Evaluation of String Distance Algorithms for Dialectology

Wilbert Heeringa, Peter Kleiweg, Charlotte Gooskens & John Nerbonne
Humanities Computing, University of Groningen
{W.J.Heeringa, P.C.J.Kleiweg, C.S.Gooskens, J.Nerbonne}@rug.nl

Abstract

We examine various string distance measures for suitability in modeling dialect distance, especially its perception. We find measures superior which do *not* normalize for word length, but which *are* are sensitive to order. We likewise find evidence for the superiority of measures which incorporate a sensitivity to phonological context, realized in the form of n-grams—although we cannot identify which form of context (bigram, trigram, etc.) is best. However, we find no clear benefit in using gradual as opposed to binary segmental difference when calculating sequence distances.

1 Introduction

We compare string distance measures for their value in modeling dialect distances. Traditional dialectology relies on identifying language features which are common to one dialect area while distinguishing it from others. It has difficulty in dealing with partial matches of linguistic features and with non-overlapping language patterns. Therefore Seguy (1973) and Goebl (1982; 1984) advocate using aggregates of linguistic features to analyze dialectal patterns, effectively introducing the perspective of DIALECTOMETRY.

Kessler (1995) introduced the use of string edit distance measure as a means of calculating the distance between the pronunciations of corresponding words in different dialects. Following Seguy's and Goebl's lead, he calculated this distance for pairs of pronunciations of many words in many Irish-speaking towns. String edit distance is sensitive to the degrees of overlap of strings and allows one to process large amounts of pronunciation data, including that which does not follow other isoglosses neatly. Heeringa (2004) examines several variants of edit distance applied to Norwegian and Dutch data, focusing on measures which involve a length normalization, and which ignore phonological context, and demonstrating that measures using binary segment differences are no worse than those using feature-based measures of segment difference.

This paper inspects a range of further refinements in measuring pronunciation differences. First, we inspect the role of normalization by length, showing that it actually worsens non-normalized measures. Second, we compare edit distance measures to simpler measures which ignore linear order, and show that order-sensitivity is important. Third, we inspect measures which are sensitive to phonetic context, and show that these, too, tend to be superior. Fourth, we compare versions of string edit distance which are constrained to respect syllable structure (always matching vowels with vowels, etc.), and conclude that this is mildly advantageous. Finally we compare binary (i.e., same/different) treatments of the segments in edit distance to gradual treatments of segment differentiation, and find no indication of the superiority of the gradual measures.

The quality of the measures is assayed primarily through their agreement with the judgments of dialect speakers about which varieties are perceived as more similar (or dissimilar) to their own. In addition we inspect a validation technique which purports to show how successfully a dialect measure uncovers the geographic structure in the data (Nerbonne and Kleiweg, 2006), but this technique yields unstable results when applied to our data. We have perception data only for Norwegian, so that data figures prominently in our argument, and we evaluate both Norwegian and German data geographically.

The results differ, and the perceptual results (concerning Norwegian) are most easily interpretable. There we find, as noted above, that non-normalized measures are superior to normalized ones, that both order and context sensitivity are worthwhile, as is the vowel/consonant distinction. The (geographic) results for German are more complicated, but also less stable. We include them for the sake of completeness.

In addition we note two minor contributions. First, although some literature ends up evaluating both distance and similarity measures, because these are not consistently each others' inverses under some normalizations (Kondrak, 2005; Inkpen et al., 2005), we suggest a normalization based on alignment length which guarantees that similarity is exactly the inverse of distance, allowing us to concentrate on distance.

Second, we note that there is no great problem in applying edit distance to bigrams and trigrams, even though recent literature has been sceptical about the feasibility of this step. For example Kessler (2005) writes:

> [...] one major shortcoming [in applying edit distance to linguistic data, WH et al] that is rarely discussed is that the phonetic environment of the sounds in question cannot be taken into account, while still making use of the efficient dynamic programming algorithm (p. 253).

Somewhat further Kessler writes: "Currently, the predominant solution to this problem is to ignore context entirely." In fact Kondrak (2005) applies edit distance straightforwardly using n-gram as basic elements. Our findings accord with Kondrak's, who also found no problem in applying edit distance using n-grams, but we evaluate the technique in its application to dialectology.

1.1 Background

Heeringa (2004) demonstrates that edit distance applied to comparable words (see below for examples) is a superior measure of dialect distance when compared to unigram corpus frequency and also that it is superior to both the frequency of phonetic features in corpora (a technique which Hoppenbrouwers & Hoppenbrouwers (2001) had advocated) and to the frequency of phonetic features taken one word at a time. Heeringa compares these techniques using the results of a perception experiment we also employ below. Heeringa shows that word-based techniques are superior to corpus-based techniques, and moreover, that most word-based techniques perform about the same. We therefore ignore measures which view corpora as undifferentiated collections below and study only word-based techniques.

A further question was whether to compare words based on a binary difference between segments or whether to use instead phonetic features to derive a more sensitive measure of segment distance. It turned out that measures using binary segment distinctions outperform the feature-based methods (see Heeringa, pp. 184–186), even though a number of feature systems and comparisons of feature vectors were experimented with. We likewise accept these results (at least for present purposes) and focus exclusively on measures using the binary segment distinctions below.

Kondrak (2005) and Inkpen et al. (2005) present several methods for measuring string similarity and distance which complement Heeringa's results nicely. We should note, however, that these papers focus on other areas of application, viz., the problems of identifying (i) technical names which might be confused, (ii) linguistic cognates (words from the same root), and (iii) translational cognates (words which may be used as translational equivalences). Inkpen et al. consider 12 different orthographic similarity measures, including some in which the order of segments does not play a role (e.g., DICE), and others which use order in alignment (e.g. edit distance). They further consider comparison on the basis of unigrams, bigrams, trigrams and "xbigrams," which are trigrams without the middle element. Some methods are similarity measures, others are distance measures. We return to this in Section 2.

1.2 This paper

In this paper we apply string distance measures to Norwegian and German dialect data. As noted above, we focus on word-based methods in which segments are compared at a binary (same/different) level. The methods we consider will be explained in Section 2. Section 3 describes the Norwegian and German data to which the methods are applied. In Section 4 we describe how we evaluate the methods, namely by comparing the algorithmic results to the distances as perceived by the dialect speakers themselves. We likewise aimed to evaluate by calculating the degree to which a measure uncovers geographic cohesion in dialect data, but as we shall see, this means of validation yields rather unstable results. In Section 5 we present results for the different methods and finally, in Section 6, we draw some conclusions.

2 String Comparison Algorithms

In this section we describe a number of string comparison algorithms largely following Inkpen et al. (2005). The methods can be classified according to different factors: representation (unigram, bigram, trigram, xbigram), comparison of n-grams (binary or gradual), status of order (with or without alignment), and type of alignment (free or forced alignment with respect to the vowel/consonant distinction). We illustrate the methods with examples, in which we compare German and Dutch dialect pronunciations of the word *milk*.[1]

2.1 Contextual sensitivity

In the German dialect of Reelkirchen *milk* is pronounced as [mɛlkə]. The bigram notation is [–m mɛ ɛl lk kə ə–] and the trigram notation is [—m –mɛ mɛl ɛlk lkə kə– ə—]. The same word is pronounced as [mɛləç] in the German dialect of Tann. The bigram and trigram representations are [–m mɛ ɛl lə əç ç–] and [—m –mɛ mɛl ɛlə ləç əç– ç—] respectively.

In the simplest method we present in this paper, the distance is found by calculating 1 minus twice the number of shared segment n-grams divided by the total number of n-grams in both words. Inkpen et al. mention a bigram-based, a trigram-based and a xbigram-based procedure, which they call DICE, TRIGRAM and XDICE respectively. We also consider an unigram-based procedure which we call UNIGRAM. The two pronunciations share four unigrams: [m, ɛ, l] and [ə]. There are $5+5=10$ unigram tokens in total in the two words, so the unigram similarity is $(2 \times 4)/10 = 0.8$, and the distance $1 - 0.8 = 0.2$. The two pronunciations share three bigrams: [–m, mɛ] and [ɛl]. There are $6+6=12$ bigram tokens in the two strings, so bigram similarity is $(2 \times 3)/12 = 0.5$, and the distance $1 - 0.5 = 0.5$. Finally, the two pronunciations have three trigrams in common: [—m, –mɛ] and [mɛl] among $7+7=14$ in total, yielding a trigram similarity of $(2 \times 3)/14 = 0.4$ and distance $1 - 0.4 = 0.6$.

Our interest in this issue is linguistic: longer n-grams allow comparison on the basis of phonic context, and unigram comparisons have correctly been criticized for ignoring this (Kessler, 2005).

2.2 Order of segments

When comparing the German dialect pronunciation of Reelkirchen [mɛlkə] with the Dutch dialect pronunciation of Haarlem [mɛlək], the unigram procedure presented above will detect no difference. One might argue that we are dealing with a swap, but this is effectively an appeal to order. The example is not convincing for n-gram measures, $n \geq 2$, but we should prefer to separate issues of order from issues of context sensitivity. We use edit distance (aka Levenshtein distance) for this purpose, and we assume familiarity with this (Kruskal, 1999). In our use of edit distance all operations have a cost of 1.

2.3 Normalization by length

When the edit distance is divided by the length of the longer string, Inkpen et al. call it normalized edit distance (NED). In our approach we divide “raw edit distance” by alignment length. The same minimum distance found by the edit distance algorithm may be obtained on the basis of several alignments which may have different lengths. We found that the longest alignment has the greatest number of matches. Therefore we normalize by dividing the edit distance by the length of the longest alignment.

We have normally employed a length normalization in earlier work (Heeringa, 2004), reasoning that words are such fundamental linguistic units that dialect perception was likely to be word-based. We shall test this premise in this paper.

Marzal & Vidal (1993) show that the normalized edit distance between two strings cannot be obtained via “post-normalization”, i.e., by first computing the (unnormalized) edit distance and then normalizing this by the length of the corresponding editing path. Unnormalized edit distance satisfies the triangle inequality, which is axiomatic for distances, but the quantities obtained via post-normalization need not satisfy this axiom. Marzdal & Vidal provide an alternative procedure which is guaranteed to produce genuine

[1] Our transcriptions omit diacritics for simplicity’s sake.

distances, satisfying all of the relevant axioms. In their modified algorithm, one computes one minimum weight for *each* of the possible lengths of editing paths at each point in the computational lattice. Once all these weights are calculated, they are divided by their corresponding path lengths, and the minimum quotient represents the normalized edit distance.

The basic idea behind edit distance is to find the minimum cost of changing one string into another. Length normalization represents a deviation from this basic idea. If a higher cost corresponds with a longer path length so that quotient of the edit costs divided by the path length is minimal, then Marzal & Vidal's procedure opts for the minimal normalized length, while post-normalization seeks what one might call "the normalized minimal length" (see Marzal & Vidal's example 3.1 and Figure 2, p. 928).

Marzal & Vidal's examples of normalized minimal distances which are not also minimal normalized distances all involve operation costs we normally do not employ. In particular they allow INDELS (insertions and deletions) to be associated with much lower costs than substitutions, so that the longer paths associated with derivations involving indels is more than compensated by the length normalization. Our costs are never structured in this way, so we conjecture that our post-normalizations do not genuinely run the risk of violating the distance axioms. We use 0 for the cost of mapping a symbol to itself, 1 to map it to a different symbol, including the empty symbol (covering the costs of indels), and ∞ for non-allowed mappings[2] We maintain therefore that (unnormalized) costs higher than the minimum will never correspond to longer alignment lengths. If this is so, then the minimal edit cost divided by alignment length will also be the minimal normalized cost. If the unnormalized edit distance is minimal, we claim that the post-normalized edit distance must therefore be minimal as well.

We inspect an example to illustrate these issues. We compare the Frisian (Grouw), [mɔlkə], with the Haarlem pronunciation [mɛlək]. The Levenshtein algorithm may align the pronunciations as follows:

1	2	3	4	5	6
m	ɔ	l		k	ə
m	ɛ	l	ə	k	
	1		1		1

The one pronunciation is transformed into the other by substituting [ɛ] for [ɔ], by deleting [ə] after [l], and by inserting [ə] after [k]. Since each operation has a cost of 1, and the alignment is 6 elements long, the normalized distance is $(1 + 1 + 1)/6 = 0.5$. The Levenshtein distance will also find an alignment in which the [ə]'s are matched, while the [k]'s are inserted and deleted. That alignment gives the same (normalized) distance. Levenshtein distance will not find an alignment any longer than the one shown here, since longer alignments will not yield the minimum cost. This also holds for the examples shown below.

2.4 n-gram weights

In the dialect of the German dialect of Frohnhausen *milk* is pronounced as [mɪljə], and in the German of Großwechsungen as [mɛlɪç]. If we compare these using the techniques of Section 2.2, using bigrams, we obtain the following:

1	2	3	4	5	6
-m	mɪ	ɪl	lj	jə	ə-
-m	mɛ	ɛl	lɪ	ɪk	k-
	1	1	1	1	1

Since n-grams are compared in a binary way, the normalized distance is equal to $(1 + 1 + 1 + 1 + 1)/6 = 0.83$. But [mɪ] and [mɛ] (second position) are clearly more similar to each other than [jə] and [ɪk] (fifth position). Inkpen et al. suggest weighting n-gram differences using segment overlap. They provide a formula for measuring gradual similarity of n-grams to be used in BI-DIST and TRI-DIST. Since we measure distances rather than similarity, we calculate n-gram distance as follows:

$$s(x_1...x_n, y_1...y_n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, y_i)$$

where $d(a, b)$ returns 1 if a and b are different, and 0 otherwise. We apply this to our example:

1	2	3	4	5	6
-m	mɪ	ɪl	lj	jə	ə-
-m	mɛ	ɛl	lɪ	ɪk	k-
	0.5	0.5	0.5	1	0.5

obtaining $(0.5+0.5+0.5+1+0.5)/6 = 3.0/6 = 0.5$ distance after normalization.

[2] For example, in some versions of edit distance, the value ∞ is assigned to the replacement of a vowel by a consonant in order to avoid alignments which violate syllabic structure.

2.5 Linguistic Alignment

When comparing the Frisian (Grouw) dialect pronunciation, [mɔlkə], with that of German Großwechsungen, [mɛlɪç], using unigrams, we obtain:

1	2	3	4	5
m	ɔ	l	k	ə
m	ɛ	l	ɪ	ç
	1		1	1

The normalized distance is then $(1 + 1 + 1)/5 = 0.6$. But this is linguistically an implausible alignment: syllables do not align when e.g. [k] aligns with [ɪ], etc. We may remedy this by requiring the Levenshtein algorithm to respect the distinction between vowels and consonants, requiring that the alignments respect this distinction with only three exceptions, in particular that semivowels [j, w] may match vowels (or consonants), that the maximally high vowels [i, u] match consonants (or vowels), and that [ə] match sonorant consonants (nasals and liquids) in addition to vowels. Disallowed matches are weighted so heavily (via the cost of the substitution operation) that the algorithm always will use alternative alignments, effectively preferring insertions and deletions (indels) instead. Applying these restrictions, we obtain the following, with normalized distance $(1+1+1+1)/6 = 0.67$:

1	2	3	4	5	6
m	ɔ	l		k	ə
m	ɛ	l	ɪ	ç	
	1		1	1	1

In comparisons based on bigrams, we allow two bigrams to match when at least one segment pair matches, the first, the second, or both. Two trigrams match when at least the middle pair matches. Comparing the same pronunciations as above using bigrams without linguistic conditions, we obtain the following alignment:

1	2	3	4	5	6
-m	mɔ	ɔl	lk	kə	ə-
-m	mɛ	ɛl	lɪ	ɪç	ç-
	1	1	1	1	1
	0.5	0.5	0.5	1	0.5

The normalized distance is $(1 + 1 + 1 + 1 + 1)/6 = 0.83$ using binary bigram weights (costs), and $(0.5 + 0.5 + 0.5 + 1 + 0.5)/6 = 0.5$ using gradual weights. But the above alignment does *not* respect the vowel/consonant distinction at the fifth position, where neither [k] vs. [ɪ] nor [ə] vs. [ç] is allowed. We correct this at once:

1	2	3	4	5	6	7
-m	mɔ	ɔl	lk		kə	ə-
-m	mɛ	ɛl	lɪ	ɪç	ç-	
	1	1	1	1	1	1
	0.33	0.33	0.67	1	1	1

Using binary bigram weights, the normalized distance is $(1 + 1 + 1 + 1 + 1 + 1)/7 = 0.86$.

The calculation based on gradual weights is a bit more complex. Two bigrams may match even when a non-allowed pair occurs in one of the two positions, e.g., [k] vs. [ɪ] at the fourth position in the alignment immediately above. The cost of this match should be higher (via weights) than that of an allowed pair with different elements—e.g., the pair [ɔ] versus [ɛ] at the second or third position—but not so high that the match cannot occur.

We settle on the following scheme. Two n-grams $[x_1...x_n]$ and $[y_1...y_n]$ can only match if at least one pair (x_i, y_i) matches linguistically. We weight linguistically mismatching pairs (x_j, y_j) twice as high as matching (but non-identical) pairs. Since we have at most $n - 1$ matching pairs, and at least 1 mismatching pair, we set the most expensive match of two n-grams to 1, and we assign the weight of $2/(2n - 1)$ to a mismatching pair, and $1/(n - 1)$ to a matching (but non-identical) one. Indels cost the same as the most costly (matching) n-grams, in this case 1.

In our bigram-based example, we obtain a weight of $2/(2 \times 2 - 1) = 0.67$ at position 4, since the pair [k] vs. [ɪ] is a linguistic mismatch. At positions 2 and 3 we obtain weights of $1/(2 \times 2 - 1) = 0.33$ since [ɔ] and [ɛ] are (non-identical) matches. Note that a segment (vowel or consonant) versus '-' (boundary) is processed as a mismatch. Therefore the weight at position 6 is equal to 0.33 ([k] vs. [ç]) $+0.67$ ([ə] versus [-]), summing to 1.

2.6 Similarity vs. distance

Theoretically, similarity and distance should be each others' inverses. Thus in Section 2.1 we suggested that similarity should always be $(1 -$ distance$)$. This is not always straightforward when we normalize.

Inkpen et al. use both similarity and distance measures. Similarity measures are LCSR (Longest Common Subsequence Ratio), BI-SIM and TRI-SIM (LCSR generalized to bigrams and trigrams), and the corresponding distance measures are NED, BI-DIST and TRI-DIST. The measures are further distinguished in the way n-gram weights are compared: as binary weights in the similarity measures, and as gradual weights in the distance measures. When comparing the pronunciations of Frisian Hindelopen [mɔəlkə] with German Großwechsungen, [mɛlɪç], and respecting the linguistic alignment conditions (Section 2.5) we obtain:

m	ɔ	ə	l		k	ə
m	ɛ		l	ɪ	ç	
0	1	1	0	1	1	1

The non-normalized similarity is equal to 2, and the non-normalized distance is equal to 5. Inkpen et al. normalize "by dividing the total edit cost by the length of the longer string" which is 6 in our example. Other possibilities are dividing by the length of the shorter string (5), the average length of the two strings (5.5) or the length of the alignment (7). Summarizing:

	shorter string	longer string	average string	alignment
sim.	0.4	0.33	0.36	0.29
dist.	1.0	0.83	0.91	0.71
total	1.4	1.17	1.27	1.00

Only the normalization via alignment length respects the wish that we regard similarity and distance as each others' inverses. [3] We can enforce this requirement in other approaches by first normalizing and then taking the inverse, but we take the result above to indicate that normalization via alignment length is the most natural procedure.

3 Data Sources

The methods presented in Section 2 are applied to Norwegian and German dialect data described in this section. We emphasize that we measured distances only at the level of the segmental base, ignoring stress and tone marks, suprasegmentals and diacritics. We in fact examined measurements which included the effects of segmental diacritics, which, however resulted in decreased consistency and no apparent increase in quality.

3.1 Norwegian

The Norwegian data comes from a database comprising more than 50 dialect sites, compiled by Jørn Almberg and Kristian Skarbø of the Department of Linguistics of the University of Trondheim.[4] The database includes recordings *and* transcriptions of the fable 'The North Wind and the Sun' in various Norwegian dialects. The Norwegian text consists of 58 different words, some of which occur more than once, in which case we seek a least expensive pairing of the different elements (Nerbonne and Kleiweg, 2003, p. 349).

On the basis of the recordings, Gooskens carried out a perception experiment which we describe in Section 4.1. The experiment is based on 15 dialects, the total number of dialects available at that time (spring, 2000). Since we want to use the results of the experiment for validating our methods, we used the same set of 15 Norwegian dialects. It is important to note that Gooskens presented the recordings holistically, including differences in syntax, intonation and morphology. Our methods are restricted to words.

3.2 German

The German data comes from the *Phonetischer Atlas Deutschlands* and includes 186 dialect locations. For each location 201 words were recorded and transcribed. The data are available at the *Forschungsinstitut für deutsche Sprache - Deutscher Sprachatlas* in Marburg. The material is from translations of *Wenker-Sätze*, taken from the famous survey by Georg Wenker in the 1879–1887 among teachers from ≈ 40.000 locations in Germany. The transcriptions are made on the basis of recordings made under the direction of Joachim Göschel in the 1960's and 1970's in West Germany (Göschel 1992, pp. 64–70). After the German reunification similar surveys were conducted in former East Germany.

The data were transcribed by four transcribers, and each item was transcribed independently by at least two phoneticians who subsequently consulted to come to an agreement. In 2002 the data was digitized at the University of Groningen.

4 Validation Methods

When we apply a measurement technique to a specific problem we are interested both in the consistency of the measure and in its validity. The consistency of the measurement reflects the degree to which the independent elements in the sample sample tend to provide the same signal. Nunnally (1978, p.211) recommends the generalized

[3] We have no proof that normalization by alignment length always allows this simple relation to similarity, but we have examined a large number of calculations in which this always seems to hold.

[4] The database is available at `http://www.ling.hf.ntnu.no/nos/`.

form of the Spearman-Brown formula for this purpose, which has come to be known as the CRONBACH'S α value. It is determined by the inter-item correlation, i.e. the average correlation coefficient for all of the pairs of items in the test, and the test size. The Cronbach's α measure rises with the sample size, and it is therefore normally used to determine whether samples are large enough to provide reliable signals.

The validity of a measure, or more precisely, the application of a measure to a particular problem is much more difficult and controversial issue (Nunnally, 1978, Chap. 3), but the basic issue is whether the procedures in fact measure what they purport to measure, in our case the sort of pronunciation similarity which is important in distinguishing similar language varieties. In examining our measures for their validity in identifying the sort of pronunciation similarity which plays a role in dialectology we compare the measures to other indications we have that pronunciations are dialectally similar. We discuss these below in more detail. We consider the correlation with distances as perceived by the dialect speakers themselves (see Section 4.1) and the local (geographic) incoherence of dialect distances (see Section 4.2).

4.1 Perception

The best opportunity for examining the quality of the measurements presents itself in the case of Norwegian, for which we were able to obtain the results of a perception experiment (Gooskens and Heeringa, 2004). For each of 15 varieties a recording of the fable 'The North Wind and the Sun' was presented to 15 groups of Norwegian high school pupils, one group from each of the 15 dialects sites represented in the material. All pupils were familiar with their own dialect and had lived most of their lives in the place in question (on average 16.7 years). Each group consisted of 16 to 27 listeners. The mean age of the listeners was 17.8 years, 52 percent were female and 48 percent male.

The 15 dialects were presented in a randomized order, and each session was preceded by a (short) practice run. While listening to the dialects the listeners were asked to judge each of the 15 dialects on a scale from 1 (similar to native dialect) to 10 (not similar to native dialect). This means that each group of listeners judged the linguistic distances between their own dialect and the 15 dialects, including their own dialect. In this way we get a matrix with 15 $\times$ 15 perceived linguistic distances. This matrix is not completely symmetric. For example, the distance which the listeners from Bergen perceived between their own dialect and the dialect of Trondheim (8.55) is different from the distance as perceived by the listeners from Trondheim to Bergen (7.84).

In order to use this material to calibrate the different computational measurements, we examine the correlations between the 15×15 computational matrices with the 15×15 perceptual matrix. In calculating correlations we excluded the distances of dialects with respect to themselves, i.e. the distance of Bergen to Bergen, of Bjugn to Bjugn, etc. In computational matrices these values are always zero, in the perceptual matrix they vary, but are normally greater than zero. This may be due to non-geographic (social or individual) variation, but it distorts results in a non-random way (diagonal distances can only be too high, never too low), we exclude them when calculating the correlation coefficient.

We calculated the standard Pearson product-moment correlation coefficient, but we interpret its significance cautiously, using the Mantel test (Bonnet and Van de Peer, 2002). In classical tests the assumption is made that the observations are independent, which observations in distance matrices emphatically are not. This is certainly true for calculations of geographic distances, which are minimally constrained to satisfy the standard distance axioms (non-negativity, symmetry, and the triangle inequality). We have argued above (§ 2.2) that the edit distances we employ are likewise genuine distances, which means that sums of edit distances are likewise constrained, and therefore should not be regarded as independent observations (in the sense need for hypothesis testing).

The Mantel test raises the standards of significance a good deal— so much that it will turn out that our small (15 $\times$ 15) matrices would need to differ by more than 0.1 in correlation coefficient in order to demonstrate significance. We will nonetheless urge that the results should be taken seriously as the data needed is difficult to obtain, and the indications are fairly clear (see below).

4.2 Local Incoherence

It is fundamental to dialectology that geographically closer varieties are, in general, linguistically more similar. Nerbonne and Kleiweg (2006) use this fact to select more probative measurements, namely those measurements which maximize the degree to which geographically close elements are likewise seen to be linguistically similar. Given our emphasis on distance it is slightly more convenient to formulate a measure of LOCAL INCOHERENCE and then to examine the degree to which various string distance measures minimize it. The basic idea is that we begin with each measurement site s, and inspect the n linguistically most similar sites in order of decreasing linguistic similarity to s. We then measure how far away these linguistically most similar sites are geographically, for example, in kilometers. *Good* measurements show that linguistically similar sites are geographically close better than *poor* measurements do.

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. We 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. We restrict our attention to the eight most similar linguistic varieties in calculating local incoherence.

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

$d_{i,j}^L$, $d_{i,j}^G$: geo. dist. between i en j

$d_{i,1\cdots n-1}^L$: geo. dist. sorted by increasing ling. diff.

$d_{i,1\cdots n-1}^G$: geo. dist, sorted by increasing geo. dist.

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.

We suppress further discussion of the calculation in the interest of saving space here, noting, however, that we used two different notions of geographic distance. When examining measurements of the German data, we measured geographic distance “as the crow flies”, but since Norway is very mountainous, we used (19th century) travel distances (Gooskens,).

5 Experiments and Results

In this section we present results based on the Norwegian and German data sources in 5.1 and Sections 5.3.

For each data source we consider 40 string comparison algorithms. We distinguish between methods with a binary comparison of n-grams and those with a gradual comparison of n-grams (see Section 2.4). Within the category of binary methods, we distinguish between three groups. In the first group, strings are compared just by counting the number of common n-grams, ignoring the order of elements, see Section 2.1). In the second group the n-grams are aligned (see Section 2.2). We call this ‘free alignment’. In the third group we insist on the linguistically informed alignment of n-grams (see Section 2.5), dubbing this ‘forced alignment’. Within the category of gradual methods, we distinguish between ‘free alignment’ (see Section 2.6) and ’forced alignment’. Finally, for each of these methods, we consider both an unnormalized version of the measure as well as one normalized by length (see Section 2.3).

A measure can only be valid when it is consistent, but it may be consistent without being valid. Since consistency is a necessary condi-

	binary			gradual	
	no alignment	free alignment	forc. alignment	free alignment	forc. alignment
uni	0.69	0.66	0.66	0.66	0.66
bi	0.70	0.69	0.69	0.66	0.68
tri	0.71	0.70	0.72	0.66	0.73
xbi	0.70	0.69	0.72	0.67	0.73

Table 1: Correlations between perceptual distances and *unnormalized* string edit distance measurements among 15 Norwegian dialects. Higher coefficients indicate better results.

	binary			gradual	
	no alignment	free alignment	forc. alignment	free alignment	forc. alignment
uni	0.66	0.66	0.66	0.66	0.66
bi	0.67	0.67	0.67	0.66	0.66
tri	0.68	0.68	0.70	0.66	0.70
xbi	0.68	0.68	0.70	0.69	0.70

Table 2: Correlations between perceptual distances and different *normalized* string edit distance measurements among 15 Norwegian dialects. Higher coefficients indicate better results.

tion for validity, we check the consistency of phonetic distance methods. For each of the methods we calculated Cronbach's α values, which is based on the average inter-correlation among the words (Heeringa, 2004, pp. 170–173). A widely-accepted threshold in social science for an acceptable α is 0.70 (Nunnally, 1978). After the consistency check, we discuss validation results.

5.1 Norwegian Perception

In this section we first discuss results of unnormalized string edit distance measures, and will compare them with their normalized counterparts farther onwards in this section.

The Cronbach's α values of the unnormalized measurements vary from 0.84 to 0.87. The Cronbach's α values of the methods with 'forced alignment' are a bit lower than the α values of the other methods. An outlier arises when using the 'forced alignment' and gradual bigram distances: α=0.78, but these all indicate that the measurements are quite consistent.

We calculated correlations to the perceptual distances which are described in Section 4.1. Results are given in Table 1. Let's note that the effect size, i.e., the r value itself, is quite high, $0.66 < r < 0.73$, meaning that the various distance measure are accounting for 43.6–53.3% of the variance in the perception measurements. All of the correlation coefficients are massively significant ($p < 0.001$), but given the stringency of the Mantel test, they do not differ significantly from one another.

The correlations are quite similar. The maximal difference we found was 0.07, so that we conclude that none of the methods is strikingly better or worse in operationalizing the level of pronunciation difference that dialect speakers are sensitive to.

The small flood of numbers in Table 1 may seem confusing. Therefore we calculated averages per factor which are presented in Table 4. We invite the reader to refer to both Table 1 and Tablee 4 in following the discussion below. Table 4 shows systematic differences. For example, contextually sensitive measures (bigrams, trigrams, and xbigrams) are usually better (and never worse) than unigram measures. The differences among the different means of operationalizing context (bigrams, trigrams and xbigrams) seem unremarkable, however. Third, measures which are sensitive to linear order are slightly worse than those which are not (variants of DICE) on average[5]. But when comparing the first column in Table 1 with the others, we see that the highest correlations (0.73) are found among the order sensitive methods. Fourth, forcing alignment to respect vowel/consonant differences yields a modest improvement in scores. Fifth, we see no clear advantage in measurements which weight n-grams more sensitively to those binary comparison methods which distinguish only same and different.

Sixth, and most surprisingly, we can compare Table 1 which provides the correlation of edit distances which were *not* normalized for length, with Table 2, which provides the results of the measurements which *were* normalized. For some normalized measurements the Cronbach's α value are minimally higher (0.01). But comparison of the correlation coefficients shows that normalization never improves measurements, and often leads to a deterioration. In Table 4 averages for the normalized measurements are given. Normalized mea-

[5] When using the unnormalized versions of the 'DICE' family, the distance is just equal to the number of non-shared n-grams.

	binary			gradual	
	no alignment	free alignment	forc. alignment	free alignment	forc. alignment
uni	0.41	0.37	0.37	0.37	0.37
bi	0.37	0.35	0.37	0.36	0.35
tri	0.37	0.33	0.35	0.36	0.35
xbi	0.36	0.35	0.35	0.37	0.35

Table 3: Local incoherence values based on travel distances for the *unnormalized* string edit distance measurements between 15 Norwegian dialects. The lower the local incoherence value, the better the measurement technique.

surements display the same systematic differences that unnormalized measurements show, except for the differences between methods which consider the order of segments and methods which do not. Measures which are sensitive to linear order are slightly better than those which are not (variants of DICE).

5.2 Norwegian Geographic Sensitivity

As we mentioned in Section 4.2, Norway is very rugged. Therefore we based our local incoherence values on travel distances rather than on geographic distances "as the crow flies". We computed local incoherence values for both unnormalized and normalized string edit distance measurements. The comparison confirms the findings of Section 5.1: unnormalized methods always perform better than normalized ones. The unnormalized results are presented in Table 3.

Recall that lower local incoherence values should reflect better measurement techniques. When we examine the table as a whole, we note again that the various techniques are not hugely different—they perform with similar degrees of success.

In Table 4, we find average local incoherence values for the factors under investigation. We find first that contextually sensitive measures (bigrams, trigrams, and xbigrams) are again superior to unigram methods, and second, measures which are sensitive to linear order are superior to the DICE-like methods (unnormalized versions). Third, linguistically informed alignments, which respect the vowel/consonant distinction, perform better than uninformed ("free") alignment (for the normalized versions). Fourth, the average values do not suggest any benefit to the gradual weighting of n-grams in comparison with the binary weighting. Most surprisingly, normalization again appears to have a deleterious effect on the probity of the measurements.

We must stress again that these finer interpretations results require confirmation with a larger set of sites.

5.3 German Geographic Sensitivity

When checking the consistency of the German measurements we find Cronbach's α values of 0.95 and 0.96 for all methods without alignment or with 'free alignment' and for all unigram based methods. The higher Cronbach's α levels for this data set reflect the fact that it is larger. We find lower α values of 0.83–0.85 for the methods with 'forced alignment'. This accords with the consistency results for the Norwegian measurements.

When using bigrams, α is equal to 0.80 (binary, normalized), 0.51 (gradual, normalized), 0.74 (binary, unnormalized) and 0.45 (gradual, unnormalized). These low values are striking, and we found no explanation for them, but they suggest that we should not attach much significance to this combination of measurement properties. On average, the unnormalized α's are the same as the normalized α's.

Since consistency values are higher than 0.70 (with one exception), we validated the methods by calculating the geographic local incoherence values. We would have preferred to use perceptions, but we have no such data in the German case.

Since we found unnormalized string edit distance measurements superior to normalized ones in the Sections 5.1 and 5.2, we focus in this section on the unnormalized methods. Unnormalized results are shown in Table 5.

Recall that the lower the local incoherence value, the better the measurement technique. We include this table for the sake of completeness, but it is clear that the results do not jibe with the results obtained from the Norwegian data. Unigram-based processing appears to be superior, and context inferior; order-sensitive processing is inferior to order-insensitive processing, and linguistically informed ("forced") alignment appears to offer no advantage.

We leave the contrast between the Norwegian and German results as a puzzle to be addressed in future work, but it should be clear that we have

Factor	Correlation with perception		Local incoherence		Number of measurements
	raw	normalized	raw	normalized	
no order	0.70	0.67	0.38	0.45	4
order	0.69	0.68	0.36	0.46	16
unnormalized	0.69		0.36		20
normalized		0.68		0.43	20
binary	0.69	0.68	0.36	0.43	8
gradual	0.68	0.67	0.36	0.43	8
free	0.67	0.67	0.36	0.43	8
forced	0.70	0.68	0.36	0.42	8
unigram	0.67	0.66	0.38	0.45	5
bigram	0.68	0.67	0.36	0.45	5
trigram	0.70	0.68	0.35	0.42	5
xbigram	0.70	0.69	0.36	0.41	5

Table 4: Average correlations between perceptual distances and *raw*, i.e., *unnormalized* string edit distance measurements among 15 Norwegian dialects. Higher coefficients and lower local incoherence values indicate better results.

	binary			gradual	
	no alignment	free alignment	forc. alignment	free alignment	forc. alignment
uni	0.94	0.88	0.87	0.88	0.87
bi	1.00	0.98	2.09	0.92	5.71
tri	1.09	1.05	2.45	0.93	2.09
xbi	0.96	0.95	2.45	0.98	2.45

Table 5: Local incoherence values based on geographic distances for for the *unnormalized* string edit distance measurements 186 German dialects. The lower the local incoherence value, the better the measurement technique.

rather more confidence in the Norwegian than in the German results. This is due on the one had to the availability of independently behavioral data we can use to independently validate our computations, but also to the more stable set of values we see in the case of the Norwegian data. Exactly *why* the German data is so much more variable is also a question we must postpone to future work.

6 Conclusions and Prospects

In this paper we examined a range of string comparison algorithms by applying them to Norwegian and German dialect comparison. The Norwegian results suggest that sensitivity to linguistic context in the form of n-grams, and to linguistic structure in alignment improves measurement techniques, but they do not confirm the value of differential weighting for n-grams. The results mostly suggest that sensitivity to order of segments improves the measurements.

The larger German data likewise is unfortunately more recalcitrant (as are other data sets we have examined, but in which we have less confidence). A disadvantage of the German data may be that several transcribers were involved, working over a period of twenty years, and that two types of surveys were used, having different orders of sentences. There may be subtle differences in pronunciation as a result of subjects' becoming more relaxed or more impatient in the course of a survey interview.

On the other hand, the Norwegian data set is small (15 dialect sites). Our conclusions rely on assumptions of its quality and transcriber consistency, but this warrants further examination. We also cannot exclude the possibility that optimal measurements depend on features of the language and/or data set.

It is tempting to wish to redo this study using a large, antiseptically clean data set, transcribed reliably by a minimal number of phoneticians, but the more important practical direction may be to try to understand which properties of data sets are important in selecting variants of pronunciation distance measures. Atlases of material on language varieties simply are not always clean and reliable, and if we wish to contribute to their analysis, we must keep this in mind.

Acknowledgments

We are grateful to Therese Leinonen, Jens Moberg and Jelena Prokič for comments on this work, and in particular for their suggestion that one should also examine the length normalization. We also thank the workshop reviewers, in particularly one who was productively harsh about the treatment of normalization in an earlier version, and also pointed out literature we had insufficiently taken note of. Finally, we are indebted to the Netherlands Organization for Scientific Research, NWO, for support (project "Determinants of Dialect Variation, 360-70-120, P.I. J. Nerbonne)

References

Eric Bonnet and Yves Van de Peer. 2002. zt: A software tool for simple and partial Mantel tests. *Journal of Statistical Software*, 7(10):1–12. Available via: `http://www.jstatsoft.org/`.

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

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

Charlotte Gooskens. Traveling time as a predictor of linguistic distance. *Dialectologia et Geolinguistica*. submitted, 3/2004.

Charlotte Gooskens and Wilbert Heeringa. 2004. Perceptual evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, 16(3):189–207.

Joachim Göschel. 1992. Das Forschungsinstitut für Deutsche Sprache "Deutscher Sprachatlas. Wissenschaftlicher Bericht, Das Forschungsinstitut für Deutsche Sprache, Marburg.

Wilbert Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Ph.D. thesis, Rijksuniversiteit Groningen.

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

Diana Inkpen, O. Frunza, and Grzegorz Kondrak. 2005. Automatic Identification of Cognates and False Friends in French and English. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nicolai Nicolov, editors, *International Conference Recent Advances in Natural Language Processing*, pages 251–257, Borovets.

Brett Kessler. 1995. Computational dialectology in Irish Gaelic. In *Proc. of the European ACL*, pages 60–67, Dublin.

Brett Kessler. 2005. Phonetic comparision algorithms. *Transactions of the Philological Society*, 103(2):243–260.

Grzegorz Kondrak. 2005. N-gram similarity and distance. In *Proceedings of the Twelfth International Conference on String Processing and Information Retrieval (SPIRE 2005)*, pages 115–126, Buenos Aires, Argentina.

Joseph Kruskal. 1999. An overview of sequence comparison. In David Sankoff and Joseph Kruskal, editors, *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, pages 1–44. CSLI, Stanford. [1]1983.

Andrés Marzal and Enrique Vidal. 1993. Computation of normalized edit distance and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):926–932.

John Nerbonne and Peter 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.

John Nerbonne and Peter Kleiweg. 2006. Toward a dialectological yardstick. *Quantitative Linguistics*, 13. accepted.

Jum C. Nunnally. 1978. *Psychometric Theory*. McGraw-Hill, New York.

Jean Séguy. 1973. La dialectometrie dans l'atlas linguistique de gascogne. *Revue de Linguistique Romane*, 37:1–24.