Identifying Linguistic Structure in Aggregate Comparison

John Nerbonne

Rijksuniversiteit Groningen

9700 AS Groningen

The Netherlands

j.nerbonne@rug.nl

Tel.: +31 (50) 363 58 15

FAX: +31 (50) 363 68 55

www.let.rug.nl/nerbonne

submitted to J. Nerbonne and W. Kretzschmar, Jr. (eds.)

Progress in Dialectometry, special issue of

Literary and Linguistic Computing,

selected proceedings of a workshop at

Methods in Dialectology XII, Moncton, Aug. 5, 2005.

February 20, 2006

Abstract

Dialectometric techniques for analyzing variation in the aggregate are maturing rapidly, but there is still little agreement on how to extract linguistic structure from aggregate comparison. The present paper explores one means of comparing aggregate analyses in order to determine linguistically concise restrictions, essentially the use of factor analysis. Using the Southern states data which Guy Lowman collected as part of LAMSAS, we apply factor analysis to the vowels involved in aggregate analyses in order to determine which alternations in pronunciation tend most to co-occur.

1 Introduction

Dialectometric techniques for analyzing variation in the aggregate are maturing rapidly, allowing us to measure linguistic differences at various levels (Heeringa & Nerbonne 2006), and to investigate the relations between language and other culturally and biologically transmitted markers of human affinity (Manni, Heeringa & Nerbonne 2006). Nonetheless there is still too little agreement on how to extract linguistic structure from aggregate comparison. The present paper explores one means of comparing aggregate analyses in order to determine linguistically concise restrictions, essentially the use of factor analysis. Using the Southern states data which Guy Lowman collected as part of the *Lingustic Atlas of the Middle and South Atlantic States* (LAMSAS), we apply factor analysis to the vowels involved in aggregate analyses in order to determine which

alternations in pronunciations tend most to co-occur. The choice of LAMSAS as a testing ground is further motivated by the fact that it has been studied extensively. We may immediately compare our findings to the existing scholarly research.

The following sections explore in succession the motivation for this investigation (2), including earlier related work, the data on which the current study is based (3), the assessment of vowel pronunciation differences (4), the setup of the experiments (5), and the results (6). Finally we offer some conclusions and our views on prospects for such work in the future in a seventh section.

2 Motivation

Let us begin by sketching a caricature of dialectometric work on language variation in contrast to other work. Non-dialectometric work on language variation, whether inspired (and instructed) by traditional dialectology, or by sociolinguistic work on variation, aims to identify individual features from all linguistic levels with interesting geographic or social distributions. Traditional dialectologists map the distribution of these features using isoglosses or frequency gradients, and work inspired by sociolinguistics adds to this tests of the significance of frequency differences, most often using logistic regression of the sort found in VARBRULE (Paolillo 2002). A strength of the non-dialectometric work is its clear identification of the distribution of individual linguistic features.

Dialectometric work has arisen partially in response to this tradition, and in

particular in response to the criticism that the choice of features studied in the earlier, non-dialectometric work suffered from arbitrariness (Goebl 1982, Goebl 1984, Nerbonne & Kretzschmar 2003). Normally, no criteria are identified as to which features are to be studied for their geographic (or social) distribution, and yet the conclusions depend greatly on their choice. It is a standard remark in all sorts of dialectological work that features seldom, if ever, overlap perfectly, and moreover that even boundaries for single features tend to "vanish" in the face of variability (Chambers & Trudgill 1998, [[1]1980], p.104). The improved statistical sophistication of the sociolinguistic work remedies this in part, since it allows the demonstration that features chosen indeed differ significantly, but in the highly multidimensional world of language variation, this is a small consolation: one suspects that any number of features will demonstrate significant association with extralinguistic variables (including geography), and indeed, this is true for **all** of the features we examine below in more detail.

Dialectometry has therefore focused, not on the distribution of individual features, but rather on the relations between aggregates involving large numbers of features. The idea is that large numbers of variables, even though they will contain a great deal of variation irrelevant to questions of geographic or social conditioning, will nonetheless provide the most accurate picture of the relations among the varieties examined. And dialectometric techniques are eminently successful in assaying these aggregate relations among language varieties, as earlier studies have shown (Goebl 1982, Goebl 1984, Nerbonne & Kretzschmar 2003), and as several of the other contributions to this volume further demonstrate.

The key to their success is their application to entireties of available data (for example, entire linguistic atlases or the entire collected records of a field worker). By focusing on such aggregates of data, these techniques attempt to undercut the criticism of other dialectological work (above) that it proceeds too rapidly to characterizations, and that it has no way of identifying which linguistic distinctions are most important in distinguishing varieties.

In fact dialectometrical analyses generally make quite minimal linguistic assumption, recording only whether the pronunication of /ɛ/ in 'pen' was the same in one site as opposed to another. This feature need not be linked to the pronunciation of 'bed', 'lend' or even 'den'. While it is methodically sound not to assume that the same phoneme system will be used from one site to another, still it is clearly interesting to check on whether this is the case, but dialectometry has failed to take this step. The "linguistic structure" mentioned in the title of this paper refers primarily to the structure provided by the phonemic inventory of the language, and we shall be successful if we can provide a link from rather parsimonious perspective of dialectometry to evidence for this level of linguistic structure.

It is the goal of this paper to improve the link between these two traditions, in particular to show how to proceed from the aggregate characterizations of dialectometry to the identification of the linguistic factors repsonsible for the aggregate differences, e.g. the different pronunciations of a single vowel phoneme. If both traditions have contributed to the understanding of linguistic variation, then it should be worthwhile to see what the connections between the two are

like. The current paper is at least somewhat successful in this respect. A further goal will not be realized in this paper, but has perhaps been brought a step closer. Ideally, we would like to say not only that the variation of a single linguistic variable contributes to the signal of geographic or social provenance, the contribution of the single-variable studies, but also *how important* it is, i.e. how much of the aggregate signal is born by a single linguistic variable. Naturally, this will require a dialectometric approach in order to characterize the aggregate signal.

2.1 An Example

We said above that a closer look at linguistic atlases inevitably reveals numerous exceptions to virtually all of the simpler characterizations of dialect differences. Figure 1 shows the mapping of a frequent characterization of the American South, the monophthongal pronunciation of the vowel in *night* [nat] (the standard pronunciation is diphthongal, [naɪt]. If the generality of this feature is much less in the data used here than many linguistic characterizations would have it, let us note that Kurath and McDavid's (1961) summary likewise suggests an imperfect tendency (Map 47).

Even though these features are shibboleths in American English, constantly being exploited by entertainers, their distribution clearly does not distinguish the south, at least not in the LAMSAS data from the 1930s.

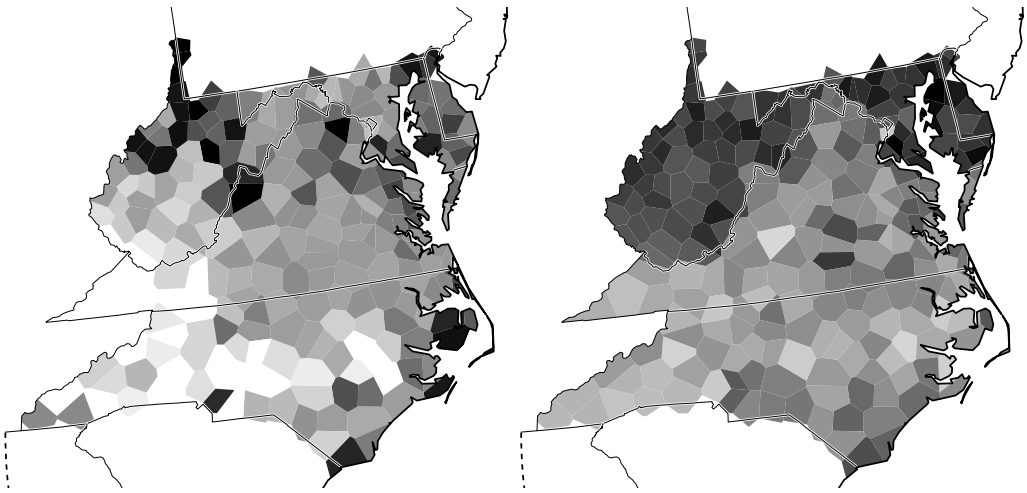


Figure 1: The darker polygons in the map on the left above show how frequently the vowel in *night* (and similar words) is pronounced [aɪ], and the light ones where it is pronounced [a]. The dark areas in the map on the right show monophthongal pronunciations of the first vowel in afternoon (and similar words) [æ] as opposed to diphthongal ones [æ^{ə}]. The idea behind dialectometric aggregation is to sum over all such differences, and this indeed gives a reliable indication of dialect differences. We note in passing that especially the [a/aɪ] shibboleth does not have the clean distribution in American speech which dialect maps sometimes suggest, at least not in the LAMSAS data from the 1930's.

2.2 Components of Aggregate Variation

The strength of the aggregating techniques used in dialectometry has not been obtained without a cost. While traditional and sociolinguistic work can characterize variation in terms of a small number of parameters (for example, adjustments in the pronunciation of segments throughout a lexicon as in the case where /æ/ is raised in some words in one area and not in another, or even more ambitiously, in terms of entire sets of segments affected by a change for which an insightful linguistic characterization exists (for example, the affrication of the German stop series in the south of Germany or the vowel shifts that Labov has identified in American pronunciation [1994, Ch.6]), dialectometric work has, for the most part, remained at the aggregate level, missing concise linguistic descriptions.

This paper explores a new way of linking dialectometric characterizations to more detailed linguistic characterizations, which is to apply factor analysis to the results of dialectometric analysis.

It is clear that linguistically informed characterizations of the association of linguistic with extralinguistic variables are more general and economical (succinct) descriptions, and it is a shortcoming of dialectometric work that it has normally neglected this part of analysis. To provide a concrete example, imagine that /æ/ is always pronounced as raised, i.e. as [æ̝] (written in LAMSAS as [æˆ]), for nearly all the pronunciations in a variety. While this will play a role in dialectometric analyses, still these analyses have not been successful in extracting such facts from the large scale, aggregate measurements. Dialec-

tometric analyses are then left in the position of characterizing the relations among varieties reliably, but failing to adduce the linguistic bases of these, at least not succinctly.

But more ambitiously, the dialectometric take on variation, emphasizing, as it does, the aggregate relations among varieties, likewise holds the promise of improving on the succinct characterizations of traditional and sociolinguistic accounts. While these latter approaches do not relate the linguistic features of their analyses to aggregate relations (which they do not produce), the dialectometric account can in principle link aggregate and single-feature characterizations.

2.3 Other Work

Heeringa (2004) computed aggregate pronunciation distances in the Netherlands, using 125 words taken from a standard atlas of Dutch dialects. He then subjected his aggregate distances to multidimensional scaling (MDS), allowing him to draw a significantly novel dialect map of the Netherlands. In order to illustrate the linguistic content of his analyses, he then examined each of the 125 words, in turn, for the degree to which they correlated with the most important dimensions of the MDS solution (pp.268–270). He was able to suggest that the most important dimension of difference was associated with the treatment of weak syllables, illustrated in the word *waren*, Eng. 'were' (pl. 3rd of the verb 'to be'), the second the Frisian/Non-Frisian distinction in the word for father, and the third the alveolar vs. uvular pronunciations of /r/ as well as, less signif-

icantly, the tendency of intervocalic /d/ to lenite to [j]. Heeringa's approach is limited because he asks the degree to which pronunciation differences for entire words correlate with aggregate differences. We would like to examine individual segments for the degree to which their variation aligns with dialectal gradients or dialectal borders.

Shackleton (2005) quantified pronunciation differences between English and American East Coast varieties with an eye to identifying the English source of the American dialects. He used a compilation of the information found on the one hand in Kurath & McDavid (1961) for American data and found on the other in Lowman's posthumously published survey of Southern English dialects (Kurath & Lowman 1970). The data was analyzed categorically, but some categories included abstract characterizations of linguistic differences. Although Shackleton's primary goal was to identify the sources of American varieties, he also extracted principal components from his findings, which enable him to identify the linguistic features which play the strongest roles in his analysis. As principal components analysis and factor analysis are statistically very similar, Shackleton's analytical approach is also very similar to the one employed below. There are differences, however. First, we have statistical reasons to prefer factor analysis (see below, § fact-anal). Second, Shackleton extracted the principal components of a matrix comparing two sets of dialects, British and American, while the present study extracts factors from a square matrix comparing all varieties one with the other. Third, Shackleton's approach relies on the availability of data which has already been analyzed into appropriate categories. We

shall extract common factors from phonetic transcriptions directly.

Nerbonne (2006) experimented with identifying linguistic information by aggregating differences, not over all of the data, but rather only over a linguistically interesting subset, in fact, just the vowels. He focused on the same subset of Lowman's data in the *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) which we also take as a case study here (see next section, and also the map in Fig. 1). Although the correlation between the pronunciation differences based on the entire data and that based on the vowel data alone was very high ($r = 0.936$), so that one could conclude that the vowel differences account for 87.6% of the variance in the data ($r^2 = 0.876$), still the technique relied on *a priori* identification of the elements suspected to be important (in this case, the vowels), which is less than optimal, and it failed to reach the level of linguistic sensitivity one would like; for example, there was no attempt to assess the contribution of individual segments.

2.4 Why Factor Analysis?

As we noted above, principal components analysis (PCA) would also be a candidate for use in the extraction of common linguistic factors. Statistically, the two are quite similar. PCA accepts as input a set of arbitrary vectors of high dimensionality and attempts to replicate their relations using new vectors of low dimensionality. The dimensions in the solution are regarded as the "principal components" of the data. It is crucial to note that PCA aims to reconstruct the original data matrix, including aspects which may not be related to common

features among the variables.

Labov (2001, pp. 286ff,354ff) has demonstrated the utility of principal component analysis in variationist studies, in particular as it may be applied to formant frequencies of vowels undergoing changes, while Clopper & Paolillo (2006) judge factor analysis superior to principal components analysis for the same sort of data because the former ignores error variances. See below (§ 5) for further comment.

While many statistical packages do not highlight the difference, allowing one to apply factor analysis (FA) to the same sorts of data as PCA, in fact, FA proceeds (internally) rather differently. FA first analyzes the data to obtain a matrix of covariances, refecting the degrees to which the component variables correlate with one another. The dimension reduction step is then aimed at reconstructing this set of correlations with a small numbers of factors. This means that FA ignores both the statistical noise in the data as well the contributions of individual variables which are not shared with others.

We are convinced that dialect data is quite noisy, and also that dialect speakers are sensitive enough to perceive signals even in isolated variables. We are therefore sceptical about including this sort of data in our attempt to isolate linguistic variables. We thus prefer factor analysis to principal component analysis as we wish to concentrate on the degree to which the individual expressions correlate with one another. Tabachnik & Fidell ([1]1996, 2001, p. 585) is an excellent resource for understanding PCA, FA and their differences.

3 LAMSAS Data

The LAMSAS material is readily accessible for reanalysis (see `http://us.english.uga.edu/lamsas/`, (Kretzschmar 1994)) and contains the responses of 1162 informants who were interviewed in 483 communities. The responses to 151 different items is included in the web distribution, which formed the basis for the work here.

LAMSAS comprises dialect material collected on the Eastern seaboard of the United States from 1933 through 1974. Our focus here is on the pronunciation of vowels in part of the data from the South, namely the part collected by Guy Lowman in 1933-1936.

We focus here on Lowman's data from North Carolina, Virginia, West Virginia, and the District of Columbia. We likewise include data from Maryland and Delaware in order to provide context for our comparisons. The map in Fig. 1 indicates the area within which the sites included in this study are found. This subset of the data included 238 field work sites, and 57,833 phonetic transcriptions of words and brief phrases or roughly 243 per site. Since we shall focus on vowels below, let us note that there is a total of 1,132 different vowels (different combinations of basic segment plus one or more diacritics) in this data.

From this totality of data, we have extracted the vowels, for example, the first vowel in the word *afternoon*, which we indicate below 'afternoon1,' and the second (and last) vowel in *Wednesday*, which we indicate 'Wednesday2.' In total we investigate 204 such vowel types (vowels in different words). Because some vowel types were not instantiated in the data, and because factor analysis (see

below) requires a complete matrix for application, we grouped the 238 sites into 30 areas with roughly 8 sites per area. The areas were determined by clustering on the basis of geographic distance alone. The pronunciation distance between two areas was then taken to be the average pronunciation distance between the pairs of sites in the respective areas, using all the data that was available.

4 Vowel Differences

We assessed the difference between different vowel pronunciations using a variant of the feature system described in Kretzschmar (1994, p.116) which we summarize in Table 1. The table notes each feature together with the range of values it may take. Even though the features are those suggested by Kretzschmar (1994), and the number of values is determined by the number of different distinctions we found in the database (see also Kretzschmar et al. p.118 on the number of distinctions), still we are responsible for the relative weights assigned to the different features. Heeringa (2004, Ch.7) finds that the segment measurements are robust with respect to small changes in relative weighting of features, and this is fortunate since it is probably impossible to set the relative weights in a non-arbitrary way (see also below).

If v_1 and v_2 are vowels, then we begin assessing the distance between the two as the sum of the absolute differences of all the features, $d(v_1, v_2) = \sum_f |f_{v_1} - f_{v_2}|$, which will be modifying slightly below. Diphthongs are represented not via particular feature configurations, but rather by two segments so that differences

v-advanced	-3, -2, -1, 0, 0.4, 1, 1.4, 2, 2.4, 3
v-high	-1.75, -1.5, -1.25, -1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75
v-rounded	-1, -0.5, 0, 0.5, 1
v-long	-0.5, 0, 0.5, 1
v-stress	0, 0.35, 0.7
v-nasal	0, 1
v-rhotic	0, 1
v-pharyng.	0, 1
v-voice	0, 1

Table 1: The set of vowel features (and feature values) used to distinguish the LAMSAS vowels.

between them are therefore effectively analyzed as the sum of differences between the first and second parts, respectively.

The feature names reflect their normal phonetic interpretation. The stress which is marked on a syllable is interpreted as a property of the vowel, which is why it appears on the list in Table 1. Vowels receive either stress, secondary stress, or no stress. Vowels were interpreted as voiced except when explicitly marked as voiceless, in which case they bore the feature [- voice]. Lowman rarely added a diacritic indicating the "pharyngealization" of a vowel, and the [v-pharyng.] feature interprets that. Vowels written as superscripts (e.g., the second parts of laxing diphthongs) are not interpreted through a feature [± super]—but rather through a weighting. Comparisons involving superscripted vowels count only 50% of what they would cost if the segments compared were both non-superscripted. The idea behind this naturally is that such minor articulations should contribute less to pronunciation difference.

The range of values reflects the number of distinctions made in the data, where we have occasionally taken the liberty of simplifying. We found 15 height

distinctions in vowels, all of which may be represented in the values here. But we could simplify the six degrees of rounding distinguished in the LAMSAS handbook to only five, as we did not find more than five in the data analyzed here.

We calculate the distance between two vowels first by simply summing the differences of all the feature values, $\sum_f |f_{v_1} - f_{v_2}|$). In order to emphasize the importance of even slight differences, we work with the logarithm of that sum $\log(1 + \text{sum})$, relecting also our view that large phonetic differences are not perceived as large dialectal differences, at least not in proportion to the differences. Finally, we wish to work with a scale with a genuine zero, leading to the following characterization:

$$d(v_1, v_2) = \log(1 + \sum_f |f_{v_1} - f_{v_2}|)$$

The use of the logarithm to de-emphasize large differences follows Heeringa & Braun (2003, 264–265), and accords with the idea that we are dealing with a psychophysical regularity (Stevens 1975). Heeringa also provides empirical analyses to justify this step. Since the values of some features may differ more than those of others, the scheme in Table 1 effectively weights some features as more important than others. Advancement may differ by as much as 6, while rounding can not differ by more than one. Diacritics representing stress, rhotism, pharyngealization and devoicing were each capable of adding maximally one unit of difference, and intermediate differences, including those indicated

by diacritics, were interpolated. The differential weightings of the features are given implicitly by the difference in the extreme values which the feature can take on.

This is the same manner of measuring vowel pronunciation differences reported on in Nerbonne (2006), so that results are comparable.

5 Factor Analysis

Factor analysis proceeds from a matrix of correlations among variables, and, based on these, postulates common factors, which may be responsible for the correlations. It is commonly used in social science as a means of detecting common factors, e.g., those which might influence the reactions to a questionnaire.

Conceptually we needed to obtain correlation matrices from the (place $\times$ place) distance matrices we are used to dealing with. These are implicitly available in the following way. First, for each vowel we obtain a distance matrix of the familiar dialectometric sort. That is, for each vowel, such as 'Alabama1' and each pair of areas, we determined the distance between the two areas $d(a_1, a_2)$—by averaging the distance between all the pairs of sites $< s_i, s_j >$, where s_i is from a_1 and s_j from a_2. This resulted in roughly 200 distance matrices, one for each vowel (type):

	Wheeling	Winston	Raleigh	Richmond	Charlotte
Wheeling	0	41	44	45	46
Winston		0	16	34	36
Raleigh			0	37	38
Richmond				0	20
Charlotte					0

We suppress the redundant $d(b, a)$ distances above where $d(a, b)$ is already present. We then derive for each pair of vowels, the correlation coefficient, i.e., the degree to which they indicate the same distance between sites. The correlations are based on the roughly 200 distance matrices calculated above.

Per vowel-pair we obtain a correlation coefficient (the vowel $\times$ vowel correlation):

	morning1	Tuesday2	pallet2	thunderstorm2	first1
morning1	1	0.02	-0.01	0.73	0.056
Tuesday2		1	0.23	-0.03	0.02
pallet2			1	0.006	0.09
thunderstorm2				1	0.043
first1					1

This correlation matrix is in the form needed for factor analysis, which we then applied, using in the open source UNIX statistics facility *R* (see `www.r-project.org/`). We used varimax as an estimation procedure, thus limiting the solutions we sought to those in which factors were orthogonal, and thereby

ignoring so-called "oblique rotations." We tested that the variables we examined were sufficiently distinct, using the KCM/Bartlett's test of sphericity, which indicated that factor analysis was applicable ($p < 0.001$).

It should also be clear that the present analysis differs from Labov's and Clopper and Paolillo's (discussed above) in proceeding from segment distances which play a role in aggregate comparison. Our goal is likewise different: we are also attempting to reduce the dimensionality of the data, but most specifically via the identification of recurrent linguistic elements which contribute to aggregate pronunciation distance.

6 Results

Using a scree plot, we could see that the first three factors are disproportionately important, accounting respectively for $16.5\%, 11\%$, and 8.5% of the variance. The fourth factor accounts for less than 4% of the variance, and subsequent factors for less. The total amount of variance explained by the three factors is low (35%), but this is common in applications with hundreds of variables.

To interpret the factors linguistically we examine the variable loadings, i.e. the correlations between the factors and the individual variables. Particularly interesting are variables which correlate fairly purely with a single factor, and we shall attend to these in what follows. Figure 2 shows the loadings with respect to each of the three pairs of factors among the most important three. We are interested in variables which have loadings close to one with respect to

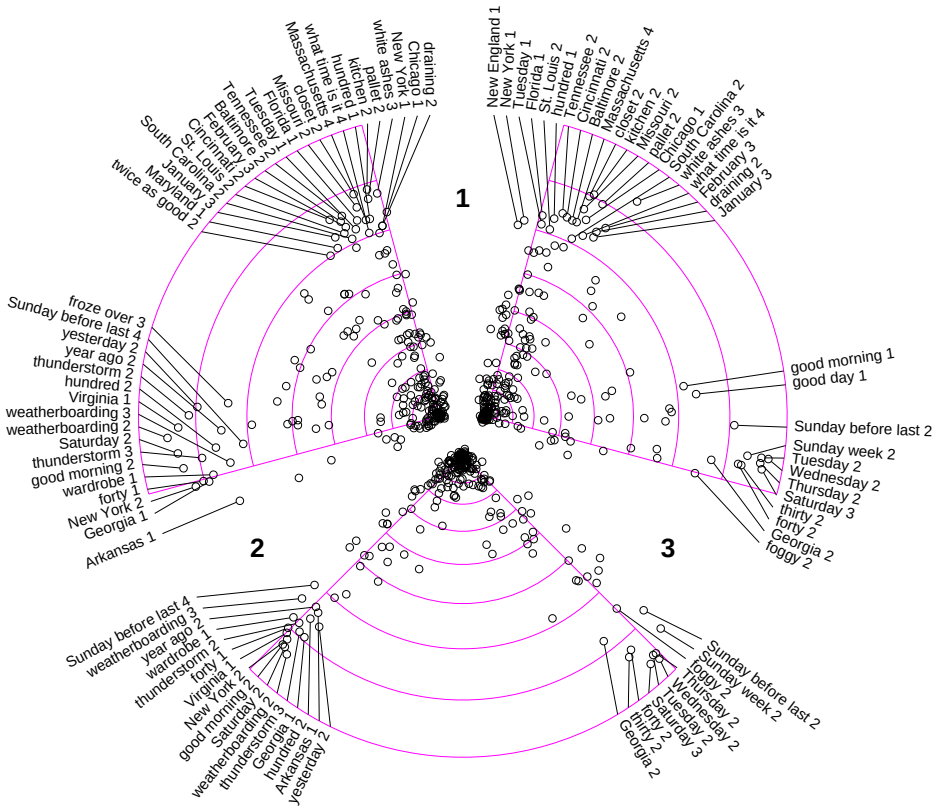


Figure 2: We graph the loadings of variables with respect to factors one and two (top left), one and three (top right), and two and three (bottom center). The labeled data points are variables which have high loadings on one factor, and low on the other.

one of the factors, and loadings close to zero otherwise.

We turn now to the identification of the linguistically dominant tendencies which factor analysis identifies. It is convenient to discuss each factor in turn.

6.1 Factor One

We first examine a selection of vowels with very high loadings with respect to the first factor.

closet2	0.884	kitchen2	0.880
pallet2	0.874	white_ashes3	0.869
Tennessee2	0.856	Cincinnati2	0.851
Baltimore2	0.844	Massachusetts4	0.830
Chicago1	0.816	draining2	0.812

It is clear that we are looking at the realization of reduced vowels. In examining the LAMSAS data directly, we find that we are looking at a distinction of [ə] vs. [ɨ] (the latter including [ɪ]). Kurath & McDavid (1961) likewise comment on the "clear regional pattern " between [ə] vs. [ɨ] (p.168, and Map 148), but we find it nonetheless surprising that an alternation among reduced vowels emerges as the strongest association in the data. Shackleton (2005) likewise identifies this (in his second PC) and shows that the dominant variant in the American south is likewise dominant in southwest England.

There are several further vowels with unusually high loadings with respect to the first factor. We summarize these in the table below, together with the interpretation of the alternation we find by examining the data directly:

Florida1	0.842	[ɔ] vs. [ɑ]	St._Louis2	0.821	[u] vs. [ʉ]
hog_pen1	0.585	[ɔ] vs. [ɑ]	Tuesday1	0.796	[u] vs. [ʉ]
			Missouri2	0.857	[ʉ^{ə}] vs. [ʉ^{ᵊ}]

In each case we have identified the major variants, in general first listing the variant spoken more frequently in the North. The fronting of [u] to [ʉ] is well-known (Kurath & McDavid 1961, Map 17), including the existence of the intermediate

form noted here in the pronunciation of *Missouri*. Kurath and McDavid likewise discuss the distribution of [ɔ] vs. [ɑ] (Maps 22-24), where they distinguish several further variants, including diphthongs. See Shackleton (2005) for the identification of some correlates in southwest England.

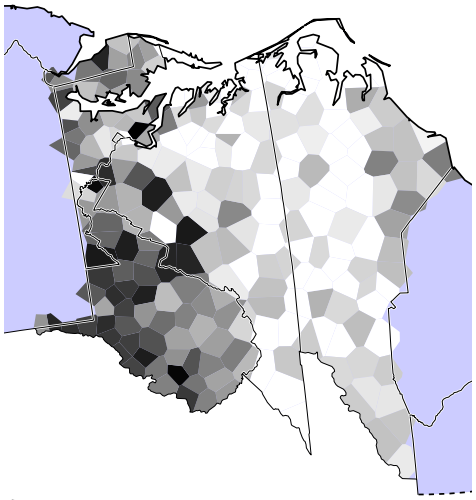
An intriguing possibility is that the treatment of reduced vowels might be somehow linked to the realization of [ɔ] vs. [ɑ], or to the degree of fronting in the [u]. We shall not consider this possibility further here. Figure 3 illustrates the geographic distribution of these contrasts, confirming that the variables are not only strongly associated, but also that they serve to distinguish the Northern and Southern parts of the area we are examining. This is confirmed by clustering the sites after using the first factor loadings as a weighting (not shown).

6.2 Factor Two

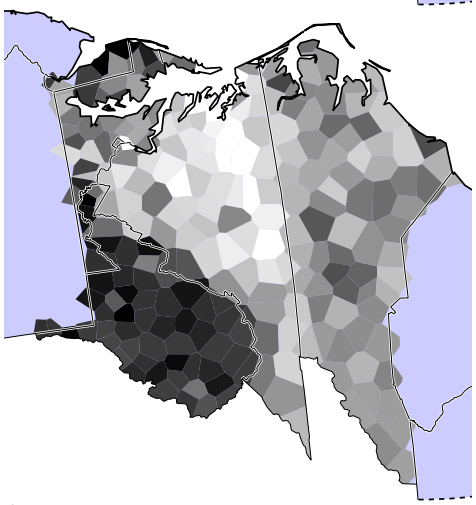
We turn now to the second factor, where we note many loadings such as the following:

weatherboarding2	0.936	Saturday2	0.926
Virginia1	0.905		

An inspection of further examples confirms that this factor is keying in on the fate of vowels before /r/, especially whether the /r/ is realized or not (including realization as r-coloring, as in [ɚ] vs. [ə]). The realization of vowels before /r/ is notoriously complex in the LAMSAS area, and Kurath and McDavid devote the entirety of their Chap. 4 to the description of the alternations, including twenty-five maps (34 through 58). It is remarkable here to see factor analysis not only



[ɯ] vs. [ʉ]



[ɔ] vs [ɑ]

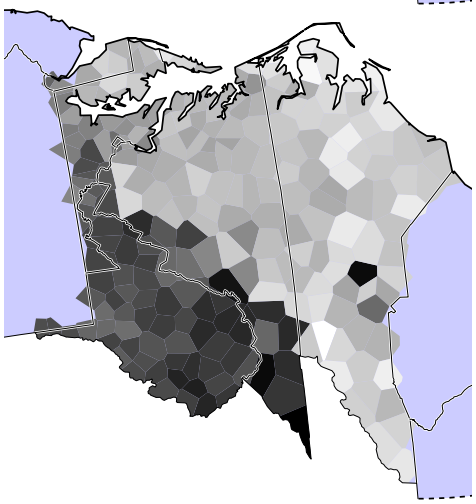


Figure 3: [ə] vs [ɨ]

single this phonetic environment out (by assigning various vowels before /r/ high loadings), but also by treating a number of the vowels as the same, including /o/ (Sunday_before_last4, weatherboarding3), /ɔ/ (New_York2, forty1), /ə,ɪ/ (Virginia1, Saturday2), and /ɑ/ (Arkansas1). (See Figure 2 for these loadings.)

In the same vein we note examples such as the following:

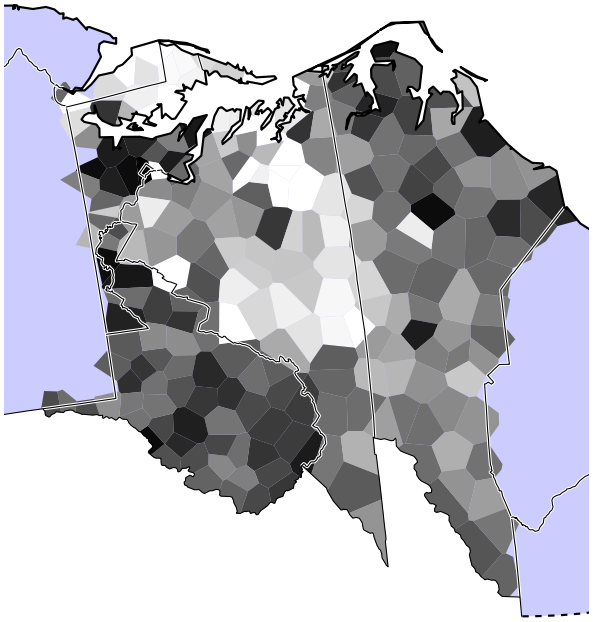
good_morning2	0.929	New_York2	0.922
forty1	0.906	thunderstorm3	0.893

In these examples we see a contrast between [ɔə] on the one hand vs. [ɔ˷ ə] on the other. This is a further indication that factor analysis is keying in on interesting, and in this case, fairly subtle alternations. It is perhaps also astounding that such a slight distinction is noted consistently enough in the data for it to play a role in factor analysis.

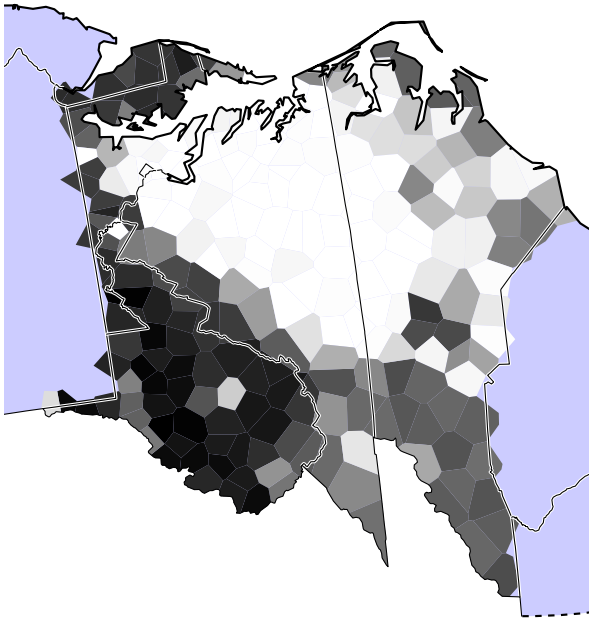
The alternations to which the second factor is sensitive distinguish the Piedmont area geographically, especially the absence of syllable final [r]. The question of a link between the two phonetic alternations is less frivolous here, since it seems quite plausible that [r]-lessness could promote the lowering of [ɔ]. Whether or not a linguistic explanation of the lowering is plausible, we note once again that Kurath and McDavid had observed the geographic distribution (Map 45).

6.3 Factor Three

The third factor concerns *inter alia* an alternation between a raised and unraised [ɨ] in the words such as the following. The LAMSAS database records one as [ɨ˄] (IPA [ɨ̝]) and the other as [ɨ], and Kurath and McDavid discuss the example



[ɔə] vs. [ɔ̣] (= LAMSAS [ɔ˷ ə])



[ər] vs. [ə] (etc.)

Figure 4:

in their Map 150 ('Missouri').

Wednesday2	0.967	Saturday3	0.961
thirty2	0.928	foggy2	0.854

Several further alternations are likewise loaded quite highly with respect to the third factor:

Georgia2	0.876	Tennessee1	0.766
sofa2	0.760	good_day1	0.775
Russia2	0.751	good_morning1	0.738

In the case of the alternations on the left, an inspection of the data reveals that these are pronounced in some places with a [ə] and at other places with [ɨ]. We note, however, that this is distinct from the treatment of reduced syllables we examined in connection with the first factor, which concerned weak syllables which were closed, i.e., which included a final consonant. The alternation here appears to be restricted to open weak syllables, essentially the one discussed by Kurath and McDavid in their Map 149 on the final vowel in *sofa*.

The first vowel in *Tennessee* is occasionally raised, as is well known from the literature; in fact, it raises well not only from [ɛ] to [ɛ̝] (LAMSAS: [ɛˆ]), but even beyond it to [ɪ] (Kurath & McDavid 1961, Map 9).

The fronted vs. non-fronted versions of [ʊ] (vs. [ʊ˖]) maybe found in Kurath and McDavid's Map 6, but the discussion there focuses on the issus of whether the [ʊ] in *wood* is diphthongized.

Figure 5 illustrates the geographic distribution of the vowel groups with high loadings for this third factor. The last, rightmost map does not suggest the sort

of north/south distinctions which the first two maps sketch, but we include it to provide a sense of the difficulty in applying these methods. We conjecture that it is less coherent geographically because we have extrapolated from too few examples in the data.

6.4 Discussion

The factors which emerged during analysis might be viewed from various perspectives. On the one hand, we are interested in identifying the linguistic base of the aggregate comparison, and from this perspective the experiment must be regarded as successful. We examined a number of variables (individual vowels) which share high loadings on the various factors, and in all cases we were able to show that these had geographically distinct distributions within the digitized LAMSAS data, and also that Kurath and McDavid had noted them in their authoritative commentary on the pronunciations in LAMSAS and LANE (*Linguistic Atlas of New England*). This suggests that we may use this techniques in areas for which no such authoritative text exists.

The technique is not fully integrated with the Levenshtein technique of measuring dialect pronunciation difference (Heeringa 2004), however, and, what is more, it relies on a separation of the data into comparable segments, not merely comparable words (as is the case with Levenshtein distance as applied to obtain aggregate analyses). A full integration of the factor analysis possibility into the software for aggregate analysis would be challenging and also time consuming.

From the perspective of seeing which variables behave similarly, the task to

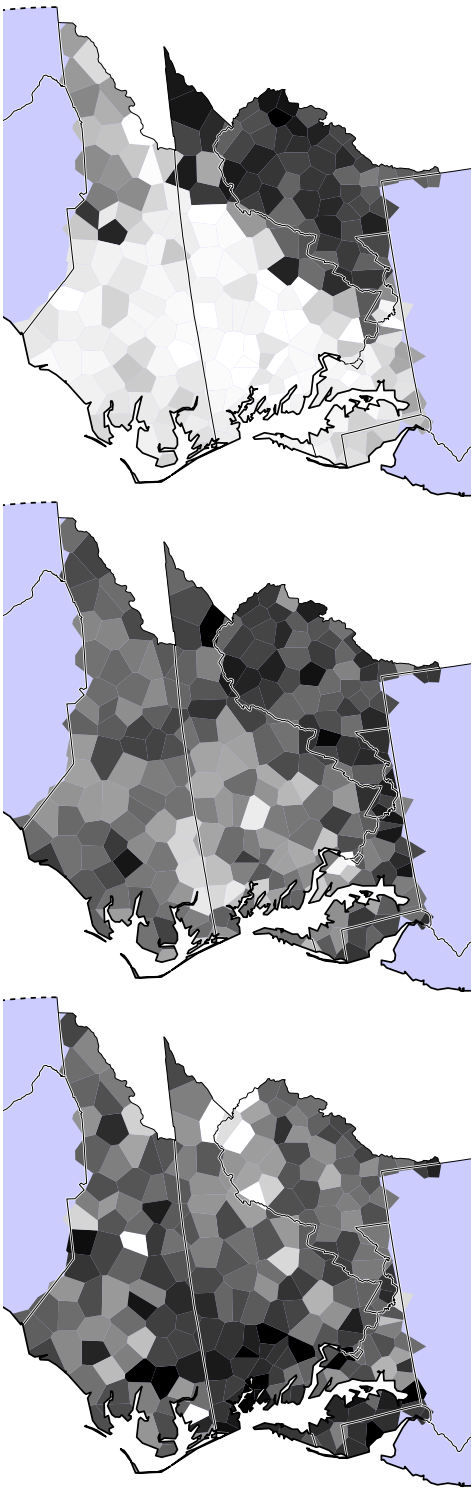


Figure 5: [i̞] vs. [i̟] [ɛ̞] vs. [ɛ̟] [u̞] vs. [u̟]

he_died_with1	April2	seven2	kitchen1	Chicago3
he_died_with3	France1	twelve1	January2	Louisiana3
New_England2	Missouri3	bureau1	St._Louis1	February1
Sunday_week3	attic2	ten1	second2	all_at_once1
half_past_seven1	backlog1	bottom2	froze_over1	Alabama2
what_time_is_it1	chimney1	driven1	dry_spell1	dry_spell2
New_Orleans2	fourteen2	broom1	froze_over2	Tennessee3
half_past_seven2	eleven2	mantel1	hog_pen2	Charleston2
Sunday_before_last5	my_wife2	night1	northeast2	northwest2
steady_drizzle1	quilt1	rose1	second1	a_little_ways2
twenty-seven1	seventy1	sofa1	tomorrow1	Washington3
twenty-seven2	three1	pallet1	January1	Baltimore1
twenty-seven3	thirteen2	twenty1	wardrobe2	bureau2
white_ashes2				

Table 2: Vowels which never received a factor loading about 0.4, i.e. dialectologically "noisy vowels."

which factor analysis is often put in social science applications, we saw, as we had hoped, that the same and similar phonemes tended to be grouped together, especially when they occurred in the same phonological environments. It would have been interesting to find examples where different phonemes were treated the same under factor analysis in a way which suggests a more uniform trend, but we certainly did not see more than a trace of this.

Finally it is worth noting, once again, that dialectological data is noisy. Sixty-six of the vowels received no high factor loading whatsoever, and these are shown in Table 2. Naturally, there may be excellent reasons why such data was included in the sample. In the case of polysyllabic words, the presence of other vowels may be an excellent reason, and in all cases the consonants might show interesting patterns. But in many cases we suspect that the patterning simply did not emerge as hypothesized.

7 Conclusions and Prospects

We set ourselves the task in this paper of adding to the set of techniques used in dialectometry, which has been successful in delineating global trends among dialects. We identified a need to interpret global tendencies adduced between entire varieties in terms of more detailed linguistic tendencies. Linguists' claims about dialect delineations may be overeager or even inexact about what characterizes a dialect area, but they are unquestionably superior in the degree to which they attempt generalization over the data, a property which makes them scientifically interesting.

This paper has been successful in comparison to earlier attempts to adduce the linguistic base of aggregated differences. Heeringa (2004) examined individual words which correlated highly with MDS results, which was instructive, but, because it relied on entire words, too coarse. Nerbonne (2006) simply aggregated over a subset of the data, which demonstrated that the subset indeed reliably correlated with the aggregate, but which left much too coarse a characterization. Shackleton (2005) used the very similar principal components approach to isolating linguistic factors, to which this paper adds the demonstration that factor analysis can be successful, as well as the application to a more standard dialectological problem, that of the analysis of a single area. Finally, we argued that factor analysis is statistically preferably to the very similar principal components analysis in that it restricts the search for factors to those capable of explaining the variance explained by the original variables.

7.1 Next steps

Nerbonne (2006) suggested two other techniques which might be applied to identify important linguistic factors in aggregate comparison.

Agrawal, Imielinski & Swami (1993), and others involved in "data-mining", have proposed that one examine all of the correlations between database elements, and among sets of these. Until Agrawal et al.'s work, there was concern that the number of combinations would make such an indiscriminant procedure infeasible, but Agrawal et al. have shown that this need not be the case. This sort of technique, like the one applied in this paper, might need to be combined with some restrictions on the data set, e.g., using phonetic segments rather than entire words.

A second promising tack is illustrated by Kondrak (2002), (Nakleh, Ringe & Warnow 2005) Gray & Atkinson (2003), who address the historical question, seeking (partially) automatic means of carrying out historical reconstruction in linguistics. An application of their techniques to data sets of dialectal data would seem to be straightforward, but dialectologists in the field record a level of phonetic detail which neither of these works is likely to have encountered thus far.

Acknowledgments

The Netherlands Organization for Scientific Research (NWO) subsidized the work reported on here, NWO grant 360-70-120 (P.I. J.Nerbonne). My thanks are

due especially to Peter Kleiweg for all of the programming on which this paper is based, including all the work on the maps. Wilbert Heeringa and two anonymous referees added very helpful remarks. Further thanks for very useful feedback to audiences in Moncton at the "Methods in Dialectology XII" conference (8/2005), in Marburg (11/2005) and to my courses at the LSA Linguistics Institute at M.I.T. (7/2005) and at Tübingen (11/2005). As always, shortcomings rest with the author.

References

Agrawal, R., T. Imielinski & A. Swami. (1993). Mining Association Rules between sets of Items in Large Databases. In *Proc. ACM SIGMOD International Conference on Management of Data*, (ed.) P. Buneman & S. Jajodia. New York: ACM pp. 207–216.

Chambers, J. & P. Trudgill. (1998, [1980]). *Dialectology*. Cambridge: Cambridge University Press.

Clopper, C. & J. Paolillo. (2006). "Northern American English Vowels: A Factor-Analysis Perspective." *Literary and Linguistic Computing* 21. special iss. *Progress in Dialectometry: Toward Explanation*, ed. by J.Nerbonne and W.Kretzschmar (this volume).

Goebl, H. (1982). *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Österreichische Akademie der Wissenschaften.

Goebl, H. (1984). *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. 3 Vol.* Tübingen: Max Niemeyer.

Gray, R. D. & Q. D. Atkinson. (2003). "Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin." *Nature* 426(27 Nov.):435–439.

Heeringa, W. (2004). Measuring Dialect Pronunciation Differences using Levenshtein Distance PhD thesis Rijksuniversiteit Groningen.

Heeringa, W. & A. Braun. (2003). "The Use of the Almeida-Braun System in the Measurement of Dutch Dialect Distances." *Computers and the Humanities* 37(3):257–271. Special Issue on Computational Techniques in Dialectometry, ed. by John Nerbonne and William Kretzschmar.

Heeringa, W. & J. Nerbonne. (2006). "De analyse van taalvariatie in het Nederlandse dialectgebied: methoden en resultaten op basis van lexicon en uitspraak." *Nederlandse Taalkunde* . Accepted.

Kondrak, G. (2002). Algorithms for Language Reconstruction PhD thesis University of Toronto.

Kretzschmar, W. A., (ed.). (1994). *Handbook of the Linguistic Atlas of the Middle and South Atlantic States.* Chicago: The University of Chicago Press.

Kurath, H. & G. Lowman. (1970). *The Dialectal Structure of Southern England: Phonological Evidence*. Number 54 *in* "Publication of the American Dialect Society" Tuscaloosa: University of Alabama.

Kurath, H. & R. McDavid. (1961). *The Pronunciation of English in the Atlantic States : Based upon the Collections of the Linguistic Atlas of the Eastern United States*. Ann Arbor: University of Michigan Press.

Labov, W. (1994). *Principles of linguistic change. Vol. 1: Internal factors*. Oxford: Blackwell.

Labov, W. (2001). *Principles of Linguistic Change. Vol.2: Social Factors*. Malden, Mass.: Blackwell.

Manni, F., W. Heeringa & J. Nerbonne. (2006). "Are Family Names just Words? Comparing Geographic Patterns of Surnames and Dialect Variation in the Netherlands." *Literary and Linguistic Computing* 21. Submitted, 10/2005.

Nakleh, L., D. Ringe & T. Warnow. (2005). "Perfect Phylogentic Networks: A New Metholodology for Reconstructing the Evolutionary History of Natural Languages." *Language* 81(2):382–420.

Nerbonne, J. (2006). Various Variation Aggregates in the LAMSAS South. In *Language Variety in the South III*, (ed.) C. Davis & M. Picone. Tuscaloosa: University of Alabama. accptd for pblctn.

Nerbonne, J. & W. Kretzschmar. (2003). "Introducing Computational Methods in Dialectometry." *Computers and the Humanities* 37(3):245–255. Special

Iss. on Computational Methods in Dialectometry ed. by John Nerbonne and William Kretzschmar, Jr.

Paolillo, J. C. (2002). *Analyzing Linguistic Variation: Statistical Models and Methods*. Stanford: CSLI.

Shackleton, Jr., R. G. (2005). "English-American Speech Relationships: A Quantitative Approach." *Journal of English Linguistics* 33:99–160.

Stevens, S. S. (1975). *Psychophysics: Introduction to its Perceptual, Neural and Social Prospects*. New York: John Wiley.

Tabachnik, B. G. & L. S. Fidell. ([1]1996, 2001). *Using Multivariate Statistics, 4th ed.* Boston: Allyn & Bacon.