Toward a Dialectological Yardstick

John Nerbonne and Peter Kleiweg

Rijksuniversiteit Groningen

9700 AS Groningen

The Netherlands

{nerbonne,kleiweg}@let.rug.nl

Telephone: +31 (50) 363 58 15

FAX: +31 (50) 363 68 55

http://www.let.rug.nl/~kleiweg/lamsas/

Dec. 8, 2003, accpt'd May 11, 2005, revised May 14, 2005

Abstract

Dialectometry measures the differences between dialects in ways which may involve many independently varying parameters which must be specified in combination in order to arrive at measures of difference. The existence of many parameters of measurement and possible interaction introduces the problem of how to choose parameter values and combinations of them intelligently. This paper proceeds from the assumption that dialectology proper must reveal geographic coherence in language variation in order to propose a yardstick with which to compare measurements made using various parameter settings, and it presents some results of its application.

1 Introduction

Dialectometry has developed out of the need to provide more objective criteria for distinguishing language varieties (Séguy 1971, Goebl 1984, Nerbonne, Heeringa & Kleiweg 1999), and in part in reaction to criticism that traditional dialect divisions had taken an arbitrary point of departure in choosing which linguistic features to base classifications on. The promise of dialectometry is that of providing classifications based on aggregates of many, potentially *all* linguistic features.

This paper notes in Section 2, however, that the range of dialectometrical techniques now available is large, and that they may lead to different results. We cannot simply take a technique from the shelf and expect its results to mesh with those from other techniques. Section 3 reviews literature aimed at validating dialectometric methods, concluding that there is a need for techniques which function in the absence of corroborating evidence such as the subjective judgments of dialect speakers on how “foreign” or “native” they judge test material to be.

Section 4 argues that it is a fundamental postulate of dialectology that geographically close varieties ought to be linguistically similar, and proposes a measure of this similarity which can be used to compare different dialectometrical measurements of the some region and the same dialect material. Section 5 tests this measure on the issue of how infrequent features should be treated, noting that this issue separates dialectologists. Finally Section 6 discusses limitations of the current proposal and highlights where improvements might be made.

2 Embarrassment of Dialectometrical Riches

Dialectometry provides techniques for assessing the linguistic distance between an arbitrary pair of sites from within a sample studied in a dialect atlas project. The range of techniques suitable for measuring linguistic distance and available to the modern dialectologist is very large. We can examine phonetic, phonological, morphological, lexical, syntactic, semantic and pragmatic levels of language. At any one of these levels, a choice of phenomena and sample material is available, and even then a myriad of further variations in techniques suggest themselves. We shall illustrate these with reference to pronunciation (phonetics and phonology) and lexis, probably the two most popular foci of dialectological work.

2.1 Phonetic Distance

The most successful techniques for measuring the distance between two pronunciations are variants of Levenshtein distance (Kruskal 1999, Kessler 1995, Nerbonne, Heeringa & Kleiweg 1999, Heeringa 2004). Levenshtein distance, which is also known as edit distance, proceeds from a notion of distance between phonetic segments, and induces a notion of sequence distance from this. Suppose we begin we a simple feature system for vowels which assigns numerical values to features as they are realized in different segments. Then, as Table 1 illustrates, it is trivial to derive a notion of segment distance as the Manhattan or City-Block distance between the two feature assignments, where $d(s^1, s^2) = \sum_{i=1}^{|s|} |s^1_i - s^2_i|$.

This illustration suggests several points at which researchers may reasonably differ, and indeed have differed, e.g., in which feature system to use (Vieregge-Cucchiarini, Almeida-Braun, Ladefoged, Chomsky-Halle's (*Sound Pattern of English*. Heeringa (2004, pp. 27–119) is an extensive discussion of these and other systems. Given a feature system, one needs to decide how the features combine to produce a single segment distance, perhaps using Manhattan, or "city-block" distance (as above) or perhaps using a Euclidean definition. Should some features be weighted more heavily than others, perhaps using an

	i	e	u	|i-e|	|i-u|
advancement	2(front)	2(front)	6(back)	0	4
high	4(high)	3(mid high)	4(high)	1	0
long	3(short)	3(short)	3(short)	0	0
rounded	0(not rounded)	0(not rounded)	1(rounded)	0	1
Total				1	5

Table 1: An assignment of values to phonetic features to vowels which allows the measurement of distance between vowels. This system was developed by Vieregge, Rietveld & Jansen (1984) and refined by Cucchiarini (1993) and was used for the evaluation of transcription quality. According to this assignment, and a Manhattan notion of feature distance, d([i],[e]) < d([i],[u]). It is obvious that many alternatives are possible.

information-gain weighting (Mitchell 1997, p. 73)? Should features indicated by diacritics in phonetic transcription be weighed less heavily than those indicated by segmental signs? Should one impose a ceiling on the distance reflecting perhaps insensitivity to relatively large distances?

It would hardly be reasonable to assay the distance between linguistic varieties as a set of phonetic segment distances. It does seem reasonable, however, to estimate the distance between varieties as the sum (or average) of the phonetic distance between words. For this we employ an additional notion of sequence distance; in particular, Levenshtein distance has proven useful in this regard. It is illustrated in Table 2. The basic idea is that we examine various ways of transforming one sequence into another, keeping track of the costs as the distances between corresponding segments (and making appropriate specifications for insertions, deletions and swaps between elements, see Kruskal (1999) and Heeringa (2004)). We then define the sequence distance as the least ex-

Standard American	sɔəgIrl	delete r	0.5
	sɔəgIl	replace I/ɜ	0.1
	sɔəgɜl	insert r	0.8
Bostonian	sɔrəgɜl		
		Sum distance	1.4

Table 2: Levenshtein distance provides a way to *lift* a definition of segment distance to a notion of sequence distance (Kruskal, ^{1}1983, 1999), but it introduces a further set of choice points for phonetic distance measures, notably the relative costs of replacements vs. insertions and deletions.

pensive means of transforming one sequence into another. In this way we see that Levenshtein distance *lifts* a notion of segment distance to one of sequence distance.

The added step of measuring the distance between sequences is sensible, but it likewise adds severals degrees of freedom to the measurement procedure. The standard Levenshtein procedures fixes the cost of insertions and deletions to be one-half that of replacements, but Heeringa (2004) argues that the appropriate notion for measuring dialect pronunciation difference is obtained by setting the cost of an insertion or deletion to be the phonetic distance between silence and the segment which is inserted (or deleted). Kondrak (2003) argues that vowels should count less heavily than consonants in sequence comparison (albeit for the purpose of aligning cognates).

Other questions arise quickly: should diphthongs be represented as single segments, figuring at the segment level, or as a pair of segments whose difference is calculated at the sequence level? How should one incorporate the effect of frequency—should one make it the primary focus of a difference measurement, as

Hoppenbrouwers & Hoppenbrouwers (2001) do in their feature-frequency technique, or should one rather ignore it?

We conclude this section on pronunciation distance by summarizing that the dialectometrist is confronted with an embarrassment of riches when he sets out to measure pronunciation differences. There are myriad plausible techniques, and we need some way of choosing between them.

2.2 Lexical Differences

Lexical differences among dialects are often examined by collecting survey material in which respondents are asked to name an object or activity (including situations in which respondents choose from a list of possible answers). For example, respondents in the survey conducted by the American *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) 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?"* (a question which elicited *clearing up*, *fairing off* and forty other dialectal variants) (Kretzschmar 1994). This is undoubtedly a simpler situation, but nonetheless several parameters can still affect measurements.

First, there is a question of how to deal with morphological variants. We might wish to abstract away from the concrete form of a response to a questionnaire so that we compare lemmata or stems (see Section 5 for an example of how this arises). In the absence of lemmatizing software, we might alternatively try using a notion such as edit distance to obtain sensitivity to similar forms (Nerbonne & Kleiweg 2003). Second, there is a question of how to deal with infrequent words. Carver (^{1}1987, 1989, p.17) suggests that infrequent data be ignored, while Goebl (1984) suggests weighting matches of infrequent words more heavily. These first two factors are presented in more detail in Section 5.3 and Section 5.4, below. Third, there is a question as to how one should handle multiple responses: should every match count and should mismatches in addition be "punished"? Finally, we note here, just as in the case of measuring pronunciation differences, we must inspect combinations: There is no *a priori* reason why the treatment of infrequent elements should be independent of the treatment of morphological variation or that of multiple responses.

In other areas such as syntax, morphology or pragmatics there is every reason to suspect that the situation will be equally rebarbative. Our measurements may involve many independently varying parameters which must be specified in combination in order to arrive at indications of differences. In this computational age, it is possible not only to imagine various combinations of parameter values, but also to examine the results of calculations using them. This benefit leads, however, to the problem of how to choose parameter values and combinations of them intelligently. We need a "yardstick" with which to compare the various parameter settings, and to decide which are preferable.

3 Previous Work

Heeringa, Nerbonne & Kleiweg (2002) examined several techniques for validating dialectometric measures with respect to expert consensus. Heeringa et al. examined a range of measurements, and after discussing and dismissing purely metric techniques such as Fischer's linear discriminant (Schalkoff 1992, p. 90), which proved to be overly sensitive to borderline areas, then turned to ways of using distance measures to classifying dialects. In particular they proceeded from the measurement proper to obtain a matrix of linguistic dissimilarities among varieties, which were then classified by clustering the linguistic dissimilarities in the matrix. The authors then proposed a procedure in which one validates a measurement (and a classification) against the consensus classifications of expert dialectologists, ignoring points of dispute.

Classifications obtained by clustering the results of various linguistic measures were then assessed for the degree to which they reflected agreement with expert opinion using an index of classification agreement (Fowlkes & Mallows 1983). Heeringa, Nerbonne & Kleiweg (2002) acknowledged that this technique suffers in its reliance on the exploratory technique of clustering, and Heeringa (2004) added to this the note that the data points ignored (in the step of restricting one's attention to elements for which there is consensus) often are classified in ways which indicate problems. In such cases there may be a dispute about whether a borderline data point belongs to A or B, but classifying it as C is a clear error. We note further here that the technique is simply inapplicable to cases for which there is no body of expert analysis, i.e., the situation in most dialect areas of the world.

Heeringa (2004) proposed that one examine the correlations of dialectometric measurements with the results of psychoacoustic judgments of similarity. He obtained data from Norwegian dialect speakers who were asked to rate a range of dialect recordings for their similarity to their own native dialects. As Heeringa noted, this basis has the advantage that the notion of “similarity” is then anchored in the perceptions of dialect users—rather than in the analyses of experts, who, as we’ve seen, are capable of construing “similar” in a rather too many ways. We do not take issue with Heeringa on this, or on the utility of the overall technique, but we do note that one normally does not have such data available. It is worthwhile to develop a technique which does not rely on an extensive and expensive set of experiments being conducted. We turn our attention below to an alternative, less demanding possibility, which is therefore more generally applicable.

4 Geographic Coherence

Let us begin by noting that dialectology proper, i.e. the study of varieties from the perspective of their geographic distribution, assumes that geography determines dialectal variation to some extent. This is important enough for us to suggest a name for the principle.

> **Fundamental Dialectological Postulate**: Geographically proximate varieties tend to be more similar than distant ones.

It is clear that this is no absolute law, but only a statistical tendency, since otherwise neither sharp boundaries nor distributed varieties could exist—both of which, however, are very real. Modern politics, standard languages and general schooling have contributed to making the contemporary Dutch-German border sharp, even though it was earlier just one point in a gradual continuum. As an example of a geographically distributed variety, consider town Frisian, which is spoken in the Frisian towns, surrounded by a distinct, rural Frisian variety. There is consensus that the town variety is is geographically distributed (Weijnen 1941, Daan & Blok 1969).

The basic idea of the fundamental dialectological postulate has also been proposed in historical linguistics (Dyen 1956). Campbell (1995) formulates a related generalization for historical reconstruction, saying that "[...] neighboring languages often turn out to be related."

We propose to use the fundamental postulate of dialectology to select more probative measurements, namely those measurements which maximize the degree to which geographically close elements are likewise seen to be linguistically similar. It is slightly more convenient to focus on the converse, and so we provide a formulation of LOCAL INCOHERENCE and suggest that dialectometry profits from choosing measures which minimize local incoherence. The basic idea is that we begin with each measurement site s, and inspect the n linguistically most similar sites and their degrees of dissimilarity to s. We then measure how far away these linguistically most similar sites are, for example in kilometers. For the purpose of this paper we measured geographic distance "as the crow flies," but we discuss reasons why we might wish to deviate from this (§ 6.1). We sum these geographic distances in D^L_s. Since *good* measurements reflect the relative similarity of geographically near sites, D_s^L should be small on average when compared to the D_s^L resulting from *poor* measurements.

The details of the formulation reflect the results of dialectometry that dialect distances certainly increase with geographic distance, leveling off, however, so that geographically more remote variety-pairs tend to have more nearly the same linguistic distances to each other. Figure 1 presents a typical such result. Our formulation will reflect this leveling off of more distant effects by discounting their contribution to local incoherence. In fact we shall sort variety pairs in order of decreasing linguistic similarity and weight more similar ones exponentially more than less similar ones.

Given this disproportionate weighting of the most similar varieties, it also quickly becomes uninteresting to incorporate the effects of more than a small number of geographically closest varieties. Fig. 2 shows that the linguistically most similar varieties are normally among the closest varieties geographically, too. We used this observation as grounds for restricting our attention to the eight most similar linguistic varieties in calculating local incoherence.

We sum the geographic distances to these linguistically most similar elements and normalize this with respect to the measurement which would always find that the geographically closest elements were likewise linguistically most similar, D_s^G. The essential calculations are as follows:

$$\mathrm{I}_l \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\frac{D_i^L - D_i^G}{D_i^G}$$

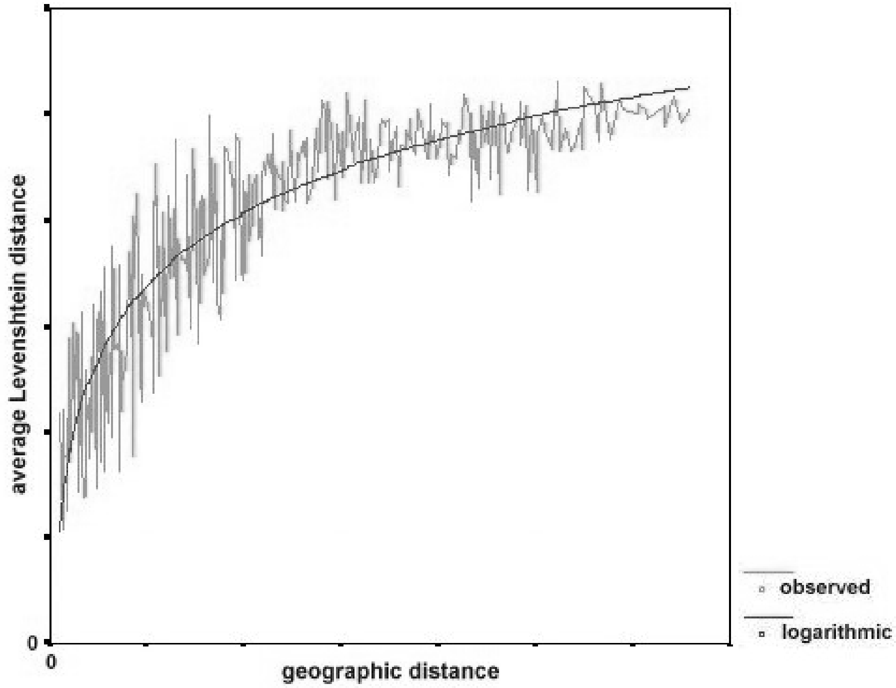

Figure 1: Dialect distance as a logarithmic function of geographic distance, from Heeringa & Nerbonne (2002). Note that the linguistic distance to more remote varieties appears to level off to a fairly flat level, suggesting that such distant points no longer reflect distance in linguistic dissimilarity as reliably. Séguy (1971) provided a similar analysis.

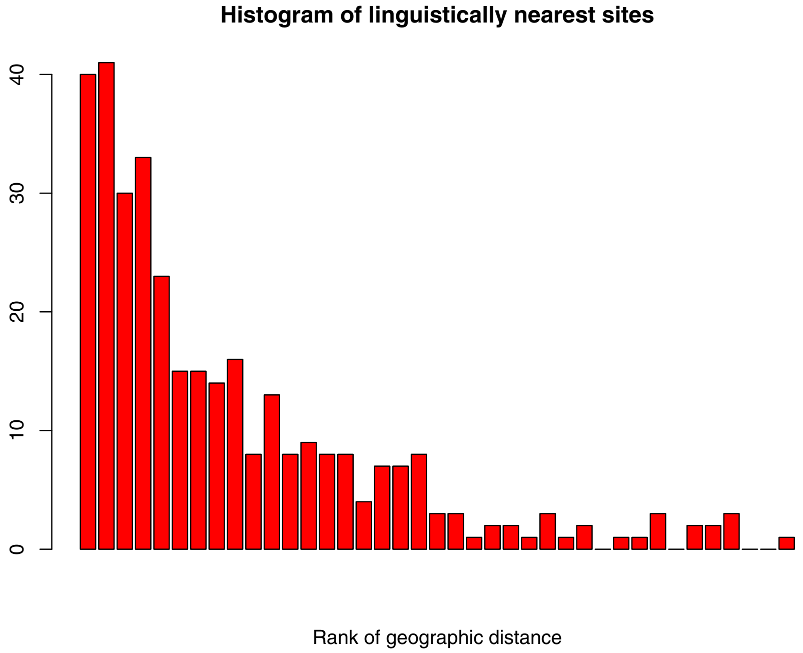

Figure 2: Histogram showing where the linguistically most similar site is found in the list of sites ordered by geographical proximity. For example, we see that the linguistically most similar site is most frequently the second closest geographically. It is clear that we indeed normally find the linguistically most similar sites quite nearby.

$$\begin{aligned} D_i^L &= \sum_{j=1}^{k} d_{i,j}^L \cdot 2^{-0.5j} \\ D_i^G &= \sum_{j=1}^{k} d_{i,j}^G \cdot 2^{-0.5j} \end{aligned}$$

$d_{i,j}^L$, $d_{i,j}^G$: geographical distance between locations i en j

$d_{i,1\cdots n-1}^L$: geographical distance sorted by increasing linguistic difference

$d_{i,1\cdots n-1}^G$: geographical distance sorted by increasing geographical distance

Several remarks may be helpful in understanding the proposed measurement. First, all of the $d_{i,j}$ concern *geographic* distances. $d_{i,1\cdots n-1}^L$ (summed in D_i^L) range over the geographic distances, arranged, however, in increasing order of *linguistic* distance, while $d_{i,1\cdots n-1}^G$ (summed in D_i^G) ranges over the geographic distances among the sites in the sample, arranged in increasing order of *geographic* distance. We examine the latter as an ideal case. If a given measurement technique always demonstrated that the neighbors of a given site used the most similar varieties, then D_i^L would be the same D_i^G, and I_l would be 0. Second, we have argued above that it is appropriate to count most similar varieties much more heavily in I_l, and this is reflected in the exponential decay in the weighting, i.e., $2^{-0.5j}$ where j ranges over the increasingly less similar sites. Given this weighting of most similar varieties, we are also justified in restricting the sum in $D_i^L = \sum_{j=1}^{k}[\ldots]$ to $k = 8$, and all of the results below use this limitation, which likewise improves efficiency.

Third, we use a very primitive notion of geographic distance here (Euclidean distance between longitude-latitude coordinates). See the final discussion for a criticism of this and a suggestion for alternatives. Fourth and finally, because we are using concrete geographic distances, this measurement cannot be applied transparently to different dialectometric situations, including different language areas or even different samplings of the same area. To overcome the difficulty that concrete geographic distances influence on I_l, we have also experimented with rank-based characterizations, which overcome this problem, but which are also naturally less sensitive.

5 Experiment

To appreciate the value of a dialectometric yardstick, we examine several options in dialectometric measures using I_l and note how the proposed yardstick functions. We shall restrict our attention to the simpler lexical case, examining the lexical data in LAMSAS.

5.1 LAMSAS

LAMSAS 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. 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 are included in the web distribution, which formed the basis for the work here. 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, as illustrated in Section 2.2 above.

We focus on lexical variation here in order to keep the illustrations simple. Nerbonne & Kleiweg (2003) explains the care that was taken to ensure that the data used for analysis are indeed comparable. The analyses presented here involved only the data collected by Guy Lowman, who gathered 71% of the LAMSAS data. Furthermore, we used only words which appeared on all three of the worksheets Lowman used (Nerbonne & Kleiweg 2003).

5.2 Lexical Distance

Séguy (1971) suggested measuring dialect differences in a way we can apply to lexical variation fairly simply: 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. 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 is 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*	—

Nerbonne & Kleiweg (2003) provides further motivation and discussion on the manner of treating missing data. We examine one sort of extension to Seguy's method below and one question on the treatment of scarce data.

5.3 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.* It would appear sensible to find a way of recognizing these inflectional variants as similar, and to measure differences while ignoring inflectional variation. This would reflect the usual view in linguistics that such forms "are just variants of one and the same word" (Spencer 1991, p.9), which, however, has also been challenged (Halle 1973, Jackendoff 1975).

One solution to this problem is to apply Levenshtein distance (see Table 2 above) to orthographic strings and to recognize deeper similarity this way. Simple string distance (Levenshtein) will count *bore* and *born* as just as distant as *bore* and *bare.* While one might argue (following de Saussure) that accidentally

form	**stem**
cease	ceas
ceased	ceas
ceases	ceas
ceasing	ceas
a hundred year	a hundr year
a hundred years	a hundr year
blew	blew
blewed	blew

Table 3: The results of applying Porter's stemmer to the LAMSAS responses. Note that the variously inflected forms of *cease* are all correctly reduced to a single form. It is dialectologically interesting to note that *year* is sometimes single and sometimes plural in construction with a numeral, and likewise that double past tense markings may be found on *blow.* But recalling that our focus here is on *lexical* differences, and that these are morphological differences, it is surely correct to abstract away from these interesting differences here.

close variants are rare since the form of words is ultimately arbitrary, still we would prefer to avoid the assumption if possible. And the orthographic similarity 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.* Since we did not have a lemmatizer at hand, we employed instead a public domain version of the Porter stemmer (Porter 1980), whose effects are illustrated in Table 3.

While we shall compare the use of stemming below to the use of a Levenshtein distance on strings, our first preference—if we are focused on lexical affinity—should be for stemming.

5.4 Sparse Data

In several areas of quantitative linguistics it is commonly remarked that very infrequent words are *noise*, unreliable evidence of linguistic structure (Manning & Schütze 1999, p. 199). Carver ([1]1987, 1989, p. 17) takes this position in particular with respect to dialectology based on lexical differences. But in is often unclear how frequent an event must be in order to be used profitably. Exactly where should the cut off be? Words that occur twice, three times, ..., ten times? Words that occur with less than 1% of the frequency of the most frequent words?

Goebl (1984) opposes this general tendency, noting that as evidence of relationship, infrequent words should in fact count more heavily. Goebl introduced *gewichteter Identitätswert*, a weighted similarity, counting overlap in infrequent words more heavily.

For concept i with n responses $w^i_1, w^i_2, \ldots, w^i_n$, let $f(w^i_j)$ be the frequency of w_j as response to query about i.

$$S(w^i_j, w^i_{j'}) = 1 - \frac{f(w^i_j) - 1}{n \cdot c}$$

where Goebl foresees experimentation with c, always $= 1$ here. The quantity $S(w^i_j, w^i_{j'})$ is a weight to be applied to the value 0 in case two sites differ in response (so that different responses always contribute nothing to the similarity measure), and applied to 1 in case the resonses are the same, in which case the contribution is simply $1 - (f(w^i_j) - 1)/n$.

To appreciate Goebl's similarity weighting, consider two cases: one where two sites share a form which occurs in 98 of 100 sites surveyed, in which their weighted similarity is $1 - (98 - 1)/100 = 0.03$; and a second in which two sites share a form which occurs only there, and at none of the other 98 sites surveyed, so that their weighted similarity is $1 - (2 - 1)/100 = 0.99$. This *emphasizes* rather than ignores infrequent words. We try an inverse of Goebl's similarity, $1 - S(w, w')$, as the corresponding dissimilarity measure.

5.5 Results

We examine various combinations of the options noted above for recognizing dialectal similarity even while the data may contain inflectional variation, and even in the face of the question of how to treat infrequent data. We contrast two solutions to handling inflection: using the Porter stemmer, and using the Levenshtein string similarity measure on orthographic strings. We also include measures ignoring the confounding of morphology (using string identity) for the sake of completeness. We then combine these measurements sometimes using Goebl's weighted similarity and sometimes eschewing it. Figure 3 summarizes the results.

As was expected, all of the measures suffer when a large number of infrequent words is omitted (all measures where $x \geq 25$). It is also striking that Goebl's weighted similarity metric is consistently lower in local incoherence (x's and triangles), and that measures that do not incorporate Goebl's weighted similarity indeed tend to benefit from the omission of infrequent words (the falling lines for

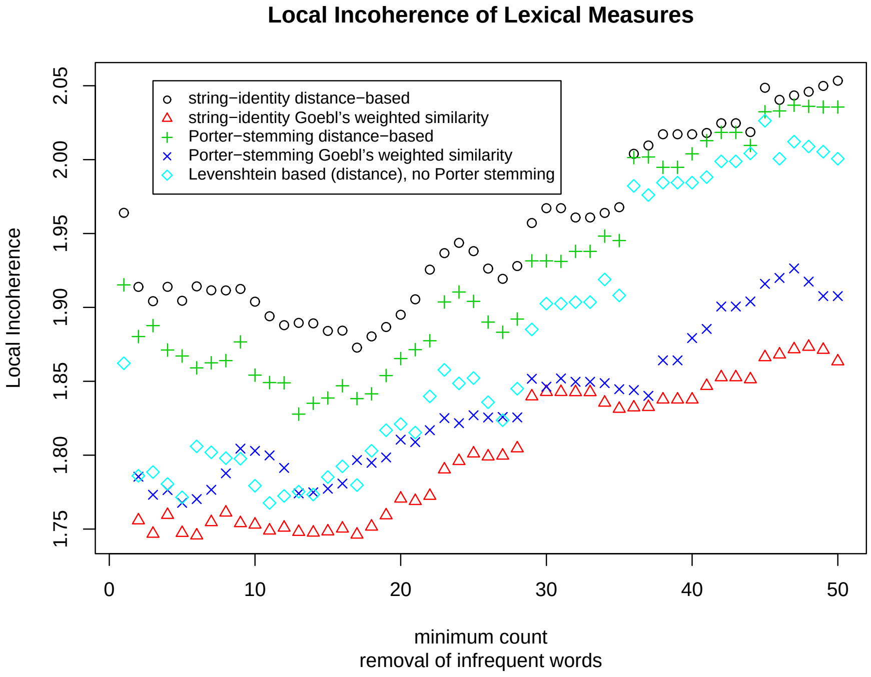

Figure 3: The x axis tracks the minimum count needed by a response in order to play a role in calculations (so indirectly how many of the least frequent words were omitted from consideration) in calculating dialectal similarity, and the y value is local incoherence, I_l. Inverse frequency weighting ("Goebl weighting") is consistently superior.

small values of x for the circles, the +'s and the diamonds). It is likewise noteworthy that, while in general both Levenshtein distance and Porter stemming seem to improve the capacity of the measurements to detect geographic coherence, still the very best measurement was obtained using string identity and Goebl weighting. We suspect that this means that we are measuring sublexical effects, e.g., the likelihood of using one or another inflected form.

If we recall that it is best to insist on Porter stemming (or a more faithful form of lemmatization), then the most interesting comparisons in Figure 3 is that between the measuring using Porter stemming with and without Goebl's weighted similarity—and this comparison (between the x's and the +'s in Figure 3, respectively) indicates the superiority of Goebl's treatment of infrequent items. Goebl's disproportionate weighting of the overlap of similar items is more sensitive in uncovering the linguistic affinities between sites.

The most sensitive measure examined was simple string identity when applied in conjunction with Goebl's weighted similarity. But as we have noted above (§ 5.7), we have good reason to suspect that this measure is tapping into sources of linguistic affinity unrelated to lexical choice.

6 Conclusion and Future Work

Our positive conclusion is that local incoherence lays bare how well a dialectometric measure is attuned to those aspects of linguistic structure which are shared locally—and which therefore should be the basis of dialectology, e.g., for the purpose of dialectological classification, for investigations into the determinants of dialectal variation, and for identifying phenomena such as dialect islands and transition zones. We suspect that many similar and improved techniques could likewise be developed provided they emphasize the features we have incorporated into local incoherence, in particular the whole-hearted recognition of geography as *the* organizing factor in dialectology, and the focus on a small number of most local varieties to avoid the "drift into noise" we find at larger distances.

6.1 Limitations and opportunities

We have worked with an intuitive notion of geographical distance in this paper, which we have consistently calculated simply as the Euclidean distance between points described by longitude and latitude coordinates. Arguably, this notion needs to be refined particularly when applied to situations where geography influences the chance of social contact—e.g., cases of variation in mountainous terrain. The idea that the diffusion of linguistic innovation is influenced by contact (or conversely, isolation), and therefore in turn by geography is perhaps so obvious that no source is cited even in handbooks (Wolfram & Schilling-Estes 2003), but there have not been extensive quantitative studies of the influence of geography on variation. On exception is Gooskens (2004), who examines examines the effect of geography on dialect variation in Norway, where the central mountain range prevented direct travel until recently. She demonstrates that travel time is a much better predictor of Norwegian linguistic distance than distance "as the crow flies." In such cases local incoherence as it is defined above does not in fact work poorly—but only because it considers only a small number of relatively similar varieties. Nonetheless, it is clear that more interesting geographic notions are more promising as superior indicators of linguistic distance, e.g. travel time.

The techniques examined in this paper may also be applied to the study of linguistic variation dependent on social status, age or gender, even if we focused here on dialectology proper, i.e. variation dependent on geography. It is clear that that any such application would require that an appropriate notion of distance (social, chronological, or gender-based) be agreed on, however, and that could be a challenging task.

Finally, it would be worthwhile, but also challenging to explore the development of measures which would allow comparison across dialectal areas, e.g. a comparison of the New England and the Middle and South Atlantic states on the American East coast, or a comparison of the Netherlands and the United States with respect to the question of which variety is greater. Given our reliance on concrete geography and the density of sampling, such questions cannot be addressed with the techniques presented here.

Acknowledgments

We are grateful to Dr. Charlotte Gooskens, Prof. Dr. Hermann Niebaum, Ms. Christine Siedle, and the audience at *Qualico 2003*, including especially

Dr. Harald Baayen, for discussion of the ideas presented here. A particular word of thanks to Dr. Wilbert Heeringa, whose stimulating work on validation partially inspired the ideas presented here, and with whom we spoke at length.

References

Campbell, L. (1995). "The Quechumaran Hypothesis and Lessons for Distant Genetic Comparison." *Diachronica* XII(2):157–200.

Carver, C. M. (^{1}1987, 1989). *American Regional Dialects: A Word Geography.* Ann Arbor: The University of Michigan Press.

Cucchiarini, C. (1993). Phonetic Transcription: A Methodological and Empirical Study PhD thesis Katholieke Universiteit Nijmegen.

Daan, J. & D. P. Blok. (1969). *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde.* Amsterdam: Noord-Hollandsche Uitgevers.

Dyen, I. (1956). "Language Distribution and Migration Theory." *Language* 32:611–626.

Fowlkes, E. & C. Mallows. (1983). "A Method for Comparing Two Hierarchical Clsuterings." *Journal of the American Statistical Association* 78:553–569.

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

Gooskens, C. (2004). Norwegian Dialect Distances Geographically Explained. In *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe ICLAVE 2, June 12-14, 2003*, (ed.) B.-L. Gunnarson, L. Bergström, G. Eklund, S. Fridella, L. H. Hansen, A. Karstadt, B. Nordberg, E. Sundgren & M. Thelander. Uppsala, Sweden: Uppsala University pp. 195–206.

Halle, M. (1973). "Prologomena to a Theory of Word-Formation." *Linguistic Inquiry* 4(1):3–16.

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

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

Hoppenbrouwers, C. & G. Hoppenbrouwers. (2001). *De indeling van de Nederlandse streektalen: Dialecten van 156 steden en dorpen geklasseerd volgens de FFM (feature frequentie methode).* Assen: Koninklijke Van Gorcum.

Jackendoff, R. (1975). "Morphological and Semantic Regularities in the Lexicon." *Language* 51(4):639–671.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. In *Proc. of the European ACL*. Dublin: pp. 60–67.

Kondrak, G. (2003). "Phonetic Alignment and Similarity." *Computers and the Humanities* 37(3):273–291. Special Iss. on Computational Methods in Dialectometry ed. by John Nerbonne and William Kretzschmar, Jr.

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. (1999). An Overview of Sequence Comparison. In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, (ed.) D. Sankoff & J. Kruskal. Stanford: CSLI pp. 1–44. ^{1}1983.

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

Mitchell, T. M. (1997). *Machine Learning*. Singapore: McGraw-Hill.

Nerbonne, J. & P. Kleiweg. (2003). "Lexical Variation in LAMSAS." *Computers and the Humanities* 37(3):339–357. Special Iss. on Computational Methods in Dialectometry ed. by John Nerbonne and William Kretzschmar, Jr.

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

Porter, M. (1980). "An Algorithm for Suffix Stripping." *Program* 14(3):130–137.

Schalkoff, R. (1992). *Pattern Recognition: Statistical, Structural and Neural Approaches.* New York: John Wiley.

Séguy, J. (1971). “La relation entre la distance spatiale et la distance lexicale.” *Revue de Linguistique Romane* 35:335–357.

Spencer, A. (1991). *Morphological Theory.* Oxford: Blackwell.

Vieregge, W. H., A. Rietveld & C. Jansen. (1984). A Distinctive Feature Based System for the Evaluation of Segmental Transcription in Dutch. In *Proc. of the 10th International Congress of Phonetic Sciences*, (ed.) M. P. van den Broecke & A. Cohen. Dordrecht: pp. 654–659.

Weijnen, A. (1941). *De Nederlandse Dialecten.* Groningen: Nordhoff.

Wolfram, W. & N. Schilling-Estes. (2003). Dialectology and Linguistic Diffusion. In *The Handbook of Historical Linguistics*, (ed.) B. D. Joseph & R. D. Janda. Malden, Massachusetts: Blackwell pp. 713–735.