Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data[1]

Charlotte Gooskens and Wilbert Heeringa

Abstract

The Levenshtein dialect distance method has proven to be a successful method for measuring phonetic distances between Dutch dialects. The aim of the present investigation is to validate the Levenshtein dialect distance with perceptual data from a language area other than the Dutch, namely Norway. We calculate the correlation between the Levenshtein distances and the distances between 15 Norwegian dialects as judged by Norwegian listeners. We carry out this analysis in order to see the degree to which the average Levenshtein distances correspond to the psycho-acoustic perception of the speakers of the dialects.

1. Introduction

In 1995 Kessler introduced the use of the Levenshtein distance as a tool for measuring linguistic distances between language varieties. He applied the algorithm to the comparison of Irish dialects. The Levenshtein distance is a string edit distance measure. On the basis of linguistic distances between dialectal varieties dialect areas can be found. More innovative is the possibility of drawing dialect maps which reflect the fact that dialects areas should be considered as continua and not as areas separated by sharp borders. Its application to the Dutch language area has produced convincing results (see Nerbonne and Heeringa 1998 and Heeringa 2004). The results are partly similar to the map of Daan and Blok 1969, which may be considered as the most authoritative Dutch dialect map up to now. Still, it is desirable to validate the method further.

In this paper we validate the Levenshtein distance. We will investigate to what extent dialect distances found with Levenshtein distance correlate with distances as perceived by the dialect speakers themselves. We will try to find an answer to the following question: may Levenshtein distance-based dialect distances be considered as a good approximation of the perceptual distances? To answer this question we will use a set of 15 Norwegian varieties. Results for Dutch may be impressive, but the Dutch dialect area is a flat, regularly populated landscape. In contrast with this, the Norwegian dialect area is less regular, due to the mountains. This may make the test harder but more revealing.

In Section 2 the data is described on the basis of which both the perception experiment and the Levenshtein distance measurements were performed. In Section 3 the perception experiment will be presented, which was carried out in order to calculate the psycho-acoustic distances between 15 Norwegian dialects. In Section 4 the Levenshtein distance will be presented and applied to data from the same 15 Norwegian dialects. In Section 5 the results of the two kinds of distance measures will be compared and explanations for the results will be suggested in Section 6. Finally some general conclusions will be drawn in Section 7.

2. Material

In order to carry out our investigation we needed to obtain suitable material. This means that we needed to have access to recordings of the same the same text in a fair number of dialects from one language area in order to carry out the perception experiments described in Section 3. At the same time we need digitized transcriptions in a form which could be used in already existing computer programs for calculating the Levenshtein distances described in Section 4.

Dialects

We chose to focus on the Scandinavian language area since the Scandinavian countries have a strong tradition of research in the area of dialectology. This has resulted in maps similar to the traditional Dutch dialect maps (for an overview see Skjekkeland 1997). These maps are useful for the interpretation of the results. Norway seems to be particularly interesting because of the strong position of the dialects in this country. In contrast to many European countries the dialect is used by people of all ages and social backgrounds not only in the private domain but also in official contexts (Omdal 1995). This makes it easy to use recent recordings of young people from all over the country without the risk that some of the speakers might use a more standardized variant of their dialect or a variety which is no longer being used in every day life. Also, it does not feel unnatural for Norwegian people to read aloud a text in their own dialect. This allowed us to use read texts, which was necessary since we needed the same text in different dialects. In Figure 1 the fifteen dialects which were used in the investigation are shown. The dialects are spread over a large part of the Norwegian language area, and cover most major dialect areas as found on the traditional map of Skjekkeland (1997, p. 276). On this map the Norwegian language area is divided into nine dialect areas. In our set of 15 varieties six areas are represented.

Figure 1 here

Figure 1. *Map of Norway showing the 15 dialects used in the present investigation. The abbreviation after the name of each location indicates the dialect area to which the variety belongs according to Skjekkeland (1997). The same abbreviations are used in other figures in this paper. Skjekkeland (1997) also gives a more global division in which Norwegian dialects are divided in Vestnorsk (covering No, Sv and Nv) and Austnorsk (covering Mi, Au and Tr).*

Text

It is time consuming to make recordings of dialects and to transcribe the texts phonetically. Fortunately, we were able to use already existing recordings of Norwegian dialect speakers. The speakers all read aloud the same text, namely the fable ‘The North Wind and the Sun’.[2] This text has often been used for phonetic investigations; see for example the *International Phonetic Association* (1949 and 1999) where the same text has been transcribed in a large number of different languages.

Speakers

There were 4 male and 11 female speakers. Thirteen of the speakers had filled in a questionnaire about their background. From this we know that the average age of these speakers was 30.5 years, ranging between 22 and 35, except for one speaker who was 66. All thirteen speakers attended university or already had a university degree. No formal testing of the degree to which the speakers used their own dialect was carried out. However, they had lived at the place where the dialect is spoken until the mean age of 20 (with a minimum of 18) and they all regarded themselves as representative speakers of the dialects in question. All speakers except one had at least one parent speaking the dialect.

Recordings
The recordings were made in a soundproof studio in the autumn of 1999 and the spring of 2000. The speakers were all given the text in Norwegian beforehand and were allowed time to prepare the recordings in order to be able to read aloud the text in their own dialect. Many speakers had to change some words of the original text in order for the dialect to sound authentic. The word order was changed in three cases (see Section 3). When reading the text aloud the speakers were asked to imagine that they were reading the text to someone with the same dialectal background as themselves. This was done in order to ensure a reading style which was as natural as possible and to achieve dialectal correctness.

The microphone used for the recordings was a MILAB LSR-1000 and the recordings were made in DAT format using a FOSTEX D-10 Digital Master Recorder. They were edited by means of Cool Edit 96 and made available at the World Wide Web.

The recordings were used in the perception experiment, which is described in Section 3.

Transcriptions
On the basis of the recordings phonetic transcriptions were made of all 15 dialects. The transcriptions were made in IPA as well as in X-SAMPA (Speech Assessment Methods Phonetic Alphabet). This is a machine readable phonetic alphabet which is still human readable. Basically, it maps IPA-symbols to the 7 bit printable ASCII/ANSI characters. All transcriptions were made by the same person which ensures consistency. Most Norwegian dialects distinguish between two tonal patterns on the word level, often referred to as tonemes. Some dialects even have a third toneme, the circumflex (e.g. Kristoffersen 2000). In our material four dialects (Bjugn, Fræna, Verdal and Stjørdal) have circumflex tonemes on one word (*mann* meaning ‘man’). In the transcriptions toneme transcriptions were included, i.e. it was indicated when the different tonemes occurred in the text. We know from the literature that the realization of the tonemes can vary considerable across the Norwegian dialects. However, no information was given about the precise realization of the tonemes in the transcriptions.

The Levenshtein distance measurements are based on the transcriptions (see Section 4).

3. Perceptual distance measurements

Perceptual data have often been used in dialectology (e.g. Daan and Blok 1969, Gooskens 1997, Preston 1999) and have proved that listeners without linguistic training are quite able to make judgments about for example distances between dialects. Perceived linguistic distances are likely to be at least partly based on objective linguistic distances. However, a number of factors other than linguistic distances might influence the way which listeners perceive distances between dialects.

We will return to this point in Section 6. In order to be able to investigate how well the Levenshtein distances correspond to the perceived linguistic distances we carried out a perception experiment on the basis of 15 Norwegian dialects. In Section 3.1 we will describe the listening experiment, and in Section 3.2 the results will be presented.

3.1 Experiment

Manipulations
In order to be able to investigate the dialect distances between 15 Norwegian dialects as perceived by Norwegian listeners, for each of the 15 varieties the corresponding recording of the fable 'The North Wind and the Sun' was presented to Norwegian listeners in a listening experiment. The running text provides the listeners with more kinds of information than the information used for the calculation of the Levenshtein distances. One important difference is that the listeners based their judgments on spoken material which contains prosody while this is not the case for the Levenshtein distances. For this reason we decided to include a monotonized version of all fragments. Since in these fragments the pitch contour is not present like in the Levenshtein distances we expect the correlation of these two distance measures to be higher than when Levenshtein distances are correlated with the original fragments.

In the listening experiment described below each of the 15 dialect recordings were presented in the following two versions:

1. **Monotonized version.** By means of electronic monotonization the intonation (including word tones) is removed from the signal.
2. **Original version.** This version has the original prosodic and verbal information, but is processed in the same way as the monotonized version.

The manipulations were carried out with the program PRAAT[3]. In order to monotonize the fragments the pitch contours were changed into flat lines. The recordings of female speakers were monotonized at 224 Hz, which is the mean pitch of the 11 female speakers. The recording of the male speakers were monotonized at 134 Hz. This was the mean pitch of the three male speakers.

Listeners
The listeners were 15 groups of high school pupils, one group from each of the places where the 15 dialects are spoken (see Figure 1). Each group consisted of 16 to 27 listeners (with a mean of 19). Their mean age was 17.8 years, 52 percent were female and 48 percent male. Only the responses of listeners who had lived the major part of their lives in the place where the dialect is spoken were used for the analysis. On average these listeners had lived in the place in question for 16.7 years. Nine of the 290 listeners (3%) said that they never speak the dialect, the rest spoke the dialect always (60%), often (21%), or seldom (16%). A large majority of the listeners (83%) had one or two parents who also spoke the dialect.

Procedure
The two versions (monotonous and original) of the 15 dialects were presented in two blocks, with the dialects randomized within each block. First the block with the monotonized version was presented and after a short break the block with the original version was presented. Each block was preceded by a practice recording (a speaker

from Stjørdal, but not one of the 15 recordings used in the real experiment). Between each two recordings there was a pause of 3 seconds.

While listening to the dialects the listeners were asked to judge each dialect on a scale from 1 (similar to own dialect) to 10 (not similar to own dialect). The whole experiment lasted approximately 20 minutes followed by a questionnaire. In this questionnaire the listeners were asked questions about their individual characteristics, such as language background, age and sex. The listeners were paid for their participation.

3.2 Results

Distances

In Table 2 the mean perceptual distances between the 15 Norwegian dialects are presented, obtained on the basis of the experiment in which the original, non-manipulated recordings were presented. 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 × 15 distances. The fact that the listeners also had to judge their own dialect resulted in varying diagonals (between 1.0 and 3.4). Some groups of listeners judged the recorded sample of the own dialect to be more than minimally distant. This might be explained by the fact that the recorded speakers were not equally representative for the dialect in question. It might however also be the case that some dialects show more variation than others. Finally, the differences can also be caused by the fact that the groups of listeners differ in some respect. For example, some groups might be more familiar with their own dialects than others or more tolerant as to what they are willing to accept as a good representation of their dialect.

There are two mean distances between each pair of dialects. For example the distance which the listeners from Bergen perceived between their own dialect and the dialect of Trondheim (mean judgment is 7.8) is different from the distance as perceived by the listeners from Trondheim (mean judgment is 8.6). Different explanations can be given for the fact that different groups perceive the same linguistic distances differently. For example it is likely that the attitude towards a dialect influence the perception of the linguistic distance. We will return to this point in Section 6.

Table 1. *Mean perceptual distances between all pairs of 15 Norwegian dialects as perceived by 15 groups of listeners when listening to the non-manipulated recordings (judged on scale from 1 = similar to own dialect 10 = not similar to own dialect).*

Table 1 here

Classification

On the basis of the distance matrix the dialects can be classified with cluster analysis. The goal of a cluster analysis is to identify the main groups. The groups are called *clusters*. Clusters may consist of subclusters, and subclusters may in turn consist of subsubclusters, etc. The result is a hierarchically structured tree in which the dialects are the leaves (Jain and Dubes, 1988). Several alternatives exist. We used the *Unweighted Pair Group Method using Arithmetic averages* (UPGMA), since we found that dendrograms generated by this method reflected distances which correlated most strongly with the original distances in the distance matrix (see Sokal and Rohlf 1962).

Since the cluster program expects only one value for each pair of different elements, distances of dialects with respect to themselves are not used, and the average of the two mean distances is used when classifying the varieties. E.g. the average of the distance Bergen-Trondheim and Trondheim-Bergen is used.

The dendrogram (Figure 2) is obtained on the basis of Table 1 and accords rather well with the map of Skjekkeland (see Figure 1). Sørvestlandsk, Austlandsk and Trøndsk groups can clearly be identified. However, the Midlandsk dialects, Bø and Lesja, do not form a close cluster. Geographically they are rather distant, so they may be rather different although they should be in the same group according to the traditional division. The Nordvestlandsk dialects (Fræna and Herøy) seem to be very different from each other, although they are geographically rather close. Possibly the fact that these dialects belong to the same group on the map of Skjekkeland may be explained by the fact that Skjekkeland based the characterization on limited number of phenomena which are (partly) different from those found in the text 'The North Wind and the Sun'. In our sample the Nordlandsk area is represented by only one variety (Bodø). This variety is grouped with the varieties of the Trøndsk area, which is not unexpected geographically.

Figure 2 here

Figure 2. *Dendrogram derived from the 15 × 15 matrix of perceptual distances showing the clustering of (groups of) Norwegian dialects. On the horizontal scale distances are given in the scale as used by the listeners.*

4. Levenshtein distance measurements

4.1 Method

Traditional dialectology has aimed to divide language areas into dialect areas mostly by drawing sharp borders between the areas on a map. The choice of the borders has often been based on the knowledge and intuition of the investigators of the areas in question. The application of isoglosses has been another widely used means of dividing language areas into dialect areas. Coinciding isoglosses are interpreted as borders. However, the use of isoglosses gives rise to a number of problems. First isoglosses do not always coincide. They can run parallel, forming vague bundles, or even cross each other, describing contradictory binary divisions. In practice, well-known isoglosses which form bundles are selected, but this makes this aspect of the method subjective. Second, the use of isoglosses gives a very categorical view of dialect differences: either a dialect is different from another dialect or it is not, no degrees of differences can be expressed. Finally, dialects might be dispersed by migration or war so that closely related dialects are no longer adjacent to each other. This causes problems when drawing the isoglosses and borders on the dialect map (see further Chambers & Trudgill 1998, pp. 89-103, Kessler 1995).

In order to solve some of the problems outlined above, several (computational) methods for measuring the linguistic distances between language varieties has been developed since the beginning of the seventies of the previous century (Heeringa 2004, pp. 14-24). In the present investigation we wish to evaluate one of the methods, the Levenshtein distance method, which has been applied successfully to Irish Gaelic

(Kessler 1995) and Dutch dialects (Nerbonne and Heeringa 1998, Heeringa 2004, pp. 213-278). The basic algorithm had been described in detail in Kruskal 1999. Compared to traditional methods (for instance the isogloss method), this approach has the advantage that varieties are compared and classified in an objective way and on the basis of the aggregate of many phenomena rather than on the basis of just single phenomena. In contrast to other computational methods, Levenshtein distance yields gradual word pronunciation differences, and the method uses the data exhaustively, which makes it most sensitive.

Algorithm

Using the Levenshtein distance, two dialects are compared by comparing the pronunciation of words in the first dialect with the pronunciation of the same words in the second. It is determined how one pronunciation is changed into the other by inserting, deleting or substituting sounds. Weights are assigned to these three operations. In the simplest form of the algorithm, all operations have the same cost, e.g. 1. Assume *afternoon* is pronounced as [ˈæəftəˌnʉˑn] in the dialect of Savannah, Georgia, and as [ˌæftərˈnuˑn] in the dialect of Lancaster, Pennsylvania[4]. Changing one pronunciation into the other can be done as follows (ignoring suprasegmentals and diacritics for this moment)[5]:

æəftənʉn	delete ə	1
æftənʉn	insert r	1
æftərnʉn	subst. ʉ/u	1
æftərnun		
		3

In fact many sequence operations map [ˈæəftəˌnʉˑn] to [ˌæftərˈnuˑn]. The power of the Levenshtein algorithm is that it always finds the cost of the cheapest mapping.

Comparing pronunciations in this way, the distance between longer pronunciations will generally be greater than the distance between shorter pronunciations. The longer the pronunciation, the greater the chance for differences with respect to the corresponding pronunciation in another variety. Because this does not accord with the idea that words are linguistic units, the sum of the operations is divided by the length of the longest alignment which gives the minimum cost. The longest alignment has the greatest number of matches. In our example we have the following alignment:

æ	ə	f	t	ə		n	ʉ	n
æ		f	t	ə	r	n	u	n
	1				1		1	

The total cost of 3 (1+1+1) is now divided by the length of 9. This gives a word distance of 0.33 or 33%.

Gradual weights

The simplest versions of this method are based on a notion of phonetic distance in which phonetic overlap is binary: non-identical phones contribute to phonetic distance, identical ones do not. Thus the pair [a,p] counts as different to the same degree as [b,p]. In more sensitive versions phones are compared on the basis of their feature values, so the pair [a,p] counts as more different than [b,p]. However, it is not always clear which weight should be attributed to the different features. The version which we use in this paper is based on the comparison of spectrograms of the sounds. A spectrogram is the visual representation of the acoustical signal and the visual differences between the spectrograms are reflections of the acoustical differences. When using spectrograms it is not necessary to make decisions about the weight of the different features. The spectrograms were made on the basis of recordings of the sounds of the International Phonetic Alphabet as pronounced by John Wells and Jill House on the cassette *The Sounds of the International Phonetic Alphabet* from 1995[6]. The different sounds were isolated from the recordings and monotonized at the mean pitch of each of the two speakers with the program PRAAT (see note 3). Next, with PRAAT a spectrogram was made for each sound using the so-called Barkfilter which is a more perceptual oriented model. On the basis of the Barkfilter representation, segment distances were calculated. The way in which this was done is described extensively in Heeringa 2004 pp. 79-119, and more briefly in Gooskens and Heeringa 2004.

Logarithmic weights

In perception, small differences in pronunciation may play a relatively strong role in comparison to larger differences. Therefore we used logarithmic segment distances. The effect of using logarithmic distances is that small distances are weighed relatively more heavily than large distances. Since the logarithm of 0 is not defined, and the logarithm of 1 is 0, distances are increased by 1 before the logarithm is calculated. To obtain percentages, we calculate:

$$(ln(distance + 1) / ln(maximum\ distance + 1)) \times 100$$

Allowed matches

To reckon with syllabification in words, the Levenshtein algorithm is adapted so that only a vowel may match with a vowel, a consonant with a consonant, the [j] or [w] with a vowel (or opposite), the [i] or [u] with a consonant (or opposite), and a central vowel (in our research only the schwa) with a sonorant (or opposite). In this way unlikely matches (e.g. a [p] with a [a]) are prevented.

4.2 Results

Distances

The Norwegian text consists of 58 different words, which proved to be a sufficient basis for a reliable Levenshtein analysis (Cronbach's α =0.86, see Heeringa 2004, pp. 170-173). Some words occur more than once in the text. In these cases the mean distance over the variants of the word is used for calculating the Levenshtein distances (see Heeringa 2004, pp. 134-135 for more details). So when comparing two dialects we get 58 Levenshtein distances. Now the dialect distance is equal to the sum of 58 Levenshtein distances divided by 58. When the word distances are presented in terms of percentages, the dialect distance will also be presented in terms of percentages. All distances between the 15 language varieties are arranged in a 15 $\times$ 15 matrix. In Table 2 the average Levenshtein distances between the 15 dialects are presented. The diagonal is always zero and the lower half is the mirror image of the upper half.

Table 2. *Average Levenshtein distances between all pairs of 15 Norwegian dialects given as percentages.*

Table 2 here

Classification

Just as we did on the basis of the perceptual distances, we performed cluster analysis on the basis of the average Levenshtein distances as well. Since the matrix is symmetric, only one half is used, while the zero values on the diagonal from upper left to lower right are not used.

Comparing our computational dendrogram (Figure 3) with the perceptual dendrogram (see Figure 2), both show an Austlandsk group which contains the varieties of Larvik, Halden, Lillehammer and Borre, and a Trøndsk group which contains the varieties of Verdal, Bjugn and Stjørdal. Although the two dendrograms do not cluster the Midlandsk varieties (Bø and Lesja) as one group, in the perceptual dendrogram they appear to be more related than in the computational dendrogram. In the perceptual dendrogram the Midlandsk dialect of Lesja is clustered with the Austlandsk varieties, although not very close. In the computational dendrogram this dialect belongs to the Trøndsk varieties. Geographically the variety is located about midway between the two areas. In both the perceptual and computational dendrogram Bø is clustered with the Austlandsk varieties, but in the perceptual dendrogram the relation appears to be stronger. The Sørvestlandsk varieties of Bergen and Time form one (rather loose) cluster in the perceptual dendrogram. In the computational dendrogram they do not form one cluster. In the two dendrograms the two Nordvestlandsk varieties do not form one cluster. In both, Fræna is clustered with the Trøndsk varieties. However, Herøy is clustered with the Sørvestlandsk varieties in the perceptual dendrogram, while in the computational dendrogram it belongs to none of the groups, but appears to be distinct from all the other varieties. In both dendrograms Bodø is clustered with the Trøndsk varieties. However, in the computational dendrogram Bodø looks as if it were closer to the Trøndsk varieties than in the perceptual dendrogram. However, the cluster with Verdal, Bjugn and Stjørdal is geographically not impossible. A striking difference can be found with regard to the dialect of Trondheim, which is clustered with the Trøndsk varieties in the perceptual dendrogram, but in the computational dendrogram it is clustered with Austlandsk varieties. Possibly the listeners recognized the recording of Trondheim as the dialect of Trondheim and let geography influence their judgments. However, the dialect of larger cities may be in contrast with their surrounding and more related to geographically more distant varieties. We conclude that the two dendrograms are rather similar, due to the fact that especially the closer clusters in the one dendrogram are also found in the other one.

Figure 3 here

Figure 3. *Dendrogram derived from the 15 × 15 matrix of average Levenshtein distances showing the clustering of (groups of) Norwegian dialects. The scale distance is given as a percentage.*

5. Perceptual versus Levenshtein distances

The aim of the present study was to validate the Levenshtein method by investigating the degree to which the Levenshtein distances between 15 Norwegian dialects correlate with the same distances as perceived by groups of listeners from the 15 places where the dialects are spoken. As already became clear above, the dendrograms (Figures 2 and 3) show many similarities.

A measure of the degree of similarity is the correlation coefficient between the perceptual distances and the Levenshtein distances. For finding the correlation coefficient, we used the Pearson's correlation coefficient (Sneath and Sokal, 1973, pp. 137-140). We correlated the average Levenshtein distances with both perceptual distances based on monotonized recordings and perceptual distances based on the original recordings. In the first case we found a correlation r=0.78, and in the second case we found a correlation r=0.80. For finding the significance of a correlation coefficient we used the Mantel test (see Heeringa 2004, pp. 74-75 for more details). For both cases we found that p<0.001, so the correlations are significant, at least at the level α=0.001. In Section 6 we will explain why the average Levenshtein distances correlate better with the original recordings-based perceptual distances than with the monotonized recordings-based perceptual distances.

Figure 4 here

Figure 4. *Scatterplot showing perceptual distances versus average Levenshtein distances, including distances of dialects with respect to themselves (r=0.80, p<0.001).*

In Figure 4 a scatterplot is shown in which the average Levenshtein distances are plotted against the sorted perceptual distances based on the original recordings. In the lower left corner of this graph, we find 15 dots on a line, for which the average Levenshtein distance is equal to 0. The dots correspond with the values in the diagonal from upper left to lower right in Table 1 en 2. In the graph the 15 dots appears to be outliers, which may increase the correlation coefficient mistakenly. Therefore we calculated correlation coefficients again excluding distances of varieties with respect to themselves. In that way we get a correlation of 0.62 (perceptual distances based on monotonized fragments) and 0.67 (perceptual distances based on original distances). In both cases p<0.001. Although these correlation coefficients are lower, they are still quite significant. This shows that the Levenshtein distances are a good representation of the distances between dialects as perceived by listeners. On the other hand it also shows that listeners base their judgments of dialectal distances on the linguistic information used in the algorithm to a great extent.

Figure 5 here

Figure 5. *Scatterplot showing perceptual distances versus average Levenshtein distances, excluding distances of dialects with respect to themselves (r=0.67, p<0.001). The linear regression line is shown as well.*

6. Factors influencing the correlation between perceptual distances and Levenshtein distances

Even though the correlation between the Levenshtein distances and the perceptual distances turned out to be high, it is still interesting to look for explanations for the fact that the correlation is not perfect. On the one hand there are several non-linguistic factors which might influence the perceptive judgments of the distances and which might result in a negative influence on the correlation. Such factors could be the attitude of the listeners towards the different dialects and their knowledge about the geographical position of the dialects.

On the other hand there are also several linguistic factors which might influence the correlations. When judging the dialects the listeners had all linguistic information (lexical, phonetic, intonational, morphological and syntactical) at their disposal because they are confronted with recordings of spoken texts. However, the Levenshtein distances are calculated only on the basis of lexical, phonetic and morphological material. Therefore we discuss the intonation and syntax in more detail in Section 6.1 and 6.2 respectively.

6.1 Intonation

Intonation is one of the most important characteristics of the different Norwegian dialect areas (Christiansen 1954; Fintoft and Mjaavatn 1980; Leitre, Lundeby and Torvik 1981; Sandøy 1993) and it can be expected to play an important role in the perception of the distances between the 15 dialects. Minimal word pairs can be distinguished by means of tonemes (toneme I, toneme II, and in some dialects circumflex) at the accented syllables. The use of tonemes and the precise pitch contour of the tonemes may differ per dialect.

Using Levenshtein distance, tonemes are not processed. Therefore listeners in the perception experiment were first asked to give judgments on the basis of monotonized recordings, and next on the basis of original recordings. It is striking that the Levenshtein distances correlate stronger with the perceptual distances based on unmodified recordings than with the perceptual distances based on monotonized recordings (see Section 5). When looking at the two perceptual matrices, it appeared that the mean judgments are almost the same (7.19 for the monotonous fragments and 7.25 for the original fragments). However, the standard deviation is smaller in the case of the monotonous fragments (1.38) than in the case of the original fragments (1.68). Three explanations suggest themselves.

First the absence of intonation yields unnatural speech. In particular the absence of intonation makes tonemes imperceptible in Norwegian which makes the fragments even more unusual. The consequence may be that this makes listeners insecure. This leads to ‘safe’ judgments, resulting in values which are found closer to the middle of the scale.

Second the lower standard deviation for the monotonous distances may have to do with the setup of the experiment. After the first session the listeners know the extremes, i.e., the most similar and most different varieties. This knowledge may be used when judging distances in the second session.

Third it is also possible that the results do indeed reflect the distances as perceived by the listeners, dialects close to the listeners own dialect being perceived as more deviant and the dialects which are very deviant being perceived as less deviant when there is no information present about intonation.

However, no matter which explanation is correct we can establish that the dispersion of the data is smaller in the case of the monotonous fragments than the original fragments. The representation on a smaller scale is less precise. This seems to us the explanation for the fact that the correlation with the Levenshtein distances is lower for the monotonous fragments.

6.2 Syntax

As far as the syntactical differences are concerned there are hardly any differences between the 15 dialects. In a number of cases an adverb has been moved from the beginning to the end of the sentence. Little research has been carried out on syntactic differences between Norwegian dialects. However, the placement of the adverbs in our material does not seem to be characteristic of the dialects in question. Therefore the fact that syntactic differences are not reflected in the average Levenshtein distances is probably not the main explanation for the fact that no perfect correlation between perceptual and average Levenshtein distances is found.

7. Conclusions

The aim of the present investigation was to validate the Levenshtein distance in a language area other than the flat Dutch area by comparing the Levenshtein distances with comparable distances as perceived by listeners from the places where the dialects are spoken. 15 Norwegian dialects were included in the study. Perceptual distances were obtained on the basis of a perception experiment, and comparable distances were calculated using Levenshtein distance.

On the basis of both the perceptual and the Levenshtein distances the 15 varieties were classified. Although differences can be found, in general the two classifications are rather similar. In both a north-south division is found. The northern cluster is dominated by a group of central varieties, and the southern cluster by a group of southeastern varieties.

In order to validate the Levenshtein distances, they were correlated with perceptual distances. Prosody plays an important role in Norwegian dialects but is not processed when using Levenshtein distance. Therefore the Levenshtein distances were correlated with perceptual distances which were obtained on the basis of an experiment in which monotonized recordings were used, and with perceptual distances obtained on the basis of an experiment in which original, non-manipulated recordings were used. In both cases, we got a high, strongly significant correlation ($r=0.62$ and $r=0.67$ respectively, $p<0.001$ for the two cases). This shows that dialect distances calculated with Levenshtein distance approximate perceptual distances rather well. This we see as a confirmation of the usefulness of the Levenshtein method as has been shown before for Dutch dialects. Now we know that the method is also applicable in a language area with a less simple geographic situation than the Dutch one.

Intonation has been repeatedly mentioned as a very important cue for the perceptual differentiation between Norwegian dialects (see Section 3). Intonational cues are not represented in the Levenshtein distances and therefore correlation with perceptual data might be expected to be higher when intonation is removed from the data. However, we found that the Levenshtein distances correlate stronger with the original recordings-based perceptual distances than with the monotonized recordings-based perceptual distances. We argued that this might be attributed to methodological deficiencies.

References

Chambers, J. K. & Trudgill, Peter. (1998). *Dialectology*. 2nd edition. Cambridge: Cambridge University Press.

Christiansen, Hallfrid. (1954). Hovedinndelingen av norske dialekter. In: *Maal og Minne*: 30-41.

Daan, Jo & Blok, D. P. (1969). Van randstad tot landrand. Toelichting bij de kaart: dialecten en naamkunde. In: *Bijdragen en mededelingen der dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*, 37, 9-41. Amsterdam: Noord-Hollandse Uitgeversmaatschappij.

Fintoft, Knut & Mjaavatn, Per-Egil. (1980). Tonelagskurver som målmerke. In: Maal og Minne 1980: 66-87.

Gooskens, Charlotte. (1997). On the role of prosodic and verbal information in the perception of Dutch and English language varieties. Doctoral dissertation, University of Nijmegen, the Netherlands.

Gooskens, Charlotte & Heeringa, Wilbert. (2004). The position of Frisian in the Germanic language area. In: Gilbert, Dicky & Schreuder, Maartje & Knevel, Nienke (eds.), *On the boundaries of phonology and phonetics*, 61-87. Groningen: Klankleergroep, Faculty of Arts, University of Groningen. Dedicated to Tjeerd de Graaf. Available at: http://www.let.rug.nl/~heeringa/dialectology/ papers/.

Heeringa, Wilbert. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Doctoral dissertation. University of Groningen. Available at: http://www.let.rug.nl/~heeringa/dialectology/thesis/.

International Phonetic Association. (1949). *The principles of the International Phonetic Association: being a description of the International Phonetic Alphabet and the manner of using it, illustrated by texts in 51 languages*. London: International Phonetic Association.

International Phonetic Association. (1999). *Handbook of the International Phonetic Association : a guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

Jain, Anil K. & Dunes, Richard C. (1988). *Algorithms for Clustering Data*. Englewood Cliffs, N.J. : Prentice Hall.

Kessler, Brett. (1995). Computational dialectology in Irish Gaelic. In: *Proceedings of the European ACL*, 60-67. Dublin: ACL.

Kristoffersen, Gjert. (2000). *The phonology of Norwegian*. Oxford: Oxford University Press.

Kruskal, Joseph B. (1999). An overview of sequence comparison. In : David Sankoff & Joseph Kruskal (eds.). *Time Warps, String edists, and Macro molecules; The Theory and Practice of Sequence Comparison*. Stanford: CSLI, 1-44.

Leitre, Arild; Lundeby, Einar & Torvik, Ingvald. (1981). *Språket vårt før og no. Språkhistorie, norrønt, dialektar, nyislandsk*. Hestholm: Gyldendal Norsk Forlag.

Nerbonne, John & Heeringa, Wilbert. (1998). Computationele vergelijking en classificatie van dialecten. In: *Taal en Tongval, Tijdschrift voor Dialectologie*, 50(2), 164-193. Available at: http://www.let.rug.nl/~heeringa/dialectology/ papers/.

Omdal, Helge. (1995). Attitudes toward spoken and written Norwegian. In: *International journal of the sociology of language*, 115, 85-106.

Preston, Dennis. (1999). *Handbook of perceptual dialectology*. Amsterdam: Benjamins.

Sandøy, Helge. (1993). *Talemål*. Oslo: Novus Forlag.

Skjekkeland, Martin. (1997). *Dei norske dialektane. Tradisjonelle særdrag i jamføring med skriftmåla.* Kristiansand: Høyskoleforlaget.

Sneath P. H. A. & Sokal R. R. (1973). *Numerical Taxonomy, A Series of Books in Biology*. W. H. Freeman and Company, San Francisco.

Sokal, R.R. & Rohlf, F.J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11, 33-40.

Tables

	Ber	Bju	Bod	Bø	Bor	Fræ	Hal	Her	Lar	Les	Lil	Stj	Tim	Tro	Ver
Bergen	1.79	9.07	8.25	8.00	7.75	7.70	8.20	6.95	8.06	8.95	8.57	8.42	4.88	8.55	8.05
Bjugn	9.16	3.44	6.44	8.26	9.29	5.80	8.30	8.05	8.44	7.32	9.10	2.21	8.00	3.30	2.85
Bodø	8.79	7.93	1.50	8.32	8.35	6.60	7.90	7.84	7.39	8.05	8.76	6.63	8.19	6.20	6.30
Bø	8.11	7.81	7.56	1.00	7.76	8.10	4.95	7.89	5.39	6.00	5.19	7.16	6.31	8.25	8.65
Borre	6.11	8.85	7.81	6.53	1.76	8.55	1.80	7.58	1.61	7.53	2.04	7.26	7.50	8.55	9.10
Fræna	9.00	7.59	7.13	8.47	8.82	3.10	8.10	7.89	8.50	7.26	9.00	6.68	7.44	6.10	7.65
Halden	7.00	8.22	8.00	6.84	4.00	8.15	2.80	7.95	2.89	6.63	3.00	7.47	7.06	8.05	8.32
Herøy	8.63	9.37	8.44	8.53	9.18	7.05	8.65	1.26	9.33	9.32	9.48	8.58	7.50	7.50	8.22
Larvik	7.47	8.70	7.69	4.05	4.06	7.75	3.25	5.61	3.44	7.16	4.67	8.21	6.88	8.35	7.55
Lesja	8.58	7.63	7.88	7.42	8.24	7.30	7.60	7.79	7.67	1.00	7.10	6.95	7.25	7.70	8.22
Lillehammer	6.78	8.33	8.13	6.26	4.47	8.05	3.10	7.53	4.11	7.32	2.76	7.68	6.88	8.70	8.16
Stjørdal	8.74	3.73	6.81	7.79	8.18	6.05	7.55	7.79	8.35	7.16	8.38	2.05	7.75	3.85	3.42
Time	7.00	9.33	8.44	8.11	8.47	8.30	8.05	7.22	8.22	9.11	8.81	8.89	1.81	8.80	9.05
Trondheim	7.84	5.89	6.75	7.53	6.47	7.35	6.05	7.16	5.94	7.94	6.33	4.47	7.63	3.35	6.84
Verdal	8.89	3.41	6.44	8.26	8.41	5.70	7.25	7.95	7.94	7.42	8.48	1.89	7.94	3.15	2.63

Table 1. *Mean perceptual distances between all pairs of 15 Norwegian dialects as perceived by 15 groups of listeners when listening to the non-manipulated recordings (judged on scale from 1 = similar to own dialect 10 = not similar to own dialect).*

	Ber	Bju	Bod	Bø	Bor	Fræ	Hal	Her	Lar	Les	Lil	Stj	Tim	Tro	Ver
Bergen	00.0	34.9	31.3	35.6	27.5	35.1	26.6	41.7	28.7	39.8	24.7	39.2	27.7	31.2	37.8
Bjugn	34.9	00.0	23.2	32.1	29.4	26.1	28.4	32.6	28.1	25.9	28.5	20.0	37.0	23.8	16.9
Bodø	31.3	23.2	00.0	33.1	28.6	30.8	27.8	34.9	23.1	30.2	26.7	27.2	34.2	27.5	27.7
Bø	35.6	32.1	33.1	00.0	28.5	37.9	27.0	31.3	27.9	30.9	28.8	39.3	33.0	30.6	34.0
Borre	27.5	29.4	28.6	28.5	00.0	38.8	17.5	39.7	21.3	36.3	15.0	39.0	31.1	25.6	32.8
Fræna	35.1	26.1	30.8	37.9	38.8	00.0	36.1	31.7	35.2	29.6	37.2	28.5	37.9	33.1	29.5
Halden	26.6	28.4	27.8	27.0	17.5	36.1	00.0	39.5	14.4	33.2	11.8	37.6	31.4	22.1	30.2
Herøy	41.7	32.6	34.9	31.3	39.7	31.7	39.5	00.0	37.7	35.4	38.1	39.7	38.3	36.7	37.1
Larvik	28.7	28.1	23.1	27.9	21.3	35.2	14.4	37.7	00.0	32.9	15.2	35.0	31.6	23.1	30.1
Lesja	39.8	25.9	30.2	30.9	36.3	29.6	33.2	35.4	32.9	00.0	32.1	24.9	35.9	34.7	29.6
Lillehammer	24.7	28.5	26.7	28.8	15.0	37.2	11.8	38.1	15.2	32.1	00.0	35.3	29.3	23.1	31.3
Stjørdal	39.2	20.0	27.2	39.3	39.0	28.5	37.6	39.7	35.0	24.9	35.3	00.0	42.0	32.4	25.9
Time	27.7	37.0	34.2	33.0	31.1	37.9	31.4	38.3	31.6	35.9	29.3	42.0	00.0	34.6	38.7
Trondheim	31.2	23.8	27.5	30.6	25.6	33.1	22.1	36.7	23.1	34.7	23.1	32.4	34.6	00.0	22.6
Verdal	37.8	16.9	27.7	34.0	32.8	29.5	30.2	37.1	30.1	29.6	31.3	25.9	38.7	22.6	00.0

Table 2. *Average Levenshtein distances between all pairs of 15 Norwegian dialects given as percentages.*

Figures

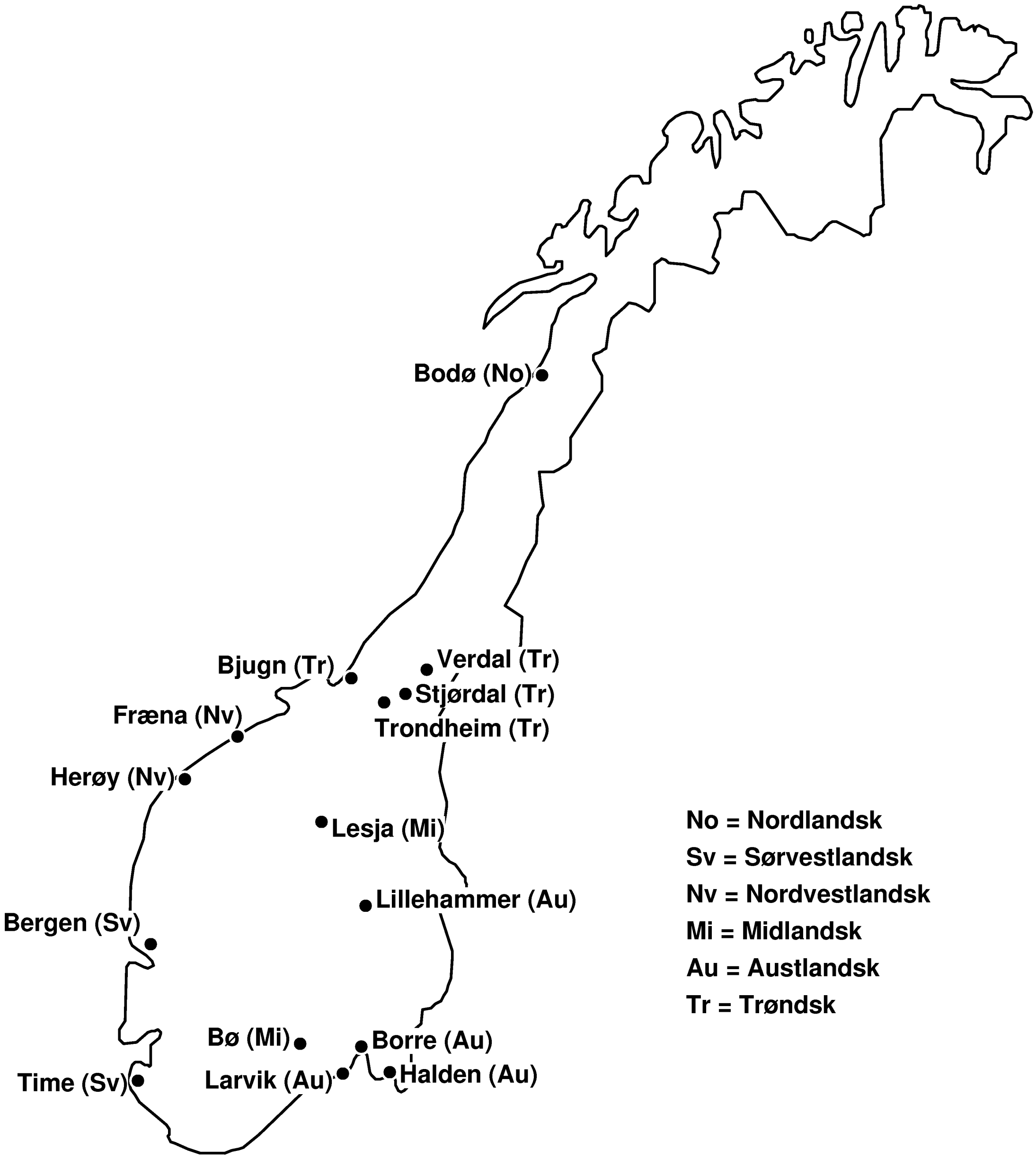

Figure 1. *Map of Norway showing the 15 dialects used in the present investigation. The abbreviation after the name of each location indicates the dialect area to which the variety belongs according to Skjekkeland (1997). The same abbreviations are used in other figures in this paper. Skjekkeland (1997) also gives a more global division in which Norwegian dialects are divided in Vestnorsk (covering No, Sv and Nv) and Austnorsk (covering Mi, Au and Tr).*

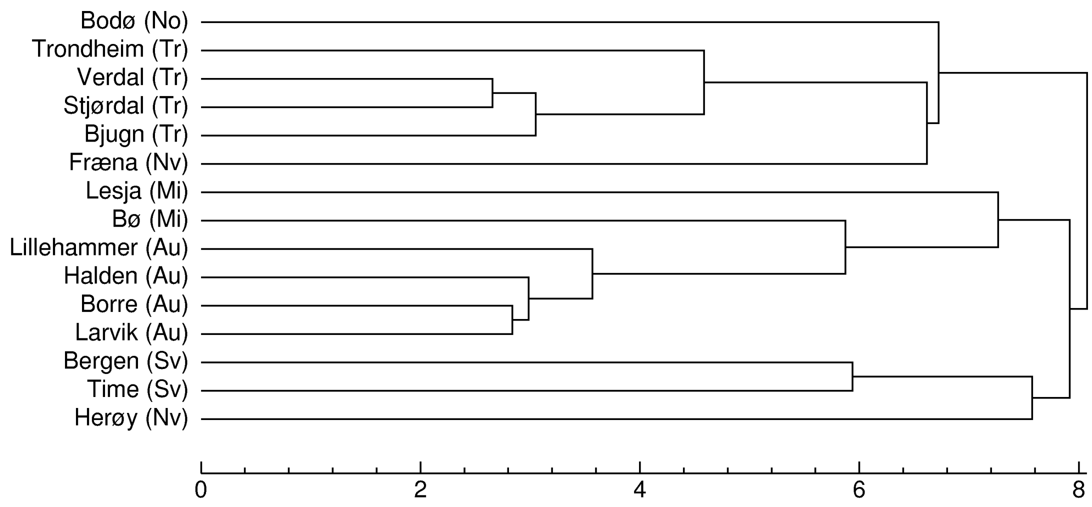

Figure 2. *Dendrogram derived from the 15 × 15 matrix of perceptual distances showing the clustering of (groups of) Norwegian dialects. On the horizontal scale distances are given in the scale as used by the listeners.*

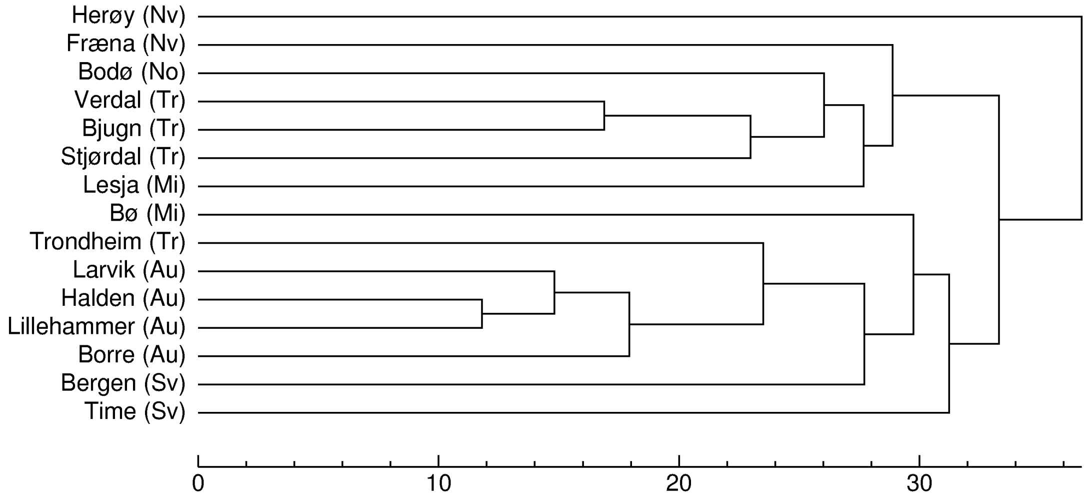

Figure 3. *Dendrogram derived from the 15 × 15 matrix of average Levenshtein distances showing the clustering of (groups of) Norwegian dialects. The scale distance is given as a percentage.*

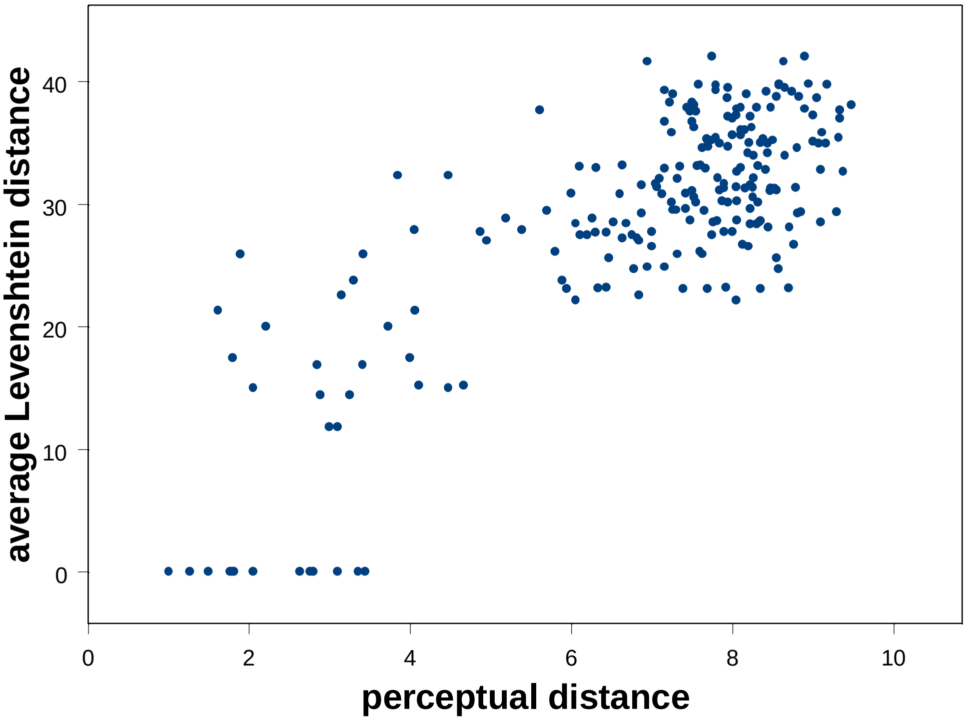

Figure 4. *Scatterplot showing perceptual distances versus average Levenshtein distances, including distances of dialects with respect to themselves (r=0.80, p<0.001).*

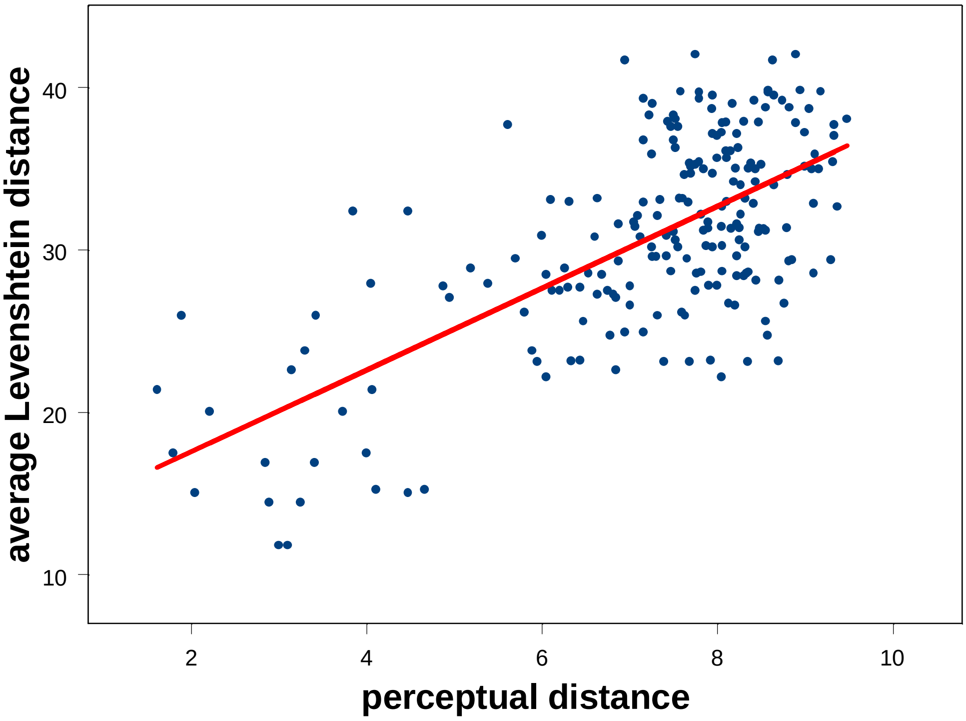

Figure 5. *Scatterplot showing perceptual distances versus average Levenshtein distances, excluding distances of dialects with respect to themselves (r=0.67, p<0.001). The linear regression line is shown as well.*

[1] The present article reports on part of a study supported by NWO, the Netherlands Organization for Scientific Research. We thank Saakje van Dellen for her obliging help with the data entry and Peter Kleiweg for letting us use the programs which he developed for the visualisation of the maps and dendrograms in the present article. Finally we would like to thank John Nerbonne for valuable comments.

[2] The recordings and the transcriptions (in IPA as well as in X-SAMPA) were made by Jørn Almberg in cooperation with Kristian Skarbø at the Department of Linguistics, NTNU, Trondheim and made available at http://www.ling.hf.ntnu.no/nos/. We are grateful for their permission to use the material and for the help of Jørn Almberg during the whole investigation.

[3] The program PRAAT is a free public-domain program developed by Paul Boersma and David Weenink at the Institute of Phonetic Sciences of the University of Amsterdam and available at http://www.fon.hum.uva.nl/praat .

[4] The data is taken from the *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) and available via: http://hyde.park.uga.edu/lamsas/.

[5] The example should not be interpreted as a historical reconstruction of the way in which one pronunciation changed into another. From that point of view it may be more obvious to show how [ˌæftərˈnuˑn] changed into [ˈæəftəˌnʉˑn]. We just show that the distance between two arbitrary pronunciations is found on the basis of the least costly set of operations mapping one pronunciation into another.

[6] See http://www.phon.ucl.ac.uk/home/wells/cassette.htm.