Computers and the Humanities **37:** 293–315, 2003.

Norwegian Dialects Examined Perceptually and Acoustically

WILBERT HEERINGA and CHARLOTTE GOOSKENS
Faculty of Arts, Humanities Computing and Scandinavian Languages, University of Groningen, the Netherlands

Abstract. Gooskens (2003) described an experiment which determined linguistic distances between 15 Norwegian dialects as perceived by Norwegian listeners. The results are compared to Levenshtein distances, calculated on the basis of transcriptions (of the words) of the same recordings as used in the perception experiment. The Levenshtein distance is equal to the sum of the weights of the insertions, deletions and substitutions needed to change one pronunciation into another. The success of the method depends on the reliability of the transcriber. The aim of this paper is to find an acoustic distance measure between dialects which approximates perceptual distance measure. We use and compare different representations of the acoustic signal: Barkfilter spectrograms, cochleagrams and formant tracks. We now apply the Levenshtein algorithm to spectra or formant value bundles instead of transcription segments. From these acoustic representations we got the best results using the formant track representation. However the transcription-based Levenshtein distances correlate still more closely. In the acoustic signal the speaker-dependent influence is kept to some extent, while a transcriber abstracts from voice quality. Using more samples per dialect word (instead of only one as in our research) should improve the accuracy of the measurements.

Key words: Barkfilter, cochleagram, dialect, dialectology, dialectometry, phonetic (dis)similarity, spectrogram

1. Introduction

Kessler (1995) introduced the use of the Levenshtein distance as a tool for measuring dialect distances. The Levenshtein distance is a string edit distance measure, and Kessler applied this algorithm to the comparison of Irish dialects. Later on, this approach was taken up by Nerbonne *et al*. (1996) and applied to Dutch dialects. The technique was also applied to Sardinian dialects by Bolognesi and Heeringa (2002). In all cases the use of the Levenshtein distance was based on phonetic transcriptions, where transcription segments were aligned by the algorithm. A detailed description of the basic algorithm is given by Kruskal (1999).

Gooskens and Heeringa (2003) calculated Levenshtein distances between 15 Norwegian dialects and compared them to the distances as perceived by Norwegian listeners. This comparison showed a high correlation between the Levenshtein distances and the perceptual distances. This investigation was based on existing recordings and corresponding phonetic transcriptions of the same text read aloud

in 15 Norwegian dialects.[1] The recordings were made in a soundproof studio in the autumn of 1999 and the spring of 2000. 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. There were 4 male and 11 female speakers. The average age of these speakers was 30.5 years. The speakers all read aloud the same text, namely the Norwegian version of the fable "The North Wind and the Sun". Further details about the material are given by Gooskens and Heeringa (2003). The same material is used for the present investigation. In Figure 1 the geographical distribution of the dialects is 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 in nine dialect areas. In our set of 15 varieties six areas are represented.

The Levenshtein distance measurements used in previous studies are based on phonetic transcriptions. However it is time-consuming to make phonetic transcriptions and furthermore the quality of the transcriptions varies sometimes greatly, depending on the skills of the transcriber. Hunt *et al*. (1999) and Ten Bosch (2000) present methods with which pronunciations are compared on the basis of the acoustic signal, without intervention of a transcriber.

Hunt *et al*. (1999) present a syllable-based speech recognition system in which unknown syllables are acoustically recognized by matching them against stored syllable templates. Syllables are represented as a sequence of acoustic-parameter vectors, each vector corresponding to one time-frame. A Levenshtein algorithm finds the optimum frame-to-frame correspondence between the template syllable and the unknown syllable and calculates the distances between them over that optimum frame correspondence.

Ten Bosch (2000) describes research in which an Automatic Speech Recognition (ASR) based distance measure is used to find the acoustic distances between dialects. Words are represented as a series of frames where each frame contains acoustic features. Words are compared by aligning the frames by a Viterbi alignment procedure, a technique roughly comparable to how phonetic segments are aligned when using transcriptions. Alignment is done by matching the frames with trained ASR Hidden Markov Models (HMMs).

In this paper a related acoustic measure is presented. The aim is to find an acoustically-based distance measure which approximates the perceptual distances well, i.e. one that does (almost) not rely on the phonetic transcriptions of segments for measuring the distances between dialects. We will experiment with different representations of the acoustic signal to investigate which representation approximates the perceptual distances the most.

In Section 2 we will show how the perceptual distance measurements were made and some overall results will be presented. The methods for measuring distances on the basis of acoustic data will be presented in Section 3. In Section 4

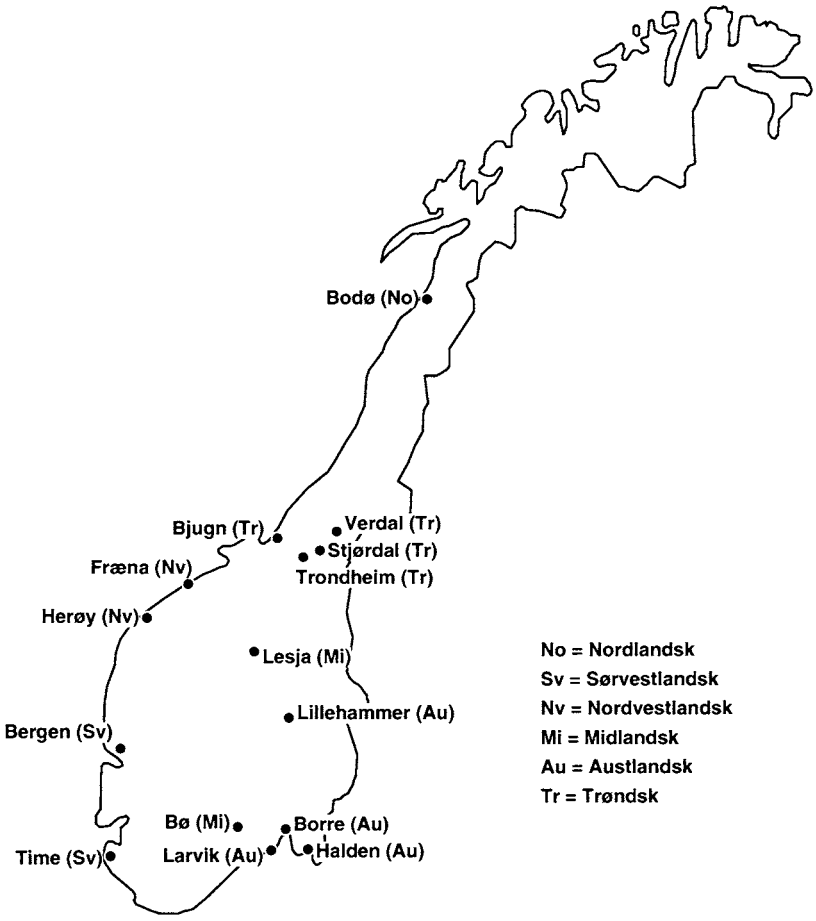

Figure 1. Map of Norway showing the 15 dialects in the present investigation. The abbreviation after the name of each location indicates the dialect area to which the variety belongs according to Skjekkeland (1997). The same abbreviations are used in the other figures in this paper. Skjekkeland (1997) 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).

the perceptual distance measurements will be compared to the acoustic measurements and in Section 5 some general conclusions will be drawn.

2. Perceptual Distance Measurements

In this section only general information about the perception experiment and some overall results will be presented. More details are given by Gooskens and Heeringa (2003).

2.1. EXPERIMENT

In order to be able to investigate the dialect distances between the 15 Norwegian dialects as perceived by Norwegian listeners, for each of the 15 varieties a recording of a translation of the fable "The North Wind and the Sun" was presented to Norwegian listeners in a listening experiment.

The listeners were 15 groups of high school pupils, one from each of the places where the 15 dialects are spoken. 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 texts of the 15 dialects were presented in a randomized order. A session was preceded by a practice recording. 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 distances. 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 is different from the distance as perceived by the listeners from Trondheim. The mean of these two distances is used when presenting the results below.

2.2. RESULTS

In order to visualize the relationship between the dialects, cluster analysis (see Jain and Dubes (1988)) was carried out on the basis of the matrices with the mean judgments of the original recordings. In Figure 2 the dendrogram produced by cluster analysis using *group average* is presented. Furthermore a multidimensional scaling analysis was carried out. The resulting plot can be found in Figure 3. In the dendrogram the two main groups are a northern group and a southern group. The southern group can be divided in a western group (Bergen, Time and Herøy) and an eastern group (the other dialects). In the multidimensional scaling plot a northern, a western and a southeastern group can be clearly identified. It is striking that the groups are rather sharply distinguished from each other. In traditional Norwegian dialectology the east-west division is often considered more important than the north-south dimension (e.g. Skjekkeland, 1997). However, the traditional division into an eastern and a western group is based on a rather limited set of phenomena. Some dialectologists therefore have suggested using more criteria which has resulted in other ways of dividing the language area. For example, Christiansen (1954) divides Norway into four dialect areas: north, south, east and west. Our data seem to support this classification. In practice many Norwegians disparage northern dialects, while seeing a certain regional unity within broad divisions, in particular East vs. West.

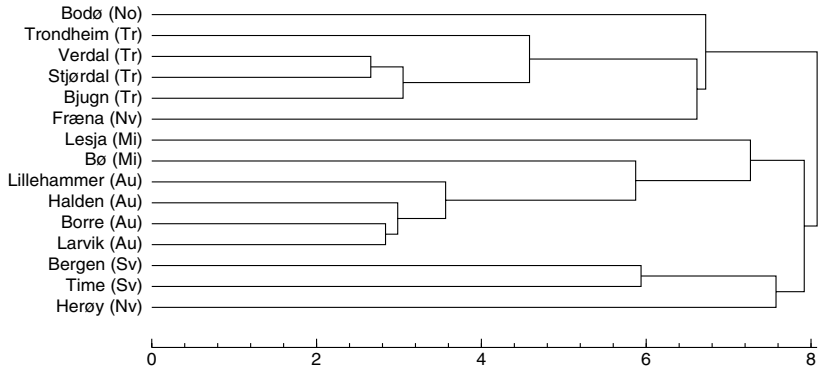

Figure 2. Dendrogram derived from the 15 × 15 matrix of perceptual distances showing the clustering of (groups of) Norwegian dialects.

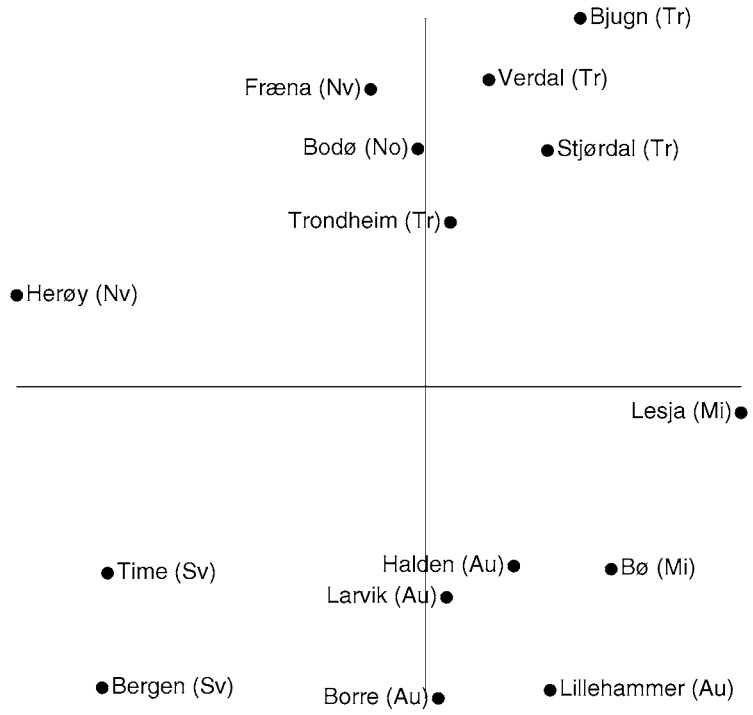

Figure 3. Multidimensional scaling of the results derived from the 15 × 15 matrix of perceptual distances.

3. Acoustic Distance Measurements

In this section we describe how acoustic measurements are made. In Section 3.1 we explain some manipulations of the samples. For the representation of the acoustic samples there are several possibilities. In Section 3.2 we account for our choice of representations. In Section 3.3 we explain how we normalize different speech rates. In Section 3.4 the application of the Levenshtein distance in the comparison of acoustic representations is explained. On the basis of the Levenshtein distances the dialects are classified. In Section 3.5 results are given for the different representations of the acoustic signal. In Section 3.6 the classification results are further examined.

3.1. SAMPLES

The Norwegian translation of the fable "The North Wind and the Sun" consists of 58 different words. Due to the free translation of some phrases for certain varieties a few of the expected words were missing. For all 15 dialects each of the 58 words were cut from the text, so we usually get 58 word samples per dialect. If the same word appears more than once in a text, we select only the first occurrence.

The voices of different speakers will have different pitches. Most obvious is the difference in pitch between male and female voices. Furthermore the intonation per speaker may vary. When two speakers read the same text aloud, the one may stress different words than the other. To make samples of different speakers as comparable as possible, all word samples were monotonized. The mean pitch of the 4 men was 134 Hz, and of the 11 women 224 Hz. The mean of the means is 179 Hz. So all word samples were monotonized on the mean of 179 Hz with the program PRAAT.[2]

We are aware of the fact that this choice removes all prosodic information about pitch and intonation contours which are known to be significant dialect markers in Norwegian. However, we found no way to exclude speaker-dependent intonation and simultaneously retain dialect-dependent intonation. Furthermore, we note that monotonizing does not remove all gender-dependent information. We have also experimented with normalizing other gender-specific properties, such as adapting the frequency scale, but found no improvement in the results yet.

The volume was not normalized because volume contains a good deal of sound specific information. For example it is specific for the [v] that its volume is greater than that of the [f].

3.2. ACOUSTIC SIGNAL

An acoustic signal can be represented by a spectrogram. A spectrogram is the representation of the acoustic intensities which are distributed over time and frequency. In our research we do not use the most commonly used type of spectrogram which

has a *Hertz*-scale, but more perceptual models. In Section 3.2.1 we describe the *Barkfilter* and in Section 3.2.2 we describe the *cochleagram*.

We reduce the representation still further when only formant tracks are used. Formant tracks represent the prominent frequency tracks in the spectrogram. In this more reduced representation the more speaker-specific information may be filtered away to some extent. We discuss this reduced representation further in Section 3.2.3.

3.2.1. *Barkfilters*

In the most commonly used type of spectrogram, the *Hertz* frequency scale is used, which is linear. The difference between 100 Hz and 200 Hz is the same as the difference between 1000 Hz and 1100 Hz. However our perception of pitch is non-linear. We hear the difference between 100 and 200 Hz as an octave interval, but the difference between 1000 and 2000 Hz is also perceived as an octave. Our ear evaluates frequency differences not linearly, but rather logarithmically. Therefore in the Barkfilter a more or less logarithmic frequency scale is used, which is called the *Bark*-scale.

To reduce the size of the intensity scale, intensity is likewise represented logarithmically, viz., using the *decibel* scale. The logarithmic scale accords with our perception of loudness.

In our research the frequencies ranges from 0 to 24.67 *Bark*. They are divided in 24 equal intervals, where for each interval the mean intensity is given. The spectrum is probed each 0.005 seconds with an analysis window of 0.015 seconds. In Figure 4 Barkfilter spectrograms are shown which are obtained on the basis of the original (not manipulated) samples of the word *nordavinden* “the northwind” in the dialects of Bjugn, Halden and Larvik. In Figure 5 spectrograms are shown which are obtained on the basis of the corresponding monotonized samples. The monotonized samples are used for the dialect comparison in our investigation.

3.2.2. *Cochleagrams*

A cochleagram is a spectrogram which models the cochlea. The spectrogram is adapted so that it gives information as it is received by the cochlea. The similarity with the Barkfilter is that it also uses the *Bark*-frequency scale. However loudness is not represented by logarithmic intensities, but with respect to a calibration at 1 kHz, and refers to the units as *phones*. If a given sound is perceived to be as loud as a 60 dB sound at 1000 Hz, then it is said to have a loudness of 60 *phon*. These relations are determined experimentally. See also Rietveld van Heuven (1997).

In a cochleagram *lateral* masking is taken into account. When sounds occur at neighboring frequencies simultaneously, one frequency may mask the other. In general, a low tone will mask a high tone rather than the opposite. Moreover, *forward* masking is modeled as it occurs in the cochlea. After hearing an intense sound our ears may be stunned for a short time. The more successive sounds are

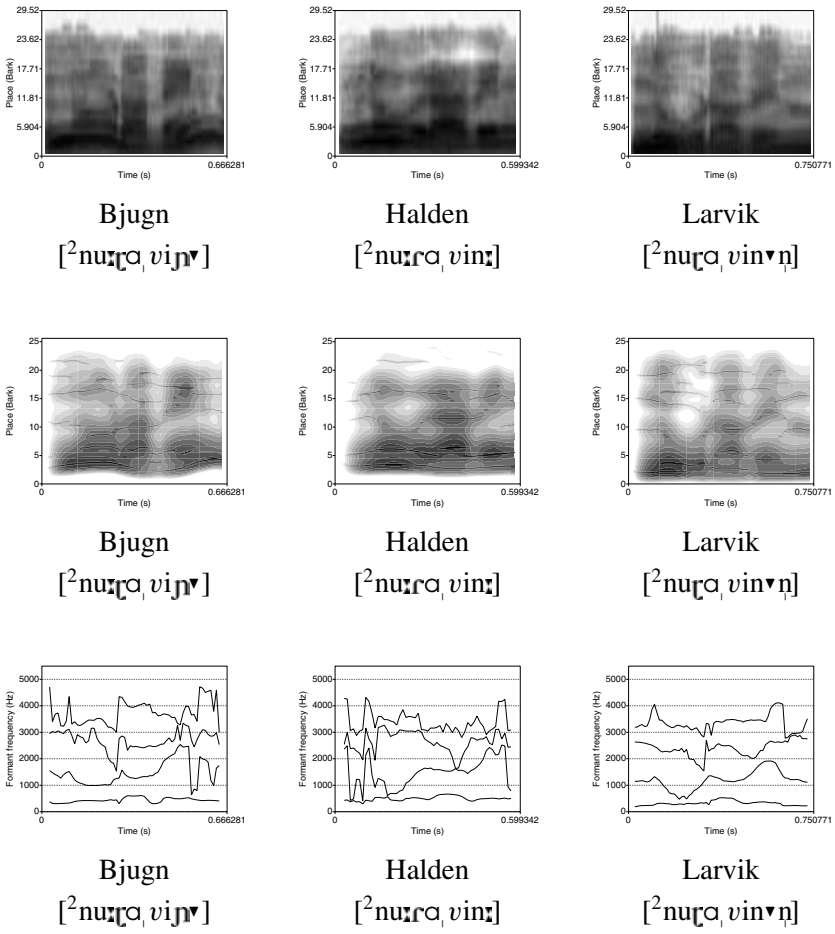

Figure 4. Different acoustic representations of three Norwegian pronunciations of *the north-wind* (nordavinden). From upper to lower we see respectively Barkfilters, cochleagrams and formant tracks obtained on the basis of the original samples.

like each other, the stronger the masking will be. This property is also incorporated in a cochleagram.

In our research the frequencies in each cochleagram range from 0 to 25.6 *Bark*. They are divided in 256 equal intervals, where for each interval the mean loudness is given. The spectrum is probed each 0.01 seconds with an analysis window of 0.03 seconds. The forward-masking time is set to 0.03 seconds. In Figure 4 cochleagrams are shown which are obtained on the basis of the original (not manipulated) samples of the word *nordavinden* "the northwind" in the dialects of Bjugn, Halden and Larvik. In Figure 5 spectrograms are shown which are obtained on the basis of the corresponding monotonized samples. As mentioned above (Section 3.2.1) only the monotonized samples are used for our investigation.

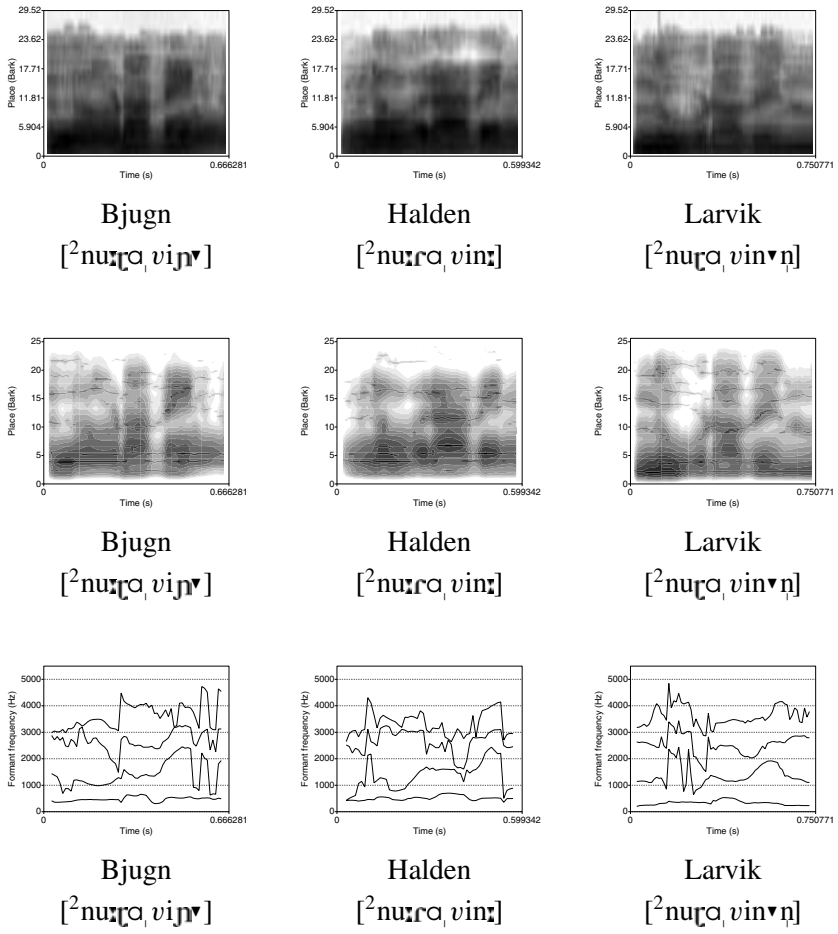

Figure 5. Different acoustic representations of three Norwegian pronunciations of *nordavinden* "the northwind". From upper to lower we see respectively Barkfilters, cochleagrams and formant tracks obtained on the basis of the monotonized samples.

3.2.3. *Formants*

Another way to study the acoustic signal is to investigate formants. When using a spectrogram with a large analysis window (about 20 ms) the frequency resolution will be high. Individual harmonics will show up as horizontal lines through the spectrogram. The lowest line represents the fundamental frequency or pitch (F0). However, when using a small analysis window (about 3 ms) the frequency resolution will be lower. Individual harmonics get smeared together. Instead of lines, bands will show up through the spectrogram. The center frequency at one time in a band is called a formant, the range of center frequencies in the course of time forms a formant track. A formant in the lowest band is called F1, a formant in the

next band F2, etc. Formants represent a frequency region that is enhanced by the resonances of the vocal tract.[3]

Having one word sample, the number of formants may vary over time. It appears that for each word sample at each time sample at least three formants can be found. Therefore the comparison of word samples is based on (the first) three formant tracks. When finding formants in PRAAT, the time step was set to 10 ms with an analysis window of 25 ms. The ceiling of the formant search range should be set to 5000 Hz for males, and to 5500 Hz for females. Because the samples on the basis of which the formants are determined are monotonized to the average of the mean pitch of the males and the females, we set this ceiling to 5250 Hz. Pre-emphasis starts at 50 Hz. In the manual which can be found in the PRAAT program pre-emphasis is explained as follows: "This means that frequencies below 50 Hz are not enhanced, frequencies around 100 Hz are amplified by 6 dB, frequencies around 200 Hz are amplified by 12 dB, and so forth. The point of this is that vowel spectra tend to fall by 6 dB per octave; the pre-emphasis creates a flatter spectrum, which is better for formant analysis because we want our formants to match the local peaks, not the global spectral slope." In PRAAT several algorithms can be chosen for finding the Linear Predictive Coding (LPC) coefficients. We chose the algorithm of Burg, which seems to be the most common one.

Before comparing formant frequencies in the comparison of words the frequencies in *Hertz* are converted to *Bark*, which is, as mentioned above, a more perceptual scale. In Figure 4 formant tracks are shown which are obtained on the basis of the original samples of the word *nordavinden* "the northwind" in the dialects of Bjugn, Halden and Larvik. In Figure 5 formant tracks are shown which are obtained on the basis of the corresponding monotonized samples.

3.3. SPEECH RATE

When we compare word samples, we have to allow for the fact that different speech rates give different sample sizes. To perform a rough normalization, first we find the number of segments per word according to the phonetic transcription and call this n. Now we regard the word spectogram as a concatenation of n equally-sized intervals. We have to arrange that each interval gets a fixed number of spectra. We call this m. When there are fewer spectra, they are expanded to m, and when there are more spectra they are reduced to m. In our research we select $m = 20$. A higher value gives no clearly different results while the computing time increases greatly. As a result each word sample is represented as a reduced spectrogram with $n \times m$ spectra. When using formants, we have formant frequency bundles instead of spectra.

We are aware of the fact that this is a rough approach, but it should be refined enough to capture significant variation.

3.4. LEVENSHTEIN DISTANCE

The Levenshtein distances calculates the cost of changing one string into another. It determines how the one string can be changed into the other in the easiest way by inserting, deleting or substituting elements. A detailed description of the algorithm is given by Kruskal (1999). Finding the distance between different pronunciations on the basis of their transcriptions, the elements are the phonetic segments. However using the acoustic signal, the elements are spectra or formant bundles.

Now a substitution is calculated as follows. Assume a spectrum or formant bundle $e1$ and $e2$ with respectively t frequencies or formants, then:

$$d(e1, e2) = \sqrt{\sum_{i=1}^{t} (e1_i - e2_i)^2}$$

For the calculation of insertions and deletions we used definitions of "silence". We defined a "silence spectrum" as a spectrum for which the intensities of all frequencies are equal to 0. A "silence formant bundle" is defined as a bundle for which all frequencies are equal to 0. This means that in absolute silence there are no vibrations.

If we used the Levenshtein distances directly, then longer words would contribute disproportionately to the estimation of distances between varieties, which does not accord with the idea that words are linguistic units. Therefore we normalize each Levenshtein distance by dividing it by the length of the alignment. Sometimes the same Levenshtein distance may correspond with different alignments having different lengths. We will illustrate this by two transcriptions, although in this paper Levenshtein is applied to spectrograms and formant tracks rather than to transcriptions. E.g the word *bee* is pronounced as [binə] "Biene" in German and as [bɛi] "bij" in Dutch. Two possible alignments are:

b	i	n	ə		b		i	n	ə
b	ɛ	i			b	ɛ	i		
0	1	1	1		0	1	0	1	1

In the example, equal sounds have a cost of 0 and different ones a cost of 1. However in our research we used the gradual weights found by the formula which is given in the beginning of this section. This example shows that the longer alignment is the more reasonable one. Therefore we divide the Levenshtein distance by the length of the longest alignment.

Using 58 words the distance between two dialects is equal to the average of 58 Levenshtein distances. When comparing two words between two dialects for which no translation is given for one or both dialects, than the distance for that word pair is taken to be the average of the distances of all word pairs for which translations in both dialects were available.

All distances between the 15 dialects were arranged in a 15×15 matrix.

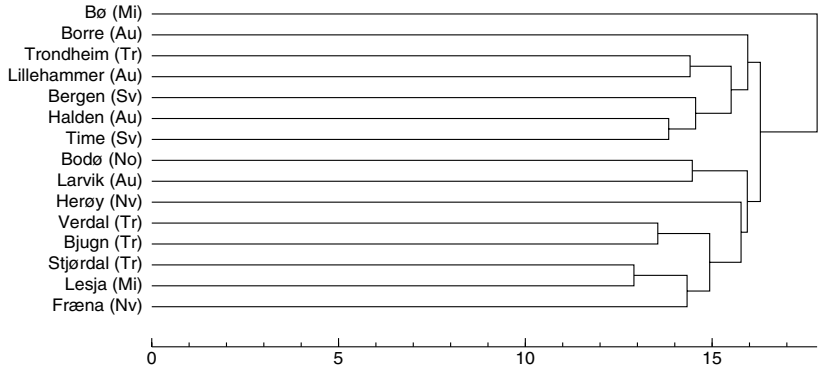

Figure 6. Dendrogram obtained on the basis of Levenshtein distances where the Barkfilter representation is used.

3.5. CLASSIFICATION

On the basis of a distance matrix of average Levenshtein distances the dialects are classified. We present results which were obtained on the basis the Barkfilter representation, the cochleagram representation and the formant track representation. For each of them we performed cluster analyses (Jain and Dubes, 1988) and multidimensional scaling (Kruskal, 1983).

3.5.1. *Barkfilters*

In Figure 6 and Figure 7, results can be found which are obtained on the basis of the Barkfilter representation. At the most significant level in the dendrogram we find Bø on the one hand, and the remaining dialects on the other hand. This group of remaining dialects is divided in a northern and southern group. It is striking that the dialect of Larvik (which is geographically located in the south) is grouped under the northern dialects, and that the dialect of Trondheim (geographically in the north) is grouped under the southern dialects. A clear division between the West and the East as in Figure 2 is not found here. In the multidimensional scaling plot the y-axis seems to correspond with the geographic north-south axis, while the x-axis seems to represent the division between male and female speakers. The texts of Herøy, Bodø, Larvik and Bø were read by male speakers, while the other texts were read by female speakers. This explains why Larvik is not grouped together with the other southeastern dialects.

In the dendrogram Bodø and Larvik appear as one cluster. In the multidimensional scaling plot Bodø is more close to Herøy. Different classification techniques sometimes give slightly different results. This shows the necessity of using both cluster analysis and multidimensional scaling.

When comparing this classification result with the results obtained from the perceptual distances (Figures 2 and 3), it is striking that the groups are less sharply

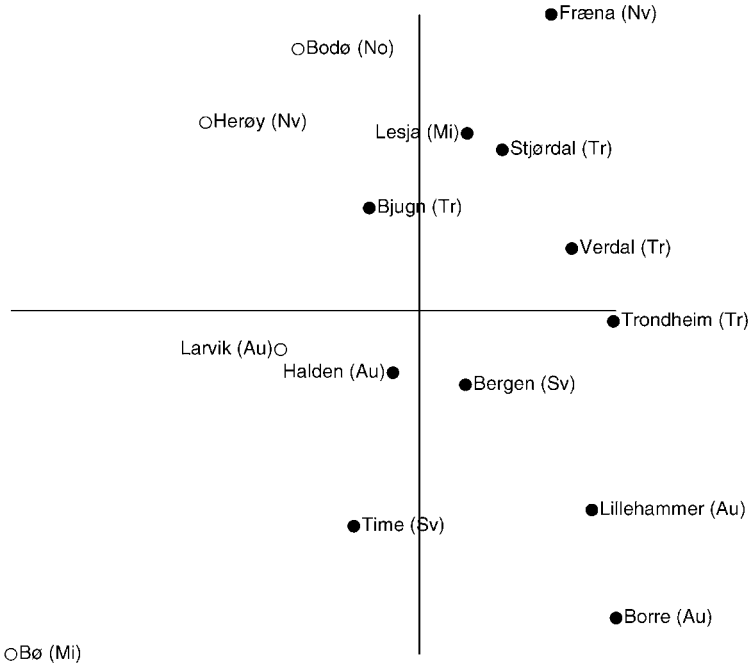

Figure 7. Multidimensional scaling plot obtained on the basis of Levenshtein distances where the Barkfilter representation is used. White dots indicate male speakers and black dots female speakers.

distinguished from each other in the acoustically based analysis. In the perception experiment subjects may judge the distance in a more categorical way while the results of the acoustic measurements differ gradually.

3.5.2. *Cochleagrams*

In Figure 8 and Figure 9 results can be found which are obtained on the basis of the cochleagram representation. In the dendrogram we see a clear division between a northern and a southern group again, apart from Bø. The dialects of Trondheim and Larvik are in the right groups now. Note the position of Bodø in the southern group. Also here a clear division between a western and eastern group as in Figure 2 is not found. In the multidimensional scaling plot the y-axis seems to be the geographic north-south axis again while the x-axis again represents the distinction between male and female speakers. The distinction between male and female speakers is even sharper than in Figure 7. Different from Figure 7, but similar to Figure 3 the southwestern group with the dialects of Bergen and Time can be found here.

Similar to the Barkfilter-based results, the cochleagram-based results show a less sharp distinction between groups than the perceptual results do. However, we

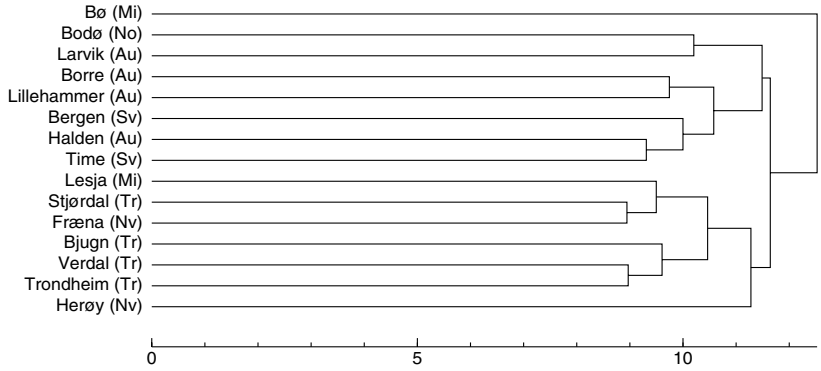

Figure 8. Dendrogram obtained on the basis of Levenshtein distances where the cochleagram representation is used.

note that the cochleagram-based results are more similar to the perceptual results than the Barkfilter results are.

3.5.3. *Formants*

In Figure 10 and Figure 11 results can be found which are obtained on the basis of the formant track representation. Again the dendrogram shows a division in a northern and a southern group (the dialect of Bø ignored). However also a southwestern group containing the dialects of Bergen and Time is found, just as in Figure 2. Similar to the dendrogram which was obtained on the basis of the cochleagram representation (Figure 8) Bodø is in the southern group. Like the multidimensional scaling plots obtained on the basis of the Barkfilter representation and the cochleagram representation (respectively Figures 6 and 8) the y-axis seems to correspond with the geographic north-south-axis, and the x-axis represents the distinction between male and female speakers. However the distinction between sexes is not as sharp here as in Figure 9. In the dendrogram the southern dialects are divided in a western group (Bergen and Time) and an eastern group (the remaining southern dialects). This division is not found so sharply in the multidimensional scaling plot.

Similar to the Barkfilter- and cochleagram-based results, also the formant-track-based results show a less sharp distinction between groups than the perceptual results do. However, we note that the formant-track-based results are more similar to the perceptual results than the Barkfilter results are. Compared to the cochleagram-based results no clear improvement can be observed.

3.6. EXPLANATION OF RESULTS

In the classification results presented thus far the north-south division is more important than the east-west division. This is different from some traditional results

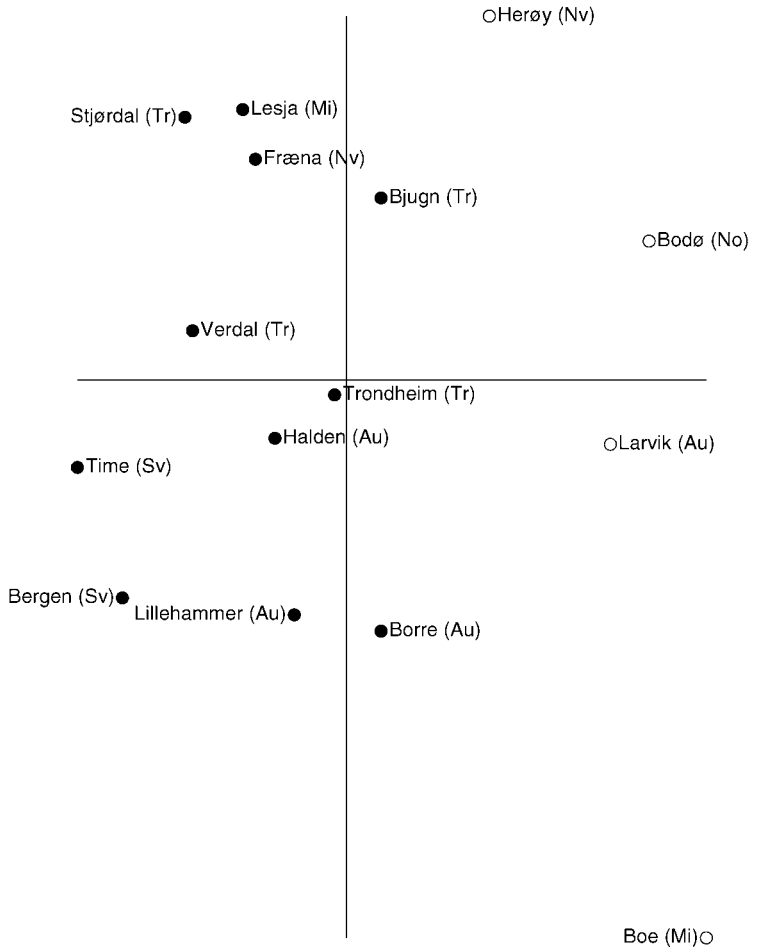

Figure 9. Multidimensional scaling plot obtained on the basis of Levenshtein distances where the cochleagram representation is used. White dots indicate male speakers and black dots female speakers.

where the east-west division is more important than the north-south division (see e.g. Skjekkeland, 1997). In Section 3.6.1 we will explore our data source and try to find which words are strongly responsible for our north-south division. Next we will examine the variation of these words to find the phenomena which contributes to the north-south dimension.

Another striking fact especially found in the multidimensional scaling plots is the separation between male and female speakers. In Section 3.6.2 we determine whether gender-specific information is still retained in the monotonized samples. We will try to find words which are clearly responsible for the male-female division, although corresponding variation can not be found in their phonetic

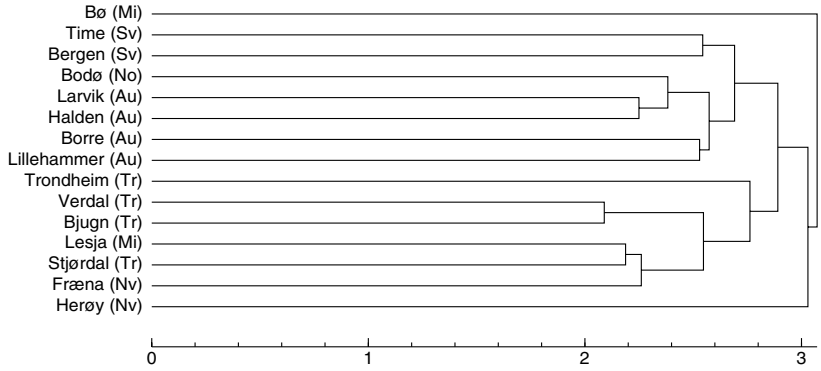

Figure 10. Dendrogram obtained on the basis of Levenshtein distances where the formant track representation is used.

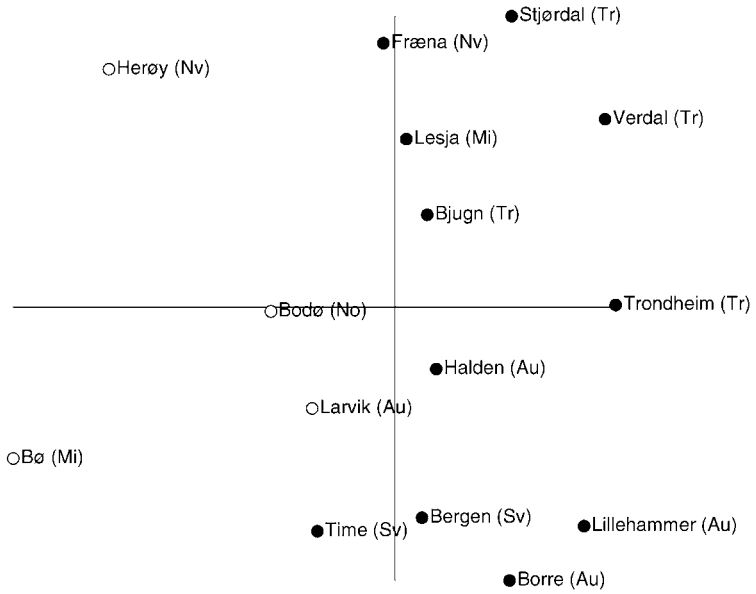

Figure 11. Multidimensional scaling plot obtained on the basis of Levenshtein distances where the formant track representation is used. White dots indicate male speakers and black dots female speakers.

transcriptions. In other words we want to find words which contain only gender variation rather than dialect variation.

3.6.1. *North versus South*

The fact that classification results show a north-south division rather than an east-west division as suggested by traditional results, may be explained from the data

source on the one hand, and the comparison method on the other hand. Therefore we want to investigate whether some particular phenomena are responsible for this north-south division.

Using the multidimensional scaling plots, we investigated the north-south dimension (the vertical dimension in the plots which is the second dimension) further. From the plot per dimension, distances between varieties can be derived. When examining the north-south dimension the distance between two varieties is equal to the absolute difference between the corresponding y-coordinates. In this way, for each pair of varieties the "north-south"-distance is found. Having 15 varieties, we get 15 × 15 distances. Since the distance begween e.g. Bjugn and Bjugn is always 0, we want to exclude the distances of varieties with respect to themselves, so we get 15 × 14 distances. Since the distance between e.g. Bjugn and Halden is equal to the distance between Halden and Bjugn, only the half of the distances are needed. So finally we only use (15 × 14)/2 distances.

In Section 3.4 we described how to calculate the Levenshtein distance between two words. Using the Levenshtein distance a distance matrix can be obtained, containing Levenshtein distances between the different pronunciations of one particular word. Also this matrix contains (15 × 14)/2 distances.

Having Levenshtein distances for one word on the one hand, and north-south-distances as found in the multidimensional scaling plot on the other hand, the two sorts of distances can be correlated. The stronger the word distances correlate with the north-south distances, the more the variation of the corresponding word contributed to the north-south dimension in the multidimensional scaling plot.

For finding the correlation coefficient, we used the Pearson's correlation coefficient (Sneath and Sokal, 1973, pp. 137–140). When having 15 varieties, a distance matrix will have 15 rows and 15 columns. The correlation coefficient between is calculated as:

$$r(X, Y) = \frac{\sum_{i=2}^{n} \sum_{j=1}^{i-1} (X_{ij} - \overline{X})(Y_{ij} - \overline{Y})}{\sum_{i=2}^{n} \sum_{j=1}^{i-1} (X_{ij} - \overline{X})^2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} (Y_{ij} - \overline{X})^2}$$

where $n = 15$. Correlation coefficients range from -1 (perfect inverse correlation) to +1 (perfect correlation). There is no correlation if $r = 0$.

For each of the 58 words we calculated the Levenshtein distances between the 15 varieties. This gives 58 matrices. Subsequently each of the matrices was correlated with the distances derived from the vertical dimension of the multidimensional scaling plot which corresponds with north-south. We did this for the multidimensional scaling plots obtained on the basis of respectively the Barkfilter, the cochleagram and the formant tracks representation.

When using the Barkfilter representation the word *the* "den" correlates strongest (r = 0.78), followed by the word *finally* "til sist"(r = 0.67). In the north, *the* is pronounced as [ðeɲ] and in the south as [ɖɳ] (or similar forms). The word *finally* is pronounced like [teʃlʉcːɪ] in the north and like [tilsist] in the south. Using cochleagrams the word *of* "av"[4] correlates strongest (r = 0.72), and next the word *the*

"den" ($r = 0.71$). In the north for the word *of* pronunciations are found like [tɔ], in the south like [ɑ]. Using formant tracks the words *blew* "blaaste" ($r = 0.56$) and *off* "av"[5] ($r = 0.55$) correlate the strongest. For the word *blew* in the north forms like [bɽɔst] are common. In the south similar forms are used; however, they are followed by a schwa, e.g. [bɽɔstə]. Examining the transcriptions of *off* we found no systematic variation which could be explained as a contribution to a north-south division.

Examining the strongly-correlating words just mentioned, we find that a wide range of phenomena contributes to the north-south division. Therefore it becomes clear that this division is not the result of a biased weighting of phenomena. In our method no choice of phenomena is made beforehand. However the map of Skjekkeland (1997) is based on a restricted set of phenomena. This may explain the difference between our results and the division as given by Skjekkeland.

3.6.2. *Male versus Female*

In the multidimensional scaling plots obtained on the basis of the acoustic Levenshtein distances a separation between male and female speakers can be found. We suspect this is caused by the fact that after monotonizing the samples still gender-specific information is retained. Therefore we will search for words which obviously contribute to the male-female division due to the fact that gender variation is retained in the acoustic samples while no dialect variation can be found in the phonetic transcriptions.

Just as we derived north-south distances from the multidimensional scaling coordinates on the basis of the vertical dimension in Section 3.6.1 we can derive male-female distances from the horizontal dimension, resulting in $(15 \times 14)/2$ distances again. As explained in Section 3.6.1 for each of the 58 words we calculated the Levenshtein distances between the 15 varieties resulting in 58 matrices.

We correlate each of the 58 matrices with the male-female distances as derived from the horizontal dimension in the multidimensional scaling plot. Examining the words corresponding with the matrices which correlates strongest with the male-female distances, it appears that for some of them a corresponding variation could be found in the phonetic transcriptions. However, we also found words which strongly correlate, but for which the variation as found in the phonetic transcriptions gives no satisfying explanation for this high correlation. Examples are *man* "mann" ($r = 0.53$) and *get* "faa" ($r = 0.48$) when using cochleagrams, *around* "rundt" ($r = 0.54$) and *get* "faa" ($r = 0.46$) when using Barkfilters, and *around* "pa" ($r = 0.74$) and *took* "tok" ($r = 0.59$) when using formant tracks. Since the dialect-specific variation in these words as noted in the transcriptions cannot explain their relatively strong correlation with the horizontal dimension, it is obvious that these words reflect gender variation to some extent. Therefore it is clear that gender variation is retained in the samples after monotonizing, which has influenced our results.

When generating multidimensional scaling plots on the basis of the eleven female speakers only, for all acoustic representations, the vertical dimension still corresponds with the north-south axis. The horizontal dimension corresponds more or less with the east-west axis. Examining the multidimensional scaling plots on the basis of all 15 speakers, the horizontal dimension may be interpreted as an east-west dimension to some extent as well, but the interpretation as male-female axis is much more obvious as may be concluded by the readers' eye.

4. Perceptual versus Acoustic Distances

In order to compare the different distance measurements the matrices resulting from the perceptual and acoustic measurements were compared by calculating the Pearson's correlation coefficient between them. In Section 3.6 correlations are calculated on the basis of $(15 \times 14)/2$ distances. Using (average) Levenshtein distances, distances of varieties with respect to itself are always equal to 0. Therefore they can be excluded. These distances are found on the diagonal in the distance matrix, containing the cells $(1, 1), (2, 2), \ldots (n, n)$. Furthermore distances are symmetric: the distance between e.g. Bjugn and Halden is equal to the distance between Halden and Bjugn. However in the 15×15 matrix of perceptual distances, the distance of Bjugn and Bjugn is not equal to 0. Furthermore the distance between Bjugn and Halden is different from the distance between Halden and Bjugn. So this may suggest that we need to use all 15×15 distances when correlating the acoustic Levenshtein distances with the perceptual distances. However it appears that the distances of varieties with respect to themselves are outliers when using the acoustic distances (they are always 0), but they are not outliers when using perceptual distances. Therefore two correlation coefficients are given for each pair of matrices, one based on the full matrices ($15 \times 15 = 225$ distances), and one based on a matrix excluding the diagonal ($15 \times 14 = 210$ distances).

Table I shows the correlation coefficients between the different acoustic Levenshtein distances and the perceptual distances. It may also be interesting to take the transcription-based Levenshtein distance into account. In this approach, the distance between two words is found by calculating the Levenshtein distance on the basis of the corresponding phonetic transcriptions. Insertions, deletions and substitutions are applied to phonetic segments instead of to spectra or formant bundles. For more details see Gooskens and Heeringa (2003). Correlations with respect to this distance are also given in Table I. All correlations in the table are significant (when $\alpha = 0.05$, the same significance level is used in the rest of the paper). All cases including the diagonal give a significantly higher correlation coefficient than those excluding the diagonal.[6]

The correlation coefficients of the different acoustic measurements with respect to the perceptual distances do not differ significantly when the diagonal is included. The greatest difference was found between the Barkfilter and the formant tracks ($r = 0.65$ vs. $r = 0.71$), but the one is still not significantly higher than the other

Table I. Correlation coefficients between matrices resulting from the different distance measurements. The values are given for the correlation coefficients including and excluding the diagonals. All correlations are significant for α = 0.05

distance based on	original perc. dist.		transcription based dist.	
	r_{incl}	r_{excl}	r_{incl}	r_{excl}
Barkfilter	0.65	0.33	0.87	0.52
cochleagrams	0.67	0.38	0.89	0.62
formants	0.71	0.50	0.93	0.77
transcription	0.79	0.64		

(z = 1.18, p = 0.119). When the diagonal is excluded, the formant track correlation coefficient is not significantly higher than the cochleagram correlation coefficient, and the cochleagram correlation coefficient is not significantly higher than the Barkfilter correlation coefficient. However the formant track correlation coefficient is significantly higher than the Barkfilter correlation coefficient (r = 0.50 vs. r = 0.33, z = 2.10, p = 0.018). Although the formant-based distances have the highest correlation coefficient with respect to the perceptual distances (for both including and excluding the diagonal), we see that the formant-based distances have a significantly higher correlation coefficient only with respect to the Barkfilter correlation coefficient when excluding the diagonal.

The table shows that all acoustic measurements correlate less-well with the perceptual distances than the transcription-based distances do. Although the formant track-based distances correlates highest of the acoustic measurements, it correlates still significantly lower than the transcription-based distances (including diagonal: r = 0.71 vs. r = 0.79, z = –1.94, p = 0.026, and excluding diagonal: r = 0.50 vs. r = 0.64, z = 2.12, p = 0.017). This may be explained from the fact that, in acoustic measurements, speaker characteristics such as voice quality play a role, while the transcriber as well as the listeners in the perception experiment are probably able to abstract from individual speaker characteristics to a great extent.

Ten Bosch (2000) correlates the ASR-based distances with transcription-based distances. He found an obvious correlation (r = 0.70). In our research we also found rather high (and significant) correlations between the acoustic distances and the transcription-based distances, where the formant-based distances had the highest correlation (r = 0.77 excluding the diagonal). Regardless whether the diagonal is included or excluded, the correlation coefficient based on the cochleagram representation is not significantly higher than the correlation coefficient based on the Barkfilter representation. However the correlation coefficient based on the formant track representation is significantly higher than both the Barkfilter correlation coefficient and the cochleagram correlation coefficient.

The fact that of the acoustic measurements the formant-based distances have the (in one case significantly) higher correlation with the perceptual distances and also have the significantly higher correlation with the transcription-based distances may indicate that the influence of voice characteristics is less strong when distances are measured on the basis of formants, rather than on the basis of the Barkfilter or cochleagrams. This seems to be confirmed by the classification results (Section 3.5) where the distinction between male and female speaker is stronger in the Barkfilter and cochleagram-based results than in the formant-based results.

5. Conclusion

The aim of this explorative investigation was to find an acoustic distance measure between dialects which approximates a perceptual distance measure. The results show that of the different acoustic measurements the formant-based distances have not only the highest correlation with the perceptual distances, but with the transcription-based distances as well. The formant-based correlation coefficient with respect to the transcription-based results is significantly higher than those of the other acoustic measurements. Since transcription based results may also be regarded as perceptual to some extent, this outcome may indicate that the influence of voice characteristics is less strong when distances are measured on the basis of formants, rather than on the basis of the Barkfilter or cochleagrams. This seems to be confirmed by the classification results where the distinction between male and female speaker is stronger in the Barkfilter and cochleagram-based results than in the formant-based results.

The correlation with the perceptual distances is higher for the transcription-based distances than for the formant-based distances (without diagonal significantly higher). Thus it is still necessary to search for a more refined method of using acoustic data for distance measurements. A disadvantage of the use of acoustic data is that it is not clear how great the influence of varying recording circumstances and the individual voice characteristics of the speakers are. To neutralize the influence of the speaker specific information, much more than one sample per dialect (as in our research) should, therefore, be used. Furthermore, it would be worthwhile to find a way in which differences in speech rates might be normalized in a more refined way than is done in our research (see Section 3.3).

We also compared both the perceptually-based results and acoustically-based results to the traditional map of Skjekkeland (1997) on which the east-west division is most important. In our results we found the north-south division to be more significant than the east-west division. In our results there was no biased weighting of phenomena. On the other hand, the traditional map of Skjekkeland is based only on a limited number of phenomena.

Acknowledgements

The present article reports on part of a study supported by a grant for the cooperation between the Departments of Linguistics in Groningen and Oslo from NWO, the Netherlands Organization for Scientific Research. We wish to thank Vincent van Heuven and John Nerbonne for advice and comments on this paper, Paul Boersma and David Weenink for help with PRAAT, Sabine Rosenhart for help with cutting the word samples, Jørn Almberg for his permission to use the recordings and transcriptions of 'The North Wind and the Sun' and for his help during the whole investigation and Saakje van Dellen for help with entering data. We thank Arnold Dalen for his help in finding a reliable dialect map and for classifying each of the 15 varieties in the right dialect group in accordance with this traditional dialect map. We thank Peter Kleiweg for his graphic programs, which we used for the visualization of the map, the dendrograms and the multidimensional scaling plots. We thank the anonymous reviewers for their valuable comments.

Notes

[1] The recordings were made by Jørn Almberg in co-operation with Kristian Skarbø at the Department of Linguistics, University of Trondheim and made available at *http://www.ling.hf.ntnu.no.nos*. At the time, the perception experiment was carried out, recordings of only 15 varieties were available. Today more than 50 recordings are available, giving much better possibilities to pick a representative selection of varieties.

[2] The program PRAAT is a free program and available via *http://www.fon.hum. uva.nl/praat/*.

[3] See also *http://www.bsos.umd.edu/hesp/newman/Newman_classes/Newman604/604.html*.

[4] In the context: "kven av dei', which means: "which of them".

[5] In the context: "ta av frakken", which means: "take the coat off".

[6] For determining whether two correlation coefficients are significantly different or not we used the website of *VassarStats* which can be found at: *http://faculty.vassar.edu/lowry/VasserStats.html*.

References

Bolognesi R., Heeringa W. (2002) De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten. *Gramma/TTT: tijdschrift voor taalwetenschap*, 9(1), pp. 45–84.

Christiansen H. (1954) Hovedinndelingen av norske dialekter. *Maal og Minne*, pp. 30–41.

Gooskens C., Heeringa W. (2003) Perceptive Evaluation of Levenshtein Dialect Distance Measurements Using Norwegian Dialect Data. *Language Variation and Change*, Submitted.

Hunt M. J., Lennig M., Mermelstein P. (1999) Use of Dynamic Programming in a Syllable-based Continuous Speech Recognition System. In Sankoff, D. and Kruskal, J. (eds.), *Time Warps, String Edits, and Macro Molecules; The Theory and Practice of Sequence Comparison*, CSLI, Stanford, 2nd edition, pp. 1–44. 1st edition appeared in 1983.

Jain A. K., Dubes R. C. (1988) *Algorithms for Clustering Data*. Englewood Cliffs, Prentice Hall, New Yersey.

Kessler B. (1995) Computational Dialectology in Irish Gaelic. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, Dublin, pp. 60–67.

Kruskal J. B. (1964) An Overview of Sequence Comparison. In Sankoff, D. and Kruskal, J. (eds.), *Time Warps, String Edits, and Macro Molecules; The Theory and Practice of Sequence Comparison*, Addison-Wesley, Massachusetts, pp. 1–40.

Kruskal J. B. (1999) An Overview of Sequence Comparison. In Sankoff, D. and Kruskal, J. (eds.), *Time Warps, String Edits, and Macro Molecules; The Theory and Practice of Sequence Comparison*, CSLI, Stanford, 2nd edition, pp. 1–44. 1st edition appeared in 1983.

Nerbonne J., Heeringa W., van den Hout E., van der Kooi P., Otten S., van de Vis W. (1996) Phonetic Distance between Dutch Dialects. In Durieux, G., Daelemans, W., and Gillis, S. (eds.), *CLIN VI, Papers from the Sixth CLIN Meeting*, University of Antwerp, Center for Dutch Language and Speech (UIA), Antwerp, pp. 185–202.

Rietveld A. C. M., van Heuven V. J. (1997) *Algemene Fonetiek*. Coutinho, Bussum.

Skjekkeland M. (1997) *Dei norske Dialektane: tradisjonelle sædrag i jamføring med skriftmåla*. HøskoleForlaget, Kristiansand.

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

Ten Bosch L. (2000) ASR, Dialects, and Acoustic/Phonological Distances. In *ICSLP2000*, Beiging.