Chapter 1

Introduction

1.1 Motivation

The work presented in this dissertation pursues two related and mutually motivated aims. It was primarily inspired in the early 1990s by the author's personal interest in the historical origins of American dialects and the differences among them. That interest sparked a search for datasets appropriate to the analysis of American and English dialects and of quantitative tools that could be applied to the data to develop a detailed understanding of dialect variation in both countries.

Linguists have made significant theoretical and empirical progress in the 130 years since Wenker's collection of German data inaugurated the systematic study of dialect variation. Systematic research on dialects of English dates to the late-nineteenth-century surveys of Ellis (1889) and Wright (1898-1905). Lowman completed a brief survey of rural dialects of southern England in 1937-38, summarized posthumously in Kurath and Lowman (1970) and Viereck (1975). The modern standard covering the oldest surviving examples of rural speech is Orton and Dieth's (1962) *Survey of English Dialects*, which provides the basis for a large number of studies including Anderson (1987), Glauser (1985), Trudgill (1999), Upton and Widdowson (1996), Upton et al. (1987), (Viereck and Ramisch, 1991, 1997), and (Wells, 1982a,b). Research on American dialects dates mainly to the inception during the 1920s of the regional *Linguistic Atlas* projects, which have now covered most of the country—see for example Kurath (1939), Kretzschmar et al. (1994), and Kurath and McDavid (1961). More recently Labov et al. (2006) have completed a North American survey based on modern methods.

---

1Wenker's linguistic atlas of Germany is available online at http://http://www.diwa.info/.
Trudgill (1986, 2004) has presented extensive, detailed studies of the British origins of colonial dialects, investigating the origins of Australian English and closely tracking the process of new-dialect development in New Zealand across several generations. Greater time depth has allowed extensive innovations to obscure the origins of American dialects, as documented by Bailey (1997) and Schneider (2003), but substantial evidence from migration patterns and settlement histories – summarized by Bailyn (1986), Fischer (1989), and Meinig (1986) – points to a largely though not exclusively southern English influence. That influence has been documented in the linguistic evidence as well: Brooks (1935), for instance, was able to use quite primitive means to uncover essentially southern English antecedents for Southern American speech; Kurath and McDavid (1961) and Kurath and Lowman (1970) compare regional distributions of English and American phonetic patterns; Montgomery (2001) traces Scots-Irish influence in the Appalachians and Upper South; Wright (2003) uncovers grammatical features currently associated with Southern American English in seventeenth-century London prisoners’ narratives; and Algeo (2003) finds “multiple lines of descent” for Southern American speech forms. Muñwene (1996) argues that the speech forms of the earliest settlers would have tended to survive and propagate through a “founder effect,” while Kretzschmar (2002) emphasizes the largely local nature of the processes by which settlers’ diverse speech developed into distinctive new patterns.

Over the past generation, moreover, quantitatively oriented researchers have laid the foundations of computational dialectology, providing a set of quantitative techniques that can be used to address a wide range of issues in language variation. Moving well beyond the isogloss methods characteristic of earlier work, researchers have developed methods of measuring differences in pronunciations (recorded impressionistically or acoustically) as numerical distances between sound segments, as described in detail by Heeringa (2004) and Nerbonne and Heeringa (2009), and have shifted their focus from studying variation in individual features to measuring distances between speakers or groups of speakers based on aggregates involving many different features, as discussed by Nerbonne (2009). Such methods can be used to compare

---

2The development of dialectometry dates to the introduction of a simple distance measure by Séguy (1973); Goebel (1982) independently developed a similar approach and has since contributed major advances in broadening the application of quantitative techniques to variation among Romance dialects. Goebel (2007) presents an English-language overview of dialectometry with applications to English data, and the Salzburg University Dialectometry Project provides a useful online discussion of aspects of dialectometry such as different measures of similarity, dispersion, and classification, at http://ald.sbg.ac.at/dm/Engl/default.htm. However, the field of computational dialectology is expanding so rapidly on independent fronts that no comprehensive introduction to the full range of quantitative techniques or their application to linguistic data currently exists.

3Linguistic and genetic systems present similar analytic problems – inferring historical relationships among information systems that are replicated with error, that are composed
1.1. **MOTIVATION**

individual variation in language use with variation among groups of speakers, to quantify the extent of systematic regional or social variation in speech, and to explore the historical development of speech diversity. Researchers have applied quantitative techniques to morphological, syntactic, and lexical variants in the traditional English dialects: Viereck and Ramisch (1997) present a number of such studies, including cluster analyses by Goebl (1997), multidimensional scaling by Embleton and Wheeler (1997), and principal component analyses by Inoue and Fukushima (1997). Other researchers, including Nerbonne (forthcoming), Labov et al. (2006) and Clopper and Paolillo (2006), apply dialectometric and related computational techniques to the classification of North American dialects.

Prior to the research reported here, however, quantitative methods had not been systematically applied to English phonetic variation; nor had dialectometric analysis been applied to data from both English and American speakers to explore how English variants might have been selected in the process of new-dialect development in the American colonies. This dissertation is comprised of three articles that, taken together, report the application of a broad set of quantitative tools to English phonetic data, and in doing so illustrate both the effectiveness and the limitations of such tools not only in describing and comparing dialects but also in uncovering aspects of the historical development, differentiation, and combination of various dialects of English. One article, "Phonetic Variation in the Traditional English Dialects: A Computational Analysis," published in the *Journal of English Linguistics* in 2007 and comprising Chapter 2 of this dissertation, uses a variety of techniques to quantify and compare speech forms from the traditional English dialects, as recorded in Orton and Dieth (1962). Where traditional methods of dialectology rely largely on subjective analysis of distributions of recorded variants to classify dialect regions and to identify the variants associated with those regions, the approach taken in that chapter attempts to introduce a greater degree of objectivity to the classification process by applying quantitative methods to patterns of usage for large numbers of variants or features for each locality, drawing on statistical properties of the data to determine regional boundaries and distinctive of units that can be favored and selected by chance, mutual interaction, or environmental pressure, that therefore gradually change over time, and that may have geographic distributions that provide insights into their historical development. Given these similarities, phylogenetic techniques used by geneticists also may be usefully applied to linguistic data despite the differences in their underlying processes of production, maintenance, and change. Linguists may therefore find useful applications for some of the algorithms used to measure genetic distances between species, to infer the historical development of groups of related species (that is, their phylogenies), and to isolate important geographic boundaries between distinctive groups. Productive collaboration between geneticists and linguists includes the analysis of Nakhleh et al. (2005) of genetic relationships among the Indo-European languages. A useful standard textbook treatment of phylogenetic techniques can be found in Nei and Kumar (2000).
features. The results provide measures of variation in phonetic usage among English localities, identify dialect regions as clusters of localities with relatively similar patterns of usage, distinguish regions of relative uniformity from others with substantially greater variation, and, to some extent, identify regionally coherent groups of features that can be associated with and used to distinguish at least some traditional English dialect regions. The results are broadly consistent with standard characterizations of traditional English dialects and regions, but they also strongly underscore the largely continuous and apparently random variation in traditional English speech forms, and differ quite noticeably in the placement of regional boundaries and the identification of regionally distinctive features, suggesting that quantitative methods can significantly improve our understanding of dialect differences and structure. At the same time, the results also suggest that, at least as applied in this study, quantitative tools have some important limitations in their ability to uncover known structural elements of traditional English dialects.

A second article, “English-American Speech Relationships: A Quantitative Approach,” published in the *Journal of English Linguistics* in 2005 and comprising Chapter 3 of this dissertation, applies similar techniques to phonetic data from Kurath and McDavid (1961) and Kurath and Lowman (1970), comparing speech forms found in southern England with those found in the regions of earliest settlement in British North America as well as in part of southern Appalachia. As in the preceding chapter, the analysis uses the statistical properties of usage of a large numbers of variants or features to identify regional boundaries and distinctive features, and to identify patterns of similarity in usage across regions of England and America. The results reported in that article suggest that American speech patterns are largely amalgams of variants brought from different - but overwhelmingly southern - English dialect regions. The patterns of similarity support a model of new-dialect formation in the American colonies involving competition within and selection from a pool of variants introduced by speakers from those different regions. Consistent with the historical evidence of seventeenth- and eighteenth-century British migrations to North America, the results reveal largely southeastern English origins for major varieties of American speech, and also suggest regional English origins for some important differences among regional American dialects, with somewhat greater southeastern influence on New England speech and southwestern influence in the American South.

A third article, “Southern American English in Perspective: A Quantitative Comparison with Other English and American Dialects,” forthcoming in *Language Variety in the South: Historical and Contemporary Perspectives* and comprising Chapter 4 of this dissertation, extends the analysis of the second article to explore regional and social variation among southern American speakers. Again, quantitative methods are applied to explore the statistical
properties of large numbers of variants or features across speakers in different regions and from different social groups within regions. The analysis demonstrates that phonetic variants used by Southern American speakers who were judged as folk speakers are much more likely to be found in southwestern England than variants used by speakers judged as common or cultivated - a finding consistent with historical evidence that West Countrymen were particularly sought after for labor supply by seventeenth-century Virginia planters prior to the widespread introduction of African slavery toward the end of the century. That finding provides evidence of regional English origins for some dialect differences of sociolinguistic relevance in Southern American English surviving well into the twentieth century.

In sum, the research presented here quantitatively characterizes differences among samples of traditional rural dialect of England and parts of the United States as they existed in the mid-twentieth century, and uses those characterizations to infer aspects of their historical development and differentiation - particularly the development of new dialects during the American colonization. Two remarkable insights emerge from this research. First, the modern dialects appear to offer rather clear evidence about the process of new dialect formation three to four centuries ago - this in spite of equally clear evidence of ongoing differentiation in the home country and the colonies, and of influence from other dialects and languages as well. Second, some of the socially relevant distinctions in twentieth-century regional American English appear to be traceable to regional dialect differences in the home country that were transported to the colonies. Both insights are made possible by the application of quantitative techniques.

The next section summarizes the quantitative tools used in the three following chapters, while a concluding chapter summarizes the insights yielded by their application to English and American phonetic data and discusses directions of study that could build on the work discussed here to yield further understanding of the varieties of English pronunciation and their ongoing evolution.

## 1.2 Quantitative Tools and Phonetic Data

Quantitative methods can be used to characterize observations of interest as variable numerical quantities, to calculate patterns of correlation among variable observations, or to identify groups of similarly varying observations or similarly varying characteristics of the observations, thereby reducing variation along a large set of dimensions to variation along a smaller set. Quantitative tools can be used simply to summarize and explore relationships among sets of variables, but they can also be used in more sophisticated ways to test (in
the statistical sense) models of such relationships.\footnote{For readers seeking non-technical summaries, Bartholomew et al. (2002) and Tabachnick and Fidell (2000) both provide straightforward introductions to the use of specific univariate and multivariate techniques. Statistical analysis of linguistic survey data is discussed in detail by Kretzschmar and Schneider (1996).}

Applied to phonetic data, such methods can be used to quantify linguistic variation by characterizing the data as numerical quantities and also to uncover patterns of variation that may be linguistically and statistically significant. In the research reported here, I use such methods to calculate measures of aggregate similarity between speakers’ linguistic usages, to group speakers on the basis of degrees of similarity, to group linguistic features on the basis of their distributions among speakers, and, to some extent, to identify linguistic features that can be said to be characteristic of particular groups of speakers or of speakers in particular regions. I also explore geographical patterns of similarity among speakers to distinguish regions of comparative uniformity of speech from regions of comparative variability.

The methods often involve a trade-off between completeness of information and simplicity and interpretability of results – that is, a trade-off between capturing broad patterns of systematic variation and preserving information about comparatively minor or random variation. In the analyses presented here, the methods reveal a great deal of apparently unsystematic variation as well as some degree of structure. The analyses also reveal important limitations in using such methods to uncover linguistic structure. Ideally, the methods would yield clear delineations of dialect groups or regions equally clearly associated with distinct linguistic forms; in practice, however, the methods may even fail to identify known structural dialect features with clear geographic boundaries.

1.2.1 Quantifying Phonetic Variation

Modern linguistics uses sophisticated methods for recording and quantifying phonetic data on the basis of characteristic acoustic patterns such as formants. For the traditional dialects of England and America, however, the most useful, readily available data consists of detailed observations of speakers’ usages in specific words, recorded impressionistically using various versions of the International Phonetic Alphabet (occasionally adapted to specific characteristics of American dialects). In the work presented here, I have used several approaches to quantify that impressionistic data:

- Where the data is presented as distinct classes of variants for particular phonemes in particular words (e.g. [e] versus [ei] in mate), I assume that each observed speaker regularly uses the variant recorded and quantify each observation as 100% frequency of occurrence of that variant in that word. A related approach uses frequencies of variant use across a group
of words considered to be of a linguistically distinct class: Anderson presents such frequencies from the Survey of English Dialects for groups of words considered to have been pronounced similarly in "standard" or "Chancery" Middle English.

- Where possible, I construct a somewhat arbitrary but linguistically defensible multidimensional, feature-based representation of the impressionistic data. For the analysis of the traditional English dialects, for example, I constructed a unique dataset of the primary observed pronunciations in each locality of 55 words that, among them, include at least one instance of each phoneme descended from each Middle English short vowel, long vowel, and diphthong, alone and followed by rhotics, as well as the variable consonants. The approach — of my construction but similar to the system of Almeida and Braun (1986) — translates over a hundred segments (including vowels, diphthongs, and consonants) into numerical values representing nearly 500 features such as degrees of length, height, backing, and rounding (or, for consonants, place and manner of articulation). For one type of analysis, I adapt the approach to represent the characteristics of glides in the diphthongs, making possible a novel analysis of variations in glide characteristics using principal component analysis.

- In some cases I combine approaches by characterizing the features of all instances of a particular class of variants as those of the "standard" variant defining the class.

Any parsing or quantification of a coding system as complex as natural language is necessarily somewhat arbitrary, and even native speakers, whose perceptions might be considered the standard against which to compare any other measure, will typically differ in their assessment of differences between dialects. The patterns of variation uncovered by the use of measures such as those described above will depend in part on the choice of segments and the choice of measure. Nevertheless, the research reported here appears to yield relatively robust patterns that repeatedly turn up under significantly different approaches. I therefore argue that those findings are likely to represent real and significant patterns of dialect variation.

1.2.2 Distribution of Variants and Features

Simple examination of the statistical distributions of variants and features — averages, standard deviations and variances, correlations, and so forth — can yield important information, including strong evidence for the randomness of
much linguistic variation. However, by summarizing aspects of the distribution of features or variants of particular phonemes among speakers or within or among geographic regions, simple analysis of variation yields useful clues into differing degrees of variability among features, geographic distributions of dialect features, and the degree of structural variation among dialects.

1.2.3 Distance Measures among Speakers

Distance measures – calculations of the degree of difference between individual speakers’ usages, preferably aggregated over a large number of words – provide the basis for most of the other quantitative tools used in the research reported here. Distance measures provide a reasonably objective gauge of whether (and which) speakers’ speech forms are dramatically different or relatively similar, allowing the researcher to distinguish degrees of difference among varieties of speech.

Any number of distance measures are available, all of which have some utility. They include: the percentage of a speaker’s total number of variants shared in common with other speakers; Pearson correlations or Euclidean distances between points representing the values of variants or features, or cosines between vectors representing those values; and various measures of linguistic distance or genetic distance. In the analyses presented here I have focused primarily on the simplest measure – the percentage of shared variants – and on a somewhat novel and more complex measure of the sum of Euclidean distances between the articulation-based numerical characterizations of a large number of segment features, a measure of linguistic distance that I believe is particularly appropriate to the data.

1.2.4 Cluster Analysis to Delineate Dialect Regions

Clustering algorithms group observations in a dataset into classes, or clusters, on the basis of measures of distance or difference between observations. Clustering, in effect, simplifies data by reducing the multitude of pairwise differences among observations to a relatively small set of relationships within and among clusters. Non-hierarchical clustering methods divide observations into an arbitrary number of unrelated groups, whereas hierarchical methods yield tree-like groups with multiple branchings, with the length and distribution of the branches reflecting degrees of similarity among observations.

There is no perfect clustering technique, and different clustering techniques can produce markedly different classifications, with the efficacy of any given technique in appropriately classifying observations depending in part on the nature of the data. One approach that tests the robustness of the results is to use a number of different clustering techniques and distance measures,
and to introduce perturbations into the data to test whether such “noise” noticeably affects the classification of observations. Patterns that consistently emerge under different approaches and noisy data appear more likely to reflect underlying real patterns in the data than patterns that do not.

I use clustering methods here to group speakers on the basis of the distance measures discussed above. Ideally (and very often in practice) the groups can be interpreted as geographically contiguous dialect regions. By distinguishing a reasonably coherent and robust set of dialect regions, the cluster analysis lays the basis for measuring degrees of similarity among speakers from different regions and for seeking patterns of linguistic usage consistent with those regions.

1.2.5 Multidimensional Scaling to Map Dialect Regions

As cluster analysis may simplify data by reducing variability among observations to a relatively small set of clusters of observations, multidimensional scaling (MDS) techniques reduce the variation in a dataset to a relatively small arbitrary number of dimensions, allowing the user to summarize fundamental relationships in the data. As with clustering, there is no perfect MDS technique. In some of the work discussed here, I apply MDS to results of a number of different cluster analyses of the same observations, further clarifying which dialect regions and transition zones appear to be robust to variations in the clustering approach and distance measure used.

1.2.6 Principal Component Analysis to Uncover Linguistic Structure

Principal component analysis (PCA) also reduces the number of dimensions in a dataset, but does so by isolating groups of variables such that the variables in any particular group are strongly correlated with each other but are uncorrelated with the rest of the variables in the dataset. PCA summarizes each group into a single “latent” variable – called a principal component, or PC – that is essentially a linear combination of the correlated variables in the group. Although the algorithms are quite different, PCA methods may be thought of as grouping variables somewhat analogously to the way clustering methods group observations. Component loadings measure the strength of specific variables in specific components; component scores measure the strength of each principal component in specific observations, and may therefore reveal important associations between groups of variables and groups of observations. “Rotation” of the principal components tends to sharpen the focus and concentration of each component by increasing its loading on smaller groups of variables. Just as
there is no perfect clustering or MDS technique, there is no perfect approach to rotation, and care must be taken to interpret the results.

In each of the articles discussed below, I apply PCA to data on the frequency of occurrence of phonetic variants, the strength of features, or both. Applied to such data, PCA identifies groups of variant frequencies or feature values that tend to occur together among speakers. PCA can thus be used to determine whether (and how strongly) groups of variants or features tend to occur together and whether (and which) groups of speakers tend to use them together. Properly performed and interpreted, PCA may therefore isolate groups of variants or features that, with any luck, will be interpretable as linguistically relevant sets of pronunciations that also have distinct geographic boundaries. In short, PCA may provide a relatively objective way of defining dialects and dialect regions in structural terms.

1.2.7 Multiple Regression and Monmonier Analysis

Multiple regression analysis quantifies the relationships between a variable of interest – a dependent variable – and a number of independent variables, allowing for interaction among the latter. In the analyses described here, I use multiple regression to analyze several relationships between linguistic and geographic variables:

- Biologists have long noted a regular correspondence between species and area – that is, a relationship between the (log of the) number of species that inhabit a geographic region and (the log of) its size. I regress the number of variants in each region on the number of localities to show a similar correspondence in the traditional dialects of England and the United States.

- I also use regression to test for a statistically significant relationship between the various measures of linguistic distance between localities and the geographic distance between them. By introducing dummy variables that represent localities' regional affiliations, I explore whether differences among localities vary systematically across the dialect regions identified by cluster and phylogenetic analyses.

- In comparing variants used by English and American speakers, I use regression to test for a statistically significant relationship between the degree of similarity between English and American speakers and the proximity of the English speakers to London. The results support the hypothesis that speech in or near the London metropolitan area played a key role in the development in American speech varieties.
The importance of geography in determining the distribution of speech forms can be further analyzed by examining the distribution of residuals or errors in regressions of linguistic distance on geographic distance: a series of large positive errors generally indicates an important dialect boundary that marks a relatively large number of linguistic changes over a short distance, whereas a series of large negative errors probably indicates unusually uniform speech over a relatively broad area. Monmonier (1973) developed a maximum difference algorithm that analyzes distance measures to map important boundaries. I show that the application of this “barrier analysis” technique yields much the same results as does cluster analysis, but that it also isolates large numbers of individual speakers in regions with particularly strong internal variation.

1.2.8 Discussion

In sum, a variety of methods can be applied to phonetic data to quantify and explore the structure and significance of linguistic variation. I use such methods in the following chapters to calculate measures of aggregate similarity between English speakers in regions of England and America, to classify those speakers on the basis of degrees of similarity, to group linguistic features on the basis of their distributions among speakers, and to identify linguistic features that can be said to be characteristic of particular groups of speakers or of speakers in particular regions. For American speakers, I explore degrees of similarity with speakers in different regions of southern England and arrive at tentative conclusions about the regional origins of some American dialect differences. The final chapter summarizes the results.