Lexical Distance in LAMSAS

John Nerbonne and Peter Kleiweg
Humanities Computing, University of Groningen

Oct 17, 2002, revised May 1, 2003

Abstract. The *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) is admirably accessible for reanalysis (see `http://hyde.park.uga.edu/lamsas/`, Kretzschmar 1994). The present paper applies a lexical distance measure to assess the lexical relatedness of LAMSAS's sites, a popular focus of investigation in the past (Kurath, 1949; Carver, 1989; McDavid, 1994). Several conclusions are noteworthy: First, and least controversially, we note that LAMSAS is dialectometrically challenging at least due to the range of field workers and questionnaires employed. Second, on the issue of which areas ought to be recognized, we note that our investigations tend to support the existences of "Midlands" area, i.e., a three-way division into North/Midlands/South-Coastal rather than a two-way North/South division, i.e., they tend to support Kurath and McDavid rather than Carver, but this tendency is not conclusive. Third, we extend dialectometric technique in suggesting means of dealing with alternate forms and multiple responses.

Keywords: dialect, dialectology, dialectometry, American English

1. Introduction

Dialectal data is notoriously complex and resistant to word-by-word or sound-by-sound analysis. This led Seguy (1971) to explore techniques which allow one to aggregate individual differences over a large amount of material. Seguy effectively invented dialectometry in this step, which Goebl (1984) was later to elaborate much more systematically. Schneider (1988) is a brief illustration and evaluation of some of these techniques. These early treatments focused on lexical variation, i.e., the question of whether the words used for given concept varied geographically, but they also included phonological and other sorts of data treated at a categorical level.

Our own work has focused on analyses of pronunciation variation in Dutch (Nerbonne et al., 1996; Nerbonne et al., 1999; Heeringa et al., 2002). The present paper represents a shift to a focus on American English and lexical variation. There are two reasons for this shift. First, in the future we should like to explore the degree to which lexical and phonetic variation coincide, testing Kurath and McDavid's 1961 claim that they "coincide fairly well." To do this, we should prefer to build on materials which record both the pronunciation and the lexical identity of dialect material, and LAMSAS does this. We should emphasize

that the present paper focuses exclusively on lexical variation, and the relation between lexical and phonological variation is subject for future study. Second, we are interested in the degree to which the techniques which have been successfully applied to Dutch indeed generalize to other languages,[1] and this motivated our exploration of the American data.

The current paper introduces LAMSAS (§ 2), and in particular the care that was needed to find a substantial and comparable subset of the data. We then turn to an explication of the technique (§ 3) and two minor extensions we propose concerning the treatment of related lexical variants and the treatment of multiple responses. Results and discussion are presented in the final sections (§§ 4-5).

2. LAMSAS

The *Linguistic Atlas of the Middle and South Atlantic States* comprises dialect material collected on the Eastern seaboard of the United States from 1933 through 1974. The area examined extends from Northern Florida northward through New York state and includes all the intermediate states with an Atlantic coast, plus West Virginia. A map is included below as Figure 2. Our focus here will be on word geography—ultimately obtained using a questionnaire in which respondents were asked for the words they used for everyday things and events, e.g., in answer to questions such as *"If the sun comes out after a rain, you say the weather is doing what?"* (used to elicit *clearing up*, *fairing off* and forty other dialectal variants).

There are good reasons for focusing first on lexical variation. First, lexical variation has been at the heart of an interesting discussion on whether there is a linguistic coherent "Midlands" in the Eastern US (in contrast to a Northern area and a Southern Coastal area), as Kurath claimed (Kurath, 1949), or whether the predominant dialect division is not simply North-South, as Carver maintains (Carver, 1989). This question presupposes that it is sensible to enquire after DIALECT AREAS, i.e., geographically delimited areas in which one finds only gradual linguistic transitions (Bloomfield, 1933, p.51), i.e., an area in which a number of linguistic variables show the same language variation and in which this coherence contrasts with other choices in variation beyond the borders of the area. Second, we originally thought that lexical responses would be a more reliable foundation for measurements, since lexical data are transcribed in a canonical way, unlike phonetic data, where transcription bias can be serious. We return to this topic below (§ 2.1.1).

The LAMSAS material is admirably accessible for reanalysis (see `http://hyde.park.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. Unfortunately, it was not all usable, a subject to which we now turn.

2.1. Data Preparation

We restrict our analyses below to the interviews conducted by Guy Lowman using two different questionnaires (LAMSAS work sheets). In this section we justify this restriction.

2.1.1. *Fieldworker Bias*

Every researcher would naturally prefer to include all available data in analysis. Our early attempts to work with the entire LAMSAS data set were consistently frustrated in this respect, however. All comprehensive measurements reflected the fieldworker source of the data rather than the expected distribution of words (or pronunciation, which we shall report on independently).[2] A further reason to focus on lexical variation was that we suspected that lexical variation would at least not be susceptible to the potentially systematic transcription inconsistencies of the various fieldworkers. But lexical variation shows a great deal of fieldworker dependency as well, as we shall see in the present section.

Since we are employing techniques that we and others have used successfully, we interpreted the "fieldworker areas" (dialect areas adduced in analysis which correspond to the areas in which a particular fieldworker collected data) as a problem in the data, but we could also be charged with *petitio principii* at this point, certainly by those who are unconvinced of the probity of dialectometric methods. But we find a strong confirmation of our suspicions if we examine the variability in the average number of responses collected by the different fieldworkers. Table I shows that the LAMSAS fieldworkers indeed differed a great deal in their elicitation practices, so much so that we suggest that this is the basis for the "fieldworker isoglosses" in lexical variation, and Figure 1 illustrates these differences graphically.

Table I shows that McDavid and the other fieldworkers collected respectively 31.5% and 38.5% more responses per interview than Lowman (on average), and moreover, that they were less consistent than Lowman in the number of responses per interview which they collected. The latter is the cause of the much higher standard deviations in their number of responses per interview. If we expected the mean number of responses per interview and standard deviation in number of

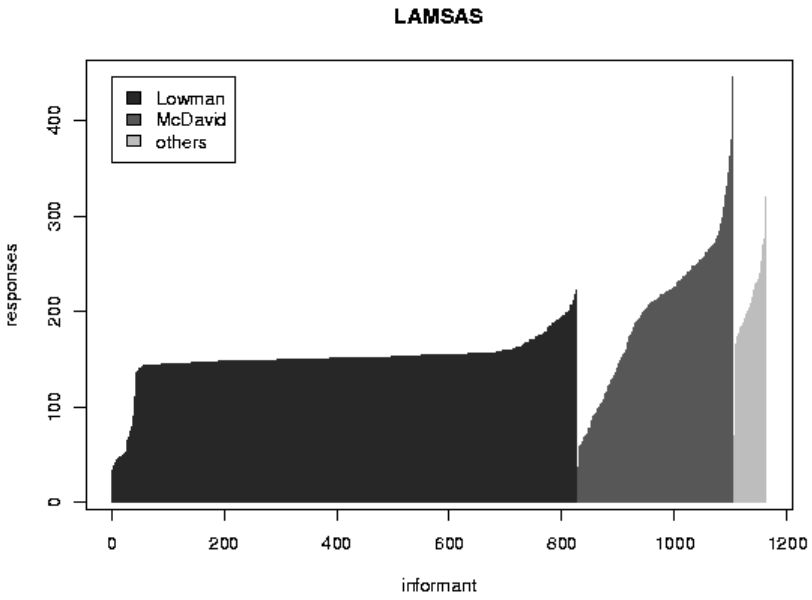

Figure 1. The number of responses per interview, sorted first by fieldworker and then in increasing order. It is clear that the LAMSAS field workers varied a great deal in their interview techniques. Lowman displayed for the most part remarkable consistency in the number of responses he elicited, which neither McDavid nor the other interviewers, taken together, attained. We return below to the slight deviations in Lowman's consistency.

Table I. Lowman conducted 71% of the LAMSAS interviews, and McDavid 24%, leaving just 5% for the remaining fieldworkers. Moreover, Lowman worked with an iron consistency, reflected in the much lower standard deviation of the number of responses per interview. Lowman's number of responses per interview differs very significantly from McDavid's $t(277) = 10, p \lll 0.001$, as does the standard deviation in number of responses per interview $F(200, 1000) = 9.4, p \lll 0.001$.

Fieldworker	**Number of Interviews**	**Number of Responses**	**Mean Responses/ Interview**	**SD Responses/ Interview**
Lowman	826	123990	150.1	25.3
McDavid	278	54855	197.3	76.8
others	58	12057	207.9	43.9
Totals	1162	190902	164.3	49.6

responses per interview to be roughly the same in fieldworkers implementing the same design, then the figures in Table I would demonstrate that Lowman and McDavid did not implement the same design (see caption).

It might be argued that the more variable response record is inherent in the LAMSAS design, which emphasized the indirect elicitation of responses, but the fact remains that different fieldworkers implemented this design in different ways. It is particularly the more variable number of responses per interview which probably confounds measurements. The variable number of responses means that the individual questionnaires do not represent the variety in the same way.

This is perhaps most easily appreciated if one imagines what would happen if the same interview were conducted three different times by people speaking the same local variety (i.e., at one site). If two interviews are conducted by the "more encouraging" interviewer, the differences in elicited vocabulary should reflect only the "noise" in the procedure. But the results of both of these longer interviews will show further and systematic differences when compared to the shorter word list which results from the interview conducted by the "less encouraging" interviewer. In particular, when the briefer interview is compared to either of the longer ones there will be fewer points of difference for our procedures to note. The chance of overlap is always greater if more responses are collected.

This guess about the differences between McDavid's and Lowman's style is not borne out simply by the records, however. In particular, LAMSAS questionnaires distinguish between NR "no response" and NA "not asked", but tracking this distinction shows that McDavid was not in every respect encouraging. Given his higher number of

responses in total, we might have expected that McDavid would record the lowest numbers of NA's and NR's, but this is not the case. While all fieldworkers failed to ask after 1% of the data (NA), McDavid failed to elicit responses 15% of the time, while Lowman (and others) obtained responses all but 10% of the time. This uneven distribution of 'no response' further strengthens our view that fieldworker techniques confound the data to some extent even if it suggests that the difference was more complex than simply "encouraging" vs. "discouraging." McDavid was apparently less encouraging in the face of no immediate response, but more encouraging about multiple answers although not consistently.[3] We have also examined the data in the LAMSAS files to see if we could determine the order of responses to a given item in a interview, reasoning that we might try analyses in which only the first or perhaps first two or three responses are used. Unfortunately, this information appears not to have been recorded. We also attempted restricting analyses to a small number of responses, in particular the most popular two or three responses, but the results were not credible.

To conclude this section, we note that, although we emphasize that the variability in fieldworkers' methods confounds our dialectometric techniques, it is likewise a problem which has the potential to vitiate other, more traditional techniques, as well. See, however, Speelman, Geeraerts and Grondelaers (2003) for techniques which complement questionnaire methodology. We should like to add that we continue to attempt various corrections to try to obtain measurements which make sense from one fieldworker to the next.

2.1.2. *Questionnaires*

Ideally, all the material from the LAMSAS questionnaires ("sets of work sheets") would be analyzed in an effort to understand the dialects of the area. As Kretzschmar (1994, pp.2, 58) notes, however, LAMSAS field workers did not consistently elicit responses from the same questionnaire ("set of work sheets"). Questionnaires were occasionally adapted to be better attuned to the variation in a given region. The LAMSAS handbook notes dozens of responses which were only found in items from a questionnaire which was used in a geographically restricted area (Kretzschmar, 1994, pp.92–102), and the LAMSAS web site explains that five different questionnaires were used (see `http://hyde.park.uga.edu/fields.html#ws`). See Table II for a summary of the frequency with which the different questionnaires were used.[4]

For the purpose of this dialectometric study, we need comparable data, e.g., data elicited using questions for which alternative answers were given. We shall ultimately analyze the alternations. Incorporating

Table II. The LAMSAS data was elicited on the basis of five different "sets of work sheets" (questionnaires). Care is needed to obtain an intersecting set of vocabulary items. The analysis in this study ignored the "combined" worksheet, which Lowman never used, and the "New England" sheet, which he used only in pilot studies.

Number of LAMSAS Interviews per Work Sheet

Work Sheets	Number
South Atlantic	565
Middle Atlantic	465
Preliminary South Atlantic	70
Combined	57
New England	5

the answers to questions which were asked in only one geographic area of the survey, on the other hand, would tend to distort the distance measures applied here, in particular, leading to exaggerated estimations of distance in those areas in which the additional information is sought. We return to this in § 3, where we introduce the measurements. So we looked for material that was commonly elicited throughout the LAMSAS area.

Since we have already decided to focus on Lowman's work, the "combined" questionnaire, which he never used, is of no interest. The New England questionnaire, which Lowman used four times in 1933 on Long Island, New York, and which another fieldworker, Bernard Bloch, used once in South Carolina, may also be eliminated because it was used only in areas geographically remote from Lowman's area of focus in LAMSAS, and only sporadically. The preliminary South Atlantic worksheet was used 70 times in 1933-34, but, as Table III shows, with a much larger mean number of responses and a much higher standard deviation in that number. Lowman and Kurath were developing the questionnaire in these years. As Table III shows, all of these earlier interviews (whether conducted with the New England or with the preliminary South Atlantic sets worksheets), are afflicted with rather higher standard deviations in the number of responses per interview, at least in contrast to Lowman's otherwise sterling reliability. This happened because these interviews were conducted as pilot studies, as the name "preliminary" suggests. Following the conclusion of the earlier section on the confusion caused by the large variability in responses/interview among the fieldworkers, we decided likewise to focus on the more reliable data here, as well.

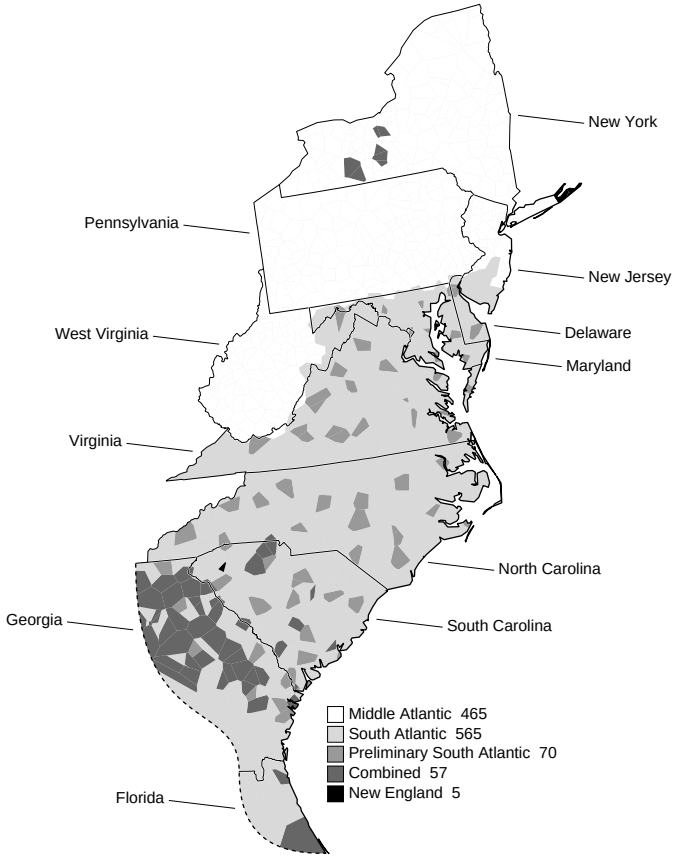

Figure 2. The LAMSAS area as it was probed by the different questionnaires. In examining results we exercise caution that our division into dialect areas is not influenced by the questionnaire used. In focusing on the 71% of data collected by Lowman, we shall ignore the data in South Carolina, Georgia, Florida and part of New York.

It turned out that the South and Middle Atlantic work sheets also did not elicit precisely the same concepts. As a way of ensuring that data be commensurable, we sought words which appeared on both of these lists. We operationalized this by using only words for which answers appeared for at least 100 interviews. This resulted in us ignoring the item *loam*.

2.2. Data Used

Of the 151 words checked in the LAMSAS area, 32 show no lexical variation at all (city names and the like), and another 42 show only variation in the form of singleton responses. The worksheets showing

Table III. Lowman first used the New England and the Preliminary South Atlantic work sheets in 1933-1934 before settling onto the other two. As the LAMSAS handbook notes (p. 58), the earlier phase, involving only 74 interviews, was experimental. The figures for mean number of responses per interview and standard deviation indicate that Lowman's methodology also varied before he went on to the 752 interviews with the South and Middle Atlantic work sheets.

Lowman's Interviews per Questionnaire

Work Sheets	Number of Interviews	Number of Responses	Mean Responses/ Interview	SD Responses/ Interview
New England	4	429	107.3	48.7
Preliminary	70	12039	172.0	43.6
South Atlantic	370	54956	148.5	22.5
Middle Atlantic	382	56566	148.1	20.6
Combined	0	0		
Totals	826	123990	150.1	25.3

no variation are obviously of no value in assessing variation, and it is a common remark in linguistics and statistical studies in general that extremely infrequent data is likely to confuse analyses (Manning and Schütze, 1999, Ch. 6). Carver (1989) confirms this for the study of dialectal variation in lexis (p.17). To counteract the confusing effect of infrequent data, we finally eliminated all responses that occurred fewer than 13 times (in only $\leq 1.5\%$ of the interviews). Once we had eliminated these infrequent responses, another 13 words showed no remaining variation, and these, too, were eliminated. Ultimately only $64 (= 151 - 32 - 42 - 13)$ response items (files) were found which served as the basis of subsequent analysis. Table IV contains a list of all the words used in the analysis.

We should remark that our elimination of words which show no variation leads to an exaggerated estimation of the lexical distance between any pair of sites across the board (at least in the degree to which lexical distance is likely in a given lexicalization). But we will only use the relative distances between sites in further analysis, so that this across-the-board increase will have no more effect than any other linear transformation on the total lexical distance. It is also worth keeping in mind that several items were included only to probe pronunciation variation, making it unsurprising that they do not prove useful here.

Table IV. The 64 concepts common to the South and Middle Atlantic work sheets which served as the basis for the analyses in this paper. These concepts elicited responses in at least 100 interviews both in the South Atlantic and in the Middle Atlantic work sheets.

New England	Sunday before last	Sunday week	Washington
a little ways	afternoon	all at once	andirons
attic	backlog	blew hard	bottom
broom	bureau	calming down	chimney
clearing up	closet	cloudburst	clouding up
dragonfly	draining	driven	dry spell
feet	first	from the south	frost
froze over	gully	half past seven	he died with
hog pen	hundred	kitchen	lightwood
mantel	marsh	miles	my wife
nice day	night	northwest	pallet
parlor	porch	quarter to eleven	quilt
rising	rose	shades	sofa
soot	southeast wind	southwest wind	stairs
sundown	sun up	three years old	thunderstorm
wardrobe	weatherboarding	what time is it	white ashes

3. Lexical Distance

We investigate refinements of a technique for uncovering common linguistic variation in a complex database of dialectal material. The basic idea is due to Seguy (1971), and is very simple: we record the responses to questions eliciting common vocabulary for a range of dialect sites. We then compare each pair of sites, recording how many answers are the same and how many are different. For this purpose we ignore questions for which there is no answer at one or both of the sites, treating LAMSAS's categories of 'not asked' and 'no response' both as missing data (see below). The proportion of answers that is the same might be referred to as the LEXICAL PROXIMITY of the sites and the proportion of answers that is different is the LEXICAL DISTANCE. For example, given the data in the table below, we should conclude that there's a lexical distance of 0.25 between Brownsville and Whiteplain since 75% of their responses was the same for the fields for which responses are available, and 25% were different.

Site	**Vocabulary Item**				
	dog	*hat*	*horse*	*toilet*	*smallest finger*
Brownsville	*dog*	*hat*	*horse*	*bathroom*	*pinkie*
White Plain	*dog*	*cap*	*horse*	*bathroom*	—

Naturally, it would be conceivable to treat missing responses differently, for example, to regard the differing responses to the question about the *smallest finger* above as contributing to lexical difference (in the current calculation, it does not). In fact, if it were certain that there were no appropriate natural response to the question in the variety being sampled, then this certainly should contribute to lexical distance. The decision not to regard such data as a reliable indication of lexical distance is motivated by several considerations: first, we carried out the analysis treating 'no response' as a category of answer with the same status as lexicalizations, and we were dissatisfied with the results. Second, we suspect that the fact that a response is missing often does *not* indicate that none is possible, but only that it did not occur to the informant promptly. Given the range of responses we find listed in LAMSAS, it seems unlikely that nonresponse may be taken as certain evidence of a lexical gap. Third, as we noted above, different fieldworkers experienced significantly different levels of 'no response,' suggesting that 'no response' is affected by fieldworker practice. This is irrelevant in the current investigation since we are concentrating on just Lowman's reports, but we think that this conservative approach to what counts as evidence of lexical difference should be followed generally.[5]

We differ from Seguy in one minor point, and we extend his method in two ways. The minor point is that, while Seguy used the absolute quantity of differing vocabulary, we normalize this over the number of comparable questionnaire items, i.e., those for which we have responses. Seguy would measure the Brownsville/White Plain difference above as 1 (or 2), while we normalize this over the number of potentially differing vocabulary items. In a survey with 100% response, our measure is a linear transformation of Seguy's and would not create differences in further analysis. If there are large differences in number of responses, our measure systematically ignores the "no response" items, which we have argued for above.

3.1. Related Lexical Items

Often the different responses elicited from informants are different forms of the same lexical item. The responses to the question "If the sun comes out after a rain, you say the weather is doing what?" resulted not only in the responses *clearing up*, *fairing off* and *breaking away*,

but also, e.g., *fair off*, *fairs off*, and *faired off*, and it seems preferable to recognize these as much more closely related to *fairing off* than to *clearing up.*

Our solution to this problem was to apply the string distance measure, Levenshtein distance, and to use this as a measure of the lexical distance of the answer (Kruskal, 1999). We have applied this extensively to measure differences in dialectal pronunciation (Nerbonne et al., 1996; Nerbonne and Heeringa, 1998; Nerbonne et al., 1999), where it has proven valid and reliable.

Naturally, this is only a rough estimate of what more correctly lemmatizing ought to do if we restrict our attention to lexical differences. That is, we ought to recover the lexeme (or lemma) from the inflected form and then count two forms as equivalent if, and only if, they are alternate forms of the same lexeme, such as *clears* and *clearing.* Our procedure will count *bore* and *born* as just as distant as *bore* and *bare.* Given the arbitrariness of the form of words (de Saussure), accidentally close variants are rare, however.

3.2. Multiple Responses

Many questions elicit multiple responses, indicating that the informant would recognize all the responses as dialectally appropriate. Multiple responses are even more common if we aggregate responses from all individual informants in a given community, and this is a natural step to take if one wishes to depress the effect of individual variation. We wish therefore to "lift" the notion of distance from a notion between strings to a related notion of distance between sets of strings where the sets represent alternative lexicalizations.

The basic idea is that we average the distances between the individual strings where we consistently choose pairs in a way that minimizes the distance measure. Consider two response sets A, B where $A = \{a_1, a_2, a_3\}$ and $B = \{b_1, b_2\}$. To calculate $d(A, B)$ we find, for each a_i in A the closest b_j in B, i.e., the b_j such that $\forall b_{j' \neq j} d(a_i, b_j) \leq d(a_i, b_{j'})$, and similarly, for each b_j in B the a_i in A such that $\forall a_{i' \neq i} d(a_i, b_j) \leq d(a_{i'}, b_j)$. We then take an average of the *set* of these minimal distances. We emphasize that we are dealing with a *set* of pairs because we wish to exclude the possibility that we would count a given distance twice. So if $d(a_2, b_3)$ is minimal with respect to alternatives for a_2 and b_3 in both A and B, it won't be counted twice.

To view this slightly differently, consider that we are interested in the cross-product of the strings in the response set, i.e., the pairs of lexical items formed when the first element comes from A and the second from B. $A \times B = \{\langle a_1, b_1\rangle, \langle a_1, b_2\rangle, \langle a_2, b_1\rangle, \langle a_2, b_2\rangle, \langle a_3, b_1\rangle, \langle a_3, b_1\rangle\}$.

First we define a natural extension of the distance function on strings to function on an arbitrary set of ordered pairs of strings, i.e. the sum of the distances between the elements of the pairs.

$$d(C) \doteq \sum_{c \in C} d(c), \qquad \text{where } C \text{ is a set of string pairs}$$

It will also be convenient to refer to the first and second projections of C, i.e., $C^1 = \{a_i | \langle a_i, b_j \rangle \in C\}$ and $C^2 = \{b_j | \langle a_i, b_j \rangle \in C\}$. So C^1 contains all the possible first elements of the relation, and C^2 all the possible second elements. We say that C COVERS $A \times B$ if, and only if $C \subseteq A \times B$, and $C^1 = A$ and $C^2 = B$. We shall seek the minimum cost COVER, and we weight this as explained earlier.

$$d(A, B) \doteq \frac{1}{|C|} \text{ Min } d(C), \qquad \text{where } C \text{ covers } A \times B$$

Let's illustrate this with an example: given $A = \{a, b, c\}, B = \{a, c, d\}$ then $C = \{\langle a, a \rangle, \langle b, d \rangle, \langle c, c \rangle\}$ covers $A \times B$, even though $|C| = 3$, while $|A \times B| = 9$. Since $d(a, a) = d(c, c) = 0$, $d(A, B) = 1/3 \cdot d(b, d) = d(b, d)/3$. We have not shown that this is the minimal cost cover, but it is.

This is the derived notion of distance between lexical dialectal alternatives which we have employed in the results reported on below.[6]

4. Results

We analyzed the 745 interviews in 357 sites conducted by Lowman between 1935 and his death in 1941 (omitting the geographically non-contiguous sites) using the notion of lexical distance defined above (in § 3), including the refinements discussed for near variants and for sets of responses. For sites at which more than one interview was conducted, we averaged individual distances to obtain site distances. We restricted our attention to those words which occurred in both of the questionnaires Lowman used, eliminating infrequently occurring words and all those words for which no lexical variation remained (87). As we noted above, this left us with 64 words on which to base the analysis.

We calculated distances for all of the more than 6×10^4 pairs ($\approx (357 \cdot 356)/2$) of sites, and we then clustered the data, using Ward's method (Jain and Dubes, 1988), which has the effect of minimizing the error introduced by the agglomerative step in clustering. See Heeringa (in prep.) for a detailed presentation of how to analyze distance matrices in dialectology, including especially clustering and multi-dimensional scaling (see below). It is important to keep in mind that hierarchical

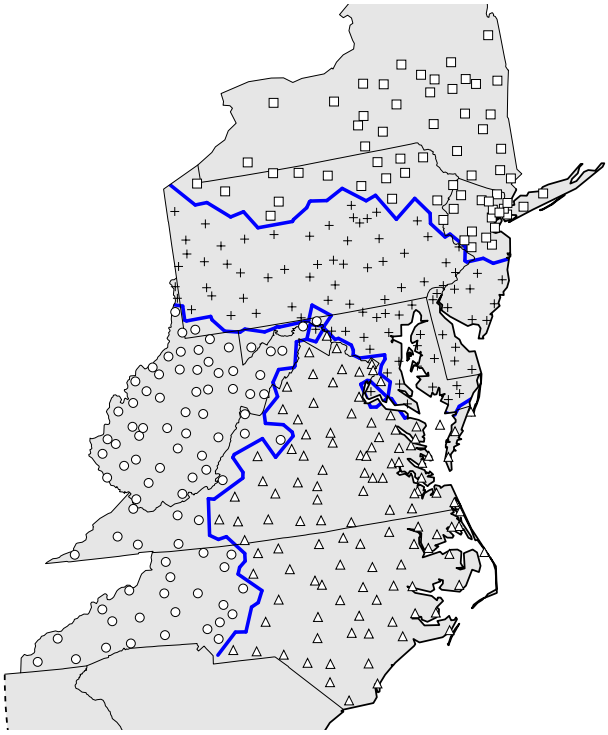

Figure 3. The four most significant dialect areas examined by Lowman, 1935-1941. If one compares the symbols in this map with those in the dendrogram, Fig. 4, it becomes apparent (i) that there is an area encompassing Kurath's "midlands", i.e., the inland South and the southern North; and (ii) that Kurath's "Route 40" boundary in northern Pennsylvania is strong—in fact the strongest division.

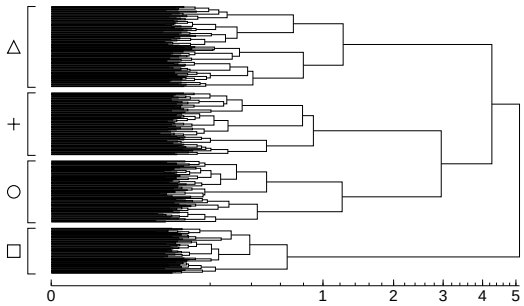

Figure 4. The dendrogram from which the map in Fig. 3 was derived. The symbols correspond with those used here. Note that all four areas are distinguished well.

Table V. The twenty responses which contributed the most to the division postulated (of the 453 elicited in the area covered). The areas refer to the map in Figure 3. See the LAMSAS web site for the exact wording of the questions in which the concept was elicited (`http://hyde.park.uga.edu/lamsas/`). Note that virtually no responses characterized any area perfectly. There are instead strong tendencies whose cumulative effect must be measured statistically.

Concept	Response	North	Midland	South Inland	South Coastal
dragonfly	darning needle	90%	8%	1%	0%
porch	stoop	88%	4%	0%	2%
frost	dew	77%	3%	40%	0%
quilt	comfort	4%	62%	83%	77%
night	evening	69%	65%	15%	8%
a little ways	a little piece	5%	69%	61%	22%
northwest	northern	0%	0%	27%	62%
pallet	pallet	0%	8%	78%	60%
afternoon	evening	28%	28%	73%	82%
Sunday week	Sunday week	6%	30%	49%	69%
lightwood	lightwood	0%	7%	1%	56%
quilt	comfortable	52%	4%	0%	0%
stairs	stairsteps	8%	30%	36%	69%
dragonfly	snake feeder	18%	43%	58%	4%
weatherboarding	weatherboarding	3%	47%	53%	51%
northwest	northwest	43%	44%	64%	93%
weatherboarding	clapboards	51%	9%	2%	2%
quarter to eleven	quarter till eleven	34%	24%	4%	47%
nice day	pretty	0%	10%	15%	52%
shades	shades	64%	24%	20%	53%

clustering by itself provides no answer to the question as to *how many* dialect areas are interesting. We can often observe large distances from one level of clustering to the next, and this in general indicates that the lower levels are quite distinct. Even in these cases the clustering technique by itself does not guarantee that the clusters chosen are much better than other alternatives. A map depicting the results is shown in Fig. 3 and the dendrogram reflecting the clustering is shown in Fig. 4.

4.1. The Differences

When we examine which responses are given in the areas we postulate, the nature of lexical variation is made clearer. To see which responses of which questionnaire items were responsible for the areas

we postulate, we collected the percentage answers of a given response per area. We then computed the standard deviation of the response percentages across the areas and sorted the results. Large standard deviations indicate words whose percentage occurrence differs a great deal in the different areas. Table V shows the twenty responses that contributed most to the borders we identified. Incidentally, we contrast this usefulness of this step with the criticism by Schneider (1988) that dialectometry fails to illuminate the link between concrete linguistic form and geography, fixed as it is, on indices of similarity.[7]

Table V is an excellent view into the nature of lexical variation. Strict association, i.e., that in which a given form is found in 0% in one area vs. 100% in another, does indeed occur, but it is infrequent. In Table V the use of *a little piece* to refer to a short distance (and also the lexicalization *snake feeder* for 'dragonfly') is restricted to the the southern North and the inland South — in accordance with the "midlands" view. In addition, the existence of dialect areas is completely compatible with there being individual words whose distribution counterindicates the dominant division. So the response *evening* for the concept 'night' characterizes the two northern areas together, and the word *pallet* is found almost only in the two southern areas, in spite of the fact that North-South is not the dominant division.

4.2. Kurath or Carver?

As Figure 5 shows, Kurath (1949) claimed that a Midlands area extends from central Pennsylvania south into West Virginia and the western parts of Virginia and North Carolina. Carver rejects this in favor of a simpler North-South divide running along the southern border of Pennsylvania. The issue is still the subject of ongoing research (Labov, 1991; Wolfram and Schilling-Estes, 1998).

When we compare our results to those of authorities on the classification of Eastern American dialects, it is important to keep in mind that we have used exactly the data available to Kurath. We do not have Carver's data, and so it would be expected that we should agree with Kurath's findings. Indeed we do agree with Kurath in all essential details about the major dialectal breaks in the Eastern United States.

Fig. 4 shows that we cluster the more southern Northern area together with the inland South: thus our reanalysis of Kurath's data contradicts Carver's central point that the North-South divide really is the most significant one. An important qualification concerns the stability of the division. The clustering technique used to produce the map in Fig. 3 is not stable: i.e., results may change greatly on the basis of a small change in input data. In order to avoid reporting an instable

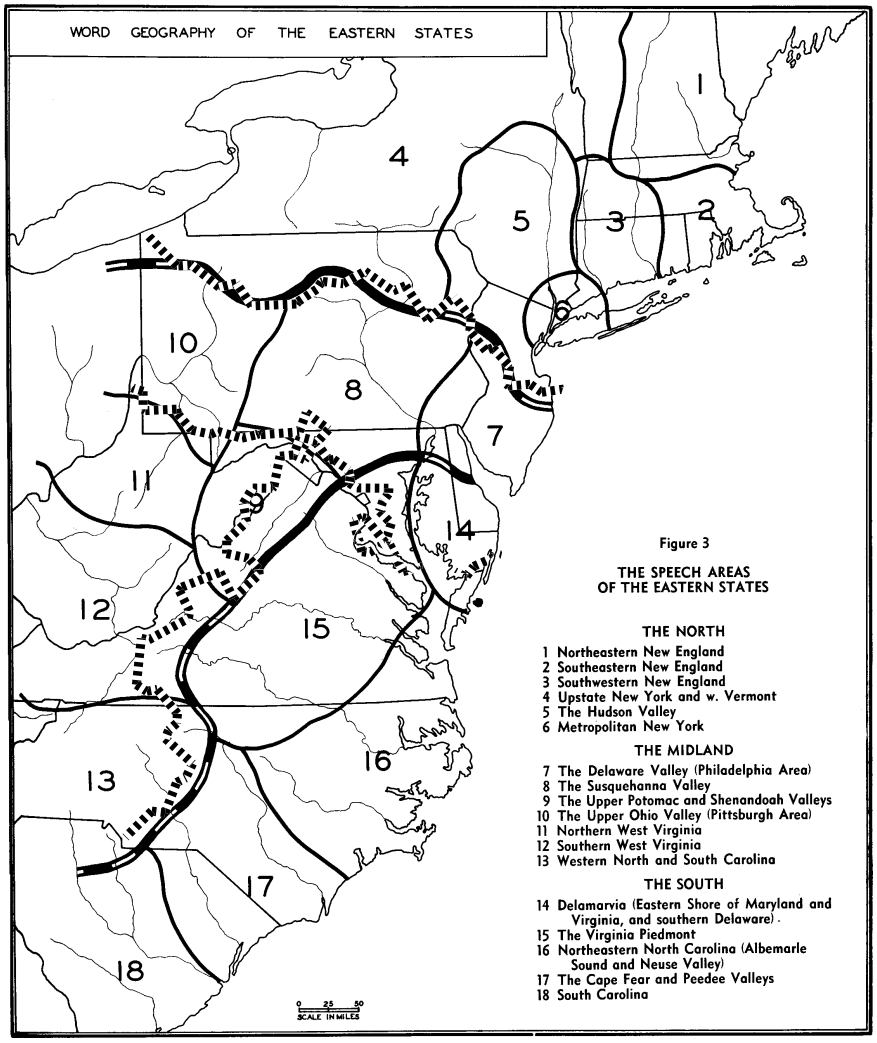

Figure 5. Kurath's dialect division in *Word Geography*, 1949, based primarily on Lowman's data. Most controversial was Kurath's postulation of the "Midlands", the area beginning with Pennsylvania in the north, West Virginia, and continuing south, but away from the coast. Our four-way division is superimposed in the broken line. The agreement with Kurath is striking, where we also see a significant North-South division, much like Carver's

result we compared clustering at approximately 40 different parameter settings, varying the number of tokens required for a word to figure in the distance measure from 1–20 and comparing the results based on the identity of strings with those based on the string distance between them. The map shown in Fig. 3 represents the result found at optimal parameter settings, but other analyses—in which notably no Midlands area emerges—are also found.

We note that we differ from both Carver and Kurath in grouping the larger part of Maryland and Delaware with the North rather than the coastal South, but this is of lesser significance. When we examine the second level of division, then we again side with Kurath in seeing something of a non-coastal Southern region (but restricted to the South), an area which Carver does not recognize (Carver, 1989, p.101), and we confirm Kurath's postulation of a major division within the North cutting East-West through Pennsylvania and New Jersey, a boundary which Carver accepts only partially (Carver, 1989, p.56).

4.3. A Finer Representation

The result of comparing each pair of elicitation sites results in a distance matrix of 357×357, which is, however, symmetric. We can imagine this as a 357-dimension space, in which each site is identified with respect to its distance to each other site. Multidimensional scaling (MDS) is a statistical technique that attempts to represent the distances in a distance matrix as economically as possible, i.e., in as few dimensions as possible. Figure 6 shows the result of applying the INDSCAL variant of MDS from R to obtain a representation in three dimensions, account for over 90% of the variance in the distances.[8] The results are visualized by coloring each point red, green, and blue in proportion to its first, second, and third MDS coordinate, respectively. We value the MDS presentation for eschewing the question of dialect areas, at least those with exact borders, but we note the concentration of red in the north, blue in the coastal south, and green in Kurath's "Midlands" section. The fact that Kurath's "midlands" area does emerge visually confirms the clustering analysis in Section 4, and suggests that the inherent instability of clustering is not a problem. But we also note that the north-south division (Fig. 3) is only marginally less successful than the Kurath division into north, "midlands" and coastal south, and this is reflected in the blue tone in the map in Fig. 6.

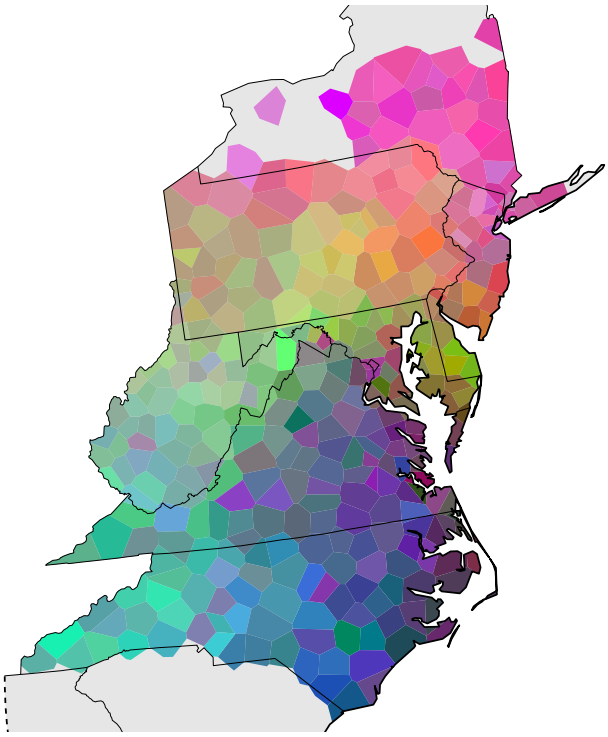

Figure 6. If we extract the most important dimensions of variation using multidimensional scaling and color the most important three dimensions red, green and blue, we obtain this map, which perhaps embodies the view that the dialects are organized on a continuum.

5. Discussion

The present paper has attempted to contribute to the understanding of how lexical variation contributes to the system of varieties. An underlying assumption has been that statistical analysis is essential if we, as dialectologists, are to avoid arbitrary selection of data and features on which to base classifications. By and large the current analyses confirm earlier non-computational analyses, but they allow us to be more specific about the bases of claims about dialect areas and natural groupings. It would undoubtedly be interesting to apply these techniques to a more homogeneous data set, a more recent one, or a data set from a larger area. The present paper has also contributed by noting that the varied techniques of the fieldworkers presents a serious problem to attempts to analyze their data together.

Incidentally we have introduced techniques to allow more sensitive measure of lexically related variants and also multiple responses. In

addition, the present study likewise sets the stage for a more detailed examination of the claim that lexical and phonological data "coincide fairly well" (Kurath and McDavid, 1961).

Finally, we note that we have made a number of potentially controversial decisions — for example, at what point to discard questionnaire items because of the suspicion that they may not have been used throughout an area, or exactly how many infrequent words to omit from analysis. We likewise introduced modifications to the basic distance measure for lexically related items and for multiple responses, without noting the effect these had on measurements. In fact we have been guided in this by a measure of the "local coherence" of the data set under a particular dialectometric setting of parameters. The size limitations of the present paper make it impossible to address this topic here, but we intend to return to it.

Acknowledgments

The Dutch Organization for Scientific Research (NWO) funded the development of the software use in the measurements here through grant 1999/11483/GW. Prof. Bill Kretzschmar of the University of Georgia has made the LAMSAS data available, and Kretzschmar, Wilbert Heeringa of Groningen, Prof. Jack Chambers of Toronto, and several anonymous CHUM referees commented usefully on one or another aspect of the material.

Notes

[1] Bolognesi and Heeringa (2002) have also applied the techniques to Sardinian, and Gooskens and Heeringa (2003) to Norwegian.

[2] Both Prof. Chambers and Prof. Kretzschmar noted that LAMSAS aficionados have long spoken of "McDavid" isoglosses.

[3] Let us take care to note that it is impossible to *prove* that fieldworkers were the source of these effects since they were in no sense assigned randomly to areas, elicitation sites or respondents. We have no reason to suspect other causes, however, so that we do suppose that fieldworkers differed substantially in the records they produced. This topic could be followed somewhat further in the LAMSAS data, but we shall not pursue it here.

[4] In fact we have also explored the question of whether the different questionnaires used confound the analyses, but it would go beyond the scope of the present paper to explore this in detail.

[5] A further reason, which plays a role in perhaps only one item is the following: we wish to guard against projecting nonlinguistic factors onto the interpretation of results. This happens in LAMSAS when informants are asked to name a resort in

North Carolina. Since almost only informants in Maryland and further south could answer this question, and since virtually everyone who answered it named *Asheville*, this is a very clear isogloss in LAMSAS—but arguably one which says little about language differences and more about the distribution of geographic knowledge.

[6] We are indebted to Wilbert Heeringa for substantial contribution to the discussion on this point.

[7] "[...] a quantitative procedure implies that the areal division is based solely upon the fact that a certain number of forms [...] is found to be different when the localities are compared—without any attention being paid to which forms these are." (Schneider, 1988, p.176). Naturally we do not claim that all dialectometric work can make the connection Schneider seeks.

[8] R is a public domain statistics package available at `http://www.r-project.org`.

References

Bloomfield, L.: 1933, *Language*. New York: Holt, Rhinehart and Winston.

Bolognesi, R. and W. Heeringa: 2002, 'De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten'. *Gramma/TTT: Tijdschrift voor Taalwetenschap*. Accepted to appear in 2002.

Carver, C. M.: [1]1987, 1989, *American Regional Dialects: A Word Geography*. Ann Arbor: The University of Michigan Press.

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

Gooskens, C. and W. Heeringa: 2003, 'Perceptual Evaluation of Levenshtein Dialect Distance Measurements using Norwegian Dialect Data'. *Language Variation and Change*. submitted, 8/2002.

Heeringa, W., J. Nerbonne, and P. Kleiweg: 2002, 'Validating Dialect Comparison Methods'. In: W. Gaul and G. Ritter (eds.): *Proceedings of the 24th Annual Meeting of the Gesellschaft für Klassifikation*. Heidelberg: Springer, pp. 445–452.

Jain, K. and R. C. Dubes: 1988, *Algorithms for clustering Data*. Englewood Cliffs, New Jersey: Prentice Hall.

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

Kruskal, J.: 11983,1999, 'An Overview of Sequence Comparison'. In: D. Sankoff and J. Kruskal (eds.): *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Stanford: CSLI, pp. 1–44.

Kurath, H.: 1949, *A Word Geography of the Eastern United States*. Ann Arbor: University of Michigan Press.

Kurath, H. and 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.: 1991, 'The Three Dialects of English'. In: P. Eckert (ed.): *New Ways of Analyzing Sound Change*. New York: Academic Press, pp. 1–44.

Manning, C. and H. Schütze: 1999, *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

McDavid, R.: 1994, 'Dialects of the LAMSAS Region'. In: W. A. Kretzschmar (ed.): *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*.

Chicago: The University of Chicago Press, pp. 147–153. (written in 1984, shortly before McDavid's death).

Nerbonne, J. and W. Heeringa: 1998, 'Computationele Vergelijking en Classificatie van Dialecten'. *Taal en Tongval* **50**(2), 164–193.

Nerbonne, J., W. Heeringa, and P. Kleiweg: 1999, 'Edit Distance and Dialect Proximity'. In: D. Sankoff and J. Kruskal (eds.): *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.* Stanford, CA: CSLI, pp. v–xv.

Nerbonne, J., W. Heeringa, E. van den Hout, P. van der Kooi, S. Otten, and W. van de Vis: 1996, 'Phonetic Distance between Dutch Dialects'. In: G. Durieux, W. Daelemans, and S. Gillis (eds.): *CLIN VI: Proc. from the Sixth CLIN Meeting*. Antwerpen: Center for Dutch Language and Speech, University of Antwerpen (UIA), pp. 185–202. Also available as `http://www.let.rug.nl/~nerbonne/papers/dialects.ps`.

Schneider, E.: 1988, 'Qualitative vs. Quantitative Methods of Area Delimitation in Dialectology: A Comparison Based on Lexical Data from Georgia and Alabama'. *Journal of English Linguistics* **21**, 175–212.

Séguy, J.: 1971, 'La Relation entre la Distance Spatiale et la Distance Lexicale'. *Revue de Linguistique Romane* **35**, 335–357.

Speelman, D., S. Grondelaers, and D. Geeraerts: 2003, 'Profile-Based Linguistic Uniformity as a Generic Method for Comparing Language Varieties'. *Computers and the Humanities* **37**. Special Iss. on Computational Methods in Dialectometry ed. by John Nerbonne and William Kretzschmar, Jr. (this volume).

Wolfram, W. and N. Schilling-Estes: 1998, *American English*. Malden, Massachusetts: Blackwell.