

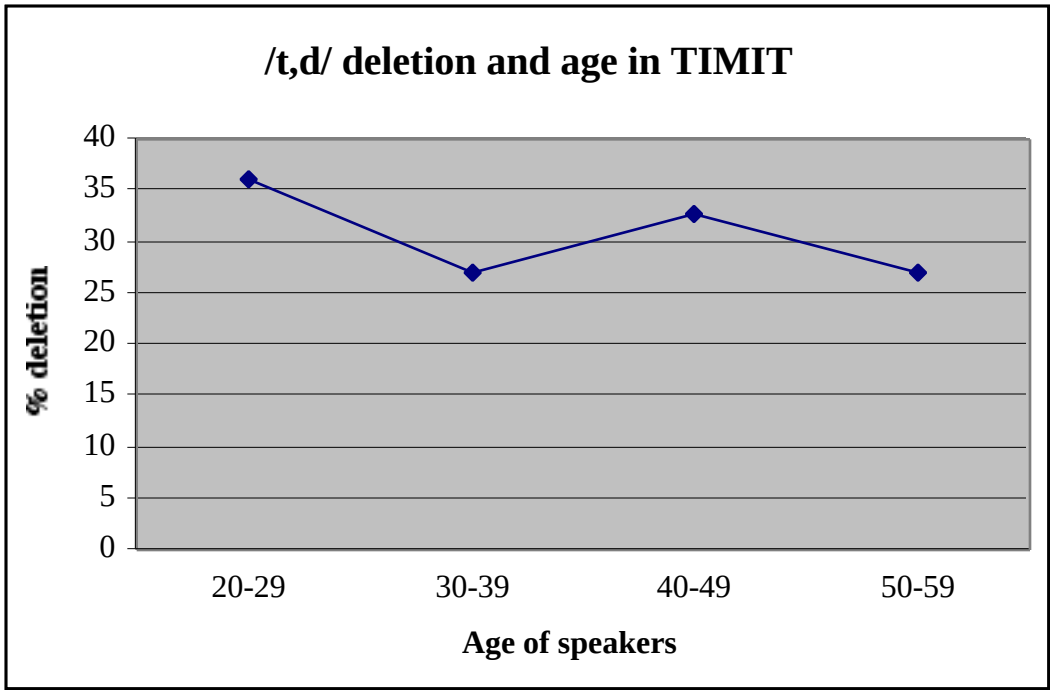

Figure 9. The correlation of age and /t,d/ deletion.

The interactions between age and /t,d/ deletion in past studies have been complicated, and these data are no exception. One of the biggest gaps is between speakers age 20-29, who have a deletion rate of 36.0% (n=968) and speakers age 30-39 who have a deletion rate of 27.0% (n=408). These data were put into a Chi-square and found to be significant at the $p < 0.001$ level. However, in our VARBRUL analysis, age was not selected as a significant factor in /t,d/ deletion overall.

I wanted to look at the deletion rate of /t,d/ in ambiguous verbs, which Guy and Boyd (1990) showed patterned differently in speakers over 45. Unfortunately, we had only two tokens of ambiguous verbs that occurred in speakers of the 45+ age group, not enough to say anything at all. Again, we see the disadvantages to using these data, which were not collected with the goal of eliciting ambiguous verb tokens from a wide age range of speakers. Otherwise, we would have made sure to use more speakers over 45, and use normal conversation, or a different set of sentences to read, that would have given us more

ambiguous verb tokens. As we move to the next three corpora, however, it seems likely that we will acquire enough data in enough age groups to be able to analyze this factor.

4.0 Discussion

In this paper, we have examined issues surrounding the sharing of linguistic corpora in quantitative sociolinguistics, now made possible by the state of digital technology and the world-wide web. There are many compelling reasons for data sharing—they are expensive and time-consuming to collect, and it is sometimes even impossible for one person to obtain data similar to someone else's. Shared, stable data sets available to the sociolinguistic community could be used to test competing theories and new methodologies, and to see whether different variables display similar patterns. Shared corpora would be useful for any linguist, but particularly for young linguists who may not have the resources to collect their own data, yet still wish to gain skills in data manipulation and interpretation. The exchange of data also gives more scientific rigor to results and annotation procedures by allowing objective outsiders to scrutinize one's data and conclusions.

However, sharing data also brings up a number of methodological concerns, since data collected within the sociolinguistic community can vary greatly— sociolinguists make different decisions about who to study, what linguistic and social variables to examine, how to classify these variables, and how to go about collecting and annotating data. There are also some key differences between sociolinguistic data pools and the corpora examined here, which were originally collected for speech technology development. Unlike most sociolinguistic data sets, which elicit a lot of data from a few representative speakers in a speech community, these corpora contain a small amount of data from a lot of speakers from a range of different speech communities.

This project seeks to assess some of the benefits and problems that arise with sharing corpora of linguistic data via a case study of a commonly studied linguistic variable, /t,d/ deletion. This variable will eventually be studied in four different corpora, but it is

examined preliminarily here in the first corpus, TIMIT. An online interface was used to sort and filter this corpus for tokens of t/d deletion and then present them in an interactive, easy-to-annotate format. Although the social information collected for the subjects in TIMIT is perhaps less than satisfactory in its lack of detail, and the diversity of the speaker group is low, the linguistic information contained in this corpus is enormous.

This is a pattern we find echoed in the results. The findings relating /t,d/ deletion to social factors in TIMIT are often less than satisfactory, whereas the results on many internal factors are both insightful and statistically sound. We don't really have enough background data to say anything conclusive about /t,d/ deletion as it relates to social class; there is a tendency for lower deletion among those of higher education, but education by itself is only a partial indicator of social class. The pattern of reverse hypercorrection among speakers with Associate degrees is fascinating, but there is not much more we can say about it, since we have no further information on these subjects. Maybe this pattern is related to occupation, or income, or even attitude, but there is no way for us to know.

Nor can we say much about geography, because the way the subject pool is broken down by region shows only slight trends that our multi-variate analysis suggests are not significant. The results with regard to age show conflicting patterns, but a multi-variate analysis also seems to indicate that age is not a significant factor on /t,d/ deletion.

There is not enough diversity in the speaker group to say anything about deletion in Latino, Asian or Native American speakers, so we are only able to compare African American and white speakers. There is a higher deletion rate in African American speakers, which agrees with past studies of /t,d/ deletion like Wolfram (1969) and Fasold (1972), and also conforms to general patterns relating race and stable variation. Although it is good that our data fit these patterns, there is nothing particularly insightful about these results.

The one social factor that is clear, statistically relevant and interesting is sex. Unlike other factors, the TIMIT corpus contained both substantial numbers of all variants (male and female) in the subject group, as well as an unambiguous way of classifying them (male

versus female). Our results found no significant difference between male and female deletion with Chi-square or with VARBRUL. This agrees with studies like Fasold (1972), while disagreeing with those that found that females deleted slightly less than males (Wolfram 1969, Neu 1980). Those studies all used much smaller subject pools than this study. These results are interesting because they fail to support broad patterns that suggest that women are less likely than men to use stable variants, possibly due to concerns about prestige (Trudgill, 69-70).

However, if the results correlating /t,d/ deletion and external factors are mostly somewhat problematic, the results correlating deletion and internal factors are mostly very valuable. These results are two-fold: some of them are interesting in how they support (or fail to support) previous studies and theories, while others give interesting new kinds of evidence.

In the former category, we have the results on preceding and following segment. Our results show a higher rate of deletion for preceding alveolar nasals and fricatives than for any other segments; this seems to provide support to the Obligatory Contour Principle proposed by Guy and Boberg (1997), but the results on preceding stop suggest that there is more going on here than mere addition of features in common (otherwise, stops would show as high a deletion rate as alveolar nasals and fricatives). This might lead us to reformulate the OCP theory to take results like this into account; perhaps certain features (like place of articulation) have more 'weight' than others.

Another set of results that are particularly interesting in light of previous theories are those on following segment, which very much fail to support Guy's theories of resyllabification. They do not show a pattern of lower deletion rates for following "resyllabifying" segments like [N] and clustering glides, with higher deletion rates for following [l] and non-clustering glides—in fact, the pattern is just the opposite. This evidence, added to the dramatically different deletion rates of following [w] and [j], makes it

seem like another explanation is needed to explain the effects of following segment, as Labov also suggested in his 1996 paper on resyllabification.

Evidence for new factors effecting /t,d/ deletion can be seen in the results obtained on the same preceding and following segment, and on the grammatical category of regular verbs. It has been noticed by linguists such as Wolfram (1969) and Neu (1980) that deletion seems to be higher when the segments surrounding the [t] or [d] were the same, in phrases like *laughed frantically* and *just saw*. Our results support this quite remarkably. Furthermore, it seems like surrounding liquids and obstruents may have very different effects; when we examined them separately, obstruents had a deletion rate of 87%, as opposed to 0% for liquids. Although this could be due to a low number of tokens for same preceding and following liquids, there might be something interesting going on here, and it is certainly worth further investigation.

The possibility that different kinds of regular verbs might influence deletion differently has been glancingly examined by Wolfram (1969) and Fasold (1972), who found that it had no significant effect. This factor is a logical extension of the "functional load" argument proposed for the different deletion rates of regular verbs, ambiguous verbs and monomorphemes in studies like Guy's (1980). If the [t] or [d] is dropped from preterit regular past tense verbs, there is a potential for ambiguity that is not present with participial forms, which might lead to less deletion of these preterit forms. This hypothesis was strongly supported by the results, which showed a significant difference between the higher deletion rate of participial forms and the lower deletion rate of preterits. This is a factor that warrants more attention; it would be interesting to see if this pattern holds up in other corpora, and also to see how far one can take the functional load argument. For example, how does the deletion rate of adjectival regular past tense forms compare to those of preterits and participials? Since they are almost monomorpheme-like, we might expect that their deletion rate would be even higher than that of participials.

The one set of results on an internal factor that is less than satisfactory is that relating /t,d/ deletion to the grammatical categories of monomorpheme, ambiguous verb, and regular verb. Almost all past studies have shown the deletion rates of these three to be on a clear continuum, with highest deletion in monomorphemes, then in ambiguous past tense verbs, then in regular past tense verbs. But when we excluded *must* from the ambiguous verb category because of its exceptionally high deletion rate, and strong verbs (*build/built*) and *went* in order to agree with past studies, the deletion rate of ambiguous verbs in our results was even lower than that of regular past tense verbs. This could be due to one of two things—the low sample size of these ambiguous verbs, or perhaps a higher incidence of participial forms in the sample of regular verbs. Either one of these shows the disadvantages of working with such a contrived data set (sentences chosen for phonetic richness) instead of natural speech, which would have a much higher number of irregular verbs and possibly of preterit forms of regular verbs. (It would be interesting to assess the proportion of preterit to participial regular past tense verbs in normal sociolinguistic interviews, and see if this factor could be influencing our results.)

Thus, in this preliminary look at using corpora of linguistic data, several positive and negative aspects emerge. TIMIT does not provide enough detailed, varied social information to allow us to say much about the more complex external factors related to /t,d/ deletion, but it does give a wealth of linguistic information that shows clear, striking patterns in several internal factors, including correlations that linguists have not previously observed. The one internal factor that presents a real problem is grammatical category; this seems to be a direct result of having only a contrived set of sentences as data.

As we move from TIMIT to the next three corpora, some of these problems will increase while others will disappear. Lack of speaker information, a major problem in TIMIT, is even worse in Switchboard, CallHome and HUB-4. CallHome provides us with education, sex and age information only, Switchboard with these three and region (classified in the same way as TIMIT). HUB-4 has no speaker information whatsoever. This

suggests that as we move to working on these corpora, we should continue to focus more on internal factors like surrounding environment and grammatical category. We may even wish to expand the scope of the internal factors we examine, perhaps beginning to code for word stress and cluster complexity as well.

The problems that were a result of TIMIT's highly contrived data set—in particular not having enough tokens of ambiguous verbs to see patterns in grammatical category—should disappear in the next three corpora. The other corpora all contain many examples of spontaneous speech, which should give us data with a higher percentage of ambiguous verbs. Thus, we should be able to examine the factor of grammatical category more thoroughly as this project continues.

TIMIT is a fairly large corpus, but the next three data sets are much bigger. With TIMIT, under 2000 tokens were annotated; a preliminary look at Switchboard gives us around 30,000 applicable tokens. These larger corpora are both daunting and exciting. It will be necessary to train additional annotators to help me with the next three corpora, which will raise questions of interannotator consistency that this study did not. Also, unlike myself and annotators used in most sociolinguistic studies, additional annotators for this project will probably have little to no background in linguistics. We will have to examine how this might affect things as well.

These larger corpora will yield much larger pools of data, which will be helpful in clarifying patterns of age and grammatical category that were confusing in TIMIT. Additional data will also help us confirm or reject other TIMIT results—for example, the claim that having the same preceding and following environment effects deletion differently with surrounding liquids (*old lady*) than with surrounding obstruents (*passed softly*) will be much stronger if it is based on more than 10 tokens of liquids.

As we move to the next three corpora, the question of style will also become much more relevant. Being able to compare the overall rates of deletion in TIMIT (the very formal environment of reading sentences aloud) to Switchboard (a spontaneous, although

somewhat formal interaction between two people who have never spoken before) to CallHome (a very casual conversation between intimate friends and family members) will give us a large amount of data showing how /t,d/ deletion varies with situational formality, often in formats that have never or rarely been studied (i.e., telephone calls, broadcast news).

Although we have chosen to examine the variable of /t,d/ deletion in this project, there are endless possibilities for other variation studies using these four corpora. There are many more phonological variables which could be studied in TIMIT, such as the occurrence of /n/ versus /E/ in words like *fishing*, and in addition to phonological studies, Switchboard, CallHome and HUB-4 also offer the possibility of variation studies in morphology, the lexicon, syntax and discourse.

Traditionally, quantitative sociolinguistics has centered around small, community-based studies. These kinds of studies are essential if a linguist is to understand the full complexity of a speech community, especially the less quantifiable factors that can play a role in linguistic variation, like attitudes, social ties, and status. But supplemental data gleaned from corpora like TIMIT can only strengthen the conclusions made in community-based studies, particularly on internal factors.

5.0 Conclusion

Ideally, sociolinguists will begin to realize that the benefits of sharing data far outweigh the potential problems, and will donate their used corpora to the sociolinguistic community. With a few modifications to their current site, the LDC is in a good position to host and regulate a website where data could be donated and accessed by sociolinguists all over the world. Access through the LDC would help donators keep a general sense of who is using their data, and why, without having to deal with the details of providing access themselves.

Actual sociolinguistic databases would provide different benefits than those of TIMIT and the other three commercial corpora that are the focus of the DASL project.

They would undoubtedly be closer to more traditional sociolinguistic studies in terms of subject pool, the quantification of social factors, and the sociolinguistic interview. This would probably help alleviate some of the problems we encountered with TIMIT, such as insufficient speaker information and a contrived data set.

In the future, sociolinguists could conduct studies with data sharing in mind. This might encourage them to collect and make available a wider range of information on their speakers and their methodology, realizing that something they might not find important might be crucial to another linguist. Perhaps sociolinguists might even be able to work together to collect a corpus as large as TIMIT or Switchboard, which would have all the benefits of those large, broad pools of data, but with sociolinguistic considerations guiding speaker selection and data collection. A website containing corpora like TIMIT, traditional sociolinguistic studies, and large-scale sociolinguistic studies would provide quantitative sociolinguists not just with more data, but with more kinds of data, than ever before.

---

[1] I would like to thank Stephanie Strassel and Chris Cieri very much for letting me be a part of this project, and for their tremendous guidance and support. Thanks also to Ted Fernald and the rest of my Senior Linguistics Seminar for their input and suggestions, particularly Jennifer Tyson for her help with an early draft.

[2] Information about this project can also be found on the project's website, www.upenn.ldc.edu/Projects/DASL

[3] Statistical tests commonly used in quantitative sociolinguistics include Chi-square, VARBRUL and ANOVA.

[4] These environments have been found to have a very high deletion rate, and their classification is somewhat unclear—they are not quite monomorphemes, but neither do they fit the other two categories. For a more complete look at what was considered and what was not, look at section 2.2.

[5] For information about the acquisition of t/d deletion and its constraints in children, see works by Julie Roberts (1995).

[6] I am particularly indebted to Stephanie Strassel for her help with this section.

[7] Our interface uses Perl.

[8] The regular expression we used was: [^aiou '][td][^A-Za-z][^td]. [^aiou '] matches anything but the list of vowels symbols, the space character and the apostrophe—we must allow preceding *e* to admit past-tense verbs; using space in the pattern avoids matching initial /t,d/; and the apostrophe avoids matching *n't*. [td]

matches either a *t* or a *d*. [^A-Za-z] matches any character but those in the English alphabet, i.e., a space. [^td] matches anything but the letters *t* or *d*.

[9] We used a grep filter: grep -iv '<socann>and' to remove all instances of *and.*

[10] We used an egrep filter: egrep -iv '<socann>[A-Za-z]+et[^A-Za-z]' to remove cases such as *pocket*.

[11] We used an egrep filter: egrep -iv '<socann>[A-Za-z]+[td]ed[^A-Za-z]' to removes cases such as *heated* and *needed*.

[12] We are still looking for the specifics of region classification in TIMIT, but the information appears to be lost. Even the collectors of these data cannot remember exactly how the regions were grouped.