A Statistical Comparison of Regional Phonetic and Lexical Variation in American English

Jack Grieve

QLVL

University of Leuven

November 22nd, 2011

Abstract

This paper presents a statistical comparison of regional phonetic and lexical variation in American English. Both the phonetic and lexical datasets were first subjected to separate multivariate spatial analyses in order to identify the most common dimensions of spatial clustering in these two datasets. The dimensions of phonetic and lexical variation extracted by these two analyses were then correlated with each other, after being interpolated over a shared set of reference locations, in order to measure the similarity of regional phonetic and lexical variation in American English. This analysis shows that regional phonetic and lexical variation are remarkably similar in Modern American English.

Acknowledgements

I would like to thank Dirk Geeraerts, Kris Heylen, Tom Ruette, Dirk Speelman, Joeri Theelen and Emily Waibel for their comments on this paper and for their help with the statistical and cartographic techniques applied in this paper. I would also like to thank Charles Boberg, Bill Kretzschmar, John Nerbonne and Martijn Wieling for their comments on the preliminary version of this paper, which was presented at *Methods in Dialectology 14* at the University of Western Ontario, August 2nd, 2011.

1. Introduction

Regional linguistic variation in American English has not been adequately compared across linguistic levels. Kurath designed the questionnaire for the *Linguistic Atlas of the United States and Canada* to elicit data from many levels of linguistic analysis, but the *Atlas* was never completed and, aside from New England (Kurath et al, 1939), Kurath and his colleagues presented their lexical (Kurath, 1949), morphological (Atwood, 1953) and phonological (Kurath & McDavid, 1961) analyses for the eastern United States separately. All three of these studies did find evidence for a three-way division between the North, the South and the Midland, but the data was never explicitly compared across these three linguistic levels. Similarly, later surveys that mapped regional linguistic variation in American English focused specifically on lexical (Cassidy, 1985; Carver, 1987), phonetic/phonological (Labov et al, 2006), or lexico-grammatical variation (Grieve, 2009; Grieve et al, 2011). Although all of these surveys compared their results to the results of previous American dialect surveys, these comparisons were always subjective and informal, as the raw data from the previous studies were not available for direct analysis.

This study presents the first formal comparison of regional linguistic variation in American English across two linguistic levels, based on the acoustic phonetic dataset from the *Atlas of North American English* (Labov et al, 2006), which consists of average formant 1 and 2 values for a series of vowels, and the lexical dataset from Grieve et al (2011), which consists of proportions for high frequency adverb and function word alternations. Both of these datasets were first subjected to an independent statistical analysis, as reported in two previous publications on lexical (Grieve et al, 2011) and phonetic (Grieve et al, submitted) variation in American English. In both studies, a series of quantitative dimensions representing common patterns of regional variation were extracted through a multivariate spatial analysis. Following a brief description of the results of these two studies, this paper

introduces a method for comparing dialect maps and describes how this method was used to compare of the dimensions of lexical and phonetic variation in American English identified in these two studies. This analysis shows that regional lexical and phonetic variation are similar in modern American English.

2. Regional Phonetic Variation

This section presents the results of the multivariate spatial analysis of the acoustic vowel formant data from the Atlas of North American English (Labov et al, 2006). This analysis is presented in greater detail in Grieve et al (submitted).

2.1 Phonetic Dataset

The phonetic analysis was based on the acoustic data from the *Atlas of North American English* (Labov et al, 2006). Data was collected during the 1990s through linguistic interviews conducted over the telephone with 762 informants from the United States and Canada. The recordings for 439 of these informants were subjected to an acoustic analysis in order to measure the average formant 1 and formant 2 values for 24 vowel variables.

This analysis is based on a subset of the acoustic data in the Atlas, specifically the formant 1 and 2 values of 19 vowel variables in the language of 402 American informants. Five vowel variables were excluded due to missing data. Canadian informants were excluded to control for national variation and Alaskan informants were excluded because as extreme geographical outliers their inclusion would confound the spatial analysis. Each vowel formant variable was then averaged over all informants from the same city, as required by the statistical analysis, resulting in a final dataset that consists of 38 vowel formant variables (formant 1 and formant 2 for 19 vowel variables) measured across 236 cities from across the contiguous United States.

Table 1 lists the 19 vowel variables using the transcription and classification systems presented in the Atlas along with their IPA equivalents and an example of the vowel variables in context. Some of these vowel variables involve the same vowel but are distinguished based on phonological context, as indicated in Table 1. The 19 vowel variables are plotted based on their average formant 1 and formant 2 values across all 236 cities in Figure 1. The values of the 38 variables were also mapped individually across the 236 locations.¹ Figure 2 maps formant 1 for /æ/, which is characterized by larger values in the Midwest. Figure 3 maps formant 1 for /eyc/, which is characterized by larger values across the North and the West. Although regional patterns are apparent in many of these maps, the underlying patterns of regional variation in the dataset appear to be highly complex. A multivariate spatial analysis was therefore applied to identify the underlying patterns of regional linguistic variation.

Figure 1 Average Formant 1 and Formant 2 Values for all Vowel Measures

¹ All maps were made in R using functions from the *maps*, *maproj* and *maptools* and *sp* packages (see Bivand et al, 2008).

Table 1 The English Vowel System

	Short		Long					
			Upglide				Inglide	
			Front Upglide		Back Upglide			
	Front	Back	Front	Back	Front	Back	Unround	Round
High	i (ɪ) <i>bit</i>	u (ʊ) <i>book</i>	iyc (i) word internal <i>beat</i>		uwc (u) word internal <i>boot</i> uwf (u) word final <i>boo</i>			
Mid	e (ɛ) <i>bet</i>	uh (ʌ) <i>but</i>	eyc (e) word internal <i>bait</i>		owc (o) before consonants boat owr (o) before r <i>bore</i>		oh (ɔ) not before r <i>bought</i> ohr (ɔ) before r <i>forest</i>	
Low	ae (æ) <i>bat</i>	o (a) <i>cot</i>	ay0 (ai) before voiceless consonants <i>bite</i> ayv (ai) before voiced consonants <i>bide</i>		aw (aʊ) <i>bout</i>		ahr (a:) before r <i>four</i>	

Figure 2 Raw Map for /ae/ Formant 1

Figure 3 Raw Map for /eyc/ Formant 1

2.2 Multivariate Spatial Analysis: Phonetic Data

Each variable was first subjected to an analysis of global spatial autocorrelation using Moran's I (Moran, 1948; Odland, 1988; Grieve, 2011) to identify significant patterns of spatial clustering.² A significant negative value for Moran's I indicates that the values of a variable exhibit spatial dispersion, where nearby locations tend to have different values to a greater degree than would be expected by chance. A significant positive value for Moran's I indicates that the values of a variable exhibit spatial clustering, where nearby locations tend to have similar values to a greater degree than would be expected by chance.

To calculate measures of spatial autocorrelation, it is necessary to define a *spatial weighting function*, which is a set of rules that assigns a weight to the comparison of every pair of locations so that comparisons between locations that are close together are given greater weight than comparisons between locations that are far apart (Odland, 1988). A simple binary spatial weighting function (see Grieve, 2011) assigns a weight of 1 to all pairs of locations within a certain distance and a weight of 0 to all other pairs of locations (Odland, 1988). The cutoff distance essentially sets the level of resolution for the analysis: a smaller cutoff is better for identifying smaller clusters, whereas a larger cutoff is better for identifying larger and less homogeneous clusters. However, because different variables can exhibit significant patterns for different cutoffs, it is useful to fit this cutoff distance for each variable. This can be achieved by calculating global Moran's I for each variable for a range of different cutoffs and by then selecting the cutoff that identified the most significant degree of spatial clustering for that variable (Grieve et al, 2011). Despite the fact that a reciprocal spatial weighting function was used in the original phonetic analysis (Grieve et al, submitted), the analysis presented here is based on a fitted binary spatial weighting function in order to maintain maximum comparability with the lexical analysis. This change in spatial

² The spatial autocorrelation analyses were conducted using Perl scripts written by the author.

weighting functions, however, has very little effect on the results of this analysis, as can be seen by comparing the factor maps presented below to the factor maps in the original study.

Moran's I and the associated z -score (Odland, 1988) were calculated for each vowel formant variable using the fitted weighting function. Based on the associated z -score, the significance of Moran's I was tested using a 1-tailed test of significance and an adjusted significance level of .0013 to correct for the number variables under analysis ($.05/38 = .0013$). Based on this analysis, 36 out of the 38 phonetic variables were found to exhibit significant levels of spatial clustering.

In addition to testing for global spatial autocorrelation analysis, a local spatial autocorrelation analysis was used to identify the location of spatial clusters in the maps for each variable (Ord & Getis, 1995). Unlike a measure of global spatial autocorrelation, which returns one value per variable that indicates the degree of clustering in the map for that variable, a measure of local spatial autocorrelation returns one value per location per variable that indicates the degree to which that location is part of a high or low-value cluster, with a significant negative Getis-Ord G_i^* z -score indicating that the location is part of a low-value cluster and a significant positive Getis-Ord G_i^* z -score indicating that the location is part of a high-value cluster.

When the results of the local spatial autocorrelation analysis were mapped, clear high and low value clusters were identified in the maps of all of the vowel formant variables. Local autocorrelation maps are provided for /ae/ on Formant 1 in Figure 4 and for /eyc/ on Formant 1 in Figure 5, corresponding to the raw maps presented in Figure 2 and Figure 3. These local spatial autocorrelations maps identify clear patterns of spatial clustering in the values of these two variables, with /ae/ on formant 1 primarily contrasting the Midwest with the Northeast and the West and with /eyc/ on formant 1 primarily contrasting the North with the Southeast. By comparing the raw maps with these smoothed maps, it is clear that the local

autocorrelation analysis has successfully identified underlying patterns of spatial clustering. It is also clear, however, that these two variables, for example, exhibit different patterns of regional variation, which is not altogether clear in the raw maps for these variables.

Finally, in order to identify common patterns of regional variation in the values of the 38 vowel formant variables, the Getis-Ord Gi^* z -scores for the complete set of variables were subjected to a factor analysis.³ Given a set of variables measured over a large number of observations, a factor analysis extracts a series of factors that each represent a common pattern in the dataset, as well as identifies the variables associated with each of these patterns and the amount of variance explained by that factor. Because the local Getis-Ord Gi^* z -scores represent the location of spatial clusters in the values of the vowel formant variables, subjecting this dataset to a factor analysis identifies common patterns of spatial clustering, as well as the specific vowel formant variables that are associated with each of these clusters.

To conduct a factor analysis it is necessary to select the number of factors to be extracted. In this analysis, four factors were extracted for two reasons. First, the factor analysis was run numerous times for different factor solutions and the cumulative and additional amount of variance explained with the addition of each factor were recorded. It was found that four factors accounted for 88.4% of the variance in the dataset, while adding additional factors would have explained relatively little additional variance. Second, the factor scores were mapped to see if the factors represent interpretable patterns of regional linguistic variation. When the four factor solution was mapped (see below), each factor was found to exhibit a clear, distinct and recognizable pattern. The factor analysis was therefore used to extract four factors. Varimax rotation was also applied to limit the number of factors onto which each of the variables load, causing the factors to more clearly reflect the spatial patterns visible in the local autocorrelation maps of the individual linguistic variables.

³ The factor analysis was conducted using the *factanal* function in R.

Figure 4 Getis-Ord Gi* Map for /ae/ on Formant 1

Figure 5 Getis-Ord Gi* Map for /eyc/ on Formant 1

The factor scores for each of the factors are mapped in Figures 6 through 9, allowing for the common patterns of spatial clustering represented by each factor to be visualized.

Factor 1 accounts for 34.8% of the variance in local autocorrelation and contrasts the Southeast with the North. Factor 2 accounts for 25.8% of the variance and contrasts the Midwest with the rest of the United States. Factor 3 accounts for 21.6% of the variance and contrasts the Northeast with the rest of the United States. Finally, Factor 4 accounts for 6.2% of the variance and contrasts the Lower Midwest and the West with the Upper Midwest and the Southeast. It is also important to note that the variables loadings, which indicate the degree to which each variable is represented by each factor, are highly interpretable, with the first two factors in particular identifying the vowels and the shifts associated with the Southern Shift and Northern Cities Shift (Labov et al, 2006).

Figure 6 Phonetic Factor 1

Figure 7 Phonetic Factor 2

Figure 8 Phonetic Factor 3

Figure 9 Phonetic Factor 4

3. Lexical Variation

This section presents the results of the multivariate spatial analysis of 40 lexical alternation variables extracted from a corpus of American letters to the editor (Grieve, 2009). This analysis is presented in greater detail in Grieve et al (2011).

3.1 Lexical Dataset

The lexical analysis was based on a corpus of 26 million words representing the letter to the editor register as written in 206 cities from across the United States (see Grieve, 2009, 2011; Grieve et al, 2011). The corpus was compiled by downloading letters to the editor written between 2000 and 2010 from the online archives of major daily newspapers from across the contiguous United States. In total, the corpus contains 26,573,826 words, spread across

159,181 letters, written by 130,659 authors, representing 206 cities from across the contiguous United States.

This analysis focused on the values of 40 quantitative lexical alternation variables, each of which consist of two synonymous high frequency function words or adverbs. These 40 alternation variables and their variants are listed in Table 2, which is organized by part-of-speech. Although most of the variants are vocabulary items that are equivalent or nearly equivalent in almost every case, a few variables involve alternations that are only equivalent in certain environments. The contexts in which these pairs of variants were counted are described in the final column of the table. Each alternation was measured as a proportion by dividing the total tokens of one variant of a variable in a city sub-corpus by the total number of tokens of all of the variants of that variable in that city sub-corpus, thereby controlling for corpus size.⁴

The values of the 40 lexical alternation variables were mapped across the 206 locations. Figure 10 maps *anyone/anybody* alternation, with *anybody* appearing to be more common in the West. Figure 11 maps *though/although* alternation, with *although* appearing to be more common in the Northeast. Although regional patterns are apparent in many of the individual maps, these patterns are not as clear as the patterns in the phonetic analysis and once again the underlying patterns of regional variation in the dataset appear to be highly complex. A multivariate spatial analysis was therefore used to identify underlying patterns of regional linguistic variation in this dataset.

⁴ All of the lexical alternation variables were measured in the city sub-corpora using Perl scripts written by the author.

Table 2 Lexical Alternation Variables

POS	Variant 1	Variant 2	Context
Adverb	actually clearly especially however maybe usually whatsoever	in fact obviously particularly (none/never)theless perhaps normally at all	
Adverb and Adjective	forward, backward, upward, downward	forwards, backwards, upwards, downwards	
Ordinal	Firstly, secondly, thirdly, fourthly, fifthly, lastly	First, second, third, fourth, fifth, last	Sentence initial before commas
Subordinator	if therefore whilst	whether thus while	Following certain verbs
Conjunction	as well as though	in addition to although	<i>Though</i> not counted following <i>as</i> or before comma or period
Preposition	about about amid amongst because of below to toward until	on around amidst among due to under toward, towards towards till, 'til	Following certain nouns Before numbers Following certain nouns
Pronoun	anyone em everyone no one someone	anybody them everybody nobody somebody	
Relative Pronoun	whom which that who who	who that which that that	Following prepositions Following commas preceded by nouns (nonrestrictive) Following nouns (restrictive) Following personal nouns Following compound pronouns
Modal Auxiliary	be going to have to may ought shall	will must might should will	Before verbs
Infinitive	so as to	in order to	
Genitive	of genitive	's genitive	Following nouns

Figure 10 Raw Map for Anyone/Anybody Alternation

Figure 11 Raw Map for Though/Although Alternation

3.2 Multivariate Spatial Analysis: Lexical Data

Each variable was first subjected to an analysis of global spatial autocorrelation using Moran's I to identify significant patterns of spatial clustering. Using the fitted weighting function, 10 out of the 40 lexical variables were found to exhibit significant levels of spatial clustering at an adjusted $(.05/40 =) .00125$ significance. It is therefore clear that the regional signal in the lexical dataset is not as strong as the regional signal in the phonetic dataset.

To identify the specific location of regional clusters in this dataset, each variable was subjected to a local spatial autocorrelation analysis using Getis-Ord G_i^* . When the results of the local spatial autocorrelation analysis were mapped, clear high and low value clusters were identified in the maps for many vowel formant variables. Local autocorrelation maps are presented for *anyone/anybody* alternation in Figure 12 and *though/although* alternation in Figure 13, corresponding to the raw maps presented in Figures 10 and 11. These local spatial autocorrelations maps identify clear patterns of spatial clustering in the values of these two variables, with *anyone/anybody* alternation primarily contrasting the East with the West and with *though/although* alternation primarily contrasting the South Central with the Northeast.⁵

To identify common patterns of regional variation in the values of the 40 lexical alternation variables, the Getis-Ord G_i^* z -scores for the complete set of variables were subjected to a factor analysis using varimax rotation. Three factors were extracted because these factors accounted for 54.4% of the variance in the dataset, with additional factors accounting for relatively little additional variance. When mapped, later factors also appeared to be largely derivative of these first three factors. The factor scores for each of the factors are mapped in Figures 14 through 16. Factor 1 accounts for 23.5% of the variance and contrasts the Northeast and the West Coast with the rest of the United States. Factor 2

⁵ There are some small differences between these maps and the local spatial autocorrelation maps presented in Grieve et al (2011) because the mean value of the variable is no longer rounded before computing Getis-Ord G_i^* .

accounts for 18.1% of the variance and contrasts the Midwest and the Northeast with the rest of the United States. Finally, Factor 3 accounts for 12.7% of the variance and contrasts the Southeast with the rest of the United States. The first two sets of factors loading are also interpretable. Factor 1 associates the use of more formal variants with the Northeast and Factor 2 associates the use of more standard American forms with the Midwest.

Figure 12 Getis-Ord Gi* Map for Anyone/Anybody Alternation

Figure 13 Getis-Ord Gi* Map for Though/Although Alternation

Figure 14 Lexical Factor 1

Figure 15 Lexical Factor 2

Figure 16 Lexical Factor 3

4. A Comparison of Regional Phonetic and Lexical Variation in American English

The multivariate spatial analysis of the phonetic and lexical datasets generated two series of factors that are each measured over a series of locations. Each of these factors represents a common pattern of spatial clustering in the phonetic or lexical dataset. The similarity between the phonetic and lexical factors was then measured by correlating these two series of factors. However, because these two series of factors are defined for two different sets of locations, before the factors were correlated, the factor scores were interpolated over a grid of evenly spaced reference locations. This section describes the interpolation and correlation of these factors.

4.1 Factor Interpolation⁶

In order to correlate the phonetic and lexical factors, the factors must be defined for the same set of locations. It was therefore necessary to geographically normalize the factors before conducting this comparison. This was achieved by interpolating each factor over a grid of evenly spaced reference locations through kriging.

Before the factor scores were interpolated, a series of centrally-symmetric grids were generated, consisting of 97, 207, 290, 408 and 506 reference locations spread out evenly across the contiguous United States. Grids of different densities were generated in order to evaluate the effect of varying this parameter based on the results of the correlation analysis. Grids containing 207 and 506 normalized reference locations are presented in Figures 24 and 25, where each horizontal line of locations has the same longitude and each vertical line of locations has the same latitude.

⁶ All of the analyses described in this section were conducted in *R* (see Bivand et al, 2008). The grids of reference locations were generated using the *map2SpatialPolygons* function (in the *maptools* package) and the *spsample* function (in *sp* package). The variograms were generated using the *variog* function, the kriging parameters were estimated using the *variogfit* function, and ordinary kriging was conducted using the *krige.conv* function (all in the *geoR* package).

Each factor was interpolated over the grid of reference locations using *ordinary kriging* (Cressie, 1993). Kriging is a geostatistical method used for estimating the values of a variable over a series of unobserved locations based on a geographically weighted average of the values of the variable at observed locations (Bivand et al, 2008). As opposed to simple kriging, where these weighted averages are based only on the distance between the unobserved and observed locations, in ordinary kriging the weighted averages are based on the empirical variogram for that variable. In particular, in order to interpolate the values of a variable using ordinary kriging, it is necessary to specify two parameters—sigma squared and phi—which can be estimated based on the empirical variogram for that variable.

A *variogram* is a function that represents the degree of spatial dependence in the values of a variable measured over a series of locations (Bivand et al, 2008). Specifically, a variogram plots the *semivariance* of a variable at every pair of locations as a function of the distance between those locations, where the semivariance of a variable at a pair of locations is equal to one half of the squared difference of the values of the variable at that pair of locations. The variogram therefore shows how the values of a variable measured over a series of locations change as the distance between locations increases. Typically, the semivariance of a variogram starts near zero, indicating that nearby locations tend to have similar values, and then increases as the distance between locations increases, indicating that as the distance between locations increases so too does the difference between the values of the variable at those locations. Often the semivariance eventually levels off so that the semivariance no longer increases with distance, although numerous other patterns are possible, some of which will be exemplified below based on an analysis of the empirical variograms for the phonetic and lexical factors.

An *empirical variogram* is a variogram that is based on the observed values of a variable. If the values of a variable are known for a sufficient number of locations, then an

empirical variogram can be used to estimate the parameters of the underlying *theoretical variogram* for that variable, which can then be used as the basis for ordinary kriging. The empirical variograms for each of the seven factors extracted by the phonetic and lexical analyses are presented in Figure 17 through Figure 23. Although none of the factors are normally distributed, and generating a variogram based on a skewed variable can be problematic, none of the factors exhibit a skewness coefficient greater than +1 or smaller than -1 and so the variograms were generated based on the raw factor scores (Oliver, 2010). For each factor, the main graph presents the empirical variogram for the complete range of distances. In the background is the *variogram cloud*, where the semivariance for every pair of locations is plotted on the y-axis and the distance between those locations is plotted on the x-axis (Bivand et al, 2008). In the foreground is a line representing the *sample variogram*, which is based on the average semivariance for all pairs of locations within a certain range of distances (Bivand et al, 2008). The sample variogram reduces the noise in the variogram cloud, allowing for the parameters of the theoretical variogram to be estimated with greater accuracy. The inset graph presents a magnified version of the sample variogram at small distances with the circles representing the average semivariance for each of the distance binnings. A fitted exponential function is also plotted in the inset (Bivand et al, 2008). The inset focuses on the value of the variogram at relatively small distances, including the approximate distance where the variogram first plateaus, because the empirical variogram becomes less reliable at larger distances. This is also the portion of the variogram that is used to estimate the parameters for ordinary kriging.

Several values must be estimated based on the empirical variogram in order to specify the values of the kriging parameters. Given an empirical variogram, the value where the variogram plateaus is known as the *sill*, the value of the variogram at distance zero is known as the *nugget*, and the distance at which the sill is reached is known as the *range*. For

example, the variogram for Phonetic Factor 1 presented in Figure 17 appears to have a sill of approximately 1.3, a nugget of approximately 0, and a range of approximately 10. The two parameters for ordinary kriging are based on these values: sigma squared is equal to the *partial sill*, which is calculated by subtracting the nugget from the sill, and phi is equal to the range (measured as Euclidean distance). By plotting and analyzing an empirical variogram for a particular variable it is therefore possible to estimate the value of sigma squared and phi so that ordinary kriging can be conducted.

In addition to estimating the values of these variables by hand, it is possible to fit a correlation function to the variogram and to then automatically extract the partial sill and range based on this fitted function. As noted above, the line plotted in the inset is an exponential correlation function that was fit to the empirical variogram (weighted based on the number of pairs per bin). Based on a fitted exponential correlation function, the variogram for phonetic Factor 1 presented in Figure 17, for example, was found to have a sill of 1.38, a nugget of 0, and a range of 8.36, which are similar to the estimates presented above. Instead of setting the parameters for each factor kriging by hand, the exponential correlation function was therefore fitted to the empirical variogram for each factor in order to automatically set the parameters for kriging.

Based on the estimations of the partial sill and the range for each of these variograms, the 7 factors were kriged over the 97, 207, 290, 408 and 506 location grids. Phonetic Factor 1 is presented in Figure 25 interpolated over 207 reference locations and in Figure 26 interpolated over 506 reference locations. By comparing these kriged factor maps to the raw map for this variable (Figure 6), it is clear that this method successfully produces normalized maps that match the original variable as well as each other.

Before presenting the comparison of the interpolated factors, it is important to note that these variograms reveal additional information about the regional patterns exhibited by

the factors. Most important, an increasing slope, at least at relatively small distances, indicates that a variable is spatially patterned. As would be expected, given the methods used to generate these factors, all of the factors exhibit an initially increasing slope. The steepness of this slope also reveals how gradually the values of the variable change across space, with a steeper slope being associated with more sudden changes and a more moderate slope being associated with more gradual changes. For example, Lexical Factor 2 has a relatively gradual initial slope and in turn has a relatively large number of middling factor scores separating the Northeast from the West, whereas Phonetic Factor 4 has a steeper initial slope and in turn has sharper changes between regions. Nonetheless, although there is variation in the steepness of these slopes, all of the variograms start at the origin, which indicates that there are no adjacent locations with very different factor scores. This is evidence that all of these factors exhibit a continuous pattern of regional variation (Oliver, 2010).

In addition to analyzing the initial section of the variogram, there is also considerable information that can be deduced from the rest of the variogram. For example, a variogram that continues to trend upward after the sill has been reached indicates that the values of the variable are changing in a consistent direction across the entire map. The only clear example of this type is the variogram for Lexical Factor 3, which when mapped exhibits a relatively steady change originating from the Northeast and moving outward. A wave-like pattern in a variogram, on the other hand, is indicative of a periodic pattern (Oliver, 2010). For example, the variogram for Phonetic Factor 4 exhibits a relatively wavy pattern and the map for this factor identifies numerous high and low value clusters that alternate across the map. Phonetic Factor 1 and Phonetic Factor 2 as well as Lexical Factor 3 exhibit a slower wave. The maps for all three of these factors are characterized by a cluster in the north central or south central States, resulting in locations on either coast that are very far apart having very similar factor scores, which is why the semivariance for these factors becomes smaller at large distances.

Similarly, Lexical Factor 1 shows one symmetrical arch because in this case the entire central region is contrasted with both coasts.

An empirical variogram can therefore be used to extract a great deal of information about the spatial patterning of a variable. In this case, the factor variograms were used primarily to estimate the sill and the range for each factor so that these variables could be interpolated across a series of unobserved locations. But the variogram can also be used to test if a variable is regionally patterned, as well as to deduce more specific information about the regional pattern exhibited by a variable, including whether the pattern is gradual or abrupt, consistent or cyclical. The variogram, however, does not allow for the similarity between the spatial clustering exhibited by two variables to be assessed. For example, Phonetic Factor 1 and Phonetic Factor 2 have relatively similar variograms, yet looking at the corresponding maps, it is clear that these factors exhibit different spatial patterns.

Quantifying the similarity between the factors is the goal of the next stage of this analysis.

Figure 17 Variogram: Phonetic Factor 1

Figure 18 Variogram: Phonetic Factor 2

Figure 19 Variogram: Phonetic Factor 3

Figure 20 Variogram: Phonetic Factor 4

Figure 21 Variogram: Lexical Factor 1

Figure 22 Variogram: Lexical Factor 2

Figure 23 Variogram: Lexical Factor 3

Figure 24 Phonetic Factor 1 Kriged for 207 locations

Figure 25 Phonetic Factor 1 Kriged for 506 locations

4.2 Factor Correlation

In order to compare the regional patterns represented by the phonetic and lexical factors, the interpolated phonetic and lexical factor scores were plotted and correlated against each other. The results of the analysis for the interpolated factor scores based on the 207 location grid are presented first, as this grid is closest to the number of locations in the original datasets. The effect of varying the grid density is then discussed.

The interpolated phonetic and lexical factors were plotted against each other in order to visualize the relationship between these two sets of factor scores. The scatterplot matrix is presented in Figure 26. Overall, these scatterplots show that there are considerable similarities between the phonetic and lexical factors. Most notably, there is a clear positive linear relationship between Phonetic Factor 1 and Lexical Factor 3, indicating that regions with high Phonetic Factor 1 scores tend to have high Lexical Factor 3 scores and regions with low Phonetic Factor 1 scores tend to have low Lexical Factor 3 scores. There is also a clear negative linear relationship between Phonetic Factor 2 and Lexical Factor 2, indicating that regions with high Phonetic Factor 2 scores tend to have low Lexical Factor 2 scores and vice-versa.⁷ Furthermore, the fact that these relationships appear to be linear (as opposed to being simply monotonic) indicates that the rate of change in the factor scores between these high and low value regions is also similar for these paired factors. In addition, there is a relatively clear positive linear relationship between Phonetic Factor 3 and Lexical Factor 1, indicating that these factors also exhibit similar patterns of spatial clustering. Phonetic Factor 4, however, does not appear to correspond to any of the factors extracted in the lexical analysis.

It is also notable that while the other factor pairs are not linearly or even monotonically related, none of these scatterplots exhibit random patterns. This is because all of these comparisons involve factors that represent clear patterns of spatial clustering. When

⁷ The sign of the correlation is largely irrelevant as it depends on which of the two variants each of the lexical alternation variables was calculated for, which was an arbitrary decision.

plotted against each other, patterns therefore emerge because these dimensions of spatial clustering are related to each other. For example, the scatterplot for Phonetic Factor 3 and Lexical Factor 3 exhibits an hourglass shape because both factors classify the West similarly, but Phonetic Factor 3 contrasts the West and the Southeast with the Northeast, whereas Lexical Factor 3 contrasts the West and the Northeast with the Southeast. While such patterns are interesting, monotonic patterns in general and linear patterns in particular are of primary interest because the goal of this analysis was to measure the similarity between the phonetic and lexical factor scores.

The correlations between the interpolated phonetic and lexical factors scores were measured by calculating the Pearson and Spearman correlation coefficients for each pair of factors. Although none of the factors are normally distributed, and despite the fact that it is inappropriate to measure the significance of a correlation between non-parametric variables based on a Pearson correlation coefficient, Pearson correlation coefficients were calculated so as to measure the strength of the linear correlation between the factor scores. It was important to assess the strength of these linear correlations because the primary goal of this analysis was to determine how similar the phonetic and lexical factors scores are to one another, which is indicated by the strength of the linear relationship between two factors. Furthermore, based on the scatterplots presented in Figure 26, it is clear that there are strong linear relationships between a number of these factors, which the Spearman correlation coefficient would underestimate. The significance of the Pearson correlation coefficient, however, was not calculated. On the other hand, the Spearman correlation coefficient was calculated in order to measure both the significance and the strength of the monotonic correlation between each pair of factors. Although a non-linear monotonic correlation indicates that the two variables are related, with high value clusters and low value clusters aligning for example, the rate of change between these high and low value clusters may differ.

Both the Pearson (r) and Spearman (ρ) correlation coefficients for the comparisons between the 4 phonetic factors and the 3 lexical factors are presented in Table 3, along with the amount of variance explained by the linear correlation (r^2). Strong linear correlations (larger than .700) are marked by a cross (†) and significant two-tailed monotonic correlations (at the .001 alpha level) are marked by an asterisk (*). Based on the Pearson correlation coefficients, of the 12 comparisons presented here, the only strong linear correlations are between Phonetic Factor 1 and Lexical Factor 3, with 83% of the variance explained, and between Phonetic Factor 2 and Lexical Factor 2, with 63% of the variance explained. Moderate linear correlations are also found between Phonetic Factor 3 and both Lexical Factor 1 and Lexical Factor 3, with 26% and 27% of the variance explained respectively. On the other hand, Phonetic Factor 4 was not found to correlate even moderately with any of the lexical factors. Phonetic Factor 4, however, only accounts for approximately 6% of the regional variance in the phonetic dataset. The lack of a strong correlation between Phonetic Factor 4 and any of the lexical factors therefore has relatively little effect on the overall match between the two sets of factors. In fact, given the considerably greater amount of variance explained by each of the first three phonetic factors, if Phonetic Factor 4 had correlated with one of the lexical factors in place of one of the higher phonetic factors then the similarity between the regional patterns in these two datasets would have decreased.

In addition, 8 out of the 12 comparisons were found to exhibit significant monotonic correlations based on the Spearman correlation coefficients, including all of the strong and moderate linear correlations. Aside from the pairs of factors discussed above, however, the monotonic correlations between these factor pairs are too weak to be interpreted as indicating that these other pairs of factor scores represent the same basic regional patterns, even though these Spearman correlation coefficients are significant.

Based on the maps for these two sets of factors, it is clear why such high correlations are obtained. Phonetic Factor 1 and Lexical Factor 3, in particular, are very similar, with both factors contrasting the Southeast with the rest of the United States. Phonetic Factor 2 and Lexical Factor 2 are also very similar, with both factors contrasting the Midwest with the rest of the United States, although Lexical Factor 2 extends the Midwest further east, which is why this correlation is weaker than the correlation between Phonetic Factor 1 and Lexical Factor 3. Lexical Factor 1 is also quite similar to Phonetic Factor 3, with both of these factors contrasting the Northeast with the Central States, although only Lexical Factor 1 groups the entire East and West Coasts with the Northeast, which is why this correlation is not stronger. This is also why Lexical Factor 1 has a moderate correlation with Phonetic Factor 1: both of these factors group the Northeast and the West together. Overall, these results therefore show that regional phonetic and lexical variation are very similar in these two American English datasets, especially due to the consistent identification of Southeastern, Midwestern and Northeastern clusters.

The correlation analysis was also repeated for the other reference grids. For all analyses, the strength of the linear correlations remained stable, even when as few as 98 or as many as 506 reference locations were used for kriging. For example, there are no notable differences between the analyses based on 207 and 506 locations in terms of the strength of the linear correlations, with the exact same correlations being assigned to the strong, moderate and weak classes. However, when the correlations were based on 506 reference locations, 11 out of 12 Spearman correlation coefficients were found to be significant. This is because the increased number of locations (i.e. comparisons) made it easier to obtain significant correlations. The increased number of locations, however, had relatively little effect on the strength of the correlations, as long as there were a sufficient number of locations to represent the patterns of spatial clustering identified by the factors.

Table 3 Correlations for Phonetic vs. Lexical Factor Scores (Kriged for 207 Locations)

Phonetic	Lexical		
	Factor 1	Factor 2	Factor 3
Factor 1	$r = -.336$ $r^2 = 11.3\%$ $q = -.499^*$	$r = -.291$ $r^2 = 8.5\%$ $q = -.389^*$	$r = .910^\dagger$ $r^2 = 82.8\%$ $q = .701^*$
Factor 2	$r = -.100$ $r^2 = 1.0\%$ $q = .054$	$r = -.796^\dagger$ $r^2 = 63.4\%$ $q = -.666^*$	$r = .100$ $r^2 = 1.0\%$ $q = .210$
Factor 3	$r = .507$ $r^2 = 25.7\%$ $q = .212$	$r = -.417$ $r^2 = 17.4\%$ $q = -.288^*$	$r = .523$ $r^2 = 27.4\%$ $q = .671^*$
Factor 4	$r = -.103$ $r^2 = 1.1\%$ $q = -.161$	$r = .413$ $r^2 = 17.1\%$ $q = .396^*$	$r = -.313$ $r^2 = 9.8\%$ $q = -.242^*$

Figure 26 Phonetic and Lexical Scatter Plots (207 Locations, Fitted Weights)

5. Discussion

This paper has presented a statistical comparison of common patterns of regional phonetic and lexical variation in American English based on the acoustic vowel data from the *Atlas of North American English* (Labov et al, 2006) and high frequency lexical alternation data from a corpus of American letters to the editor (Grieve et al, 2011). First, both the phonetic and lexical datasets were subjected to separate multivariate spatial analyses. These analyses identified the most common patterns of spatial clustering in these two datasets based on factor analyses of the locally autocorrelated phonetic and lexical variables. The patterns extracted by these two analyses were then correlated with each other, after being interpolated over a shared set of reference locations, in order to measure the similarity of regional linguistic variation in these two datasets. This is the first formal comparison of regional linguistic variation in American English across two different linguistic levels.

Overall, the comparison revealed that regional phonetic and lexical variation are remarkably similar in American English. Phonetic Factor 1 and Lexical Factor 3 were found to have the strongest correlation, with the maps for both of these factors contrasting the Southeast with the rest of the United States. Phonetic Factor 2 and Lexical Factor 2 were also found to be strongly correlated, with the maps for both of these factors contrasting the Midwest with the rest of the United States. In addition, Phonetic Factor 3 and Lexical Factor 1 were found to be moderately correlated, with the maps for both of these factors contrasting the Northeast with the Central United States. Lexical Factor 1, however, also groups the West with the Northeast, which is why Lexical Factor 1 is also moderately correlated with Phonetic Factor 1, which distinguishes the Southeast from both of these regions. Phonetic Factor 4, on the other hand, does not correlate even moderately with any of the Lexical factors, although Phonetic Factor 4 accounts for relatively little variance in the phonetic dataset. The primary patterns of regional variation in the two datasets therefore largely align, showing that

phonetic and lexical variation in Modern American English follow the same basic regional patterns.

It is possible that the small differences that do exist between the phonetic and lexical factors are due to the fact that the two datasets are not diachronically aligned. Not only is the phonetic dataset based on language gathered during 1990s, whereas the lexical dataset is based on language published during the 2000s, but the phonetic dataset was controlled for length of residence, with only long term residents of the cities under analysis being interviewed, whereas the lexical dataset included all available informants who were residents of the cities under analysis, regardless of how long they had been residents of that city. This means that the phonetic data ostensibly represents English as spoken in sampled cities before the 1990s, as so many residents of these cities at that time would not have qualified as informants. Given that the patterns of regional variation in the dataset are otherwise very similar, it is therefore possible that the small differences between these patterns are a result of language change. For example, the fact that only Lexical Factor 1 contrasts both the East and West Coasts with the Central United States may represent a change in progress, indicating that the Central United States has become linguistically more distinct in recent years (or alternatively that the coastal regions have become linguistically more similar). Similarly, the fact that only Phonetic Factor 4 identifies a Midland pattern may represent a change in progress, indicating that the Midland dialect region is in the process of disappearing.

Despite these differences between the phonetic and lexical factors, overall both of these analyses resulted in a very similar picture of American dialect regions: in both cases, the Midwest, the Southeast, the Northeast and the West were identified as important dialect regions. These regions, of course, are highly familiar, as indicated by the fact that common names for these regions are already in existence. These regions are also very similar to other geographical categorizations of the United States including the major cultural regions of the

United States (Zelinsky, 1973) and the four major Census regions of the United States. Nonetheless, these regions are notably different from traditional dialect regions, including the regions identified based on a subjective analysis of same the phonetic dataset that was analyzed here (see Labov et al, 2006). In particular, both of the analyses presented here identified distinct Midwestern and Northeastern dialect regions but neither analysis identified a strong Midland dialect region. However, as argued in Grieve et al (submitted), based on a careful analysis of the individual vowel formant variables, it is clear that the values of numerous individual vowel formant variables cluster in the Northeast and the Midwest, whereas as the values of very few vowel formant variables cluster in the Midland. In addition, the finding that both the phonetic and lexical datasets resulted in such a similar picture of regional linguistic variation in American English offers further support for the patterns of regional linguistic variation identified by both of these analyses.

It is also important to consider linguistic explanations for why the specific pairs of phonetic and lexical factors identified by this analysis are so similar. This can be achieved by considering the specific phonetic or lexical variables that load on each of the factors. The phonetic factors identify chain shifts (see Grieve et al, submitted) whereas the first two lexical factors identified functional-grammatical patterns associated primarily with formality (see Grieve et al, 2011). However, aside from the Northeast appearing to be the most linguistically conservative region of the United States in terms of both phonetics and lexis, there are no obvious connections between the phonetic and lexical variables loading on the aligned factors. For example, the Northern Cities Shift (Phonetic Factor 2) and the more frequent use of standard American lexical items in the Midwest (Lexical Factor 2) do not seem to be linguistically related. There is no reason, however, to require a linguistic explanation for these alignments, as it is presumably extra-linguistic factors, such as

historical and modern communication patterns, that result in these regional correspondences across linguistic levels.

It is also important to acknowledge that the strong correlations between the phonetic and lexical factors are not only evidence of the similarity of regional variation across these two linguistic levels, but these strong correlations are also evidence of the similarity of regional linguistic variation in speech and writing. Similarly, because these two datasets were compiled using different approaches to data collection, with the phonetic dataset being based on recordings of linguistic interviews and the lexical dataset being based on a corpus of natural language, these results also show that both of these approaches to data collection are valid, as they both produce very similar results. Because linguistic interviews are the standard approach to data collection in dialectology, this result shows in particular that a corpus-based approach to data collection is a valid approach to gathering dialect data.

This paper has also introduced a new method for statistically comparing dialect maps. In this case, the method was used to compare the aggregated factors extracted by a multivariate spatial analysis, but this method could also be used for comparing other types of dialect maps. For example, the method could be used to compare the similarity between two maps of the same linguistic variable as measured by two different dialect surveys or at two different times. Finally, because this method involves the analysis of variograms and the application of ordinary kriging, this paper has also introduced these two fundamental geostatistical concepts to the field of dialectology.

References

- Atwood EB. 1953. *A Survey of Verb Forms in the Eastern United States*. Ann Arbor: University of Michigan Press.
- Bivand RS, Pebesma EJ, Gomez-Rubio V. 2008. *Applied Spatial Data Analysis with R*. New York: Springer.
- Carver CM. 1987. *American Regional Dialects*. Ann Arbor: University of Michigan Press.
- Cassidy FG. 1985. *Dictionary of American Regional English*. Cambridge, MA.: Harvard University Press.
- Cressie N. 1993. *Statistics for Spatial Data*. New York: Wiley.
- Grieve J. 2009. *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English*. Ph.D. Dissertation. Manuscript. Northern Arizona University
- Grieve J. 2011. A regional analysis of contraction rate in written Standard American English. *International Journal of Corpus Linguistics* 16.
- Grieve J, Speelman D, Geeraerts D. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23: 193-221.
- Grieve J, Speelman D, Geeraerts D. Submitted. A multivariate spatial analysis of vowel formants in American English. Submitted to the *Journal of Dialect Geography*.
- Kurath, Hans. 1949. *Word Geography of the Eastern United States*. University of Michigan Press.
- Kurath H, Hansen L, Bloch B, Bloch J. 1939. *Handbook of the Linguistic Geography of New England*. Providence: Brown University Press.

- Kurath H, McDavid RI. 1961. *The Pronunciation of English in the Atlantic States*.
University of Michigan Press.
- Labov W, Ash S, Boberg C. 2006. *Atlas of North American English: Phonetics, Phonology, and Sound Change*. New York: Mouton de Gruyter.
- Moran PAP. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B* 37: 243- 251.
- Odland JD. 1988. *Spatial Autocorrelation*. Sage Publications.
- Oliver MA. 2010. *Geostatistical Applications for Precision Agriculture*. New York: Springer.
- Ord JK, Getis A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis* 27: 286-306.
- Zelinsky W. (1973). *Cultural Geography of the United States*. Englewood Cliffs, NJ: Prentice-Hall.