Quantitative assessment of English-American speech relationships

Shackleton Jr., Robert George; Shackleton, R.G.

Chapter 2

Phonetic Variation in the Traditional English Dialects: a Computational Analysis

Abstract. This article illustrates the utility of a variety of quantitative techniques by applying them to phonetic data from the traditional English dialects. The techniques yield 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 transitional zones with substantially greater variation, and identify regionally coherent groups of features that can be used to distinguish some dialect regions. Complementing each other, the techniques provide a reasonably objective method for classifying at least some traditional English dialect regions on the basis of characteristic features. The results largely corroborate standard presentations in the literature but differ in the placement of regional boundaries and identification of regional features, as well as in placing those systemic elements in a broader context of largely continuous and often random variation.[1]

2.1 Introduction

In the 130 years since Wenker's collection of German data inaugurated the systematic study of dialect variation, scholars have extended the repertoire of

[1]The final, definitive version of this paper has been published in the *Journal of English Linguistics*, Vol. 35, No. 1 (March 2007) by SAGE Publications, Inc. On-line version available at `http://eng.sagepub.com/cgi/content/abstract/35/1/30`.

recording and analytic techniques far beyond the impressionistic collection and compilation of dialect features and the identification of isoglosses. Over the past generation, quantitatively oriented researchers have laid the foundations of a discipline of computational dialectology, providing a set of quantitative techniques that can be used to address a wide range of issues in language variation: Can useful metrics be developed to measure differences in speakers' phonetic, lexical, and syntactic usages in different locations – or, by extension, in different social groups, or at different periods? How does individual variation in language use compare with variation among speakers of the "same" dialect, and how does the latter compare with variation among speakers of "different" dialects? Is regional dialect variation largely random, or geographically continuous, or can the continuum reasonably be divided into dialect areas with relatively distinct boundaries? Is it possible to distinguish core dialect regions with relatively uniform patterns of usage from transitional zones with greater diversity? Can dialect regions be characterized by systematic variations in features, such as chain shifts or devoicing of voiced consonants; and if so, what features distinguish a given dialect region from neighboring areas? Can a standard language be traced to origins in geographically restricted dialects? All of those issues can be explored through the application of quantitative techniques to dialect data.

As a general rule, quantitative methods are simply ways of characterizing observations of interest as variable quantities, of teasing out patterns of correlation among variable observations, or of isolating groups of similarly varying observations, thereby reducing variation along a large set of relevant dimensions to variation along a smaller set. Such methods can therefore be used to explore the questions posed above by characterizing linguistic data as quantities; by establishing measures of linguistic difference – gauges of the degree of aggregate similarity between speakers' linguistic usages; by classifying speakers into groups on the basis of similarity; and by grouping linguistic features on the basis of their distributions among speakers – in short, by quantifying linguistic variation and uncovering patterns of variation that are both linguistically and statistically significant.

Some quantitative tools, such as multiple regression and analysis of variance, can be used to explain or predict variation in a linguistic phenomenon of interest on the basis of variation in several other phenomena. Such tools are often used in conjunction with the assessment of explicit models using tests of statistical significance and so forth. Other multivariate techniques, such as cluster analysis, multidimensional scaling, or principal component and factor analysis, are used to summarize and explore interrelationships among sets of variables more generally, and are less typically used in conjunction with statistical tests of specific models.[2]

[2]For readers seeking non-technical summaries, Bartholomew et al. (2002) and Tabachnick

To the extent that quantitative methods help researchers identify dominant patterns by eliminating dimensions of variation in the data, they also result in a loss of information because generally speaking, not all of the variation can be summarized in a smaller number of dimensions. The methods therefore typically involve a trade-off between completeness of information and simplicity and interpretability of results. In the study of linguistic variation, they often involve a trade-off between capturing broad patterns of relatively systematic variation and preserving information about relatively minor ones.

Although most of the techniques discussed here can be applied to a wide range of data, a few are drawn from the study of genetic variation. Linguistic and genetic systems present similar mathematical problems despite the differences in their underlying processes of production, maintenance, and change. Like historical linguists, geneticists face the problem of inferring historical relationships among information systems that are replicated with error, that are composed 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.[3] 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.

The field of computational dialectology is expanding so rapidly on independent fronts that no current comprehensive introduction to the full range of quantitative techniques or their application to linguistic data exists.[4] To illustrate their usefulness I provide a series of applications to phonetic data from

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).

[3] A useful standard textbook treatment of phylogenetic techniques can be found in Nei and Kumar (2000). For an interesting example of collaboration between geneticists and linguists see Nakhleh et al. (2005).

[4] For the measurement of dialect differences – or dialectometry – Heeringa (2004) provides the most comprehensive English-language review. The Salzburg University Dialectometry Project provides a useful online discussion of various aspects of dialectometry, such as different measures of similarity, dispersion, and classification, at `http://ald.sbg.ac.at/dm/Engl/default.htm`. The development of dialectometry (as well as the coining of the term) dates to 1973, when Jean Séguy introduced a simple distance measure in the *Atlas Linguistique de la Gascogne*. Hans Goebl (1982) independently developed a similar approach and has since contributed major advances in broadening the application of quantitative and mapping techniques to variation among Romance dialects. In North America, Nerbonne (forthcoming) and Nerbonne and Kleiweg (2003) have applied dialectometric techniques to data from the *Linguistic Atlas of the Middle and Southern Atlantic States* (*LAMSAS*), while Labov et al. (2006) and Clopper and Paolillo (2006) have applied related computational techniques in the classification of North American dialects.

the traditional dialects of England, highlighting strengths and weaknesses of different tools.[5]

- I construct data sets that quantify variation in the dialects, and use the data to construct measures of linguistic distance, thereby establishing degrees of difference among speakers in different localities.
- I apply clustering and phylogenetic methods to those linguistic distance measures to classify localities into dialect regions of varying coherence.
- I then use regression analysis and barrier analysis to explore the relationship between geographic and linguistic distances within and among the dialect regions.
- Finally, I apply principal component analysis to identify groups of phonetic variants and features that can arguably be said to distinguish some of the dialect regions.

Any parsing or quantification of a coding system as complex as natural language is necessarily somewhat arbitrary – 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 – and the patterns of variation uncovered by the use of such measures depend in part on the choice of segments and the choice of measure. Nevertheless, the analyses yield relatively robust patterns that appear repeatedly under significantly different approaches and that are likely to represent real and significant patterns of dialect variation. The results provide strong quantitative evidence for regions of relatively uniform use of distinctive features as well as others of substantially greater than average variation, while placing both against a background of largely continuous variation.

2.2 Data: *Survey of English Dialects* and *Structural Atlas of English Dialects*

The primary data source is Orton and Dieth's (1962) *Survey of English Dialects* or *SED*, the best broad sample of the most traditional forms of rural English

[5] Other researchers have applied similar techniques to morphological, syntactic, and lexical variants in the traditional English dialects. Viereck and Ramisch (1997) include a number of such studies, including cluster analyses by Goebl, multidimensional scaling by Embleton and Wheeler, and principal component analyses by Inoue and Fukoshima. However, to my knowledge such methods have not been systematically applied to English phonetic variation. An obvious significant extension would be to apply the techniques to an extended data set that includes the lexical, morphological, and syntactic variants enumerated by (Viereck and Ramisch, 1991, 1997).

dialect that were still in use in the mid-20th century.[6] Focusing their resources on recording the most recessive features of the language, *SED* fieldworkers interviewed a handful of elderly people – mainly men, who were considered more likely to use nonstandard traditional speech – in each of 313 relatively evenly spaced, mainly small, rural agricultural communities throughout England, using questionnaires, diagrams, pictures, and spontaneous conversation to elicit responses. In choosing locations, they gave some consideration to geographic features – mountainous terrain, rivers, and so forth – that were likely to influence linguistic differences among localities. As a consequence, the *SED* data is a highly representative sample of an important dimension of variation in mid-20th century English dialects, though it should not be considered representative of variation across other equally relevant dimensions, such as age or socioeconomic status.

The *SED* responses were recorded impressionistically, using the 1951 revision of the International Phonetic Alphabet.[7] The *SED* material is presented by locality, with a county name and number – e.g., Northumberland 1 – as illustrated in Figure 2.1 and Table 2.1. Geographic coordinates for each locality in the *SED* are taken from the United Kingdom's Ordnance Survey website.[8] All the results from a given locality are presented together, making it impossible to distinguish the responses of individual informants. The responses in the *SED* thus represent an aggregated sample of the speech habits of a particular locality rather than those of a particular individual and will be referred to accordingly.[9] Where the *SED* presents more than one form of a word for a locality, I select the first unless another form is specifically referred to as "older" or "preferred."

In addition to data taken directly from the *SED*, I take derived data from Anderson's (1987) *A Structural Atlas of the English Dialects* or *SAED*, which presents a series of more than 100 maps showing the geographic distribution

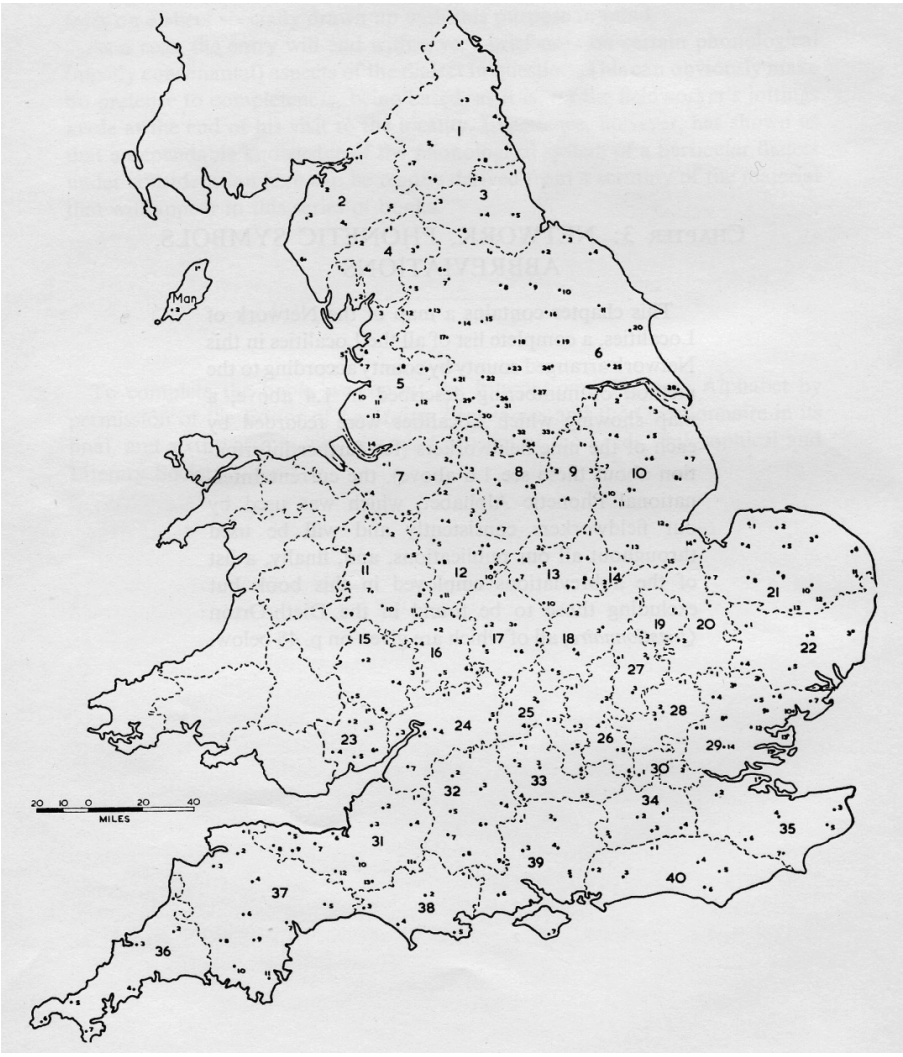
[6] The systematic study of English dialects is well into its second century, building on the foundations of Ellis (1889) and the examples of the continental European and American linguistic surveys. Many studies have drawn on the *SED* for source material, including important introductions to English dialects such as (Wells, 1982a,b) and Trudgill (1999); the atlases by Kolb et al. (1979), Anderson (1987), Upton et al. (1987), Upton and Widdowson (1996), and (Viereck and Ramisch, 1991, 1997). Sanderson and Widdowson (1985) provide a historical review of English dialect research, as well as a very useful discussion of the SED. The data used in this study is available from the author upon request.

[7] Speakers were also tape-recorded in nearly all the localities and in principle their speech could be studied through a rigorous acoustic analysis.

[8] The geographic coordinates are freely available online at `http://www.ordnancesurvey.co.uk/oswebsite/freefun/didyouknow/`.

[9] The degree of variation bears emphasis: even in localities that appear as the most representative of their respective regions in the present analysis, it is common to find two or three sometimes quite different variants in use in a given set of words in a given locality. Competition among variants is especially apparent in areas where regions of generally different usage come into contact.

Figure 2.1: Localities of the *Survey of English Dialects*



Source: Orton, Harold. 1962. *Survey of English Dialects: An Introduction.* Leeds, UK: E. J. Arnold. Reproduced by the permission of the University of Leeds.

Table 2.1: Traditional Counties in the *Survey of English Dialects*

1.	Northumberland	15.	Herefordshire	28.	Hertfordshire
2.	Cumberland	16.	Worcestershire	29.	Essex
3.	Durham	17.	Warwickshire	30.	Middlesex
4.	Westmoreland	18.	Northamptonshire	31.	Somerset
5.	Lancashire	19.	Huntingdonshire	32.	Wiltshire
6.	Yorkshire	20.	Cambridgeshire	33.	Berkshire
7.	Cheshire	21.	Norfolk	34.	Surrey
8.	Derbyshire	22.	Suffolk	35.	Kent
9.	Nottinghamshire	23.	Monmouthshire	36.	Cornwall
10.	Lincolnshire	24.	Gloucestershire	37.	Devonshire
11.	Shropshire	25.	Oxfordshire	38.	Dorsetshire
12.	Staffordshire	26.	Buckinghamshire	39.	Hampshire
13.	Leicestershire	27.	Bedfordshire	40.	Sussex
14.	Rutland				

and frequency of occurrence of different phonetic variants in groups of words found in the *SED*, where all of the words in a given group include a segment that is posited to have taken a single uniform pronunciation in an older form of the language – the "standard" Middle English dialect of the Home Counties of southeastern England. The groups include all of the Middle English short and long vowels, diphthongs, and most of the relatively few consonants that exhibit any variation in the English dialects. Although it is by no means complete or comprehensive, the data set reduces an enormous amount of phonetic information to a tractable form that makes possible a rapid and wide-ranging analysis of phonetic variation in the traditional dialects of England as they existed in the middle of the 20th century.

To provide a somewhat broader perspective on the traditional English dialects recorded in the *SED*, I expand the data set to include the pronunciations of three idealized speakers: a speaker of the Middle English "standard" on which the *SAED* classification of localities' responses is based; a speaker of mid-20th century Received Pronunciation; and a speaker of my own native speech pattern, the "Western Reserve" dialect of northern Ohio, which was chosen by American radio broadcasters in the early 20th century as most closely representing an American standard. More difficult but very enlightening extensions would include data from other regions – Scotland and other English-speaking countries – and from speakers from a wider range of ages and socioeconomic backgrounds.

2.3 Converting Linguistic Data to Quantities: Variants and Features

Phonetic data can be quantified in a number of useful ways: by classification into sets of variant phonemes, by perception- or articulation-based measurement of features such as vowel height and backing, or by measurement of acoustic features through spectrograms and formant tracks. Using such approaches, distances between two speakers' phonemes or segments can be measured as differences between their variants, articulatory features, or acoustic features. Detailed analysis by Heeringa (2004) suggests that no such approach provides an ideal quantification of speech, but that measures of differences between speakers based on them are fairly closely and similarly correlated with native speakers' perceptions of those differences, at least in the aggregate.[10]

I take two distinct approaches to quantifying the *SED* data for this analysis. One quantifies a small number of mainly vocalic phonemes in *SED* responses into sets of features, such as degrees of height, backing, and rounding. The other approach is based on the *SAED*'s classification of a much larger set of *SED* responses into sets of phonetic variants – for instance, the use of [**ou**], [**oᵊ**], [**ia**], etc., in words typically pronounced like *bone*. The feature-based approach makes possible a detailed analysis of subtle variation in a small set of responses; the variant-based approach permits a rapid analysis of a large number of responses with comparatively little effort. Comparing the results of the two approaches yields further understanding of their relative strengths and weaknesses and of dialect variation in general. Although comparison of strings of segments such as words or phrases would provide further insight, the analysis considers only short segments in the interests of economy.

On the whole, these approaches almost certainly considerably understate the full extent of phonetic variation among the localities in the *SED* – the feature-based approach because it is based on such a small set of phonemes, the variant-based approach because its classifications obscure a great deal of variation. That understatement simplifies the use of various algorithms to uncover structure, but to such an extent that a skeptic may reasonably wonder whether the results are not due mainly to the choice of classifications and words than to the use of quantitative tools. Despite their limitations, however, both data sets faithfully record a wide range of variation. Moreover, the sources of understatement do not appear likely to result in any other systematic bias in the measurement of dialect differences. As a result, the distance measures, clusters, and factors uncovered in this study appear likely to be relatively robust to improvements in the measurement of true variation.

[10] See especially Heeringa (2004), Chapter 7, Section 4.3.

2.3.1 Feature-Based Approach: Quantification

The feature-based approach closely analyzes a set of 55 words shown in Table A.1 in Appendix A. The set includes at least one example of every short vowel, long vowel, and diphthong in "standard" Middle English, alone and followed by rhotics, as well as the variable consonants. The approach translates 122 segments (including vowels, diphthongs, and consonants) into vectors of numerical values representing 483 features such as degrees of height, backing, and rounding. That translation – necessarily a somewhat arbitrary process – provides numerical characterizations that can be used to calculate a measure of perceptual or articulatory distance between segments.

According to the approach I adopt here – of my construction but similar to the system of Almeida and Braun (1986) – short and long vowels are represented as vectors of four values: 1.0 to 7.0 for height, 1.0 to 3.0 for the degree of backing, 1.0 to 2.0 for rounding, and 0.5 to 2.0 for length. Thus, for instance, [**a**] takes values of [1.0, 1.0, 1.0, 1.0] while [**uː**] takes values of [7.0, 3.0, 2.0, 2.0]. I ignore most diacritics. Diphthongs are represented by a vector of eight numbers; for example, [**ou**] and [**oːᵊ**] would take values of [5.0, 3.0, 2.0, 1.0, 7.0, 3.0, 2.0, 1.0] and [5.0, 3.0, 2.0, 2.0, 4.0, 2.0, 1.0, 0.5], respectively. For some types of analysis, I convert the values describing the second element of a diphthong to differences between the second element and the first. For the diphthongs above, for example, the values would be [5.0, 3.0, 2.0, 1.0, 2.0, 0.0, 0.0, 0.0] and [5.0, 3.0, 2.0, 2.0, -1.0, -1.0, -1.0, -1.5], respectively. If monophthongs are assigned values of 0.0 for the second set of features, monophthongs and diphthongs can both be included in a square matrix, with the second set of features representing the characteristics of the glide. This second approach makes possible a novel analysis of variations in glide characteristics in the principal component analysis described below. Consonants are represented by a value representing the presence or absence of the relevant distinctive feature; rhotics are represented by two values representing the place and manner of articulation.[11] I also tabulate data on whether multiple responses were recorded in a given locality.

[11] Rhotics are assigned values for place and manner of articulation consistent with the International Phonetic Alphabet as in the system of Almeida and Braun (1986): the uvular trill [ʁ] is assigned a value of 9 for place and 3 for manner; alveolar trill [r] is assigned values of 3 and 4, respectively; alveolar approximant [ɹ] is assigned values of 7 and 4, and retroflex tap/flap [ɽ] is assigned values of 4 and 6. The absence of a rhotic is assigned a place value intermediate between postalveolar and retroflex (5.5) and a manner value of 9. This approach thus places significant emphasis on the presence and realization of rhotics.

2.3.2 Feature-Based Approach: Distribution and Correlations

Only 447 of the features have any variance; the remaining 36 are constant across all localities. That each word has its own history is underlined by the fact that the means and standard deviations of vowel features are distributed more or less uniformly; that is, for any given feature – vowel height, for example – the mean value across localities for any given word is unique, and the mean values for all words show no distinctive pattern at all. Even features of the vowels in words as similar as *sun* and *butter* have noticeably different means and standard deviations. However, feature distributions reveal an important underlying pattern. For every feature, high standard deviations strongly tend to be associated with middling feature values, while low deviations are associated with extreme values; that is, low or high vowels tend to have fairly uniform distributions across speakers. A low (or high) vowel typically is low (or high) in most localities, and so the standard deviation of its distribution tends to be low. In contrast, vowels of medium height tend to have varied expressions across dialects and high standard deviations in their distributions. The same observation holds for backing, rounding, and length, but the pattern has different causes in different cases. For some features, including vowel height, it arises from the fact that the position of middling vowels such as [ɛ] and [ɔ] is quite variable – that is, that middling vowels tend to be rather unstable. In the case of rounding and length, however, it is due mainly to the fact that features with average values closer to the mean are simply those for which many localities use a rounded or short variant while many others use an unrounded or long one.

Relatively few features are closely correlated. Only about two percent of all Pearson correlations between features have an absolute value greater than 0.5, and only about 15 percent are greater than 0.25. The typical feature will have a correlation with absolute value of 0.5 or more with only a dozen other features. However, those averages mask a great deal of variation in the degree of correlation among features. About a quarter of the features have no more than two correlations with absolute value greater than 0.5, but roughly another quarter have correlations that high with 20 or more other features. On closer examination, the highly correlated features turn out to be composed largely of three classes – one composed of fricatives, which all tend to be voiced in the Southwest; another composed of features of second segments of diphthongs, which tend to develop inglides in the North and upglides in the Southeast; and a third composed of rhotic features or, in the non-rhotic dialects, their replacements.

2.3.3 Variant-Based Approach: Quantification

In the variant-based approach, I use data from the *SED* and *SAED* to calculate localities' frequencies of usage of groups of phonetic variants (mainly of vowels) in groups of words believed to have had uniform pronunciations in "standard" Middle English. The variant-based data set summarizes over 400 responses, grouping them into 199 variants of 39 phonemes or combinations of phonemes.[12] The full set of groups of variants is shown in Table A.2 in Appendix A; the words used appear in the index of the *SAED*. (Throughout the rest of the presentation, I write the Middle English form considered common – not to say ancestral – to the group as /**x**/ and the variants recorded in the *SED* as [**x**].)

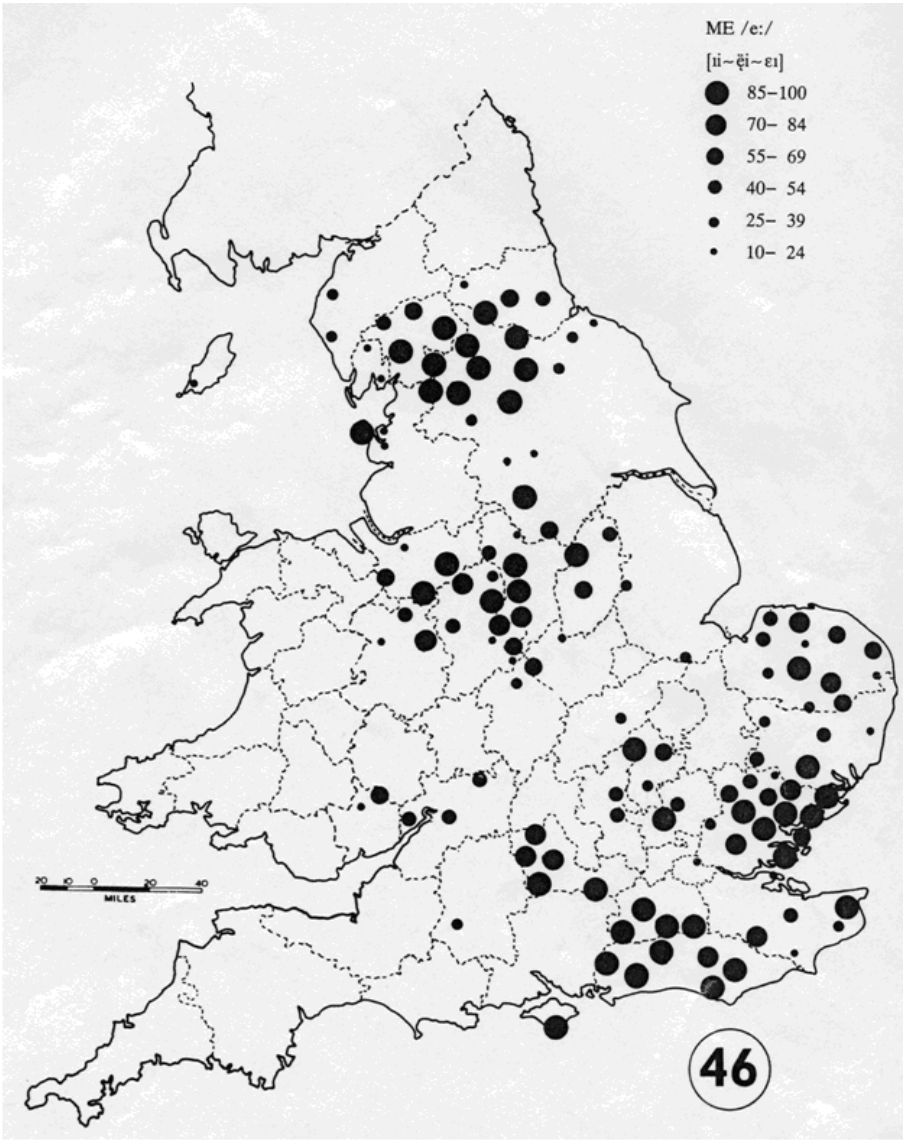
The typical map from the *SAED* shown in Figure 2.2 illustrates strengths and weaknesses of both the data and the approach. The map shows approximate frequencies of use in *SED* localities of variants closely approximating [**əi**], [**ëi**], or [**ɪi**] – that is, rising diphthongs with a centered, center-front, or slightly higher than center-front onset – in a set of 20 words, such as *cheese* and *geese*, believed to have been pronounced with /**eː**/ by speakers of "standard" Middle English.[13] A closer examination of the *SED* material reveals that the variants in question are not used uniformly in the different regions: localities in the Southeast and East Anglia uniformly use [**ɪi**]; those in the Northwest Midlands tend to use [**ɛi**] or [**ëi**] but occasionally [**ɪi**], and those in North Yorkshire typically use [**ëi**] or [**əi**] but also occasionally [**ɪi**]. In this case, then, the responses can be classified further into three separate groups on the basis of differences both geographical and linguistic. (In a few cases, localities from geographically separate regions are classified as having "different" variants even though the variants are actually the same, on the grounds that the variant is likely to have arisen independently in the two regions. For instance, localities in the North who use [**iː** ∼ **eː**] in words with /**ɛː**/ in Middle English are distinguished from those in the South and Midlands who use the same variant, since there is a large geographic gap between the regions in which the variant does not appear at all.)

I also translate the variants described in Table A.2 into vectors of numerical values representing features. In contrast to the first feature-based approach, I

[12] The addition of standard Middle English also requires the addition of 10 variants that were no longer in use in any of the localities under study in the twentieth century; the variants identified for Received Pronunciation and Western Reserve were all in use in at least a few localities.

[13] Because the maps present only ranges of frequencies, I use the averages of those ranges (i.e. 92.5 percent for speakers with frequencies between 85 percent and 100 percent). That approach also requires that each locality's frequencies be normalized to sum to 1.0 over each set of words that are held to share a phoneme in Middle English (since they may not do so automatically) and introduces a layer of error to the data that should be of relatively little consequence for the analysis.

Figure 2.2: A Typical Map from the *Structural Atlas of English Dialects*



Source: *A Structural Atlas of the English Dialects*, Peter Anderson, Copyright 1987 Croom Helm Ltd. Reproduced by permission of Taylor and Francis Books UK.

represent short monophthongs as a vector of three numbers representing height, backing, and rounding; long monophthongs and diphthongs are represented by a vector of six values. The treatment of consonants and rhotics follows the feature-based approach. Where a variant in Table A.2 actually represents a range of distinguishable sounds, its characterization generally takes the mean value for each feature over the range of sounds assigned to the specific variant in the *SAED*. As in the feature-based approach, I ignore most diacritics.

The classification of words into groups imposes several limitations. It tends to understate the true extent of variation in several ways and, by precluding the pairwise comparison of individual words or phonemes in words, limits the range of approaches that can be used to analyze that variation. A historical basis for grouping words together is as compelling as any other approach, but the words may not in fact have been pronounced the same even in "standard" Middle English, in which case they may be inappropriately grouped even on the stated basis. In addition, specified variants often comprise a range of distinguishable sounds – such as [**əi**], [**ëi**], and [**ɪi**] in the example above – thus masking a significant amount of variation. Similarly, even informants using traditional speech forms in a given locality may have substantially different usage that is masked by the presentation of the *SED* material by locality. Perhaps more importantly, two localities' usage may differ considerably across a given group of words even if they have identical frequencies of use of each variant. In the extreme, one can imagine two localities using each of two variants for 50 percent of the words, but using different variants for each word. Conversely, of course, the approach also may tend to overstate the true extent of variation in some cases, as when two localities may use the same variant (for instance, [**ɪi**] in the example above) but be classified as using different variants. Nevertheless, as mentioned above, the approach also has the advantage of allowing the rapid analysis of a large set of responses.

2.3.4 Variant-Based Approach: Distribution and Correlations

The full distribution of variants resembles the asymptotic hyperbolic (or "A-curve") distribution discussed by Kretzschmar and Tamasi (2002) as being common to dialect data: a small set of the variants in the data set accounts for most usage while the minority has relatively limited distributions. The average variant is used in about 20 percent of all localities, but 29 variants (about 15 percent of the total) are used in 50 percent or more of all localities, while 73 (about 36 percent) are used in five percent or fewer. Ignoring variations in the frequency of occurrence of different phonemes, 25 variants account for nearly half of all usage across all phonemes, while the least common 115 variants account for only 10 percent.

Most variants have a relatively unique distribution among and frequency of use within *SED* localities. The distributions of variants may overlap a great deal – even the distributions of variants of the same phoneme – but they rarely entirely coincide. This can be seen most clearly in the Pearson correlations between variants. Only about three percent of all the Pearson correlations between variants are greater than 0.5, and only about 11 percent are greater than 0.25. Those values imply that the typical variant will have a correlation of 0.5 or more with only five or six other variants and a correlation of 0.25 or more with only about 20 of them. However, those averages mask a great deal of variation in the degree of correlation among variants. Most variants have very few large correlations with others: 53 variants have no correlations as large as 0.5, and 95 variants have three or fewer. In contrast, 25 variants have 15 or more correlations of 0.5 or greater and 35 variants have 12 or more. Interestingly, most of the variants with a large number of large correlations are found either in the far Southwest or in the far North of England. That finding suggests that those two regions tend to have relatively distinctive speech forms with numbers of features that regularly co-occur in them, exemplified, for instance, by the very similar geographic distributions of voiced fricatives in the Southwest.

With the exception of the extensive correlation of a relatively small group of variants, the typically low levels of correlation among variants provide strong evidence that most variants do not tend to co-occur very regularly with many others. That observation, in turn, implies that localities in the *SED* may share specific variants in common but are unlikely to share overall patterns of usage, and provides support for the view of dialect variation as a largely continuous phenomenon.

At the same time, the correlations among variants provide a number of clues as to the extent of systematic structural variation among the speech varieties in *SED* localities. The most obvious example is fricative voicing in the Southwest; but perhaps the most notable instance is the set of positive correlations between parallel developments in the Middle English front and back low long vowels shown in Table 2.2. In much of the North of England, the vowels tend to merge with great regularity. In other parts of the North, they both tend to develop inglides. In parts of the North Midlands, they are simply raised, and the degree of raising tends to be correlated. Finally, the vowels both develop upglides in most of the South, and where they do, the heights of the initial vowels in the resulting diphthongs tend to be similar. The co-evolution of the low front and back vowels thus appears to be a genuinely structural development in the English dialects. Note, however, that in most of those cases the correlations between parallel developments, while positive, are not especially large: the structural parallels appear to be at least partly systematic in nature but may perhaps best be interpreted as statistical tendencies rather

Table 2.2: Correlations between Developments in Front and Back Long Vowels

Middle English /**aː**/ develops to:	Correlation:	Middle English /**ɔː**$_1$/ develops to:
[**iə**]	0.8940	[**iə** ∼ **eə** ∼ **ɛə**][a]
[**iɛ** ∼ **jɛ**]	0.7883	[**iɛ** ∼ **jɛ** ∼ **ji**][a]
[**ia** ∼ **ea**]	0.9114	[**ia** ∼ **ea**][a]
[**eə** ∼ **ɛə**][b]	0.6488	[**uə** ∼ **oə** ∼ **ʌuə** ∼ **ɔə** ∼ **oːə**]
[**eː**]	0.5987	[**oː**]
[**iː**]	0.4690	[**uː** ∼ **üː** ∼ **ʏː**]
[**ei** ∼ **eːi**]	0.3075	[**ou**]
[**ɛi** ∼ **ɛːi**]	0.5003	[**ɔu** ∼ **ɒo** ∼ **u**]
[**æi** ∼ **ai**]	0.5032	[**əu** ∼ **ʌu** ∼ **æu**]

a - In medial position.
b - In Yorkshire and Lincolnshire.

than strict relations.

In other cases, the correlations among variants reveal some proposed dialect structures to be largely chimerical. For example, the "Potteries" dialect of north Staffordshire and surrounding counties has been characterized by the following pronunciations:[14]

- *Bait* and *bate* are pronounced [**biːt**]; that is, Middle English /**aː**/ and /**ai**/ merge and develop to [**iː**].
- *Beat* and *beet* are pronounced [**bɛit**] – Middle English /**ɛː**/ and /**eː**/ merge and develop to [**ɛi**].
- *Boat* is pronounced [**büːt**] or [**bʏːt**] – Middle English /**ɔː**$_1$/ or /**ɔː**$_2$/ develop to [**üː** ∼ **ʏː**].
- *Boot* is pronounced [**bɛʊt**] – Middle English /**oː**/ develops to [**ɛʊ**].
- *Bout* is pronounced [**bait**] – Middle English /**uː**/ develops to [**ai**].
- *Bite* is pronounced [**bɑːt**] – Middle English /**iː**/ develops to [**ɑː**].
- *Bought* is pronounced [**bout**] – Middle English /**ou**/ remains or reverts to [**ou**].

[14] See for example Trudgill (1999), p. 41. Trudgill attributes this characterization to the speech of all of Staffordshire plus parts of Cheshire, Shropshire, Derbyshire, Warwickshire, and Worcestershire.

- *Caught* is pronounced [**kɔʊt**] – Middle English /**au**/ develops to [**ɔʊ**].

All of the variants are indeed found in *SED* localities in north Staffordshire, Cheshire, and Derbyshire in words from the 11 relevant Middle English word groups. However, a close inspection reveals that no *SED* locality uses them all; only two localities use the variants in 9 of 11 groups, and 6 other localities use 7 or 8. Taking into consideration frequency of use over the range of words in each group, the highest-scoring locality, Staffordshire 3, uses the "Potteries" variants in 60 percent of the possible occurrences, and only 8 localities use them in more than one-third of possible occurrences. Moreover, further analysis of the variants' frequencies of occurrence reveals that many of their correlations are relatively weak, even in the relevant region. Viewed in terms of correlations between variants and frequencies of use in the *SED* material, actual "Potteries" usage, even among traditional dialect speakers of the mid-twentieth century, appears to be more of a tendency to use certain variants with greater frequency rather than a coherent, distinct linguistic structure.

Taken altogether, the patterns of correlation among features and among variants suggest two important insights into the patterns of variation in the traditional English dialects. First, with the exception of the distinctive shifts of the far North and the far Southwest, phonetic variation in the dialects is simply not very systematic, but instead tends to involve largely uncorrelated variations that, in some areas, coalesce into patterns that appear more systematic. Second, most of the phonetic variation tends to involve single features rather than combinations of features. Even when groups of correlated features appear to indicate greater structural variation, the structural shift involved tends to be fairly simple: rhotic type, voicing or devoicing, or upglides versus inglides in diphthongs.

2.4 Calculating Linguistic Distances

A variety of measures can be used to quantify the difference between specific usages of one speaker or locality and another and to aggregate large numbers of such differences into a single quantity, although none of them should be taken as a perfectly accurate gauge. For variants, for example, distances between specific phonetic segments can be taken as 0 for speakers with the same variant or 1 for speakers with different variants. For feature-based characterizations of segments, distances can be calculated using such measures as Manhattan "city-block" distance or Euclidean distance. (For instance, the Manhattan distance between [**ɛ**] and [**u**] is 7.0; the Euclidean distance 4.6.) Such measures can be converted to logarithms – in this case, 1.95 and 1.53, respectively – on the argument that small differences have a relatively greater perceptual distance and should be given greater weight relative to larger ones.

Any measure of distance between segments can be extended to whole words or utterances using the more complex Levenshtein distance, which is defined as the minimum cost of changing one word or utterance into another by means of insertions, deletions, and substitutions of one segment for another.[15] The Levenshtein distance is extremely useful for comparing dialects but requires the coding of entire words rather than individual segments. I therefore focus on segment-based measures instead.

2.4.1 Segments

To calculate a linguistic distance between segments, I introduce a somewhat novel measure of Euclidean distance between the articulation-based numerical characterizations of segment features. The variant-based and feature-based approaches differ slightly. In the variant-based approach, the distance calculation is generally intended to reflect the number of changes that have occurred since the variants diverged from an ancestral monophthong or diphthong, so that the approach taken depends on the nature of both descendant variants and that of the inferred ancestral vowel. If both variants are of the same type, the distance calculation is straightforward. If one variant is short and the other long or a diphthong, the distance is generally calculated over the characteristics of the short variant and of the first element of the long variant, plus 1.0 for the lengthening or the diphthong's additional element in the diphthong. For example, the difference between [**e**] and [**iː**] is 1.414 – the square root of the sum of 1.0 for the distance between [**e**] and [**i**] plus 1.0 for lengthening. Matters get more complicated still if some descendents of an ancestral short or long phoneme have developed inglides and others offglides, as for example with Old English [**oː**] developing to [**iə**] in some northern dialects and to [**ʊi**] in parts of West Yorkshire. In this case, I calculate the distance between the features of the second element of the first diphthong and those of the first element of the second one, adding 2.0 for the addition of an inglide in the former and an upglide in the latter. I make analogous adjustments for other complex comparisons, such as between shortened and diphthongized descendants of a lengthened ancestral vowel. I generally give a value of 1.0 to distances between variants of consonants (which usually involve voicing or devoicing) except in the case of the various rhotics, where I calculate a Euclidean distance between two-element vectors representing place and manner of articulation.

The feature-based approach differs from the variant-based approach in

[15] First introduced into dialectology by Kessler (1995) to measure dialect distances among speakers of Irish Gaelic, the Levenshtein distance has become a preferred approach for many researchers because of its comprehensiveness and flexibility. The more complex Damerau-Levenshtein distance also accounts for transposition of segments and can therefore address dialect changes involving extensive metathesis.

treating all monophthongs as single segments with four variable features, with length as a feature ranging in value from 0.5 to 2.0. I calculate the distance between a monophthong and a diphthong over the characteristics of the monophthong and the first element of the long variant plus 1.0 for the difference between a monophthong and a diphthong; the distance between diphthongs is a straightforward Euclidean distance calculation over all eight features, ignoring considerations of historical development. Consonants and rhotics are treated as described for the variant-based approach.

2.4.2 Aggregation

Researchers can use a variety of different approaches to aggregate differences between localities usages – whether based on perception, articulation, or acoustics – and to calculate aggregate distance measures, with no approach likely to yield a perfect measure. The present analysis aggregates feature-based Euclidean distances by taking the average distance over all 122 segments in the data set. An aggregate measure of 1.0 thus implies that on average, two localities' phonemes in this set of segments differ about as much as do [e] and [ɛ] or [o] and [ɔ]. Note that the aggregation does not take into account the many segments that are unvarying across English localities. In that sense, the aggregation procedure greatly exaggerates the "true" degree of distance among localities.

The variant-based approach is generally similar, except for adjustments needed to accommodate localities' varying frequencies of use of several different variants. In many cases, localities may share some variants in common but not others, but by construction the data obscures the extent of word-by-word overlap. Localities with the same frequency of use of a particular variant may or may not use them in the same words. Since there is no way to determine the degree of overlap between them short of comparing them word by word, the analysis simplifies the process by assuming that shared frequencies of a given variant correspond to shared pronunciations in specific words (over which the localities have zero linguistic distance), distributes the remainder uniformly among remaining variants, and calculates distances accordingly.

Table 2.3 shows a simple example in which two speakers each occasionally use three different variants over a given set of words (by assumption, only one variant per word), but with different frequencies of use. Variants 1 and 2 are assumed to have a linguistic distance of 1.5; variants 1 and 3 a distance of 2.0, and variants 2 and 3 a distance of 2.5. For the calculation of linguistic distance, the speakers are assumed to share pronunciations where their frequencies of use overlap; that is, they are both assumed to use Variant 1 in 10 percent of the words, Variant 2 in 10 percent, and Variant 3 in 30 percent. For the remaining 50 percent of words, Speaker 1 is assumed to use Variant 2, while

Table 2.3: Calculation of Variant-Based Linguistic Distance: An Illustration

	Speaker 1 Frequencies (%)	Speaker 2 Frequencies (%)	Common to Both Speakers (%)	Remainder (Speaker 1 – Speaker 2; (%)
Variant 1	10.0	40.0	10.0	–30.0
Variant 2	60.0	10.0	10.0	50.0
Variant 3	30.0	50.0	30.0	–20.0
Total	100.0	100.0	50.0	50.0
Linguistic Distances				
Variants 1 & 2	1.5			
Variants 1 & 3	2.0			
Variants 2 & 3	2.5			

Speaker 2 is assumed to use Variant 1 in 30 percent and Variant 3 in 20 percent. The distance calculation between the two speakers is thus 30 percent times 1.5 (the distance between Variants 1 and 2) plus 20 percent times 2.5 (the distance between Variants 2 and 3), for a total of 0.95. Thus, even though the speakers use three different variants with distances between the variants ranging from 1.5 to 2.5, the reasonable assumption that they very likely share at least some of those variants in specific words reduces their average linguistic distance over this set of words to somewhat less than 1.0. The approach yields the minimum possible average distance over this set of words, and therefore almost certainly further understates the actual degree of variation between any two *SED* localities. However, in this respect as with the use of word groups, the approach appears unlikely to result in any other systematic bias in the distance measurement, and the relative distances among localities are therefore probably robust to improvements in the measurement of true variation. I also adjust the variant-based linguistic distance by weighting distances for particular sets of phonemes by the number of words taken by Anderson from the *SED* for each group of words, on the grounds that the relative frequency in that selection may be a fairly reasonable approximation of the frequency of occurrence in traditional general speech.[16] With over 400 responses in total and 39 vowels, diphthongs and consonants, the average phoneme is represented by about 10 tokens, with the number of tokens per phoneme ranging from one to 33.

In addition to the measures discussed here, one may calculate Pearson correlations for vectors of variant frequencies – more accurately thought of as a similarity measure than a distance measure – or a distance measure used in

[16]The distance measures are calculated using Fortran-based programs developed by and available from the author upon request.

genetics, such as Nei's distance, which tends to be rather closely correlated with the Pearson correlation. Although such approaches have less linguistic foundation than the feature-based results presented here, they tend to yield similar results.

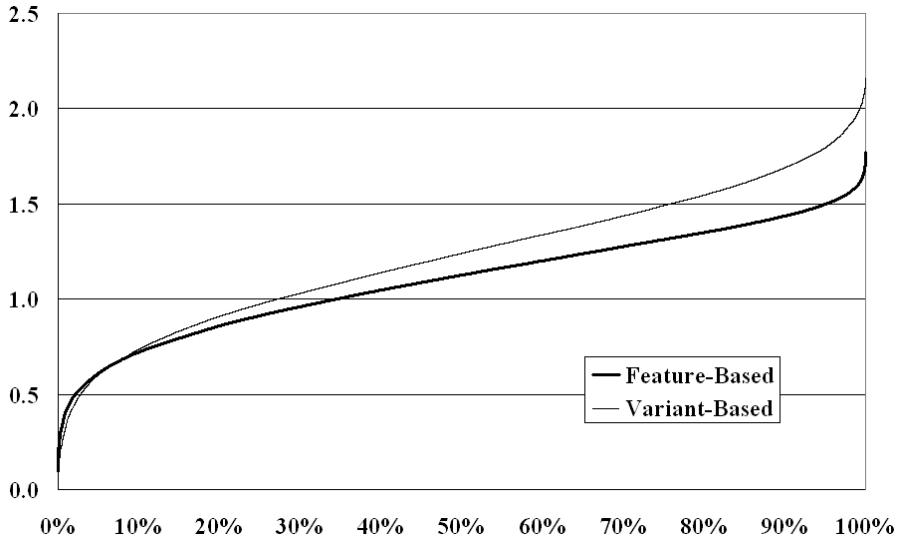
2.4.3 Results: Feature-Based Distances

The feature-based linguistic distances reveal a wide range of variation among the localities in the *SED*, but as shown in Figure 2.3, the vast bulk of the measured similarities are rather uniform, with nearly 85 percent of the distances between localities falling between 0.7 and 1.4. The average distance across all localities is about 1.1. (To put those values in perspective, consider that the maximum distance between short vowels is about 7.3 and between long vowels or diphthongs is about 10.4, and the distance between two randomly chosen vowels or diphthongs – or between two randomly selected vectors of vowels – is thus likely to be around 4.4.) Relatively few localities are radically different from each other, with the largest distances of nearly 1.8 pointing to differences between Cumberland 1 on the Scottish border and a number of localities in East Anglia and the Southwest. Nevertheless, even though neighboring localities are occasionally very similar in their speech forms – the smallest distance points to a pair of localities in Somerset that are very nearly identical in their speech – similarity appears to be the exception rather than the rule even at relatively short geographic distances. Only about two percent of the distances are less than 0.5, implying that the typical locality will be that similar to no more than 6 or 7 other localities. Thus there are relatively few localities that are extremely similar in their speech but also relatively few (mainly those in Northumberland and Devonshire) that are quite different from the rest of England.

Localities have quite diverse patterns of close similarity. Some localities – mainly those centered in broad regions with relatively uniform speech patterns, such as North Yorkshire, Leicestershire, and western Essex – have linguistic distances of less than 0.5 with as many as 25 other localities. In contrast, nearly 40 localities, which generally appear to be either genuine outliers such as Cumberland 1 on the Scottish border or in transition zones between regions of relative uniformity (Worcestershire 1, Herefordshire 7, Hertfordshire 3, Oxfordshire 2, and Northamptonshire 5 being notable examples) have no feature-based distances of less than 0.5.

The distance measures are fairly closely correlated with geographic distance, with a Pearson correlation of 0.70 – a strong indication that geographic separation has played an important role in the development of the traditional English dialects. One consequence of this fact is that localities in the middle of the country – localities that often tend to have relatively large distances

Figure 2.3: Distribution of Linguistic Distances



with neighboring localities because they are in or near a major transition zone between North and South – also tend to have more in common with localities at both ends of the country than do speakers at the ends with each other.[17] Most of the localities with the lowest average distances with all other localities, ranging just above 0.9, are in the Midlands. In contrast, most of the localities that are relatively distinctive from all other localities are in the Far North or the Southwest, with Cumberland 1 having the highest average of over 1.4. Standard deviations of distances tend to be highest for localities in the Southwest, suggesting that those localities tend to have a wide range of distances with other localities, while skewness tends to be highest for localities both in the Far North or the Southwest.

[17]Goebl and Schiltz (1997) describe degrees of "skewness" for *SED* localities on the basis of lexical and morphological characteristics. They show that localities with the same average similarity with other localities may differ in the distribution of distances from low values to high values, so that one is relatively similar to most other localities but very different from a few, while the other is somewhat dissimilar to most but not terribly similar to any. They show that for the linguistic variables they examine, a swath of south-central England has the lowest degree of skewness and thus the greatest degree of integration into the dialect continuum. In the present study, in contrast, the lowest degree of skewness for localities' phonetically-based linguistic distances occurs in the Midlands, mainly but not invariably among those localities that have the lowest average distances.

Feature-based linguistic distances provide another prism through which to view the "Potteries" dialect region, which remains as indistinct as it was when analyzed in terms of variants. If we calculate linguistic distances between idealized "Potteries" pronunciations and *SED* responses for words from the relevant groups – *gate*, *meat*, *cheese*, *white*, *loaf*, *moon*, *house*, *daisy*, *daughter*, and *snow* – we find that localities around north Staffordshire have the smallest distances from the ideal, but that even the localities with the smallest distances – most of them slightly north of Staffordshire in Derbyshire, Cheshire, and Lancashire – have distances that are about 40 percent as large as the largest distances.

Data on the occurrence of multiple responses for features yields no strong patterns. On average, about 11 percent of localities provide multiple responses to a given word, but the frequency varies by word from zero to over 60 percent. No obvious geographic pattern emerges, although localities in three regions – the Thames Valley, Leicestershire, and East Yorkshire – appear to have the lowest frequencies of multiple responses, suggesting that they are areas of relatively uniform speech.

2.4.4 Results: Variant-Based Distances

As shown in Figure 2.3, the variant-based measure is more widely distributed compared with the feature-based distance, with smaller low values and larger high values. (The source of that difference is not obvious, but I speculate that it has to do with the much larger range of responses used in the variant-based analysis and the lower emphasis placed on differences in rhoticity.) Nevertheless, the feature-based and variant-based linguistic distances are closely correlated, with a Pearson correlation of 0.817, and quite similar in patterns of similarity, difference, and correlation with distance. Several broad regions appear to have relatively uniform speech patterns reflected in large numbers of relatively small linguistic distances, while localities in other regions have much higher shortest distances. Largely the same set of distinctively different localities as in the feature-based analysis appear to lie in transition zones between regions of relative uniformity. The localities with the lowest average variant-based distances with all other localities are nearly all from the Midlands, while those with the highest average distances are nearly all from the Far North or the Southwest. Patterns of standard deviations and skewness are also similar to those uncovered using the feature-based approach.

2.4.5 Middle English, Received Pronunciation, Western Reserve, and the *Survey of English Dialects*

Comparison of linguistic distances between the *SED* localities and the three outliers – Middle English, Received Pronunciation, and Western Reserve – also yields several useful insights. As might be expected, no locality has speech patterns that are very similar to Middle English. The mean average feature-based linguistic distance is about 1.7; the smallest is just under 1.3, and the largest is just under 1.9. For the variant-based distances those values are nearly 1.8, just under 1.4, and over 2.6. No clear geographic pattern of similarity emerges, except that localities in the Southwest are consistently the least similar to the Middle English standard. Otherwise, a band of rather uniform dissimilarity with Middle English runs from the Upper North to the southeastern coast. Thus the diachronic distance of roughly six centuries between the "standard" Middle English of southeastern England and the mid-twentieth century traditional dialects is, on average, about 45 percent to 50 percent greater than the average synchronic distance between the traditional dialects. Although the evidence may be too slim to support the weight of the proposition, it may be appropriate to infer that speakers of traditional English dialects in the mid-twentieth century typically differed as much in their speech as speakers in a given locality separated by roughly four centuries of time, and those with the greatest linguistic distances differed as much as speakers separated by roughly 7 to 9 centuries.

Received Pronunciation has an average linguistic distance from *SED* localities that is roughly the same as the overall average – slightly higher for the feature-based distance and slightly lower for the variant-based distance. It tends to have its lowest feature-based distances with eastern and East Midlands localities, particularly with those in Cambridgeshire. It has its lowest variant-based distances with a somewhat more diverse group of localities near the Home Counties – Huntingdonshire, Buckinghamshire, and Hertfordshire in particular – but also with various other localities in Herefordshire, Monmouthshire, and Norfolk. Those patterns may reflect a historical origin of Received Pronunciation in the Home Counties, but it may equally well reflect variations in the extent to which informants in *SED* localities used elements from the standard in their speech. In the absence of a compelling method of distinguishing between those two possibilities, the data do not appear to provide a great deal of insight into the regional sources of Received Pronunciation – except of course that they do not come from Oxford.

Western Reserve has an average feature-based distance from the *SED* localities of about 1.25 and an average variant-based distance of roughly 1.31 – consistent with a time distance of four to four-and-a-half centuries. Even considering only the southern half of England, from which most of the early set-

tlers who influenced the development of American speech immigrated, Western Reserve has average distances consistent with a time distance of about three-and-a-half to four centuries – noticeably larger than the time span from even the earliest English settlements in America to the mid-twentieth century.[18] Although Western Reserve's feature-based distances are clearly lowest for localities in the Southwest and particularly in Somerset – in part because of their shared rhoticity – its variant-based distances are generally lowest for localities spread through southern England from Shropshire to Kent. Western Reserve's lowest feature-based distance, 0.79 with Somerset 1, makes it more similar to that locality than all but 50 English localities – more similar, in fact, than some localities in neighboring counties – and places it rather squarely in the southwestern family of dialects.

Perhaps a more revealing insight is that Received Pronunciation and Western Reserve are in many respects quite similar. Although their feature-based distance is rather significant – 0.99 – they have a variant-based distance of roughly 0.53, making them more similar to each other by that measure than they are to any of the *SED* localities, despite the non-rhotic nature of Received Pronunciation and the strongly rhotic nature of Western Reserve. Closer examination reveals that the difference in distance measures is accounted for in very large part by the feature-based measure's greater weighting of differences in rhoticity.[19] Nonetheless, even by that measure, Western Reserve's distance from Received Pronunciation is only about 25 percent greater than its closest distances, making the American variety closer to the English standard than it is to all but 32 localities. Conversely, about 90 localities are closer to Received Pronunciation than is Western Reserve. The variant-based dis-

[18] Interestingly, regional American informants analyzed in Shackleton (2005) [i.e. Chapter 3] had an average linguistic distance from Lowman's southern English informants of about 1.18, while the average distance among the English informants was about 1.00. For the *SED* localities in the same geographic region as Lowman's informants, the average variant-based linguistic distance with Western Reserve speech is about 1.23, while the average distance among the *SED* localities is about 1.08. Despite the fact that the two studies measure pronunciations of different variants by different informants, they yield fairly similar ranges of similarity among speakers in southern England and between those speakers and American speakers, tending to corroborate these comparisons of synchronic and diachronic distance. However, the similarity between linguistic distances measured from Lowman's data and those measured from *SED* data obtains only for the full data sets; for any given countym the two sets of distances can differ considerably.

[19] The importance of rhoticity as the source of this difference in patterns becomes evident when one compares the patterns of similarity between the two speech forms and those of the *SED* localities. Using the variant-based linguistic measure, Received Pronunciation and Western Reserve have very similar patterns of distance with *SED* localities, such that the patterns have a correlation coefficient of 0.84. Using the feature-based measure, their patterns remain similar except that Western Reserve's distances with the fully rhotic localities of the Southwest are all rather uniformly shifted lower, compared with Received Pronunciation. As a result of that shift, the two patterns become slightly negatively correlated.

tance implies a time distance between the English and American "standards" of roughly two centuries, consistent with a separation around the time of the American Revolution; but the feature-based distance implies a much greater time-distance. Taken together, those findings seem to suggest that the development of an American "standard" may have been fairly strongly influenced by English norms, but that distance measures at their current state of development provide only very rough gauges for the comparison of synchronic and diachronic differences.

2.5 Grouping Speakers: Cluster Analysis and Multidimensional Scaling

Cluster analytic techniques are algorithms that divide the observations in a data set into classes, or clusters, based on relationships within the data – generally some measure of distance or difference between observations.[20] Clustering can thus be said to simplify the data by reducing the differences among observations to a relatively small set of relationships within and among clusters. I use clustering methods here to classify localities into groups whose speech is relatively similar, as gauged by the distance measures discussed above. Clustering techniques include non-hierarchical methods, in which the data is divided into an arbitrary number of groups and each observation is assigned to a particular group, and hierarchical methods, in which groups may be divided into subgroups. Non-hierarchical methods exclude any relation among clusters, while hierarchical methods allow subclusters to be more or less closely related as members of larger clusters. "Divisive" hierarchical methods divide and subdivide a data set into subsets on the basis of distances between data points until some predetermined limit is reached; "agglomerative" hierarchical methods start with each observation as a separate cluster, join the most similar ones, and continue to join the resulting clusters until all clusters have been united. Hierarchical methods can be used to produce phenograms – figures that resemble trees, in which the length and distribution of the branches represent the degree 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 correctly classifying observations depending in part on the nature of the data.[21] One approach that tests the robustness of the results is to

[20] Bartholomew et al. (2002), Chapter 2, provides a clear introduction to cluster techniques, including an illustration involving Midlands data from the *SED*. Romesburg (2004) provides a more detailed overview of cluster analytic techniques, as well as a clear discussion of the strengths and weaknesses of different algorithms.

[21] Kleinberg (2003) shows that no clustering method can simultaneously achieve three

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 localities. Patterns that consistently emerge under different approaches and noisy data are likely to reflect underlying patterns in the data.

For the feature-based approach, I apply seven different hierarchical methods – Ward's Method, Weighted Group Average or Unweighted Pair Group Method with Arithmetic mean (UPGMA), Unweighted Group Average, Single Linkage or Nearest Neighbor, Complete Linkage or Furthest Neighbor, Weighted Centroid, and Unweighted Centroid – to the linguistic distance measures. For the variant-based distance measures, I apply the same hierarchical methods to four different distance measures: Pearson correlation, Nei's distance, and unweighted and weighted linguistic distance. For each method and distance measure, the clustering is carried out 100 times with random perturbations to the data, all using the Rug/L04 software developed by Peter Kleiweg.[22]

Geneticists use similar methods to infer relatedness among distinct populations, using data on the frequency of occurrence of different genetic variants in different populations to calculate estimates of genetic distance such as Nei's distance. One particularly useful approach is to calculate a family tree that minimizes the squared errors between the genetic distances between each pair of observations and the distances between them along the branches of the estimated tree. Here I apply two such estimations, called Kitsch and Fitch, to the variant-based measures of linguistic distance, using programs from the Phylogeny Inference Package (PHYLIP version 3.65) developed by Joseph Felsenstein. Kitsch uses the matrix of linguistic distances among observations to construct a rooted tree that minimizes the squared errors between the linguistic distances and the distances between the observations along the branches of the calculated tree; Fitch uses the same approach to calculate an unrooted tree.[23] (Other approaches used in genetics are generally less appropriate to distance data.)

2.5.1 Multidimensional Scaling

Multidimensional scaling (MDS) refers to a set of mathematical techniques that reduce the variation in a data set to a manageably small arbitrary number of dimensions, allowing the user to uncover fundamental relationships in

simple, desirable properties (scale invariance, consistency, and richness), and that every method involves unavoidable trade-offs among these desirable properties.

[22] Rug/L04 is freely available online at `http://www.let.rug.nl/~kleiweg/indexs.html`.

[23] PHYLIP is freely available online at `http://evolution.genetics.washington.edu/phylip.html`.

the data.[24] As cluster analysis may simplify data by reducing variability to a relatively small set of clusters, multidimensional scaling simplifies the interrelationships in the data by reducing as much of the variation as possible to a relatively small number of dimensions.

MDS techniques are similar to the principal component techniques described below, but involve weaker assumptions about the data. The techniques can be applied to any measure of distance or difference between observations in the data set, and can be used to develop two- or three-dimensional maps in which distances between points reflect differences among the observations. Here, I apply MDS to the full set of results of the cluster analyses described above, reducing the variation to a set of points in three dimensions that can be represented as colors on a standard map, again using the Rug/L04 software. The results reveal relatively clear dialect regions and transition zones that appear to be robust to variations in the clustering approach and distance measure used.

2.5.2 Results

Different clustering methods tend to classify the speakers into a variety of different regional patterns. Nevertheless, taken as a whole the clustering methods produce a rather clear picture of the traditional English dialect regions. Drawing on a range of algorithms and distance measures, and introducing multiple perturbations to test the robustness of the results, the cluster analysis consistently yields the pattern of clustering shown in the "honeycomb" maps in Figure 2.4, one of which presents the output of a multidimensional scaling of the aggregated clustering results using all of the feature-based distance measure, and the other of which shows results using the variant-based measures.[25]

The maps reveal a pattern of 7 more-or-less distinct major regions and roughly twice as many minor ones, appearing as differently shaded regions. The regions vary quite a bit in the degree of uniformity, measured by the average linguistic distance between a region's localities. The most diverse major region's average distance is nearly 50 percent greater than that of the most uniform; for the minor regions, average distances within regions vary by a factor of nearly four.

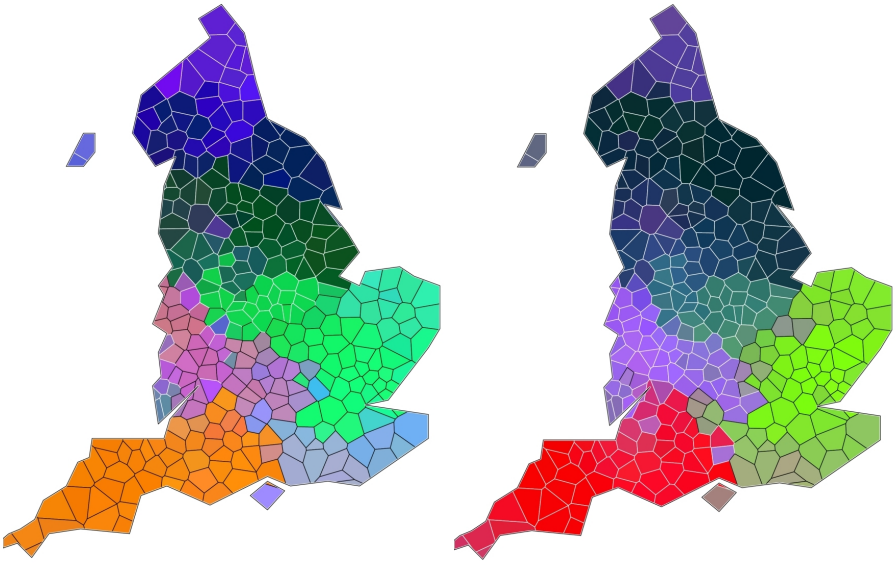
- The ***Far North*** encompasses all of the old counties of Northumberland and the northern part of old Durham. The localities of this region consistently cluster together to the exclusion of all others, under both variant- and feature-based approaches and using any distance measure.

[24] Both Kruskal and Wish (1978) and Bartholomew et al. (2002), Chapter 3, provide straightforward introductions to multidimensional scaling.

[25] The polygons in these maps represent individual localities, each of which is located in the middle of its polygon.

Figure 2.4: English Dialect Regions

(a) Feature-based (b) Variant-based



- By any linguistic measure, the localities just south of the Far North tend to cluster together along an east-west axis rather than with localities to their north or south. The ***Upper North*** includes Cumbria (i.e. the old counties of Cumberland, Westmoreland, and northern Lancashire), southern Durham, and most of North and East Yorkshire. However, under the variant-based approach, a more restricted area encompassing only northern Lancashire, southern Westmoreland, and northeastern Yorkshire appears as a somewhat distinct subregion – here designated the *Upper Northwest*. I designate the remainder of the region *Cumbria-North Yorkshire*.

- The localities in the ***Lower North*** also cluster along an east-west axis into a broad, linguistically relatively diverse region, including not only West and South Yorkshire but also Lancashire (including Mersey and Greater Manchester), Cheshire, Derbyshire, Nottinghamshire, and most of Lincolnshire. Under any linguistic measure, *Lincolnshire* in the east forms a distinct subregion that appears to bear affinities not only with the

rest of the Lower North but also with the Upper North as well. Under the feature-based measure, *Cheshire-Derbyshire* forms a quite distinct subregion as well. Under the variant-based measure, I designate the Lower North excluding Linconshire as the *Lower Northwest*.

- Linguistically speaking, the ***Central Midlands*** region is the most internally uniform of the broad regions. The *Staffordshire* subregion to the west, including nearly all of Staffordshire and the northern tip of Worcestershire, is rather more diverse, while the *East Central Midlands*, which includes the southeastern edge of Staffordshire, the northern half of Warwickshire, all of Leicestershire and Rutland, most of Northamptonshire, and most southerly section of Lincolnshire, is very uniform.

- The most linguistically diverse broad region is the ***Upper Southwest***, even the subregions of which are more internally diverse than most of the major regions. A subregion, designated the *West Midlands*, includes all of Shropshire, Hereford, and Monmouth, as well as most of Worcestershire and northwestern Gloucestershire. A second subregion, designated the *Central South*, includes the eastern edges of Worcestershire and Gloucestershire, the southern half of Warwickshire, the southwestern corner of Northamptonshire, all of Oxfordshire, most of Buckinghamshire, and western Bedfordshire.

- The ***Southeast***, including all of East Anglia and the Home Counties, is more diverse than all other regions but the Upper Southwest, but its variation is more uniform from locality to locality. The region splits into three subregions: *North Anglia*, including all of Norfolk and most of Suffolk; the *Central Southeast*, including the southwestern corner of Suffolk, Cambridgeshire, Huntingdonshire, most of Bedfordshire, Hertfordshire, Middlesex, Essex, and the areas of Kent and Surrey nearest to London; and the *Southeast Coast*, which includes Berkshire, Sussex, most of Surrey and Kent, and the Isle of Wight.

- The ***Lower Southwest*** also splits into three regions, even though it is nearly as uniform as the Central Midlands despite including twice as many localities. The rather diverse *Central Southwest* includes nearly all of Hampshire, all of Wiltshire and Dorsetshire, most of Somerset, and the southern half of Gloucestershire. *Devonshire*, which takes in all of that county as well as western Somerset and eastern Cornwall, is the most linguistically uniform subregion except for Lincolnshire, while *Cornwall* encompasses the rest of that county.

The maps in Figure 2.5 illustrate the variant-based distances between the 15 most typical regional localities and all of the other *SED* localities, with

darker coloring denoting localities with greater similarity to the most typical locality in the relevant region.[26] Feature-based linguistic distances between the most typical localities and the other localities in their regions yield very similar patterns. Some localities are quite similar to all of the others in their designated region, revealing a fairly large, relatively uniform dialect region. That pattern shows up quite clearly in the Far North, Lincolnshire, Leicestershire, and the subregions of the Southwest. In other cases, the most typical localities do not appear to be strongly similar to many of the other speakers in their region at all – for instance, in the Lower Northwest, the West Midlands, and the Central South. Those regions appear to be considerably less uniform in their speech patterns, and are perhaps better thought of as transition zones than as distinct dialect regions. Other regions – notably in the Upper North and the Southeast – are neither as uniform as Lincolnshire nor as diverse as the Central South.

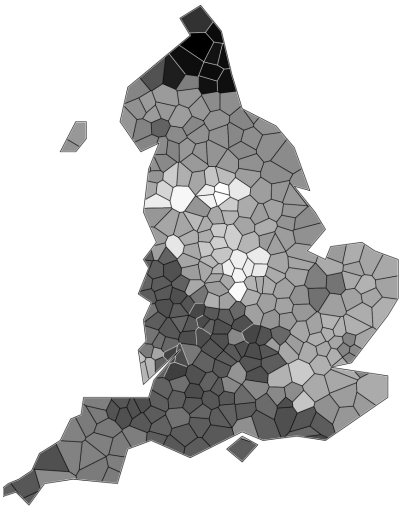
Under most approaches and measures, the most important boundary separates the South and Midlands and the second most important separates the Southeast from the Southwest and West Midlands; other important boundaries separate the Midlands from the North, the Lower North from the Upper North, and the Upper North from the Far North. However, the application of multidimensional scaling to the aggregated results reveals subtle gradations within clusters as well as outliers within each region – for instance, Hampshire 4 and Somerset 1 in the Southwest, Bedford 1 and Cambridgeshire 1 in the Southeast, and Oxfordshire 2 in the West Midlands. (The fact that those localities take the colors of more distant regions does not necessarily imply that they cluster with those regions; rather, they typically are transitional localities that have an unusual mix of variants from neighboring regions.) Some regions are very distinct, but others less so. In particular, the Central South, Lower Northwest, and Lincolnshire subregions occasionally cluster into other regions entirely, suggesting that the speech patterns in those regions have somewhat more diffuse affinities than most of the other regions. On the whole, however, the boundaries are remarkably robust and clear, as are the transitional areas.

The regional clustering resulting from this approach finds corroboration in a separate approach involving average regional frequencies of variants and average regional values of features. If one compares every locality's frequencies of variant usage to regional average frequencies of variant usage, the usage patterns of the localities in that region are all nearly always more closely correlated with the region's average usage than are any other region's localities. The same pattern holds for feature values: values of localities in a region are usually more closely correlated with that region's average feature values than with those of any other region. (The exceptions tend to be precisely those localities that appear as outliers in terms of linguistic distance within their

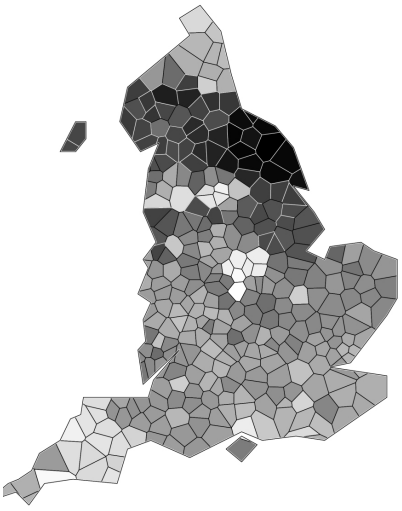
[26]Similar maps in Viereck (1985) showing lexical and morphological distances reveal similar patterns.

Figure 2.5: Variant-based Linguistic Distances for Typical Localities

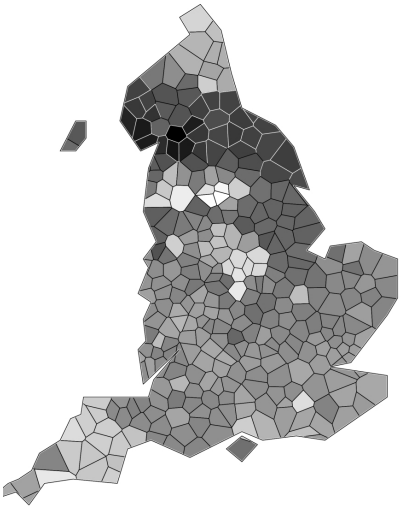
a. Far North: Northumberland 3



b. Cumbria-North Yorks.: Yorkshire 11



c. Upper Northwest: Yorkshire 5



d. Lower Northwest: Yorkshire 33

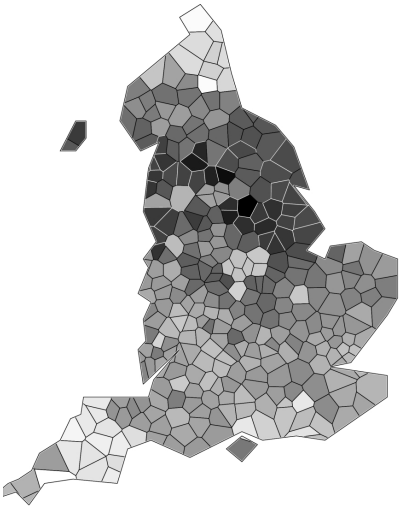
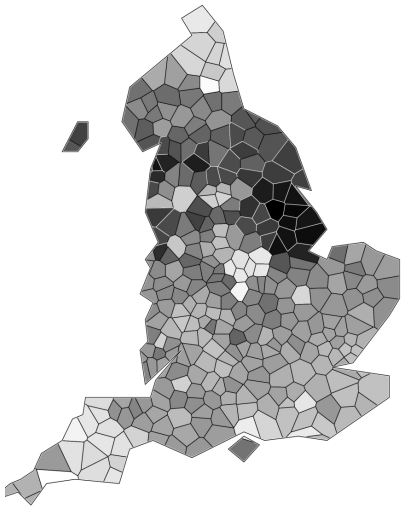
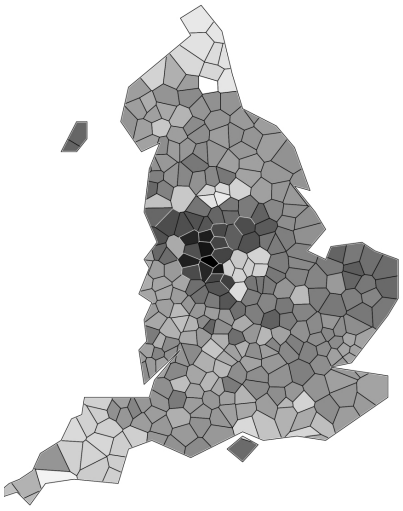


Figure 2.5: Variant-based Linguistic Distances for Typical Localities (Cont.)

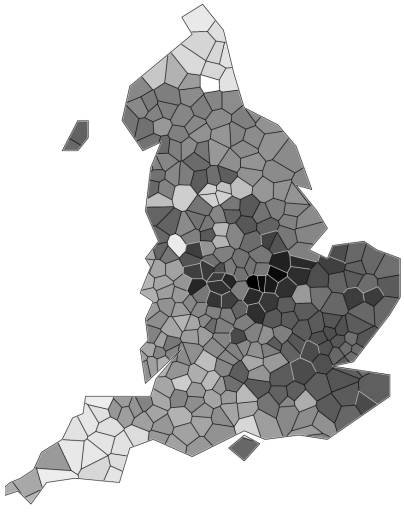
(e) Lincolnshire: Lincolnshire 4



(f) Staffordshire: Staffordshire 6



(g) E. Central Midlands: Leicestershire 9



(h) West Midlands: Monmouth 6

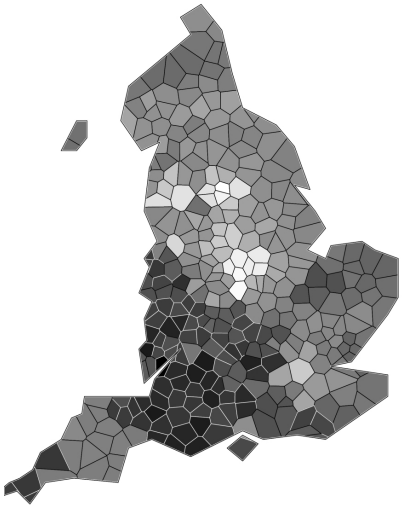
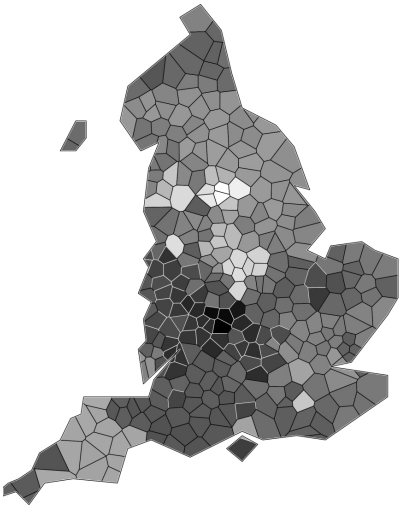
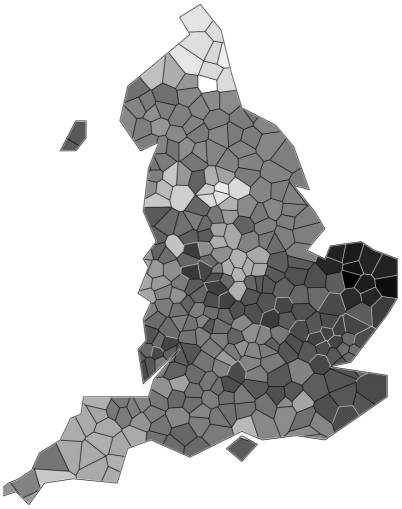


Figure 2.5: Variant-based Linguistic Distances for Typical Localities (Cont.)

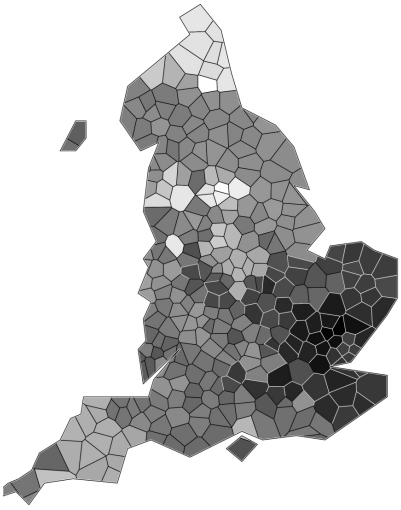
(i) Central South: Warwickshire 7



(j) North Anglia: Norfolk 9



(k) Central Southeast: Essex 2



(l) Southeast Coast: Berkshire 5

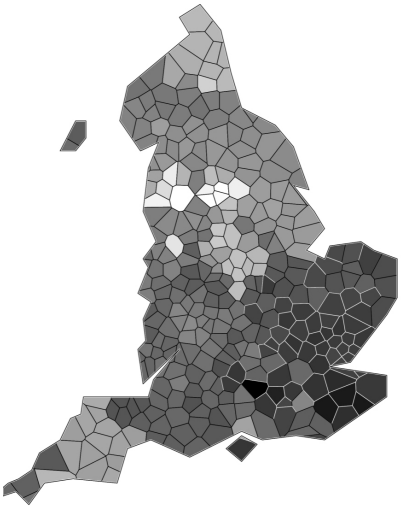
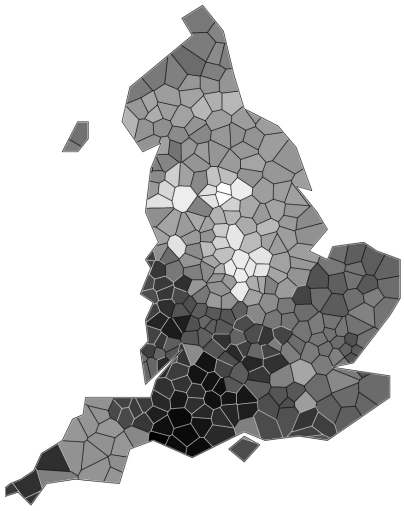
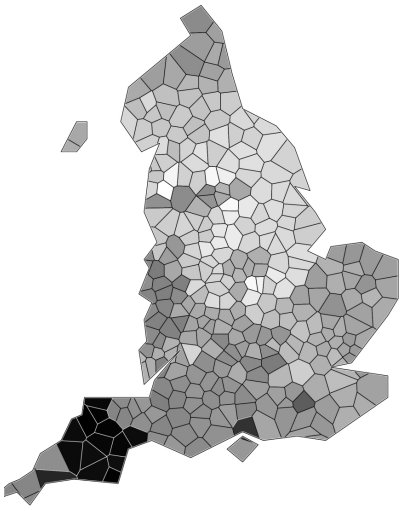


Figure 2.5: Variant-based Linguistic Distances for Typical Localities (Cont.)

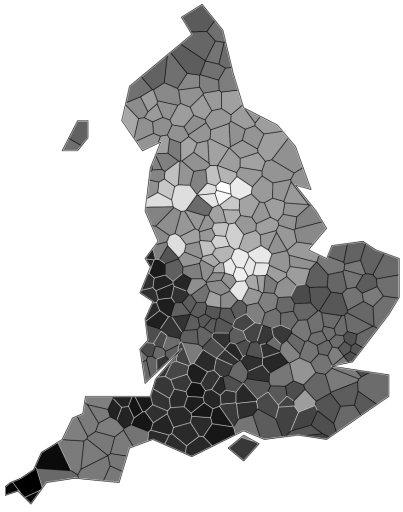
(m) Central Southwest: Dorsetshire 3



(n) Devonshire: Devonshire 3



(o) Cornwall: Cornwall 5

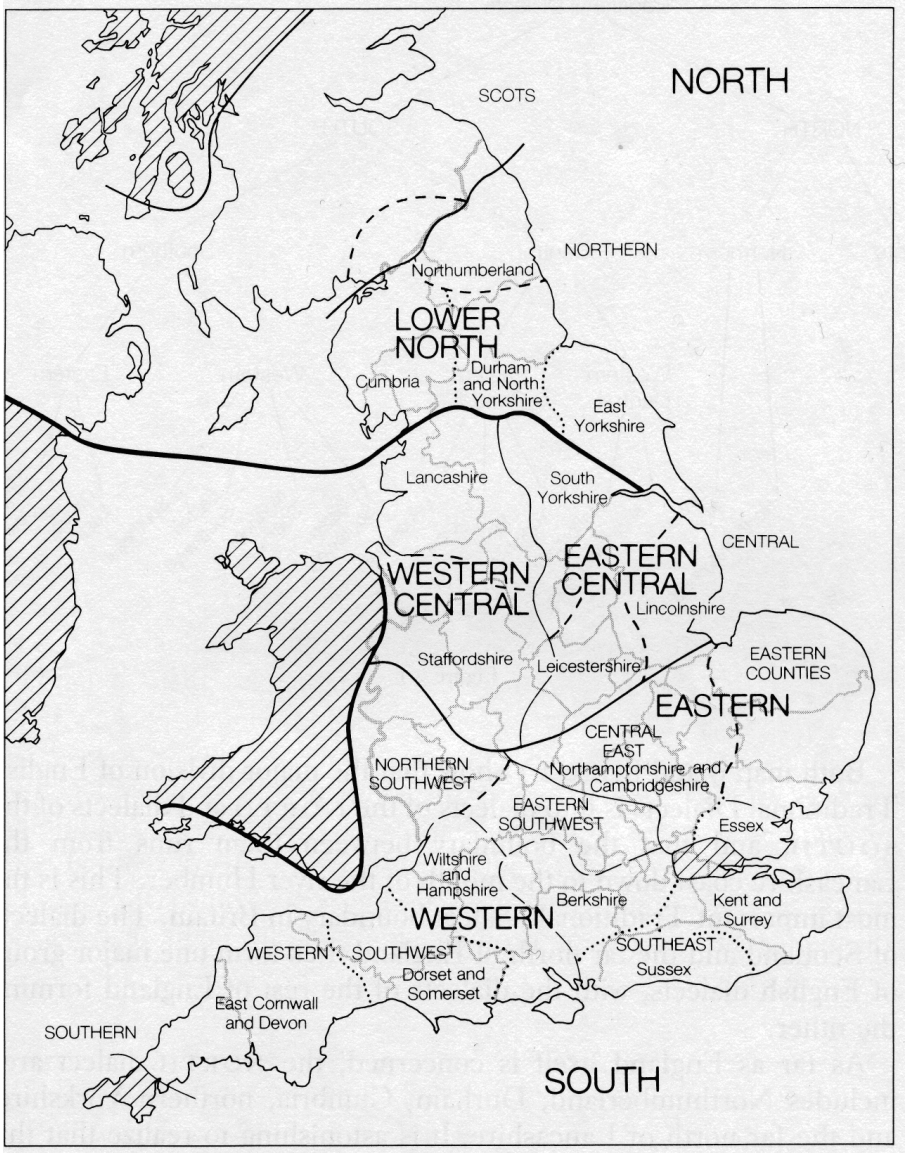


respective regions.) For the Far North, for instance, every locality's vector of frequencies has a Pearson correlation of 0.86 or higher with the vector of average regional frequencies in the Far North, while the next highest-scoring locality has a correlation of 0.80. Moreover, in a majority of cases the locality with the highest correlation with the vector of average regional frequencies is also the "most typical" locality in the sense of having the lowest average distance with all the other localities in the region. Those results reinforce the impression that the regional clustering is underlain by sets of linguistic characteristics in each region to which the localities in the respective regions tend to gravitate.

The dialect regions delineated here are very similar to those found in the dialect map of Trudgill (1999), shown in Figure 2.6, with only a few distinctive differences: the Midlands (in Trudgill's terms, the Central) dialect region is much more restricted and is associated with the northern family of dialects rather than the southern one; the Southeast Coast clusters among the Southeastern dialects rather than the Southwestern ones, and much of Trudgill's Southeast clusters into a larger Central Southeastern region that extends up to the Midlands boundary. Moreover, the analysis finds very little distinction between east and west in the northern dialect regions: parts of Cumberland consistently cluster with parts of eastern Yorkshire, and much of Lancashire consistently clusters with Darbyshire and Nottinghamshire (though not Lincolnshire).

The delineation presented here also bears a strong resemblance to that described by Ellis (1889), except that Ellis places the Southeast Coast with the Southwest and Northamptonshire in the Southeast. It bears similarities to several maps presented in some dialectometric studies using lexical and morphological variants in the *SED* but differs noticeably from others. It resembles the "centers of gravity" described by Händler and Viereck (1997), exhibits a moderate degree of similarity with the clusters identified by Goebl (1997) in a complete linkage analysis using several distance measures over the same data, and is fairly similar to Thomas's (1997) characterization as well. However, the delineation is not very similar to that proposed by Goebl and Schiltz (1997) on the basis of localities' number of shared features in that data, even though many of the relatively strong boundaries that they describe appear in Figure 2.4; nor is it similar to the dialect division presented by Inoue and Fukushima (1997) on the basis of a multivariate analysis of that data. It remains an open question whether or not an exactly analogous approach using lexical and morphological data would yield essentially the same demarcation as the present study.

Figure 2.6: The Traditional Dialect Regions from Trudgill (1999)



Source: *The Dialects of England*, 2nd Revised Edition, Peter Trudgill, Copyright 1999 Blackwell UK. Reproduced by permission.

2.5.3 Phylogenetic Inference

Applied to the variant-based distance measures, the phylogenies produced by the Kitsch and Fitch programs strongly support the delineation of dialect regions described above, including the major and minor boundaries and the various outliers. The only noticeable deviation is that the Midlands region tends to expand slightly north to include those localities in Cheshire and Derbyshire that are closest to Staffordshire. In the case of Fitch, the strongest boundary separates the Southeast from the rest of England; in the case of Kitsch, it separates the North and South. Interestingly, and consistent with the modest variant-based linguistic distance between them, both Received Pronunciation and Western Reserve speech appear as a pair of fairly closely related outliers – within the Southeast in the case of Kitsch, and separate from all the *SED* localities in the case of Fitch.

2.6 Exploring Dialect Geography: Multiple Regression and Barrier Analysis

2.6.1 Multiple Regression Analysis

Multiple regression analysis quantifies the relationships between a variable of interest – a dependent variable – and a number of other independent variables, allowing for interaction among the latter.[27] For instance, multiple regression can be applied to a set of measurements of a group of individuals' weights, heights, waistlines, age, and gender to analyze how their weights vary with their other characteristics. (In this example, gender would be represented by a 0 for one gender and a 1 for the other; an independent variable that takes this form is referred to as a "dummy" variable.)

I use regression analysis to test for a statistically significant relationship between the various measures of linguistic distance between localities and the geographic distance between them, using the Statistical Package for the Social Sciences (SPSS) for Windows Version 7.5.[28] However, by introducing dummy variables that represent localities' regional affiliations, distance regressions can be used to explore whether differences among localities can be interpreted as being more-or-less wholly a matter of geographic separation that has resulted in the gradual accumulation of differences over generations, or whether those differences also vary systematically across the dialect regions identified by cluster and phylogenetic analyses.

[27] See Tabachnick and Fidell (2000), Chapter 5, for an overview of multiple regression analysis.

[28] SPSS is available at `http://www.spss.com/`.

2.6.2 Distance Regressions

Table 2.4 shows the results of a regression of the variant-based linguistic distances on the natural logs of geographic distances and on a set of intra- and interregional dummy variables.[29] Each of the 120 dummies takes a value of 1.0 only when two localities are from two specific regions. For example, the dummy variable "Dummy, Regions 1 and 15" takes a value of 1.0 only when one locality is in the Far North and the other is in Cornwall. As a result, the parameter that is estimated on this variable is based only on linguistic distances between localities in those two regions.

Table 2.4: Distance Regressions with Regional Dummy Variables

Variable	Value	Variable	Value
Adjusted R-Square	0.774	Dummy, Regions 3 and 3	0.024*
Constant Term	0.35	Dummy, Regions 3 and 4	0.348
Distance Coefficient	0.137	Dummy, Regions 3 and 5	0.126
Dummy, Regions 1 and 2	0.385	Dummy, Regions 3 and 6	0.497
Dummy, Regions 1 and 3	0.513	Dummy, Regions 3 and 7	0.490
Dummy, Regions 1 and 4	0.686	Dummy, Regions 3 and 8	0.678
Dummy, Regions 1 and 5	0.534	Dummy, Regions 3 and 9	0.621
Dummy, Regions 1 and 6	0.697	Dummy, Regions 3 and 10	0.807
Dummy, Regions 1 and 7	0.678	Dummy, Regions 3 and 11	0.760
Dummy, Regions 1 and 8	0.603	Dummy, Regions 3 and 12	0.746
Dummy, Regions 1 and 9	0.551	Dummy, Regions 3 and 13	0.944
Dummy, Regions 1 and 10	0.809	Dummy, Regions 3 and 14	1.182
Dummy, Regions 1 and 11	0.802	Dummy, Regions 3 and 15	0.737
Dummy, Regions 1 and 12	0.819	Dummy, Regions 4 and 4	0.218
Dummy, Regions 1 and 13	0.787	Dummy, Regions 4 and 5	0.178
Dummy, Regions 1 and 14	1.250	Dummy, Regions 4 and 6	0.399
Dummy, Regions 1 and 15	0.636	Dummy, Regions 4 and 7	0.370
Dummy, Regions 2 and 2	0.058*	Dummy, Regions 4 and 8	0.583
Dummy, Regions 2 and 3	0.107	Dummy, Regions 4 and 9	0.552
Dummy, Regions 2 and 4	0.290	Dummy, Regions 4 and 10	0.700
Dummy, Regions 2 and 5	0.163	Dummy, Regions 4 and 11	0.682
Dummy, Regions 2 and 6	0.383	Dummy, Regions 4 and 12	0.812
Dummy, Regions 2 and 7	0.368	Dummy, Regions 4 and 13	0.870
Dummy, Regions 2 and 8	0.602	Dummy, Regions 4 and 14	0.995
Dummy, Regions 2 and 9	0.589	Dummy, Regions 4 and 15	0.627
Dummy, Regions 2 and 10	0.647	Dummy, Regions 5 and 5	0.251
Dummy, Regions 2 and 11	0.623	Dummy, Regions 5 and 6	0.469
Dummy, Regions 2 and 12	0.753	Dummy, Regions 5 and 7	0.425
Dummy, Regions 2 and 13	0.891	Dummy, Regions 5 and 8	0.581
Dummy, Regions 2 and 14	1.910	Dummy, Regions 5 and 9	0.557
Dummy, Regions 2 and 15	0.685	Dummy, Regions 5 and 10	0.788

Continued on Next Page

[29] That approach is consistent with the observation that increasing distance tends to have a decreasing incremental influence on linguistic differences.

Table 2.4 : Distance Regressions with Regional Dummy Variables (Continued)

Variable	Value	Variable	Value
Dummy, Regions 5 and 11	0.698	Dummy, Regions 8 and 14	0.676
Dummy, Regions 5 and 12	0.885	Dummy, Regions 8 and 15	0.201
Dummy, Regions 5 and 13	0.838	Dummy, Regions 9 and 9	0.175
Dummy, Regions 5 and 14	1.820	Dummy, Regions 9 and 10	0.420
Dummy, Regions 5 and 15	0.567	Dummy, Regions 9 and 11	0.440
Dummy, Regions 6 and 6	0.125	Dummy, Regions 9 and 12	0.467
Dummy, Regions 6 and 7	0.193	Dummy, Regions 9 and 13	0.426
Dummy, Regions 6 and 8	0.622	Dummy, Regions 9 and 14	0.600
Dummy, Regions 6 and 9	0.453	Dummy, Regions 9 and 15	0.148
Dummy, Regions 6 and 10	0.486	Dummy, Regions 10 and 10	0.105
Dummy, Regions 6 and 11	0.469	Dummy, Regions 10 and 11	0.161
Dummy, Regions 6 and 12	0.647	Dummy, Regions 10 and 12	0.329
Dummy, Regions 6 and 13	0.912	Dummy, Regions 10 and 13	0.610
Dummy, Regions 6 and 14	1.930	Dummy, Regions 10 and 14	0.946
Dummy, Regions 6 and 15	0.717	Dummy, Regions 10 and 15	0.487
Dummy, Regions 7 and 7	0.041*	Dummy, Regions 11 and 11	0.105
Dummy, Regions 7 and 8	0.568	Dummy, Regions 11 and 12	0.289
Dummy, Regions 7 and 9	0.456	Dummy, Regions 11 and 13	0.700
Dummy, Regions 7 and 10	0.404	Dummy, Regions 11 and 14	0.935
Dummy, Regions 7 and 11	0.353	Dummy, Regions 11 and 15	0.450
Dummy, Regions 7 and 12	0.538	Dummy, Regions 12 and 12	0.920
Dummy, Regions 7 and 13	0.854	Dummy, Regions 12 and 13	0.621
Dummy, Regions 7 and 14	0.995	Dummy, Regions 12 and 14	0.830
Dummy, Regions 7 and 15	0.591	Dummy, Regions 12 and 15	0.365
Dummy, Regions 8 and 8	0.262	Dummy, Regions 13 and 13	0.780
Dummy, Regions 8 and 9	0.268	Dummy, Regions 13 and 14	0.272
Dummy, Regions 8 and 10	0.446	Dummy, Regions 13 and 15	0.000
Dummy, Regions 8 and 11	0.494	Dummy, Regions 14 and 14	0.142
Dummy, Regions 8 and 12	0.442	Dummy, Regions 14 and 15	0.180
Dummy, Regions 8 and 13	0.349	Dummy, Regions 15 and 15	0.048*

* - Not significant at the 0.99 level.

According to the adjusted R-square of the variant-based regression, geographic distance and regional differences account for about 77 percent of the variation in measured linguistic distances. (Excluding the dummies, geographic distance alone accounts for about 50 percent.) All but four of the regional dummy variables have p-values of 0.01 or less – an unusually strong result in any kind of cross-section regression, and pointing to speakers' regional affiliations as an important source of their linguistic (dis)similarities. The value of the parameter for the geographic distance variable indicates that ignoring regional affiliation – itself, however, partly a function of geographic distance – 100 miles of distance between two localities increases their linguistic distance by about 0.63 (0.137 times 4.605, the natural log of 100), and the

average geographic distance of 129 miles should result in a linguistic distance of about 0.67, roughly 54 percent of the average linguistic distance.

The results of the feature-based regression (not shown) are quite similar: the adjusted R-square of the regression indicates that geographic distance and regional differences account for about 80 percent of the variation in feature-based linguistic distances; all but six of the regional dummy variables are highly significant; geographic distance parameter indicates that ignoring regional affiliation, the average geographic distance should result in a linguistic distance of about 0.68, about 62 percent of the average distance.

When dummy variables are included in a regression, one of them is left out, and the constant term is, in effect, the parameter for that dummy. (The proper values for the other dummy parameters therefore also include the constant term.) In this analysis, the constant term applies to localities within the Region 1, the Far North, indicating that all else being equal, hypothetical speakers living in the same location in Region 1, the Far North, would have a variant-based distance of about 0.035 and an average feature-based linguistic distance of 0.233 – again illustrating the fact that the smallest feature-based distances tend to be larger than the smallest variant-based distances. For each distance measure, the value of the parameter for "Dummy, Regions 8 and 8" indicates that localities in the West Midlands typically are considerably more distinct from each other than those in the Far North are from each other.[30] The interregional dummy variables take similar interpretations, except that by construction they incorporate into the value of the dummy parameter the effect of the average distance between locations in two separate regions. Thus, for instance, the values of "Dummy, Regions 1 and 14" (1.025) and "Dummy, Regions 1 and 15" (0.636) in Table 2.4 indicate that the variant-based distance between two randomly chosen localities from the Far North and Devonshire is likely to be considerably larger than that between two localities from the Far North and Cornwall, despite the fact that the Far North is very nearly the same distance from Devonshire and Cornwall. However, the corresponding values for the feature-based regression are noticeably smaller, reflecting the fact that the feature-based linguistic distances are not as widely dispersed as the variant-based linguistic distances.

The Pearson correlation between feature-based and variant-based dummies across equivalent regions is quite high – 0.82 – despite the fact that values for individual dummies are frequently very different. That strong correlation –

[30] The negative value of the parameter for "Dummy, Regions 5 and 5" in the variant-based regression has no realistic interpretation, but it indicates that localities in Lincolnshire are much more similar to each other than is typical in most regions of England. The insignificant values for four intraregional dummies imply that the effect of geographic distance on intraregional linguistic differences in those regions cannot be distinguished from its effect on such differences in the Far North.

nearly as strong as the correlation between variant-based and feature-based linguistic distances – provides another indication that the regional variations uncovered in this analysis are quite robust to different characterizations of the phonetic data and likely reflect real patterns of dialect variation.

Table 2.5 illustrates the importance of regional speech differences, showing the percentage difference between the average predicted variant-based linguistic distances between localities in different regions, with and without regional dummies. Positive values indicate that localities in the relevant regions have greater-than-average speech differences, given their geographic separation; negative values point to smaller linguistic differences. For example, one of the largest values in Table 2.5, 24.9 for interregional differences between Staffordshire and the bordering West Midlands, indicates that the dialect boundary between the two regions is a particularly significant one. Conversely, the quite negative values between North Anglia, the Central Southeast, and the Southeast Coast point to relatively uniform speech across those three regions. Generally speaking, the largest values in the table apply to differences between dialect regions in the Midlands and Upper Southwest, suggesting that there is a broad central region of England of particularly great dialect diversity, perhaps because the conditions for interaction and influence among speakers is greater in a wide transition zone between the North, Southeast, and Southwest. Note also that the highest values among intraregional dummies are for Staffordshire, the West Midlands, and the South Central region, all part of the large transition zone in the center of England; but the East Midlands dummy takes a very low value, indicating that it, like its eastern neighbor Lincolnshire, is a much more uniform region linguistically than most of the regions to its west. The variant-based and feature-based values are generally similar, with a Pearson correlation of 0.69.

The fact that the intraregional dummy variables include a distance component complicates the interpretation of their values and that of the geographic distance term as well. Put simply, the effect of geographic distance on speech differences between regions is highly variable. Other non-geographic influences also affect linguistic differences over long distances; they are not quite as important as geographic distance alone, on average, but between some regions they can be at least as important, either enhancing or offsetting the effect of geographic distance on speech differences.

Table 2.5: Percentage Difference between Calculated Distance Scores Based on Variant-based Regression Results with and without Regional Dummy Variables

Region	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
(1) Far North	-30.3	-4.0	8.3	8.7	-6.4	3.3	-0.3	-7.7	-12.4	5.3	2.7	1.1	-0.4	12.3	-13.4
(2) Upper Northwest	-4.0	-44.7	-26.8	-16.5	-29.7	-14.4	-17.6	-3.6	-5.8	-1.1	-5.2	0.3	9.8	19.6	-9.0
(3) Cumbria-North Yorkshire	8.3	-26.8	-15.7	-7.3	-34.8	-3.2	-8.5	4	-3.3	7.9	3.3	-0.3	14.2	27.1	-4.8
(4) Lower Northwest	8.7	-16.5	-7.3	-15.1	-26.0	0.8	-9.5	4.0	-1.6	5.8	3.5	8.9	14.8	19.0	-9.8
(5) Lincolnshire	-6.4	-29.7	-34.8	-26	-64.9	-0.9	1.5	-1.1	0.8	29.5	12.8	17.8	11.9	22.0	-15.1
(6) Staffordshire	3.3	-14.4	-3.2	0.8	-0.9	-2.1	-16.4	24.9	1.1	-8.4	-8.1	2.5	27	32.2	-1.4
(7) East Midlands	-0.3	-17.6	-8.5	-9.5	1.5	-16.4	-23.8	9.0	9.3	-7.7	-10.0	-1.3	23.1	22.9	-10.9
(8) West Midlands	-7.7	-3.6	4.0	4.0	-1.1	24.9	9.0	-2.0	-15.5	-14.7	-7.7	-11.4	-11.1	10.1	-33.4
(9) Central South	-12.4	-5.8	-3.3	-1.6	0.8	1.1	9.3	-15.5	-4.7	-10.2	1.2	1.6	-2.1	-0.9	-38.4
(10) East Anglia	5.3	-1.1	7.9	5.8	29.5	-8.4	-7.7	-14.7	-10.2	-15.2	-24.5	-19.1	-3.8	12.3	-20.4
(11) Central Southeast	2.7	-5.2	3.3	3.5	12.8	-8.1	-10.0	-7.7	1.2	-24.5	-20.3	-13.7	10.1	16.3	-20.8
(12) Southeast Coast	1.1	0.3	-0.3	8.9	17.8	2.5	-1.3	-11.4	1.6	-19.1	-13.7	-28.3	11.3	13.3	-24.6
(13) Central Southwest	-0.4	9.8	14.2	14.8	11.9	27.0	23.1	-11.1	-2.1	-3.8	10.1	11.3	-25.8	-17.9	-46.7
(14) Devonshire	12.3	19.6	27.1	19.0	22.0	32.2	22.9	10.1	-0.9	12.3	16.3	13.3	-17.9	-50.3	-25.0
(15) Cornwall	-13.4	-9.0	-4.8	-9.8	-15.1	-1.4	-10.9	-33.4	-38.4	-20.4	-20.8	-24.6	-46.7	-25.0	-34.4

The regression results yield several important insights. The highly significant geographic distance coefficients suggest that regional dialect diversity is closely tied to geographic separation and that such diversity is largely continuous. Distance alone accounts for at least half of phonetic variation in the traditional English dialects. However, differences in the magnitudes of the intraregional dummy variables point to regions of relative uniformity as well as regions of much greater diversity, controlling for geographic distance. Moreover, the highly statistically significant dummy parameters indicate that systematic differences between regional speech forms account for roughly a quarter of the variation in the dialects.

2.6.3 Variant-Area Regressions

A regression of the number of variants in each region on the number of localities further reinforces the sense of varying degrees of regional uniformity. Biologists have long noted a regular correspondence between the (log of the) number of species that inhabit a geographic region and (the log of) its size – the species-area relationship. Consistent with the view that speech variants inhabit a speech region and are subject to mutation and selection pressures analogous to those that affect species in the biosphere, a similar pattern tends to obtain between the number of speech variants in use and the number of speakers. In the present data set, there is a high degree of correlation between the number of variants and the number of localities in the 15 regions delineated by cluster and phylogenetic analysis; a log-log regression of the number of variants on the number of speakers suggests that an additional interview in another locality in a region will typically increase the number of variants found in the region by about 1.3, and that such a relationship accounts for about two-thirds of the variation in the number of variants by region.

However, the errors or residuals in the variant-area regression provide useful information about the relative uniformity of speech in different regions. Those errors are particularly positive in Linconshire, Cumbria-North Yorkshire, and the Central Southeast, while the most negative errors are for Staffordshire, the Central South region, and the Southeast Coast. Those errors indicate that the former group of regions – which are also regions in which linguistic distances are relatively small compared with geographic ones – tend to have lower diversity in terms of the number of variants, while the latter group – which also tend to have greater linguistic diversity over a given geographic distance – tend to have greater diversity in terms of the number of variants.

2.6.4 Barrier Analysis

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. An illuminating example of its use by Manni and Barrai (2001) tests the associations between dialect differences, gene frequencies, and the distribution of surnames in the Italian province of Ferrara. Applied to English dialect data, Monmonier's algorithm may complement regression analysis in identifying important speech boundaries. The algorithm is implemented using the Barrier Version 2.2 software package developed by Manni, Guérard, and Heyer.[31]

Applied to matrices of feature-based and variant-based linguistic distances, Monmonier's algorithm produces broadly similar results that are, in each case, largely consistent with the regional classification developed above. (Applying the algorithm to residuals of a regression of variant-based or feature-based distances on geographic distances tends to yield similar patterns.) The algorithm identifies many of the same regional boundaries as do the cluster and phylogenetic analyses, including those between the Southeast and Southwest, the Upper Southwest and the Central Midlands, the Far North and Upper North, and the subregions of the Southeast, Upper Southwest, and Lower Southwest. The algorithm also identifies several other barriers within regions: using either set of distance measures, it isolates areas of south Lancashire, Cheshire, and Derbyshire as distinctive subregions; and using one measure or the other, it isolates West Yorkshire, Devonshire, Somerset, or Essex. However, the algorithm tends not to find clear boundaries along the eastern English coast, between Yorkshire and Lincolnshire, Lincolnshire and the East Midlands, or East Midlands and the Southeast.

In addition, however, roughly half of the boundaries the algorithm identifies isolated individual localities that are particularly distinct from their neighbors. A few of those lie well away from any of the others, such as Cumberland 1 and Essex 7, and a few are in Hampshire on the boundary between the Lower Southwest and the Southeast Coast. However, pointing yet again to the region's unusual linguistic diversity, the great majority of distinctive localities

[31] Barrier Version 2.2 is freely available online at `http://www.mnhn.fr/mnhn/ecoanthropologie/software/barrier.html`.

are in or near the Upper Southwest, including Shropshire 2 and 5, several localities in Monmouth, Gloucestershire 4, Worcestershire 1 and 7, Oxfordshire 2 and 3, Buckingham 6, and Hertfordshire 3.

2.7 Uncovering Linguistic Structure: Principal Component Analysis

Like cluster analysis and MDS, principal component analysis (PCA) reduces the number of dimensions in a data set, but it does so by finding groups of strongly correlated variables that are uncorrelated with the rest of the variables, and reducing them to a single "latent" variable – called a principal component or PC – that is essentially a linear combination of the correlated variables.[32] In effect, PCA clusters groups of variables (in this case, feature values or variant frequencies) in a manner somewhat analogous to how cluster analysis distinguishes groups of observations (in this case, localities). PCA may thus allow a researcher to summarize the patterns of relationship among the variables as a smaller number of uncorrelated variables that contain most of the information in the data set.

Applied to a data set of linguistic features, PCA may isolate groups of variants or features that tend to occur together and that, with any luck, have a structural linguistic interpretation. Component scores, which measure the strength of a particular principal component in the data for each locality, may reveal associations between those linguistic structures and specific regions. In short, a properly performed and interpreted PCA may provide a relatively objective way of defining dialects and dialect regions. I use the Statistical

[32] Tabachnick and Fidell (2000), Chapter 13, provides an introduction to principal component analysis, as does Bartholomew et al. (2002), Chapter 5. Several recent studies have applied principal component analysis or the closely related factor analysis to linguistic data, including Labov et al.'s (2006) and Clopper and Paolillo's (2006) analyses of acoustic features of North American English, Nerbonne's (forthcoming) analyses of variation in Middle and Southern Atlantic American English, and Shackleton's (2005) examination of usage differences among speakers of various English and American dialects. As noted by Nerbonne, an argument can be made that principal component analysis, which is essentially an exploratory technique, is less preferable to use for the present analysis than factor analysis, which is more appropriate when the aim of the analysis is to uncover underlying structures – or "latent variables" – that are manifested in the variables under analysis. However, factor analysis requires a matrix of variables in which none of the rows are linear combinations of the others, which for the variant-based analysis requires that localities' frequencies of the various pronunciations of a given set of words cannot sum to 100 percent. Moreover, principal component analysis and factor analysis typically yield very similar results, so the advantages of being able to examine all pronunciations in the data would appear to outweigh the advantage of being able to rigorously uncover latent variables.

Package for the Social Sciences (SPSS) for Windows Version 7.5 for the variant-based analysis and SAS/STAT 9.1 for the feature-based analysis.[33]

PCA uncovers sets of variables whose values are strongly positively or negatively correlated – that is, groups of variants that tend to occur together or that tend to occur separately. (Variables that always occur separately are not independent; mathematically speaking, independence – or orthogonality – implies that there is no pattern of co-occurrence at all. In visual terms, negative correlation places variants at opposite poles of a line, such as north and south, while independence places them at ninety degrees to each other on a perpendicular line – that is, north and east.) PCs therefore typically have two poles: one pole will have large positive values, or loadings, for a group of variables that tend to be found together; the other will have large negative values for another group of variables that are also found together but never with the first group.[34]

In the standard approach, each PC is uncorrelated with – or orthogonal to – all the others. The first PC "extracts" or accounts for the maximum possible variance from the data set that can be accounted for by a single linear combination of variables; the second extracts the maximum possible amount of the remaining variance, and so on. Each variable is assigned a score in each PC, but will typically take low values in all but one or at most a few PCs. For each PC a component score for each observation can be calculated as the sum of the products of each variable's value in the observation and the variable's score in the PC. A given observation may have high component scores for several quite different PCs.

Several variations on the standard PCA approach allow for the adjustment of PCs in ways that simplify them, generally by placing more emphasis on the highest-scoring variants (in technical terms, by "rotating" them to increase the loading on those variables). Orthogonal rotations leave the PCs uncorrelated, whereas non-orthogonal rotations allow for some correlation between PCs. In the present study, orthogonal rotation typically yields more clearly interpretable sets of variables than does standard PCA – especially when applied to variants rather than to features – and the groups of variables thus identified appear also to be used by localities in more sharply defined regions. For the data used in this study, however, non-orthogonal PCA produces PCs that are very largely uncorrelated, so that non-orthogonal rotation appears to be unnecessary.

PCA yields a large number of PCs, raising the issue of how many to interpret as yielding useful information. Although there is no hard and fast rule

[33] SAS/STAT is available at `http://www.sas.com/technologies/analytics/statistics/stat/index.html`.

[34] As a rule of thumb, researchers take into consideration only variables with laodings of 0.32 and larger in absolute value as part of a PC. See Costello and Osborne (2005), p. 4.

in this regard, the most widely recommended approach is called the "scree" test, which recommends retaining PCs that account for the greatest amount of variation up to the point at which the explained variation drops precipitously from one PC to the next, and disregarding PCs below the "break."[35] Applying PCA to the feature-based and variant-based data yields obvious breaks, but several of the PCs beyond the breaks are easily interpretable in linguistic terms and have component scores that link them to geographically coherent regions that are similar, in many cases, to the regions delineated by the cluster analyses, suggesting that in some cases the scree test may be safely ignored.

2.7.1 Feature-Based Results

Because it allows for continuous variation in articulation, the feature-based PC analysis yields PCs whose values vary continuously as a function of articulation. Only a handful of unrotated PCs yield clear, easily interpreted results in the feature-based approach. Consider the first unrotated PC, whose PC scores are presented in Table 2.6 and whose component scores are illustrated in Figure 2.7a, and which captures about 12 percent of total variation in the data. The first PC clearly distinguishes southern – and particularly Southwestern – English features from northern ones, with more central regions yielding intermediate component scores. The features with the highest loadings generally involve the raising or lengthening of some short vowels, the backing and rounding of /**a**/ after /**w**/, and the lowering, backing, and unrounding of /**u**/ – all general southern English innovations. Other features associated with the Southwest also take relatively large loadings in the PC, including manner of rhoticity and fricative voicing. Southeastern [**l**]-vocalization also has a relatively strong loading, and another set of large loadings obtains for degrees of height, backing, and rounding in many of the diphthongized Middle English long vowels. (Recall that the values describing the second element of a diphthong are converted to differences between the second element and the first.) These values have a straightforward interpretation: upgliding variants like [**ɒi** $\sim$ **ɔi**] in *night* or [**æi** $\sim$ **ai**] in *gate* will tend to yield relatively large positive loadings, while long or ingliding variants like [**iː**] in *night* or [**ia**] in *gate* will yield large negative loadings.

[35] The recommendation is generally phrased in terms of highest eigenvalues rather than highest explanatory value. The widely recommended practice of retaining PCs with eigenvalues greater than 1.0 is among the least accurate of methods. See Costello and Osborne (2005).

Table 2.6: Loadings: First Feature-based Unrotated Principal Component

	First Vowel				Second Vowel				Rhotic		Consonant				
Word	Height	Back-ing	Round-ing	Length	Height	Back-ing	Round-ing	Length	Place	Manner	Frica-tive	Aspi-rate	Glottal Stop	Velar	Other
Short															
Apple	0.394	–	–	0.113	0.053	–	–	–	–	–	–	–	0.276	−0.702	–
Catch	0.531	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ask	0.266	0.230	–	0.759	–	–	–	–	–	–	–	–	–	–	–
Father	−0.057	0.261	–	0.620	−0.145	−0.176	–	0.010	−0.063	0.593	0.438	–	–	–	–
Wasp	0.181	0.690	0.686	0.119	–	–	–	–	–	–	–	–	–	–	–
Water	0.643	0.780	0.758	0.818	−0.042	−0.163	−0.147	−0.158	0.202	0.511	0.410	–	0.110	–	–
Aunt	0.087	0.140	−0.143	0.695	–	–	–	–	–	–	–	–	0.072	–	–
Yellow (1)	0.318	−0.158	−0.146	–	–	–	–	–	–	–	–	–	–	−0.498	–
Fox	−0.150	0.067	−0.091	0.056	–	–	–	–	–	–	0.481	–	–	–	–
Yolk	−0.154	0.060	−0.189	−0.099	0.393	0.117	0.402	0.096	–	–	–	–	–	0.153	–
Cross	0.470	0.096	−0.086	0.768	–	–	–	–	0.188	0.386	–	–	–	–	–
Women	0.081	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Colt	0.432	0.343	0.222	0.400	−0.339	−0.261	−0.222	–	–	–	–	–	−0.066	–	–
Sun	−0.680	0.481	−0.151	–	–	–	–	–	–	–	0.477	–	–	–	–
Butter	−0.711	0.657	−0.750	–	–	–	–	–	0.075	0.478	0.293	–	0.199	–	–
Short Rhotic															
Farm	−0.003	0.334	–	−0.437	–	–	–	–	−0.032	0.579	0.395	–	–	–	–
Forty	0.147	−0.360	−0.364	0.068	−0.158	−0.074	−0.115	−0.042	−0.002	0.593	0.430	–	−0.257	–	0.615
Hurt	0.240	−0.421	−0.520	−0.103	–	–	–	–	0.090	0.595	0.694	−0.085	0.102	–	–

Continued on Next Page

Table 2.6 : Loadings: First Feature-Based Unrotated Principal Component (Continued)

	First Vowel				Second Vowel				Rhotic		Consonant				
Word	Height	Back-ing	Round-ing	Length	Height	Back-ing	Round-ing	Length	Place	Manner	Frica-tive	Aspi-rate	Glottal Stop	Velar	Other
Long															
Gate	−0.404	0.273	–	−0.087	0.479	0.207	–	−0.043	–	–	–	–	–	–	–
Potato (1)	−0.099	0.176	–	−0.331	0.219	−0.150	–	0.058	–	–	–	–	–	–	–
Meat	0.034	−0.053	–	0.188	0.177	−0.373	–	0.119	–	–	–	–	−0.043	–	–
Peas	0.158	−0.012	–	0.199	0.098	−0.268	–	0.137	–	–	–	–	–	–	–
Cheese	0.163	−0.034	–	0.096	−0.097	−0.107	–	0.085	–	–	–	–	–	–	–
Wheel	0.195	−0.022	–	0.021	−0.459	0.508	0.335	−0.410	–	–	–	−0.263	–	−0.758	–
White	0.270	0.498	0.460	−0.121	−0.200	−0.530	−0.451	0.134	–	–	–	−0.268	0.071	–	–
Blind	−0.530	0.353	0.361	0.023	0.469	−0.414	−0.390	0.094	–	–	–	–	–	–	–
Night	−0.628	0.510	0.435	−0.451	0.538	−0.392	−0.407	−0.205	–	–	–	–	–	–	–
Both	−0.351	0.740	0.444	0.371	0.587	−0.156	0.309	−0.009	–	–	0.420	–	–	–	–
Comb	0.097	0.441	0.100	0.178	0.383	−0.301	0.207	−0.227	–	–	–	–	–	–	–
Loaf	−0.366	0.581	0.098	0.201	0.582	−0.076	0.333	0.140	–	–	0.248	–	–	–	–
None	−0.403	0.507	−0.060	0.212	0.646	−0.140	0.286	0.013	–	–	–	–	–	–	–
Cold	0.303	0.342	0.215	−0.459	0.114	−0.139	0.087	0.120	–	–	–	–	–	0.307	–
Nose	−0.461	0.497	−0.179	0.224	0.554	0.211	0.604	0.049	–	–	–	–	–	–	–
Potato (2)	−0.290	0.166	0.042	0.330	–	–	–	–	0.392	0.392	0.438	–	0.176	–	–
Moon	0.181	0.333	0.455	0.369	0.217	−0.255	−0.089	0.072	–	–	–	–	–	–	–
Wool	0.005	−0.157	0.060	−0.047	−0.181	−0.282	−0.187	−0.230	–	–	–	–	–	−0.427	−0.242
Roof	−0.059	0.313	0.372	0.166	0.250	−0.345	−0.153	−0.105	0.202	0.461	–	–	–	–	–
House	−0.368	−0.464	−0.469	−0.475	0.349	0.472	0.328	−0.162	–	–	0.166	−0.032	–	–	–
Long Rhotic															
Mare	−0.505	−0.594	–	0.028	0.422	0.345	−0.213	−0.199	0.193	0.494	–	–	–	–	–
Hear	0.327	−0.345	–	−0.120	−0.313	0.208	−0.236	0.041	0.118	0.528	–	−0.049	–	–	–
Fire	0.390	0.528	0.407	−0.023	−0.306	−0.534	−0.407	0.062	0.073	0.566	0.503	–	–	–	–
More	0.090	0.734	0.616	0.128	−0.166	−0.553	−0.458	−0.249	0.117	0.550	–	–	–	–	–
Door	−0.191	0.529	0.452	0.291	−0.259	−0.635	−0.560	−0.344	0.000	0.508	–	–	–	–	–
Hour	−0.429	−0.342	−0.390	−0.240	0.439	0.359	0.366	0.023	0.138	0.502	–	–	–	–	–

Continued on Next Page

Table 2.6 : Loadings: First Feature-Based Unrotated Principal Component (Continued)

	First Vowel				Second Vowel				Rhotic		Consonant				
Word	Height	Back-ing	Round-ing	Length	Height	Back-ing	Round-ing	Length	Place	Manner	Frica-tive	Aspi-rate	Glottal Stop	Velar	Other
Diphthong															
Daisy	−0.408	−0.096	–	−0.467	0.609	0.092	–	0.080	–	–	–	–	–	–	–
Eight	−0.296	−0.049	–	0.078	0.137	0.002	−0.070	−0.081	–	–	–	–	0.05	–	–
Voice	−0.141	0.016	0.040	−0.005	–	–	–	–	–	–	–	–	–	–	0.024
Straw	−0.373	0.435	0.383	0.344	0.413	−0.286	−0.158	−0.157	0.196	0.251	–	–	–	–	0.110
Thaw	0.366	0.564	0.403	0.584	−0.544	−0.346	−0.418	−0.004	0.183	–	0.374	–	–	–	–
Daughter	0.374	0.113	0.122	0.843	−0.858	−0.168	−0.450	−0.155	0.069	0.522	0.314	–	0.165	–	–
Dew	0.171	0.379	0.459	0.365	−0.212	−0.432	−0.440	−0.224	–	–	–	–	–	–	0.280
Tuesday	0.125	0.386	0.478	0.429	−0.177	−0.461	–	−0.295	–	–	–	–	–	–	0.266
Snow	0.118	0.284	0.060	−0.417	0.372	−0.065	0.245	0.052	–	–	0.279	–	–	–	–
Yellow (2)	−0.055	0.144	0.159	0.036	0.193	−0.134	–	–	0.496	0.514	–	–	–	–	–
Diphthong Rhotic															
Hair	−0.046	−0.162	−0.111	−0.247	0.228	0.207	−0.249	−0.085	0.321	0.484	–	0.029	–	–	–
Other															
Very	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.065
Kitten	–	–	–	–	–	–	–	–	–	–	0.266	–	0.422	–	–

Note: "Other" consonants are: voicing of **[t]** in *forty*, loss of **[w]** in *wool*, **[w]** for **[v]** in *voice* and *very*, intrusive **[r]** in *straw*, and palatalization in *dew* and *Tuesday*. *Yellow* (1) and *Yellow* (2) refer to the first and second vocalic segments, respectively, and *Potato* (1) and *Potato* (2) refer to the second and third vocalic segments.

The first PC is thus capturing the contrast between southern upglides and northern inglides or downglides, and is translating longer glides into larger PC scores. In this sense, the PC summarizes the historical developments of a group of Middle English long vowels. Nevertheless, it captures amalgams of southern and northern speech patterns that do not represent the speech patterns of any particular locality: many of the Southwestern features to which it assigns large positive loadings – fricative voicing and rhoticity in particular – are not found in the Southeast, and upgliding is generally muted in the Southwest.

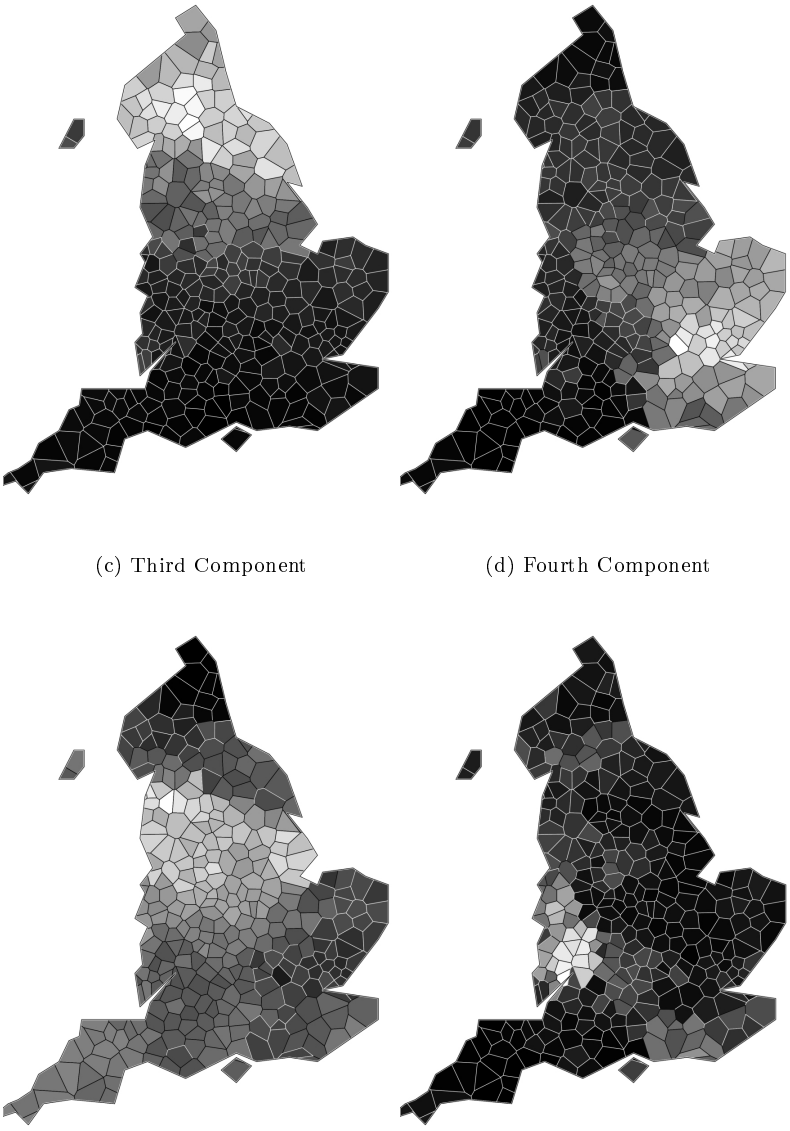
In contrast to the first PC, the second, which accounts for about 10 percent of the variation and whose component scores appear in Figure 2.7b, clearly distinguishes the Southeast and Central Midlands regions from both the Southwest and the North, assigning negative scores to the former and its strongest negative scores to localities nearest London – precisely those localities in which upglides tend to be most extended. While the PC does a good job of locating upgliding in the appropriate regions, however, it assigns positive loadings not only to ingliding features in the North but also to rhotic and fricative features found in the Southwest. In this case, in effect, negative loadings identify a group of dialect regions by the presence of upglides and by the absence of Southwestern and Northern features.

The third unrotated PC, accounting for 5 percent of the variation and shown in Figure 2.7c, isolates the Far North by its place of articulation of rhotics, by the preservation of /**h**/ and /**hw**/, and by certain ingliding features. The fourth PC (shown in Figure 2.7d) rather indistinctly isolates the Severn Valley by its negative scores, the fifth the Far North once again by certain inglides, and the sixth, localities in the Lower Northwest and Staffordshire with strongly rounded upglides in the mid-back Middle English long vowels.

Varimax rotation reallocates the variance captured in the unrotated PC into rotated ones, increasing each PC's loadings on somewhat smaller groups of variables and thus sharpening the focus and regional concentration of each PC. As shown in Table 2.7 and Figure 2.8a, the features that yield large positive loadings for the first varimax PC and the high-scoring localities are much more distinctively Southeastern rather than generally southern: the short vowel developments and upgliding diphthongization tend to receive higher positive loadings, but rhoticity and fricative voicing do not. The PC thus distinguishes Southeastern upgliding from northern ingliding much more clearly than does the first unrotated PC. Similarly, the second varimax PC – shown in Figure 2.8b – much more clearly distinguishes manner of rhoticity and, secondarily, fricative voicing as the distinctive sets of generally Southwestern features, while the third PC (Figure 2.8c) isolates fricative voicing as the distinctively Lower Southwestern features.

Figure 2.7: Feature-based Unrotated Principal Components

(a) First Component (b) Second Component



(c) Third Component (d) Fourth Component

Note: Darker (lighter) shading indicates larger positive (negative) scores.

Table 2.7: Loadings: First Feature-based Rotated Principal Component

	First Vowel				Second Vowel				Rhotic		Consonant				
Word	Height	Back-ing	Round-ing	Length	Height	Back-ing	Round-ing	Length	Place	Manner	Frica-tive	Aspi-rate	Glottal Stop	Velar	Other
Short															
Apple	0.602	–	–	0.168	0.205	–	–	–	–	–	–	–	0.467	−0.625	–
Catch	0.462	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ask	0.072	0.580	–	0.622	–	–	–	–	–	–	–	–	–	–	–
Father	−0.159	0.606	–	0.546	−0.098	−0.178	–	0.035	−0.034	0.069	0.050	–	–	–	–
Wasp	0.381	0.625	0.619	0.216	–	–	–	–	–	–	–	–	–	–	–
Water	0.542	0.734	0.719	0.664	0.005	−0.166	−0.163	−0.159	0.183	0.022	0.031	–	0.184	–	–
Aunt	−0.030	0.453	−0.128	0.675	–	–	–	–	–	–	–	–	0.041	–	–
Yellow (1)	0.427	−0.172	−0.177	–	–	–	–	–	–	–	–	–	–	−0.237	–
Fox	−0.043	−0.033	−0.086	0.104	–	–	–	–	–	–	0.019	–	–	–	–
Yolk	−0.486	−0.025	−0.299	−0.320	0.704	0.107	0.514	0.023	–	–	–	–	–	0.294	–
Cross	0.225	0.057	−0.024	0.491	–	–	–	–	−0.108	−0.025	–	–	–	–	–
Women	0.063	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Colt	0.159	0.198	0.010	0.001	−0.073	−0.194	−0.012	–	–	–	–	–	−0.005	–	–
Sun	−0.642	0.484	−0.003	–	–	–	–	–	–	–	0.015	–	–	–	–
Butter	−0.624	0.536	−0.596	–	–	–	–	–	0.078	−0.025	−0.055	–	0.288	–	–
Short Rhotic															
Farm	−0.039	0.572	–	−0.029	–	–	–	–	−0.016	0.061	0.030	–	–	–	–
Forty	0.259	0.000	−0.029	0.374	−0.056	−0.165	−0.168	−0.067	0.017	0.063	0.025	–	−0.174	–	0.156
Hurt	0.084	−0.327	−0.412	0.184	–	–	–	–	0.105	0.157	0.266	0.063	0.051	–	–

Continued on Next Page

Table 2.7 : Loadings: First Feature-Based Rotated Principal Component (Continued)

	First Vowel				Second Vowel				Rhotic		Consonant				
Word	Height	Back-ing	Round-ing	Length	Height	Back-ing	Round-ing	Length	Place	Manner	Frica-tive	Aspi-rate	Glottal Stop	Velar	Other
Long															
Gate	−0.561	0.034	–	−0.224	0.631	0.140	–	−0.018	–	–	–	–	–	–	–
Potato (1)	−0.401	0.029	–	−0.425	0.655	0.071	–	0.055	–	–	–	–	–	–	–
Meat	0.202	0.206	–	0.026	0.140	−0.578	–	0.300	–	–	–	–	−0.115	–	–
Peas	0.294	0.207	–	0.095	0.040	−0.514	–	0.297	–	–	–	–	–	–	–
Cheese	0.029	0.163	–	−0.117	0.034	−0.236	–	0.228	–	–	–	–	–	–	–
Wheel	0.025	0.145	–	−0.203	−0.269	0.409	0.424	−0.198	–	–	–	−0.156	–	−0.714	–
White	0.280	0.535	0.296	−0.093	−0.197	−0.552	−0.291	0.047	–	–	–	−0.167	0.083	–	–
Blind	−0.387	0.487	0.201	0.024	0.370	−0.519	−0.228	−0.049	–	–	–	–	–	–	–
Night	−0.548	0.591	0.283	−0.386	0.452	−0.472	−0.263	−0.124	–	–	–	–	–	–	–
Both	−0.672	0.487	0.108	−0.076	0.815	−0.128	0.541	0.065	–	–	0.022	–	–	–	–
Comb	−0.278	0.295	−0.159	−0.127	0.681	−0.152	0.483	0.004	–	–	–	–	–	–	–
Loaf	−0.712	0.367	−0.141	−0.167	0.840	−0.077	0.398	0.102	–	–	0.032	–	–	–	–
None	−0.357	0.442	−0.237	−0.052	0.569	0.120	0.450	0.075	–	–	–	–	–	–	–
Cold	0.098	0.287	0.078	−0.349	0.214	−0.195	0.120	0.079	–	–	–	–	–	0.259	–
Nose	−0.726	0.346	−0.449	−0.167	0.847	0.047	0.728	0.016	–	–	–	–	–	–	–
Potato (2)	−0.521	0.428	0.060	−0.034	–	–	–	–	0.142	0.142	0.001	–	0.210	–	–
Moon	0.064	0.140	0.183	0.156	0.248	−0.012	0.117	0.176	–	–	–	–	–	–	–
Wool	0.041	−0.055	0.072	−0.011	−0.038	−0.149	−0.093	−0.133	–	–	–	–	–	−0.444	0.049
Roof	0.008	0.209	0.200	0.088	0.133	−0.175	−0.063	−0.087	−0.105	0.002	–	–	–	–	–
House	−0.095	−0.280	−0.291	−0.229	0.110	0.315	0.256	−0.200	–	–	0.013	0.124	–	–	–
Long Rhotic															
Mare	−0.348	−0.435	–	0.025	0.224	0.040	−0.128	−0.012	0.097	0.023	–	–	–	–	–
Hear	−0.035	−0.028	–	−0.078	0.062	−0.133	−0.129	0.108	0.073	0.037	–	0.146	–	–	–
Fire	0.242	0.563	0.197	0.129	−0.309	−0.516	−0.197	0.014	−0.015	0.036	0.040	–	–	–	–
More	−0.295	0.549	0.411	0.261	0.181	−0.210	−0.093	−0.097	−0.128	0.056	–	–	–	–	–
Door	−0.475	0.380	0.277	0.282	0.075	−0.173	−0.066	−0.078	−0.154	−0.024	–	–	–	–	–
Hour	−0.144	−0.171	−0.243	−0.096	0.140	0.205	0.210	−0.018	0.066	0.005	–	–	–	–	–

Continued on Next Page

Table 2.7 : Loadings: First Feature-Based Rotated Principal Component (Continued)

	First Vowel				Second Vowel				Rhotic		Consonant				
Word	Height	Back-ing	Round-ing	Length	Height	Back-ing	Round-ing	Length	Place	Manner	Frica-tive	Aspi-rate	Glottal Stop	Velar	Other
Diphthong															
Daisy	−0.467	0.077	–	−0.489	0.654	0.091	–	0.077	–	–	–	–	–	–	–
Eight	−0.155	0.011	–	−0.011	0.088	−0.029	−0.020	0.015	–	–	–	–	−0.004	–	–
Voice	0.148	0.020	0.032	0.049	–	–	–	–	–	–	–	–	–	–	0.072
Straw	−0.208	0.296	0.254	0.147	0.326	−0.347	−0.307	−0.270	−0.102	−0.107	–	–	–	–	0.096
Thaw	0.162	0.365	0.241	0.325	−0.245	−0.203	−0.210	−0.021	0.424	–	0.398	–	–	–	–
Daughter	0.398	0.209	0.198	0.669	−0.682	−0.067	−0.316	−0.143	0.016	0.033	0.006	–	0.195	–	–
Dew	0.095	0.144	0.139	0.030	−0.053	−0.147	−0.124	0.039	–	–	–	–	–	–	−0.075
Tuesday	0.090	0.166	0.191	0.096	−0.063	−0.201	–	−0.015	–	–	–	–	–	–	−0.051
Snow	−0.187	0.168	−0.148	−0.578	0.633	−0.100	0.410	−0.032	–	–	0.023	–	–	–	–
Yellow (2)	−0.198	0.143	0.188	0.038	0.323	−0.102	–	–	0.106	0.115	–	–	–	–	–
Diphthong Rhotic															
Hair	−0.151	−0.222	−0.054	−0.150	0.159	−0.032	−0.194	0.072	0.225	−0.011	–	0.181	–	–	–
Other															
Very	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.089
Kitten	–	–	–	–	–	–	–	–	–	–	0.046	–	0.355	–	–

Note: "Other" consonants are: voicing of **[t]** in *forty*, loss of **[w]** in *wool*, **[w]** for **[v]** in *voice* and *very*, intrusive **[r]** in *straw*, and palatalization in *dew* and *Tuesday*. *Yellow* (1) and *Yellow* (2) refer to the first and second vocalic segments, respectively, and *Potato* (1) and *Potato* (2) refer to the second and third vocalic segments.

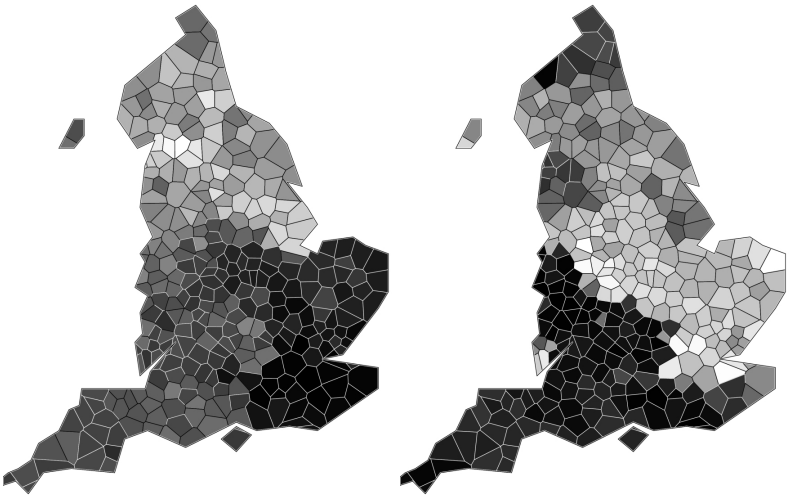
The fourth varimax PC distinguishes the Far North and Upper North as the regions that retain lengthened long high vowels, such as [**u:**] for in *house* and *hour* and [**i:**] in *blind*, and inglides in words with Middle English /**a:**/ and /**ɔ:**$_1$/. The fifth PC (not shown) isolates the Far North on the basis of placement of rhotic articulation. The sixth PC, which assigns large positive loadings to backed and rounded onsets to upglides in words with Middle English /**i:**/ such as *white*, *blind*, and *night*, yields a strangely checkered pattern throughout southern England, revealing the largely unstable nature of the onset – a pattern consistent with the fact that studies generally find relatively little acoustic or perceptual difference between low front and low back vowels. The seventh PC reveals a similarly checkered Southeastern pattern of backing and shortening of the onset in *meat*, *peas*, and *cheese*. The eighth PC isolates North Anglia on the basis of glottal stops for medial /**t**/.

The feature-based analysis yields further insights when the analysis is restricted to classes of features. Restricted to short vowels, the first varimax PC yields a north-south pattern consistent with the general southern short vowel shift, but the second PC isolates a region of particularly strong front vowel raising in *apple* coupled with particularly strong backing in *ask*, *father*, and *aunt* near London and along the Southeast Coast. Restricted to words with rhotics, the analysis yields five varimax PCs that isolate the entire Southwest and northern rhotic regions, but also isolate the Upper and Lower Southwest – as well as Far and Upper North – on the basis of place and manner of articulation of consonantal rhotics in *roof*, *cross*, and *straw*. Restricted to non-rhotic long vowels, the analysis not only distinguishes the upgliding and ingliding regions of the Southeast and north but also reveals a long swath of localities from the Southwest to southern Lancashire that tend to retain long variants in mid-back long vowels.

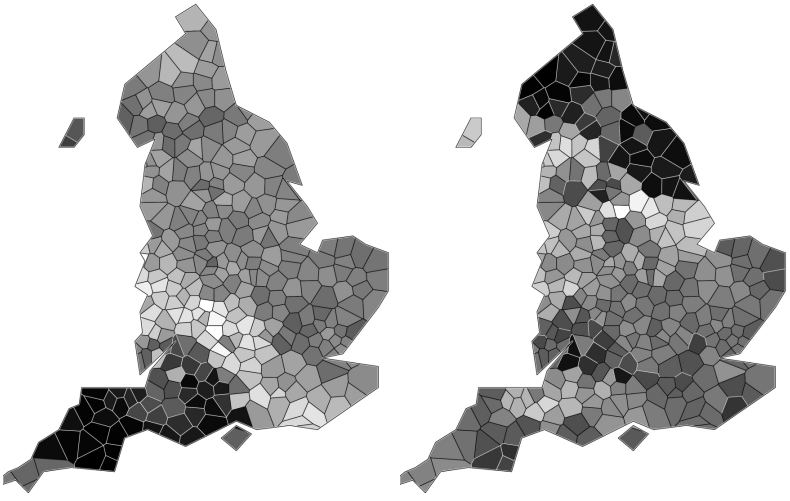
Taken together, unrotated and varimax PCA of the feature-based data isolates two types of features: one, exemplified by fricative voicing or rhoticity, is effectively a binary type, tends to give rise to sharp boundaries, and isolates regions of the Southwest and Far North; the other, a more typically vocalic and continuously varying type, reveals regions of more gradual variation in the north and Southeast. Despite the greater emphasis on articulatory detail in the feature-based analysis, the results yield a significantly less detailed picture of regional dialect variation than does the variant-based analysis discussed in the next section. (For instance, rather surprisingly, the feature-based analysis never clearly isolates Devonshire's fronting of back vowels.) It is not clear whether that outcome results from some fundamental element of PCA applied to feature-based data or from the relatively limited size of the data set. In either case, however, it appears that feature-based analysis yields results consistent with the variant-based PCA and with the cluster and regression analyses as well, while clarifying the essentially continuous nature of much

Figure 2.8: Feature-Based Rotated Principal Components

(a) First Component (b) Second Component



(c) Third Component (d) Fourth Component



Note: Darker (lighter) shading indicates larger positive (negative) scores.

dialect variation.

2.7.2 Variant-Based Results

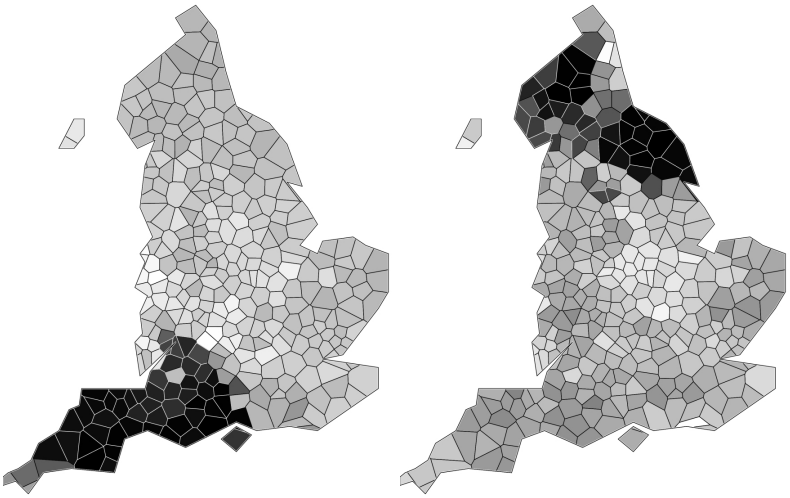
The variant-based PC analysis yields insights that differ from but are largely complementary to those derived from the feature-based approach. Applied to the variant-based data, the results of an unrotated PC analysis are difficult to interpret in linguistic terms, but varimax rotation yields a set of identifying variants for 12 regions that are largely congruent with most of the 15 regions identified by the cluster and phylogenetic analyses. With appropriate caveats, in at least some cases the variants may arguably be said to characterize a traditional English dialect region.

The first rotated PC, whose largest positive loadings are presented in Table 2.8 and whose component scores appear in Figure 2.9a, accounts for about 12 percent of the variation in the data set. The map shows that the first PC largely – though not perfectly – overlaps with the broad Southwest dialect region, while the loadings reveal that the features most closely associated with the PC are the voicing of fricatives and occasionally of medial dentals, the plausibly related voicing and dentalizing of medial fricative [**s**], and full lowering and unrounding of /**u**/.[36] The PC also assigns high loadings to strong rhoticity as well as to a set of vocalic features that nearly fully describe a nonstandard regional dialect system of vowels, mainly involving the fronting of back vowels, development of (or even monophthongization to) a low-front onset in /**iː**/, and relatively little raising of the low-front long vowels. However, those rhotic and vocalic features do not load as strongly in the PC because they are not as closely correlated with it or with the highest-scoring features: [**ɽ**] is found in a much wider area than the Southwest, while most of the vocalic features are found in narrower regions, predominantly in Devonshire – or in wider regions that only partially overlap with the Southwest. Thus the rhotic features are found uniformly but not uniquely in the Southwest, while the vocalic ones are found there uniquely but not uniformly.

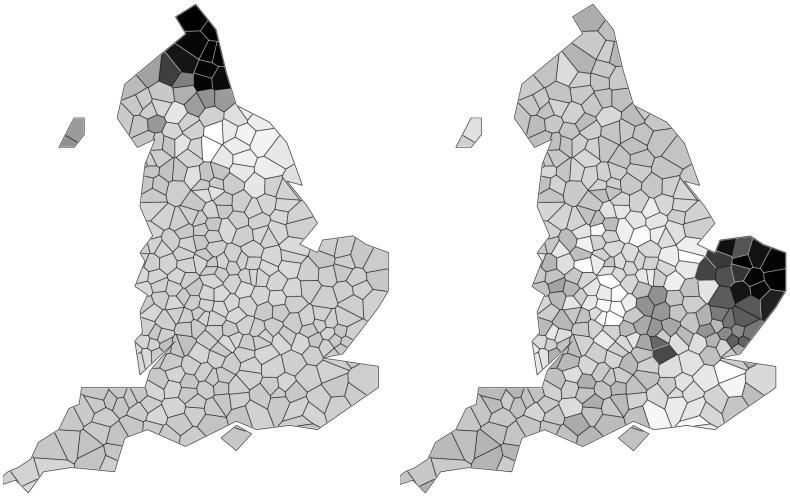
[36] The historical origins of voicing of initial fricatives are poorly understood and, as voicing is found in Germanic dialects on the continent, may even extend to the earliest period of Anglo-Saxon settlement in England. Whatever the case may be, it appears likely that voicing prevailed over much of southern England, including most localities south of a line extending from mid-Shropshire to the Suffolk-Essex border, until as late as the 17th century and may not have become considerably more restricted until the onset of industrialization in the early 19th century. See Voitl (1988).

Figure 2.9: Variant-based Rotated Principal Components

(a) First Component (b) Second Component



(c) Third Component (d) Fourth Component



Note: Darker (lighter) shading indicates larger positive (negative) scores.

Table 2.8: Loadings: First Variant-based Rotated Principal Component

Long Vowels	
ME /**aː**/ develops to [**eə** ∼ **ɛə**] in the Southwest (*SAED* Map 37)	0.645
ME /**ɛː**/ (< Old English /**œ**/, /**ea**/) shortened to [**ɛ**] in Devonshire (*SAED* Map 54)	0.648
ME /**ɛː**/ (< Old English /**œ**/, /**ea**/) develops to [**ɛi**] through the South and Midlands (*SAED* Map 51)	0.546
ME /**eː**/ is sporadically shortened to [**ɪ**] in 9 words in the South (*SAED* Map 48)	0.351
ME /**iː**/ develops to a low-front or low-center monophthong in Devonshire (*SAED* Map 21 & 22)	0.531
ME /**iː**/ develops a low-front or front onset in the Southwest (*SAED* Map 21 & 22)	0.396
ME /**ɔː**$_1$/ develops to [**ʏː**] in Devonshire (*SAED* Map 74)	0.510
ME /**ɔː**$_1$/ develops to [**wʌ** ∼ **wu**] in initial position in 3 words in the West and North (*SAED* Map 71)	0.442
ME /**ɔː**$_2$/ develops to [**ʏː**] in Devonshire (*SAED* Map 74)	0.515
ME /**oː**/ develops to [**ʏː**] in Devonshire (*SAED* Map 61)	0.589
ME /**oː**/ shortened and fronted to [**ʏ**] in Devonshire (*SAED* Map 18)	0.322
ME /**uː**/ develops to [**au**] in the West Midlands and lower North (*SAED* Map 31)	0.545
ME /**uː**/ develops to [**ɛʏ**] on the edges of Devonshire (*SAED* Map 30)	0.405
Short Vowels	
ME /**a**/ + /**f**/, /**θ**/, or /**s**/ becomes lengthened mainly in the South and Midlands	0.335
ME /**o**/ + /**f**/, /**θ**/, or /**s**/ lengthened in the South (*SAED* Map 9)	0.430
ME /**o**/ + /**f**/, /**θ**/, or /**s**/ typically lowered to [**ɔ** ∼ **ɔː**] mainly in the South	0.321
ME /**u**/ fully lowered and unrounded with tenseness to [ʌ̟] in the Southwest	0.863
Diphthongs	
ME /**au**/ develops to a closer, more rounded [o̝ː] along the South coast (*SAED* Map 13)	0.718
ME /**iu**/ develops to [**ʏː**] in 5 words in Devonshire (*SAED* Map 100B)	0.687
ME /**ɔu**/ (< Old English /**aw**/) develops to [**oː**] in the West (*SAED* Map 96B)	0.480
ME /**ɔu**/ (< Old English /**aw**/) final develops to [ɚ] in 5 words in the Southwest and Southwest Midlands	0.340
ME /**ɔu**/ (< Old English /**aw**/) final remains [**ɔu**] or develops to [**ɪ**] in 5 words sporadically	0.320

Continued on Next Page

Table 2.8 : Loadings: First Variant-based Rotated Principal Component (Continued)

Rhotics	
ME /**ar**/ [**r**] retained with [ɻ] in the South (*SAED* Map 6)	0.580
ME /**ir**/-/**ur**/ [**r**] retained with [ɻ] in the Southwest (*SAED* Map 20)	0.512
ME /**or**/ final and pre-consonantal [**r**] retained with [ɻ] in the South (*SAED* Map 12)	0.571
Consonants	
ME /**f**/ (> Old English /**f**/) develops to [**v**] in the Southwest (*SAED* Map 105)	0.970
ME /**f**/ (> Old French /**f**/) develops to [**v**] in the Southwest (*SAED* Map 106)	0.872
ME /**s**/ (> Old English /**s**/) develops to [**z**] in the Southwest (*SAED* Map 107)	0.977
ME /**s**/ (> Old French /**s**/) develops to [**z**] in the Southwest (*SAED* Map 108)	0.896
ME /**θ**/ develops to [**ð**] in the Southwest (*SAED* Map 109)	0.973
ME /**θr**/ develops to [**dɻ** ∼ **ðɻ**] in the Southwest (*SAED* Map 110)	0.962
ME medial /**t**/ develops to [**d**] along the southern coast	0.841
ME medial /**s**/ develops to [**d**] in *isn't*, *wasn't*, *doesn't*, *hasn't* (4 instances) in the Southwest	0.721
ME /**h**/ retained in the Southwest (*SAED* Map (111)	0.477
ME /**w**/ lost before /**u**/ in the Southwest (*SAED* Map 114)	0.417

Note: Only items with scores > |0.32| are shown.

The first PC suggests that the strongest candidates as identifying features of Southwestern speech are voicing of fricatives (and occasionally of medial dentals) and the full lowering and unrounding with tenseness of /**u**/. Other features with positive loadings are also diagnostic but not necessarily defining. That result is consistent with Trudgill's (1999) association of [**z**] in seven with this dialect region.

Although it clearly identifies a number of variants associated with the region, neither the PC nor its associated variants is replicated perfectly in any single Southwestern locality, and they can only rather loosely be thought of as representing a Southwestern dialect or even a group of dialects. Even the voicing of fricatives appears in only about three-quarters of possible occurrences in the Southwestern localities. Not even the most typical Southwestern locality uses all of the variants with PC loadings above 0.32 in the first rotated PC, and barely any of the localities even use a majority of them. Not much more than half of the variants associated with the PC occur even in the locality with the highest component score for the first PC; conversely, roughly 20 percent of the variants occur in several of the lowest-scoring localities, which tend to be somewhat further north in the West Midlands region. Moreover, the PC excludes a large number of variants that are widely used throughout the Southwest: strikingly, it does not include a single variant that is the most common Southwestern pronunciation of a Middle English long vowel. Because their distributions very rarely closely overlap with the boundaries of the Southwest, most variants that might constitute a vowel system common to most of the Southwest do not in fact appear as part of the Southwestern PC. Such limitations appear to be inherent in the use of principal component or factor analysis in the analysis of linguistic variation – and underline the fact that systematic variation is the exception rather than the rule in the traditional dialects.

The second rotated PC, whose component scores are shown in Figure 2.9b, accounts for about 6 percent of the variation in the data set. It appears to be strongly associated with most but not all of the Cumbria-North Yorkshire subregion of the Upper North. The defining variants in this PC all involve the development of a high front (and sometimes palatalized) onset plus schwa for /**aː**/, /**ɔː**$_1$/, and /**oː**/, leading to their near-merger or even total merger. (The remainder of Cumbria-North Yorkshire is associated with PC 22 – not shown – which has many variants in common with the rest of the region but which tends to develop onsets in /**eː**/ and /**uː**/ and less fronting of the second element of /**oː**/.) While other variants also feature relatively strongly, the fronted onsets of several long vowels – not discussed at length by Trudgill – appear to be the most defining variants of the region. In contrast to the case of the Southwest discussed previously, the second PC includes nearly all of the variants that are the most common regional pronunciations of Middle English

long vowels. Still, by no means do all localities in the region use the defining variants in the PC, and none use all of them. As with the first PC, only rather loosely may the high-scoring variants in the PC be said to define a dialect or group of dialects.

The third rotated PC, whose component scores are shown in Figure 2.9c, accounts for another 6 percent of the variation. The PC assigns high positive loadings to several variants that very clearly delineate the Far North. Many of the localities in the region also have several other generally northern variants, including the fronting and palatalization of long back vowels discussed above. However, several variants are unique to the Far North, most importantly the uvular [**ʁ**] but also the raising and fronting of the low back long vowels to [**øː**] (a full merger of both vowels to [**øː**] is found only in the coastal localities), the fronting and occasional lengthening of /**o**/ to [**œ** ∼ **œː**] and of /**au**/ to [**aː** ∼ **æː**], and the survival of [**h**] and [**hw**]. Again, although not all localities in the region use all of them (and localities in the Upper North occasionally use several) these variants together can reasonably be said to define the region – as indeed Trudgill does with [**ʁ**], [**h**], and [**hw**]. As with the second PC, the third includes nearly all of the variants that are the most common regional pronunciations of the Middle English long vowels.

The fourth PC, accounting for about 4 percent of total variation, rather weakly delineates most of East Anglia, as shown in Figure 2.9d. Although Trudgill delineates and defines this region by its continuing use of [**h**] – a variant that indeed scores fairly highly in the PC – the present analysis suggests that the development of [**w**] for /**v**/ and the development of a centered, unrounded onset in /**iː**/ are somewhat more strongly associated with the region. Several more widely distributed variants coincide in East Anglia, particularly in Norfolk, so that the region can be further associated with the unique occurrence of those variants together – glottalized medial /**t**/, a fricative in *aren't you*, the fronting and unrounding of /**o**/, and the perhaps related raising of /**a**/ – even though none of those variants score highly on the PC. The PC does not assign high loadings to any variants indicative of the *moan*/*mown* distinction, the fronting of /**o**/, the *fear*/*fair* merger, or the other variants known to be characteristic of East Anglia; nor does it isolate many of the most common East Anglian variants of Middle English long vowels.

The fifth PC (not shown), accounting for about 4 percent of total variation, isolates the Severn Valley, essentially the southern half of the West Midlands subregion of the Upper Southwest. The variant with the highest loading , the development of /**au**/ to [**ɑː**], points to a broader, related set of developments that coincide in the region: a general lowering and unrounding of short back vowels, as well as at least occasional retraction and rounding of low short front ones. The region also develops diphthongs with high onsets (as well as centering or backing) in some words of the middle long vowel classes (that is, /**ɛː**/

and /ɔː$_1$/) – a very long-standing dialect development. Otherwise, the middle long vowels show relatively little movement, as in much of the west of England, perhaps contributing in the West Midlands to the relatively slight onsets in the long high vowels. (Trudgill defined the region mainly in terms of the retraction and rounding of the vowel in *land* as well as the conservation of [**a**] in *bat*. The present analysis suggests that the former development is strongly characteristic in the region, but that [**æ**] is considerably more commonly used than [**a**].)

The sixth PC distinguishes Devonshire from the rest of the Southwest by its unique fronting of back vowels, a structural feature that appeared in the first PC as well: /ɔː$_1$/, /ɔː$_2$/, /oː/, and /**iu**/ all generally develop to [ʏː], while /**u**ː/ develops to [**ɛʏ** ∼ **œu** ∼ **œʏ**]. (Trudgill similarly points to the fronting of /**o**ː/ in *boot* as the defining characteristic of Devonshire speech.) Another variant that is nearly as important is the development of a low monophthong for /**i**ː/.

The seventh PC distinguishes, to some extent, the unusual speech around the "Potteries" zone in Staffordshire, Cheshire, and Darbyshire. Neighboring localities to the north and west tend to have raised monophthongs [**e**ː] and [**o**ː] for /**a**ː/ and /ɔː/, respectively, while southern and eastern neighbors tend to have offglides such as [**ei**] and [**ou**]. In the transition zone, localities tend to use [**i**ː] and [**u**ː], as though emphasizing both the raising of their northerly neighbors and the offglides of their southerly ones. The PC cannot be said to delineate a broad speech region or define its variants, but it does flag some of the defining variants of two neighboring regions. As mentioned previously, Trudgill notes Staffordshire and the surrounding regions as having an even more complete system that includes the variants mentioned above and several more besides. The present analysis suggests that there is so much variation among localities in the region or that there is so much overlap of elements of the system with other regions that the complete system does not appear as a PC.

In the case of the eighth PC, the negative scores delineate and define the Southeast Coast and part of the Thames estuary. Here, the primary defining variants involve a tendency to enhance the general Southern raising of short front vowels in at least some contexts. The rest of the important variants appear to be manifestations of a tendency, systematic but by no means uniform or universal throughout the Southeast, for long vowels to all develop into diphthongs with rising offglides, and many of the variants that take low to moderate positive scores in this PC further reflect that tendency. The present analysis did not include and therefore fails to note Trudgill's defining variants for parts of the Southeast Coast, namely the loss of /**l**/ at the end of words and the tendency to use [**ɔu**] in words like *old*. It also fails to yield another perfectly reasonable way of characterizing the Southeast Coast – as the only

region that combines upgliding diphthongization of long vowels with retention of rhoticity. Such a combination of variants is very unlikely to be captured by principal component analysis because rhoticity and diphthongization each occur over very large regions with only a comparatively small area of overlap.

The negative scores of the ninth PC define another region of the English Southeast – in this case the Central Southeast region around and to the north of London – by a process of diphthongization. The defining characteristic appears to be a tendency to front and lower the first elements of some of the diphthongs, such as [**æi** ∼ **ai**] for /**aː**/ and especially [**əu** ∼ **ʌu** ∼ **æu**] for /**ɔː**$_1$/. The results are consistent with the hypothesis that the diphthongization is a long-term, ongoing process centered near the metropolis, and that older forms are preserved in the periphery. (The fact that the progressive elements are typically not found in American speech but regularly appear in English-speaking countries settled more recently may provide some confirmation for that hypothesis.) The analysis fails to capture the *moan*/*mown* distinction noted in this region by Trudgill.

The tenth PC delineates southern Lancashire, a portion of the Lower North, by the same variant noted by Trudgill – its localities' use of an alveolar approximant rhotic. The eleventh PC rather poorly delineates the northern half of the West Midlands, along with parts of the Lower North, by the retention of a long monophthong for the low long vowels, as mentioned in the discussion of the seventh PC describing the transition zone between those regions and the Midlands proper. The twelfth PC, characterized not only by the use of the retroflex [**ɽ**] but by the tendency to insert it in such words as *window*, does not really delineate any dialect region, but it tends to take moderately high values through a broad swath of the Upper Southwest and the Southeast Coast, tracing the southern edge of the major boundary of rhoticity in the South of England, the Wessex Line. The PC highlights the region of competition between rhotic and non-rhotic variants, where rhotic speakers tend to overcompensate by inserting rhotics where they are otherwise absent.

The thirteenth PC clearly delineates Lincolnshire as the region which combines a tendency to develop a low-center onset to the high long vowels with a tendency to develop centering diphthongs for the low and middle long vowels (and without strongly raising them). The fourteenth PC isolates and defines south Yorkshire by its unique front upglides in the long low back vowels, and the fifteenth does the same for north Lancashire and environs – the Upper Northeast – by its tendency to merge /**ɔː**$_1$/ and /**aː**/ into a falling diphthong [**ia** ∼ **ea**]. Beyond those, the PCs appear increasingly uninformative from a linguistic point of view: the sixteenth isolates a dozen localities in the Southeast Coast region that use a rhotic approximant; the seventeenth isolates a single locality on the Scottish border in Cumberland that uses a rhotic trill; and the eighteenth isolates some western localities that use a retracted, rounded vowel

in hand.

In sum, PCA identifies sets of identifying variants for a dozen-plus regions largely congruent with many of the 15 regions identified by the cluster and phylogenetic analyses, in the process accounting for roughly half the variation in the data set. In at least some cases, the PCs appear to provide a fairly objective method for characterizing the traditional English dialect regions on a quantitative basis. However, the PCs often isolate variants that are unique to a fairly small subregion, and they often include variants that are widely used throughout the relevant region but that are not unique to it. No locality in a region uses all of the variants identified by the relevant PC, and few even use most of them. With the exception of three PCs that are composed of variants found in Devonshire, Lincolnshire, and the Far North, respectively, none of the PCs isolate extensive sets of variants that could be considered even partial systems of vowels unique to particular regions of England.

2.7.3 Variant-based Principal Components and Variant Frequency

Additional context for the results of the PC analysis can be derived by examining the frequency of the two most common modern regional variants of most of the Middle English phonemes examined in this study. Regions do indeed tend to have fairly uniform speech patterns: the two most frequent variants in each category account for at least 91 percent and as much as 98 percent of all usage in their respective regions. The most extensive intraregional variation occurs among long vowels and, to a lesser extent, diphthongs. The regions identified by distance measures, distance regressions, and variant-area regressions as having the greatest variation – the Lower Northwest, the West Midlands, and the Central South – are the regions which have a large number of relatively low-frequency variants for the long vowels. With a few exceptions, variant usage for short vowels and consonants tends to be relatively uniform in each region. Although variant usage is relatively uniform within a given region, however, very few common variants are unique to a specific region: in the great majority of cases variants are found in two or more regions. There is therefore relatively little scope for PC analysis to identify regions through variants that are unique to them. Nevertheless, even though any given variant tends to be found in a number of regions, each region has a relatively unique overall pattern of frequently used variants.

These frequency details help explain the efficacy and limitations of the PC analysis in identifying unique or important regional variants. The PCs discussed in the previous section identify many of the regionally most common variants and nearly one-third of the regionally most common long vowels and diphthongs. In three of the 15 regions – Cumbria-North Yorkshire,

Lincolnshire, and Devonshire, linguistically the most uniform regions – PCs assign high loadings to 8 of the 12 most common long vowels and diphthongs. However, in three other regions in the transitional center of the country – Staffordshire, the East Midlands, and the Central South – PCs fail to isolate a single one (though one PC identifies a number of less common variants in northern Staffordshire). Moreover, about one-third of the variants isolated by the PC analysis are forms that are fairly uncommon even in the region or regions in which they appear. Equally importantly, only about one-quarter of the variants that appear in a PC and that are among the two regionally most common variants for a Middle English phoneme can be more or less uniquely associated with a specific region: the remaining three-quarters are common in more than one region, and in a number of cases they are identified by more than one PC in more than one region. For example, the development of a low-front or front onset (i.e. [**ai** $\sim$ **ɛi**]) for Middle English /**iː**/ appears in the second PC as a characteristic of some importance in the Far North, where it is ubiquitous. However, it also features in the third PC as a feature of marginal importance in Cumbria-North Yorkshire, where it is also practically ubiquitous, and it surfaces yet again in the thirteenth PC as a feature of Lincolnshire speech, where it appears with a frequency of 69 percent.

In short, PCs rarely isolate linguistic structures, variants, or features that are unique to regions. Nevertheless they do tend to identify sets of variants or features that are associated with regions, and a robust general characterization of regional varieties of English usage emerges from the insights from the PC analysis and an examination of the regional distribution of variants.

2.8 Conclusions

In summary, the quantitative tools presented here yield reasonable measures of variation in phonetic usage among English localities, illustrate the largely continuous geographical nature of that variation (particularly with respect to vocalic features), identify a score of major and minor dialect regions as clusters of localities with relatively similar patterns of usage, distinguish regions of relative uniformity from transitional zones with substantially greater variation, and isolate regionally coherent groups of features that can be said – with appropriate caveats – to distinguish several but by no means all of the dialect regions. The results largely corroborate standard characterizations in the literature, but differ from those previous studies in several cases in the placement of several important dialect boundaries, in the association of features with dialects, and in placing those systematic characteristics against a background of largely continuous variation. The techniques reinforce each other to provide a consistent picture of the traditional English usages, tending to confirm the

applicability of computational techniques to the study of dialect variation.

All of the techniques appear useful, though they need not all be used in concert. The results suggest that even a small but judiciously chosen feature-based data set can yield a great deal of insight, and that in the absence of feature detail, variant-based data also can be used quite effectively. Clustering techniques are useful for delineating dialect regions, but they do not necessarily distinguish between regions of greater uniformity and greater diversity – a task for which distance regressions, variant-area regressions, and barrier analysis are more suitable, and any one of which is likely to be sufficient. Finally, principal component analysis appears to be useful in identifying distinctive features or variants that are more-or-less characteristic of dialect regions. The technique may be equally applicable to the study of linguistic variation across class, gender, and other dimensions as well. In some circumstances, feature-based PCA may not only capture structural features such as upgliding or ingliding, but may provide measures of their strength, while variant-based PCA may capture groups of variants that reflect structural shifts. In this respect, feature-based and variant-approaches may be substitutes rather than complements.

Taken together, the results provide a phonetic basis for the following tentative classification of the traditional English dialects and dialect regions, as reflected in the usage of older rural speakers in the mid-twentieth century.

The relatively uniform ***Far North*** is distinguished from the rest of the country most notably by its use of uvular [**ʁ**] – the Northumbrian burr. However, the region also features the survival of [**h**] and [**hw**] – features that are occasional but increasingly uncommon as one moves further south – as well as the fairly common merger, raising, and fronting of the low back long vowels to [**øː**] and the fronting and occasional lengthening of /**o**/ to [**œ ∼ œː**] and of /**au**/ to [**aː ∼ æː**]. As in most of the north of England, the Middle English qualities of short vowels tend to be conserved.

The ***Upper North*** develops centering diphthongs for most of the long vowels. *Cumbria-North Yorkshire* is characterized by the development of a high front (and sometimes palatalized) onset plus inglide for /**aː**/, /**ɔː**$_1$/, and /**oː**/, leading to a significant degree of merger. The *Upper Northwest* shows a similar development for /**aː**/ and /**ɔː**$_1$/, except that the glides are to a lower, more fronted [**a**] rather than [**ə**]. Both regions show the occasional development of /**oː**/ to [**jɤ ∼ iø**].

The ***Lower North*** combines a tendency to develop the high long vowels into upglides with a low-center onset with a tendency to develop centering diphthongs for the low and middle long vowels without strongly raising them. The *Lower Northwest* is also characterized, in the west, by use of an alveolar approximant rhotic, and in the east by fronted upglides in the long low back vowels. *Lincolnshire* has practically no unique features, save perhaps the merging of /**aː**/ and /**ai**/ to [**eə ∼ ɛə**], yet has a sufficiently unique and uniform

combination of variants that it can be clearly identified by a variant-based PC.

The ***Central Midlands*** sport such widely used and incoherently distributed variants and features that PC analysis cannot uncover definitive sets, but the region can generally be identified by the unique overlapping of northern retention of older forms for the short vowels and the southeastern innovations in the long vowels discussed below. The *East Midlands* – in many ways the heart of the English dialect world, with many of the most typical localities in the *SED* and a highly uniform pattern of speech – never scores highly in any principal component. *Staffordshire* appears as an island of relative uniformity in an important transition zone with unusual variability, but several techniques suggest that a "Potteries" system is not evident the *SED* data, though principal component analysis isolates a component of it in the northern part of the region.

The ***Upper Southwest*** sees the most northerly expressions of several southern characteristics, including rhoticity and short vowel shifts. The *West Midlands* are characterized by a general lowering and unrounding of short back vowels, occasional retraction and rounding of low short front ones, occasional high onsets in words with the middle long vowels, and little movement otherwise. The *Central South* proves to be a highly variable, transitional region whose variants are in some cases too widespread and in other cases too infrequent to identify it. An examination of common features suggests that the Central South combines southern shifts in short vowels and retention of rhoticity with minor northern tendencies to develop inglides in long vowels.

The ***Southeast*** sees extensive shifts in the Middle English short vowels as well as the development of the Middle English long vowels into coherent systems of upgliding diphthongs that are strongly evident in several of the variant-based and feature-based principal components. *North Anglia* appears best defined by a hodgepodge of unique features, including glottalization of medial /**t**/, the development of [**w**] for /**v**/, the development of a centered, unrounded onset in /**iː**/, the continuing use of [**h**], but the analysis fails to identify several other distinctively East Anglian features. The *Central Southeast* is distinguished by the tendency to exaggerate the fronting and lowering of the first elements of some of the upgliding diphthongs, such as [**æi** ∼ **ai**] for /**aː**/ and especially [**əu** ∼ **ʌu** ∼ **æu**] for /**ɔː**$_1$/, while the *Southeast Coast* is distinguished by the tendency to enhance the general Southern raising of short front vowels in at least some contexts combined with retention of rhoticity; /**l**/-vocalization appears with relatively high frequency in the latter two regions as well.

The ***Lower Southwest*** appears best defined by the voicing of fricatives and occasionally of medial dentals, the voicing and dentalizing of medial fricative [**s**], and full lowering and unrounding of /**u**/. Along with much of the Upper Southwest and the Southeast Coast, the Lower Southwest retains a

fully retroflex rhoticity after vowels. The *Central Southwest* has little that is unique about it, tending to have features that overlap with those in neighboring regions, while *Cornwall* has little to distinguish it from the Central Southwest except its lower frequency of fricative voicing. *Devonshire* is distinguished from the rest of the Southwest by its unique fronting of back vowels and the development of a low monophthong for /**i:**/.

The neatness of these tentative classifications should not obscure the largely continuous and often unsystematic nature of phonetic variation across England. Even in regions with relatively uniform speech, few localities – not even those localities that can be characterized as having speech most typical of their regions – adhere rigidly to the patterns uncovered by any quantitative approach. Regional speech patterns appear to be only rather loosely characterized either by the patterns in their associated principal components, by the speech patterns of their most typical localities, or by average frequencies of occurrence of features. In linguistics as in biology, variation abounds, and type specimens are necessarily arbitrarily chosen.