Measuring Norwegian Dialect Distances using Acoustic Features

Wilbert Heeringa* Keith Johnson† Charlotte Gooskens‡

16th January 2006

1 Introduction

Computational dialectometry has been proven to be useful for finding dialect relationships and identifying dialect areas. The first to develop a method of measuring dialect distances was Jean Séguy, assisted and inspired by Henri Guiter (Chambers and Trudgill, 1998). Strongly related to the methodology of Séguy is the work of Goebl, although the basis of Goebl's work was developed mainly in dependently of Séguy (Goebl, 1982, 1993). In 1995 Kessler used the *Levenshtein distance* for finding linguistic distances between Irish Gealic dialects, and in 1996 the same algorithm was applied to Dutch dialects by Nerbonne et al.. The The Levenshtein distance is a sensitive measure with which distances between strings (in this case transcriptions of word pronunciations) are calculated. Gooskens and Heeringa (2004) showed that linguistic dialect distances measured with Levenshtein correlate significantly with perceptual distances for 15 Norwegian varieties ($r = 0.67$, $p < 0.001$).

Pronunciation based dialect 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 sometimes varies greatly, depending on the skills of the transcriber. When several transcribers are involved, the data may not be consistent. What the result of this can be, is, for example, shown by Heeringa (2005) who found the Frisian dialect area to be divided in a northern and southern part, which reflects the work areas of the two transcribers.

In the field of *Automatic Speech Recognition* methods can be found with which pronunciations are compared on the basis of the acoustic signal, without intervention of a transcriber. Among others we mention Hunt et al. (1999) and

*University of Groningen, Department of Information Science
†UC Berkeley, Department of Linguistics
‡University of Groningen, Scandinavian Department

Ten Bosch (2000). A first attempt to measure dialect distances acoustically was made be Heeringa and Gooskens (2003). Although their methodology is mainly acoustically based, they still consult transcriptions for the purpose of speech rate normalization.

The goal of this paper is to find a fully acoustically based measure which approximates the quality of the semi-acoustically based measure of Heeringa and Gooskens (2003). We will experiment with different representations of the acoustic signal to investigate which representation gives the best results. The results are validated by comparing them to results of a perception experiment of Charlotte Gooskens. Both our computational measurements and the perception experiment are based on recordings of the same 15 Norwegian dialects. The data come from a database compiled by Jørn Almberg and Kristian Skarbø.[1]. The database comprises translations of the fable 'The North Wind and the Sun'. Both recordings and transcriptions are available. The perception experiment is based on the recordings and our computational measurements on the transcriptions. The distribution of the 15 varieties is shown in Figure 1. 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). In this map the Norwegian language area is divided in nine dialect areas. In our set of 15 varieties six areas are represented.

In Section 2 we describe the perception experiment. In Section 3 we describe our acoustic model and its parameters. In Section 4 we validate the results of our methodology and show some results. In Section 6 some conclusions will be drawn.

2 Perceptual distance measurements

In this section we briefly describe the perception experiment and show some results. A detailed description is given by Gooskens and Heeringa (2004).

2.1 Experiment

In order to obtain distances between 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

[1]Department of Linguistics, University of Trondheim. The recordings are available at `http://www.ling.hf.ntnu.no.nos`. When 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.

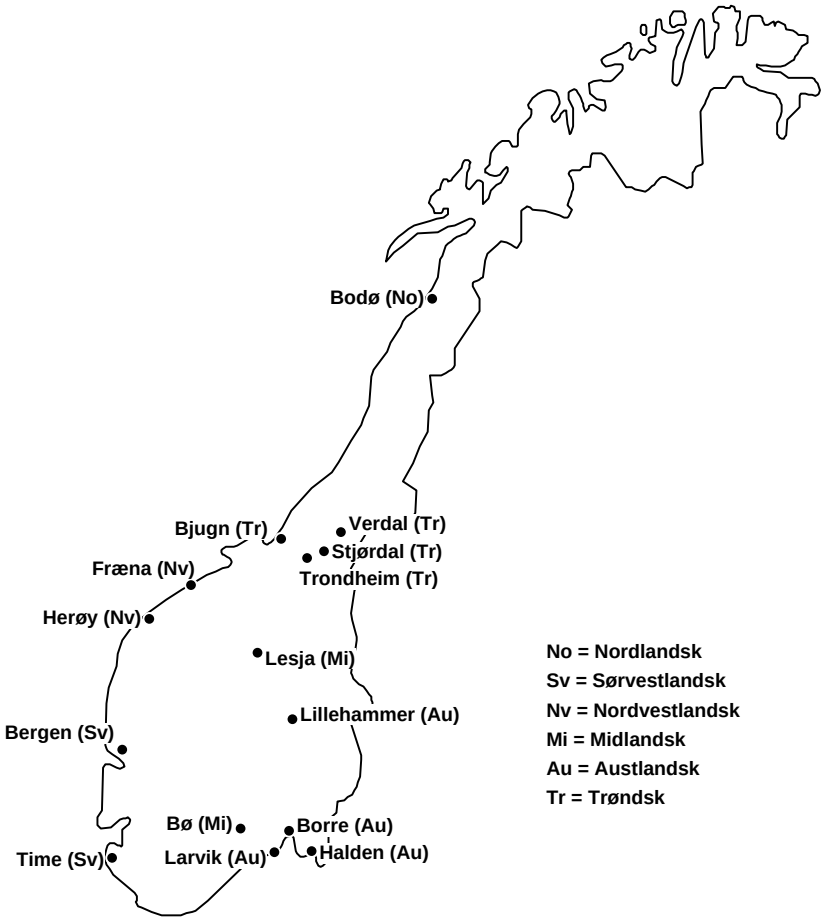


Figure 1: Map of Norway showing the 15 dialects in the present investigation. Skjekkeland (1997) distinguishes nine Norwegian dialect groups. Six groups are represented by our set of 15 dialects. The abbreviation after the name of each location indicates the dialect group to which the variety belongs. 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).

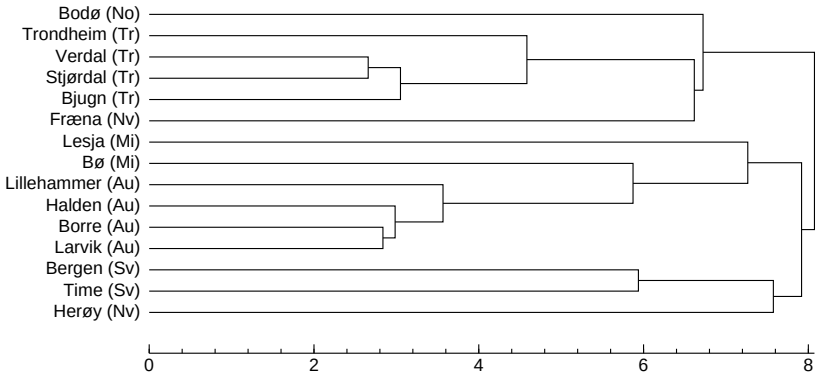


Figure 2: Dendrogram derived from the 15 × 15 matrix of perceptual distances showing the clustering of (groups of) Norwegian dialects. The tree structure explains 91% of the variance.

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 × 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 to Bergen. 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. In Figure 2 the dendrogram produced by cluster analysis using *group average* is presented.

Furthermore a multidimensional scaling analysis was carried out. In our research we used MDS routines as implemented in the statistical R package.[2] 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

[2]The program R is a free public domain program and available via `http://www.r-project.org/`.

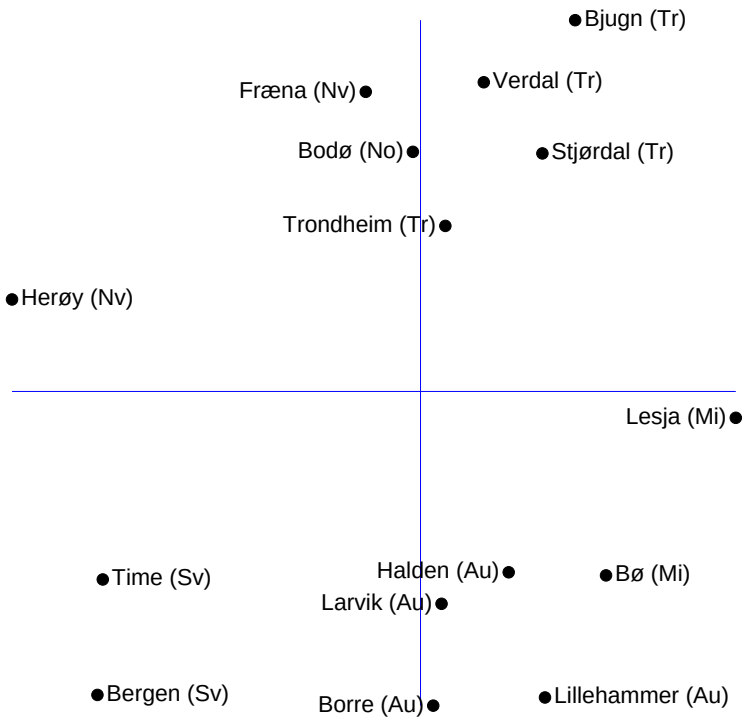


Figure 3: Multidimensional scaling of the results derived from the 15 × 15 matrix of perceptual distances. The vertical axis represents the first dimension, and the horizontal axis the second dimension. The two dimensions explain 67% of the variance.

dialects). In the multidimensional scaling plot a northern, a western and a south-eastern 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.

3 Acoustic distance measurements

In this section we decribe the acoustic model we used for calculating linguistic distances between the 15 Norwegian dialects. In cases where our model differs from the model of Heeringa and Gooskens (2003), we will make a remark about

that.

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 (nearly) 58 words were cut from the text, so we usually got 58 word samples per dialect. If the same word appears more than once in a text, we selected only the first occurrence.

Vowels are often remarked to be the more fluid bearers of varietal differences. For example Nerbonne (2005) showed that vowels are responsible for a great deal of the Southern American English dialect variation. Therefore, we also perform measurements on the basis of vowels only. We cut vowels from those words which appear in all 15 dialects and which have the same lexeme in all dialects. In this way we got 34 vowels corresponding with 34 words. The vowels show phonetic variation across the 15 dialects. The selected vowels appear in almost every dialect, but in a very few cases they were absorbed by a following nasal. In these cases the first part of the nasal is cut.

Heeringa and Gooskens (2003) did not use the orginal samples, but monotonized versions of them in order to remove gender differences. They realized that in this way prosodic information is lost as well. Pitch and intonation contours are known to be significant dialect markers in Norwegian (Christiansen, 1954; Fintoft and Mjaavatn, 1980). Therefore we use the original, non-manipulated samples and look for other ways to neutralize gender differences (see Section 3.5).

3.2 Representations

Heeringa and Gooskens (2003) only examined one acoustic feature: formant tracks. Vowels can be easily identified by their formants as can be seen in the IPA quadrilateral, where height corresponds with the first formant and advancement with the second formant (Rietveld and Van Heuven, 1997, p. 133). In addition we also consider zero crossing rates. The zero crossing rate is sensitive to the difference between voiced and unvoiced speech sections. High zero crossing rates indicate noise, i.e. frication and low values are found in periodic, i.e. sonorant parts of speech (Frankel et al., 2000).

3.2.1 Formant tracks

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 blended 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.

In PRAAT[3] several algorithms can be chosen for finding the Linear Predictive Coding (LPC) coefficients. We chose the algorithm of Burg. This algorithm may initially find formants at very low or high frequencies. However we used the version in PRAAT which removes formants below 50 Hz and formants above the maximum formant frequency minus 50 Hz.[4] The algorithm of Burg is much more reliable than the Split Levinson algorithm which always finds the requested number of formants in every frame, even if they do not exist.

The number of formants may vary over time in a word and per word. In the PRAAT program, we maintained the default value for the maximum number of formants which may be found: 5. Next, we found the minimum number of formants examining all points in time of all words which are taken into consideration. After that, on the basis of this minimum number of formants the word samples were compared. In the samples we used (see Section 3.1) for each word sample at each time sample, at least three formants could be found. Therefore, the comparison of word samples here is based on (the first) three formant tracks.

When finding formants in the computer program PRAAT, the time step was set to 0.01 seconds with an analysis window of 0.025 seconds. The ceiling of the formant search range should be set to 5000 Hz for males, and to 5500 Hz for females. Since we want to use the same ceiling for both males and females, we set it to the average of 5250 Hz.

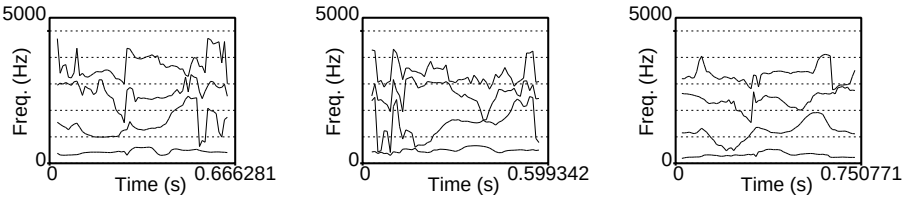
Pre-emphasis starts at 50 Hz. In the manual which can be found in the PRAAT program pre-emphasis is explained as follows:

> "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 Figure 4 we show visualizations of three Norwegian pronunciations of the word *nordavinden* 'the northwind' using formant tracks. The pronunciations of the dialects of Bjugn, Halden and Larvik are given.

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

[4]The maximum formant frequency was set to 5250 Hz, which is the average of 5000 Hz (males) and 5500 Hz (females).



Bjugn	Halden	Larvik
[²nuːɽɑˌʊjɲˑ]	[²nuːrɑˌʊinː]	[²nuɽɑˌʊinˑɳ̩]

Figure 4: Formant track representations of three Norwegian pronunciations of *nordavinden* 'the northwind'.

The PRAAT program gives formant frequencies in Hertz. We also consider frequencies in Bark, which may be a more faithful scale perceptually. For this purpose we used the formula of Traunmüller (1990) as suggested in standard works about phonetics (Rietveld and Van Heuven, 1997, e.g.):

$$Bark = \frac{26.81 \times Hertz}{1960 + Hertz} - 0.53 \tag{1}$$

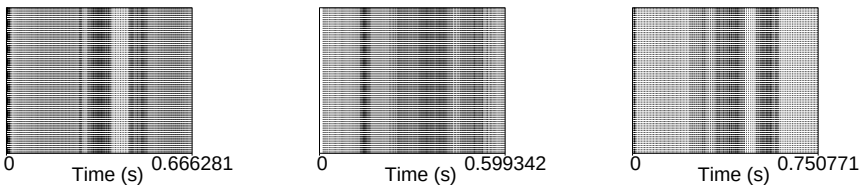
Furthermore we experimented with an approach in which z-scores instead of either the Hertz or the Bark frequency values are used. Per frame we calculated the mean and the standard deviation. Next within frame f a z-score is calculated for each frequency f_i:

$$f_i = \frac{f_i - f_{mean}}{f_{standard\ deviation}} \tag{2}$$

3.2.2 Zero crossing rates

The number of times per interval of time that the amplitude waveform crosses the zero line is called the zero crossing rate. Zero crossing raisers are the points in time when the waveform changes from negative to positive, and fallers represent the times when the amplitude goes down from positive to negative.

PRAAT offers a function which gives us the points in time of the raisers or fallers or both raisers and fallers. We used the default setting: raisers only. However, when using fallers or raisers and fallers, nearly the same results are obtained. We converted the zero crossing times to zero crossing rates using a time step of 0.01 seconds, the same as used in the formant analysis. The analysis window was set to a different size: 0.05 seconds. A larger analysis window gives more fluent estimations, but the size of our analysis window is just smaller than the shortest word sample.



Bjugn	Halden	Larvik
[²nuːɽɑˌʊjɻˑ]	[²nuːrɑˌʊinː]	[²nuɽɑˌʊinˑɳ̩]

Figure 5: Zero crossing distributions of three Norwegian pronunciations of *nordavinden* 'the northwind'. White vertical lines across the black horizontal lines represent the times of the zero crossing raisers.

In Figure 5 the zero crossing distributions are shown for three Norwegian pronunciations of the word *nordavinden* 'the northwind'. Again the pronunciations of the dialects of Bjugn, Halden and Larvik are given.

3.3 Leading and trailing zeros

The words we used are cut from a running text. Although the samples are cut as acurate as possible, some leading or trailing silence may still be included. We removed them automatically. Heeringa and Gooskens (2003) did not remove leading and trailing zeros.

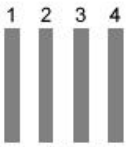
3.4 Speech rate normalization

Different samples sizes may reflect dialect variation, but can also be the result of different speech rates. Therefore we had to normalize over speech rate. Heeringa and Gooskens (2003) normalized over the number of segments of a sample according to the transcription. We describe this transcription-based approach in more detail in Section 3.4.1. Since our goal was to develop a fully transcription independent methodology, we also consider another normalization procedure where the samples of a word pair are stretched so that they get the same number of frames. That transcription independent approach is discussed in Section 3.4.2.

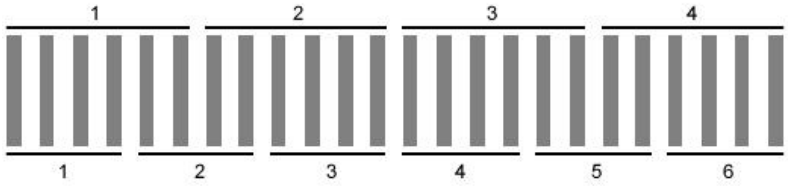
3.4.1 Transcription-based

Assume that the acoustic representation of a word sample consists of l frames. If the number of segments of this word pronunciation according to the phonetic transcription is m, and we want to represent each segment by n frames, then we represent the complete word sample by $m \times n$ frames. Changing the representation of l frames into a representation of $m \times n$ frames is realized in two steps. First

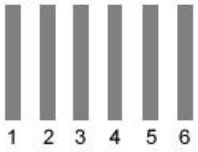
we duplicate each of the l frames $m \times n$ times. This gives $l \times m \times n$ frames in total. Second we regard the $l \times m \times n$ frames as $m \times n$ groups, each consisting of l frames, and fuse the frames in each group to one frame by averaging them. The result is a representation of $m \times n$ frames. We illustrate this by an example. Assume we have a word sample of $l = 4$ frames:



If this word pronunciation is transcribed as a sequence of $m = 2$ frames, and we want to represent each segment by $n = 3$ frames, then we represent the complete word sample by $2 \times 3 = 6$ frames. We change the representation of 4 frames into a representation of 6 frames. For this purpose first we duplicate each of the 4 frames 6 times. This gives 24 frames in total:



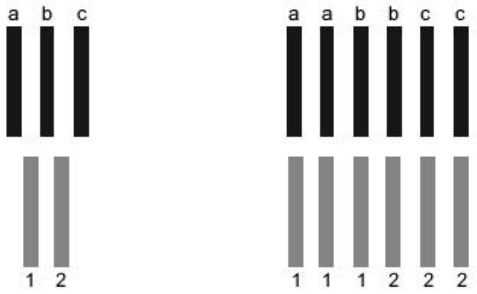
Second we treat the 24 frames as 6 groups, each consisting of 4 frames, and fuse the frames in each group to one frame by averaging them. The result is a representation of 6 frames:



In our research we chose $n = 20$, i.e. 20 frames per segment. A higher value gives nearly the same results, but the computing time increases greatly.

3.4.2 Transcription independent

When comparing one segment of m frames with another segment of n frames, each of the m frames is duplicated n times, and each of the n frames is duplicated m times. So both segments get a length of $m \times n$. Below two segments are schematically visualized, one with 3 frames (black bars) and one with 2 elements (grey bars). Now both get a length of 6 when each of the 3 frames are duplicated 2 times, and each of the 2 frames are duplicated 3 times.



3.5 Comparison of frames

Formant tracks When using the formant track representation, a sample is represented as a series of frames, each frame having three formant frequency values. When comparing a frame of a word pronunciation of one dialect with the corresponding frame of the corresponding word pronunciation of another dialect, the distance is calculated as:

$$d(f_1, f_2) = \sum_{i=1}^{n} \mid f_{1_i} - f_{2_i} \mid \tag{3}$$

where $n = 3$.

The distance measure we used is known as the *Manhattan* distance. Heeringa and Gooskens (2003) used the *Euclidean* distance: the square root of the sum of the squared differences. Since we found the better results with the Manhattan distance, this measure will be used throughout this paper.

A frame in one sample does not always correspond with another frame in the second sample. Frames can be inserted or deleted (see Section 3.6. In these cases frames are compared to a 'silence frame'. A 'silence formant frame' is defined as a frame for which all frequencies are equal to 0. This means that in absolute silence there are no vibrations. When using z-scores instead of the original Hertz or Bark values, the values are still set to 0.

Zero crossing rates When using zero crossing rates, frames consist of only one value. The distance between two frames is equal to the absolute difference of the two zero crossing rates. The value in a 'silence zero crossing rate frame' is set to 0: there are no zero crossings during silence.

Combined representation When combining formant frame distances with the corresponding zero crossing rate distances, the two distances are multiplied:

$$d(f_1, f_2) = \left(\sum_{i=1}^{n} \mid formant_{1_i} - formant_{2_i} \mid \right) \times \mid zero_1 - zero_2 \mid \tag{4}$$

where $n = 3$.

3.6 Levenshtein distance

Using the Levenshtein algorithm, the distance between two words is determined by comparing the pronunciation of a word in the first dialect with the pronunciation of the same word in the second. The algorithm determines 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. A detailed description is given by Kruskal (1999). We illustrate the algorithm by an example. Assume *gåande* or *gående* 'going' is pronounced as [goːɑns] in the dialect of Bø and as [gɔːnə] in the dialect of Lillehammer. Changing one pronunciation into the other can be done as follows (ignoring suprasegmentals and diacritics):

goːɑns	substitute o/ɔ	1
gɔːɑns	delete ɑ	1
gɔːns	insert ə	1
gɔːnəs	delete s	1
gɔːnə		
		4

In fact many sequence operations map [goːɑns] to [gɔːnə]. 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 words will generally be greater than the distance between shorter words. The longer the words, the greater the chance for differences with respect to the corresponding word in another dialect. 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 get following alignment:

g	oː	ɑ	n		s
g	ɔː		n	ə	
	1	1		1	1

In this paper, Levenshtein distance was applied to acoustic samples instead of phonetic transcriptions. Instead of phonetic segments, acoustic frames were aligned. In our example all operations have a weight of 1. However, when comparing acoustic samples, substitutions, insertions and deletions have gradual weights, calculated in the way as described in Section 3.5. Levenshtein distance was used in the same way by Heeringa and Gooskens (2003).

Using 58 words the distance between two dialects is equal to the average of 58 Levenshtein distances, and when using 34 vowels the distance is equal to the mean of 34 Levenshtein distances. When comparing two varieties on the basis of k word pairs, it may appear that for one or more of the pairs for one or both

varieties, no sample is available. This can be the result of either the fact that no translation is given or the fact that the sample was smaller than the analysis window used (in case of the vowels). In these cases, the word or vowel pair was ignored.

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

4 Validation

In this section we validate our computational results with the results of the perception experiment. For this purpose we correlate the computational distances with the perceptual distances (Section 4.1). We distinguish three types of measurements: vowel-based measurements (Section 4.2), semi-acoustic word sample-based measurements (Section 4.3) and (fully) acoustic word sample-based measurements (Section 4.4). We end with some conclusions (Section 4.5).

4.1 Correlation

In order to correlate the different computational measurements to the results of the perception experiment, the computational 15×15 matrices are correlated with the perceptual 15×15 matrix. When correlating we exclude the distances of dialects with respect to themselves, i.e. the distance of Bergen to Bergen, of Bjugn to Bjugn etc. These distances are found on the diagonal in the distance matrix, containing the cells $(1, 1), (2, 2), \ldots (n, n)$. In computational matrices these values are always 0, in the perceptual matrix they vary, usually being higher than the minimum score. This may be the result of the fact that for example the dialect of the speaker of Bergen is different from the dialect of the listeners in the same location. Since this causes uni-directional distortion for the diagonal distances (they only can be too high, not too low), we exclude them whe calculating the correlation coefficient.

For finding the correlation coefficient, we used the Pearson's correlation coefficient (Sneath and Sokal, 1973, pp. 137–140). For finding the significance of a correlation coefficient we used the Mantel test. In classical tests the assumption is made that the objects which are correlated are independent. However, values in distance matrices are usually correlated in some way, and not independent (Bonnet and Van de Peer, 2002). A widely used method to account for distance correlations is the Mantel test (Mantel, 1967). As significance level we choose $\alpha = 0.05$. With the Mantel test it is also possible to determine whether one correlation coefficient is significantly higher than another.

4.2 Acoustic vowel measurements

Table 1 shows correlation coefficients between perceptual distances and different acoustic distance measurements. The measurements are made on the basis of vowels only. Correlations are given for the complete data set of 15 dialects. Since the mean vocal tract dimensions of males differ from those of females, gender differences may influence our results. Therefore we also show correlations on the basis of a subset of 11 dialects. The recordings of these dialects are pronounced by female speakers. The dialects of Bø, Bodøe, Herøey and Larvik are pronounced by male speakers and excluded in the smaller set.

In the table we find three acoustic representations: formant tracks, zero crossing rates, and a combined representation where both formant tracks and zero crossing rates are used (see Sections 3.2 and 3.5). When formant tracks are used, we consider the Hertz scale and the Bark scale (see Section 3.2.1. Besides measurements on the basis of the original Hertz and Bark frequencies, also measurements are given where the frequencies are normalized per frame (see Section 3.2.1).

For each of the measurements we checked whether the 34 vowels are a sufficient basis for reliable analyses. We calculated Cronbach's α values for each of them. A widely-accepted threshold in social science for an acceptable α is 0.70 (Nunnally (1978), Heeringa (2004, p. 170–173)). For ten measurements we found lower α values, varying from 0.16 to 0.66. For these measurements the correlations are given in normal type setting, the other ones are printed in bold.

Considering differences in representation, the combined representation (formant tracks and zero crossing rates) has higher correlations than the formant track representation which has in turn higher correlations than the zero crossing rate representation in most cases. The measurements on the basis of the Bark scale have mostly higher correlations than those on the basis of the Hertz scale. Normalizing frequencies improves results when using the combined representation, but not when using formant tracks only. For none of the three factors (representation, frequency scale and normalization) we found significant differences. The highest correlation is obtained when using the combined representation and normalized Bark frequencies, followed by the version with the Hertz frequencies.

4.3 Semi-acoustic word measurements

Table 2 shows correlation coefficients between perceptual distances and different acoustic distance measurements. The measurements are made on the basis of words. The speech rate is normalized by counting the number of segments in the transcriptions (see Section 3.4.1). Just as for the vowel-based comparisons correlations are given for all 15 dialects and for the 11 dialects, pronounced by females.

As in Section 4.2 for each of the measurements we checked whether the 58

formant bundles	zero crossings	original frequencies		normalized frequencies	
		15 dial.	11 dial.	15 dial.	11 dial.
Hertz	no	**0.22**	0.30	0.23	0.25
Bark	no	**0.24**	0.32	0.23	0.28
Hertz	yes	**0.28**	**0.27**	**0.34**	0.32
Bark	yes	**0.31**	**0.31**	**0.35**	0.33
no	yes	**0.21**	0.15	**0.21**	0.15

Table 1: Correlations between perceptual distances and different acoustic distance measurements based on vowels only. Correlations are given for 15×15 and 11×11 matrices excluding the diagonals. All correlations are significant for α=0.05. Correlations in bold are based on measurements with Cronbach's $\alpha > 0.70$.

words are a sufficient basis for reliable analyses and calculated Cronbach's α values for each of them. Most of them were higher than the threshold of 0.70, but six of them were lower, varying from 0.63 to 0.67. For these measurements the correlations are given in normal type setting, the other ones are printed in bold.

In the table we find that both the formant track representation and the combined represention have mostly higher correlations than the zero crossing rate representation. The combined represention gives only an improvement in comparison with the formant track representation when normalized frequency values are used. Considering the frequency scale we find that the Bark scale gives the better results, but when frequencies are normalized the Hertz scale gives the better results. Frequency normalization improves results when the combined represention is used, but not when using formant tracks in most cases. Considering the three factors (representation, frequency scale and normalization) we did not find significant differences in most cases. For 15 dialects the highest correlation is obtained when using the formant track representation and original Bark frequencies. For 11 dialects the two highest candidates are the same as we found for vowels (see Section 4.2). The highest correlation coefficient is found when the combined representation is used and normalized Hertz frequencies are used. Using Bark frequencies gives the second best correlation.

4.4 Acoustic word measurements

Table 3 shows correlation coefficients between perceptual distances and different acoustic distance measurements. The measurements are made on the basis of words. When two word samples are compared, they are stretched so that they get the same number of frames before the distance between them is calculated (see Section 3.4.2). As for the vowel-based comparisons and the semi-acoustic

formant bundles	zero crossings	original frequencies		normalized frequencies	
		15 dial.	11 dial.	15 dial.	11 dial.
Hertz	no	**0.49**	**0.57**	**0.50**	**0.55**
Bark	no	**0.53**	**0.58**	**0.48**	**0.55**
Hertz	yes	0.37	0.47	**0.49**	**0.60**
Bark	yes	0.41	0.51	**0.49**	**0.59**
no	yes	**0.36**	0.50	**0.36**	0.50

Table 2: Correlation between perceptual distances and different semi-acoustic distance measurments based on words. Correlations are given for 15×15 and 11×11 matrices excluding the diagonals. All correlations are significant for α=0.01. Correlations in bold are based on measurements with Cronbach's $\alpha > 0.70$.

word-based comparisons, correlations are given for all 15 dialects and for the 11 dialects, pronounced by females.

As in Sections 4.2 and 4.3 for each of the measurements we checked whether the 58 words are a sufficient basis for reliable analyses and calculated Cronbach's α values for each of them. We found most of them to be higher than the threshold of 0.70, but four of them were lower, varying from 0.56 to 0.68. For these measurements the correlations are given in normal type setting, the other ones are printed in bold.

Looking at the table, we find that both the zero crossing rate representation and the combined representation have higher correlation coefficients than the formant track representation. The combined representation gives better results than the zero crossing rate representation when normalized frequency values are used. Considering the frequency scale, again our findings accords with those of the semi-acoustic measurements: the Bark scale gives the better results when original frequencies are used, but when frequencies are normalized the Hertz scale gives the better results. Frequency normalization improves results for both the formant track representations and the combined representations. For the semi-acoustic measurements we found improvements only for the combined represention. The highest correlation is obtained when using the combined representation and normalized Hertz frequencies, followed by the version with the Bark frequencies. For 15 dialects we found the correlations of the two measures to be significantly higher than those using the formant track representation and non-normalized frequencies (0.46 versus 0.26, p=0.02, 0.46 versus 0.27, p=0.03). For 11 dialects they are nearly significantly higher (0.56 versus 0.36, p=0.07, 0.55 versus 0.36, p=0.07). The two versions also had the highest correlations for the vowel measurements and for the semi-acoustic measurements based on the 11 female dialects.

formant bundles	zero crossings	original frequencies		normalized frequencies	
		15 dial.	11 dial.	15 dial.	11 dial.
Hertz	no	**0.26**	**0.36**	**0.35**	**0.39**
Bark	no	**0.27**	**0.36**	**0.31**	**0.38**
Hertz	yes	0.36	0.46	**0.46**	**0.56**
Bark	yes	0.38	0.49	**0.46**	**0.55**
no	yes	**0.37**	**0.49**	**0.37**	**0.49**

Table 3: Correlations between perceptual distances and different acoustic distance measurements based on words. Correlations are given for 15×15 and 11×11 matrices excluding the diagonals. All correlations are significant for α=0.01. Correlations in bold are based on measurements with Cronbach's $\alpha > 0.70$.

4.5 Conclusions

Comparing the results in Sections 4.2, 4.3 and 4.4, we found the correlations of the vowel-based measurements to be lower than those of the word-based measurements. This can easily explained by the fact that only one vowel contains less information than a complete word. Nevertheless all vowel correlations are significant for $\alpha = 0.05$.

Representation The correlations of the semi-acoustic measurements are higher than those of the (fully) acoustic measurements, but with one exception. For the zero crossing rates we did not find a clear difference: 0.36 versus 0.37 (15 dialects) and 0.50 versus 0.49 (11 dialects). This gives us the impression that zero crossing rate measurements are quite robust in the sense that they are speech rate normalization procedure-independent. Possibly zero crossing distributions represent the segmental structure to some extent. This may explain why combined measurements are better than formant track measurements for all cases of the acoustic measurements, but for only the half of the cases of the semi-acoustic measurements. In case of the semi-acoustic measurements, segmental information is already read from the transcriptions, the segmental information of the zero crossing distribution may partly be superfluous.

Frequency scale Looking at the vowel based measurements we find the tendency that the Bark scale gives higher correlations than the Hertz scale. For the word-based measurements we find the same when the original, non-normalized frequency values are used. When normalized frequency values are used, there is hardly any difference, in a few cases the Hertz measurements are just higher. Therefore the use of the Bark scale is only useful when non-normalized frequency values are used.

Frequency normalization The use of normalized frequency values gives only an improvement for the combined representation when measurements are obtained on the basis of vowels or on the basis of semi-acoustic word sample measurements. For acoustic word sample measurements normalization leads to improvement for both the formant track representation and the combined representation. The idea behind frequency normalization within a frame is that the relative positions of the F1, F2 and F3 to each other is more important than the absolute values of the three formants. But then still we cannot explain our findings here.

Our choice For all measurements, both vowel-based, semi-acoustic word-based and acoustic word-based, we found that the same two measurements outperform the other ones, namely the versions using the combined representation and normalized frequency values. To decide about the frequency scale we look at the very small difference for the 11 female dialects where the Hertz scale just gives higher results. So we choose the version which uses the Hertz scale.

5 Results

In this Section we present hierarchical clustering and multidimensional scaling analysis of the computational method which appears to be the best one in Section 4, i.e. which shows results which approximates the perceptual distances most closely: the version with the combined reresentation (formant tracks and zero crossing rates) and normalized Bark frequencies. Since it is our aim to develop a fully transcription-independent comparison method, we present results of the version which does not use any information from transcriptions. For all 15 dialects, its results correlate with $r = 0.46$ to perception, and for 11 dialects we found $r = 0.56$.

On the basis of the distances obtained with this method, the dialects are classified. As in Section 2.2 for the perceptual distances, we perform cluster analysis and multidimensional scaling. In Figure 6 the dendrogram is shown. When we compare this dendrogram with the one in Figure 2, we find in the two figures that both the Trøndsk, the Austlandsk and the Sørvestlandsk dialects are clustered together. The dendrograms disagree about the Midlandsk dialects and the Nordvestlandsk dialects. Lesja is clustered along with the Trøndsk dialects in the perceptual dendrogram and along with the Austlandsk dialects in the computational dendrogram. The disagreements about Bø, Bodø and Herøy may have to do with the fact that their recordings were pronounced by males. However the position of Larvik, which recording was also pronounced by a male, is not so deviant in comparison with the perceptual dendrogram.

We also applied multidimensional scaling. The resulting plot is shown in Figure 7. This two dimensional plot explains 74.3% of the variance of the original

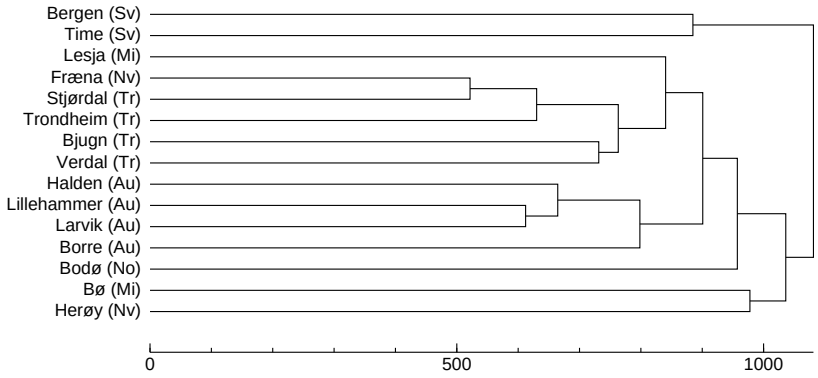


Figure 6: Dendrogram obtained on the basis of Levenshtein distances where the combined representation (formant tracks and zero crossing rates) is used. The tree structure explains 41.5% of the variance.

computational distances. When comparing this plot with the perceptual based plot in Figure 3, we find some similarities: The Sørvestlandsk dialects and the dialect of Herøy on the left, the Trøndsk dialects on top and the Austlandsk dialects in the lower right corner. However in the perceptual plot the different groups are much sharper distinguished. Furthermore, the dialects of Bø, Bodø and Herøy are located on top, and the dialect of Larvik is found much higher and more distant from the dialect of Halden than in the perceptual multidimensional scaling plot. Since these four dialects are pronounced by males, we get the impression that the first dimension, represented by the vertical axis in the plot, represents gender to a large extent.

Possibly it may be better to ignore the first dimension, and examine the second and higher dimensions. In order to find out whether higher dimensions may be interesting, we scaled our computational distances to the largest possible number of dimensions allowed by the R program: 12. Next we calculated distances between the 15 dialects per dimension, resulting in 12 distance matrices. Next we correlated each of the matrices with the perceptual distance matrix. We squared the correlation coefficients and multiplied them by 100. In this way for each dimension we got a percentage which represents the amount of variance which that dimension explains of the perceptual distances. The variances are shown in Figure 8. This figure suggests that especially the first, second and third dimension are important.

We found that the first dimension distances correlate (nearly) signifantly stronger with the perceptual distances than the fourth and higher dimension distances do (highest p was equal to 0.08). The same applies for the second dimension distances (highest p was equal to 0.09) and the third dimension distances(highest p was equal to 0.07). Therefore we focus on the first, second and third dimension.

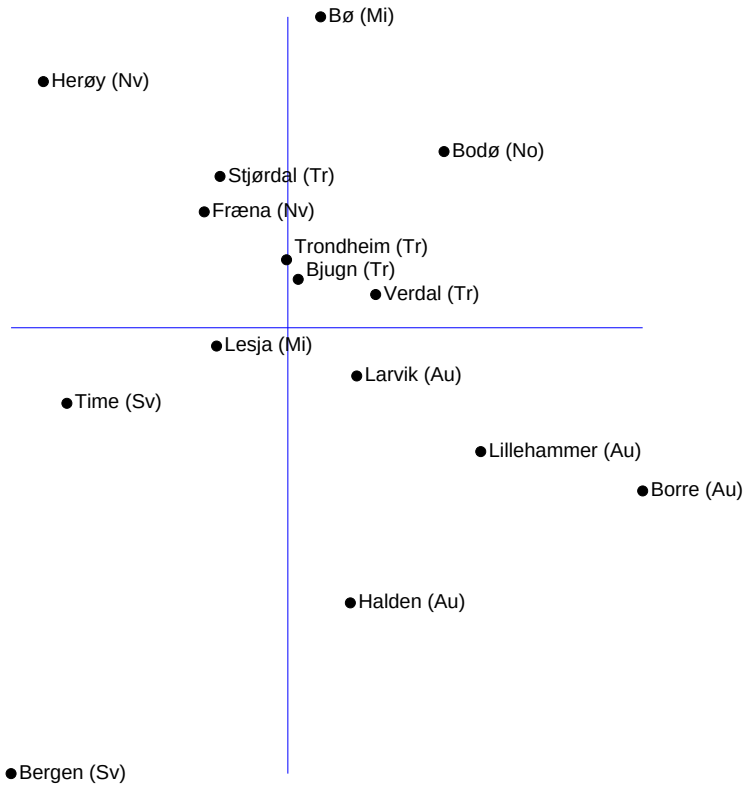


Figure 7: Multidimensional scaling plot obtained on the basis of Levenshtein distances where the combined representation (formant tracks and zero crossing rates) is used. The vertical axis represents the first dimension, and the horizontal axis the second dimension. The two dimensions explain 74.3% of the variance. The dialects of Böe, Bodöe, Heröey and Larvik were pronounced by male speakers.

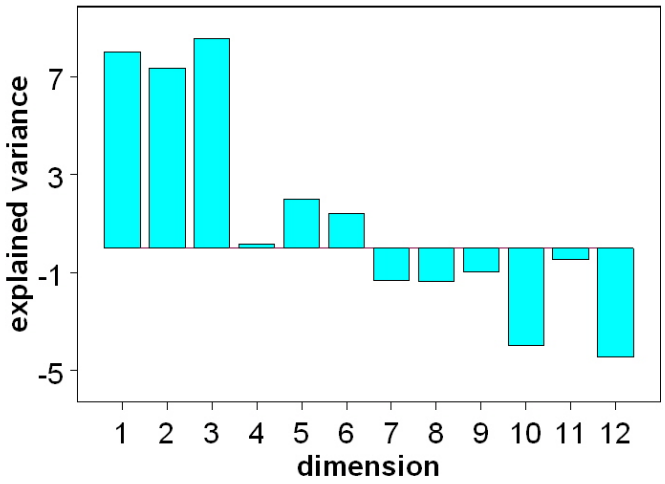


Figure 8: The computational distances are scaled to 12 dimensions. For each dimension a bar shows how much variance that dimension explaines of the perceptual distances. The variances are given in percentages.

In the results of the perceptual measurements we found no influence of gender differences (see Figures 2 and 3). Nevertheless we found a relatively high variance for the first dimension. Therefore we cannot conclude that the first dimension just represents voice quality. However, we want to exclude this dimension since it is the only way to exclude the influence of gender differences. Therefore we scaled the computational dimensions to three dimensions, and drew a plot on the basis of the second and third dimension. The plot is shown in Figure 9. The three dimensions explain 78.3% of the variance of the original computational distances, and the second and third dimension together explain 37.9%.

Different from Figure 7 and more similar to Figure 3 is that the different groups are distinguished more sharply. The western group (mainly Sørvestlandsk dialects), the northern group (mainly Trøndsk dialects) and the southeastern dialects (mainly Austlandsk dialects) can be found in both Figure 3 and Figure 9. Looking at the dialects pronounced by males, we find that they are scattered over the plot. We especially judge the position of Larvik (close to Halden) and Bø (south of the Austlandsk dialects instead of north of the Trøndsk dialects) to be better in comparison with their positions in Figure 7. However the new plot is not an improvement in every respect. The Trøndsk dialects are not as close as in Figure 7, especially Trondheim is very deviant from the other Trøndsk dialects in the new plot. We expected the dialect of Lesja to be more in between of the northern and southeastern dialects as in both Figure 3 and 7.

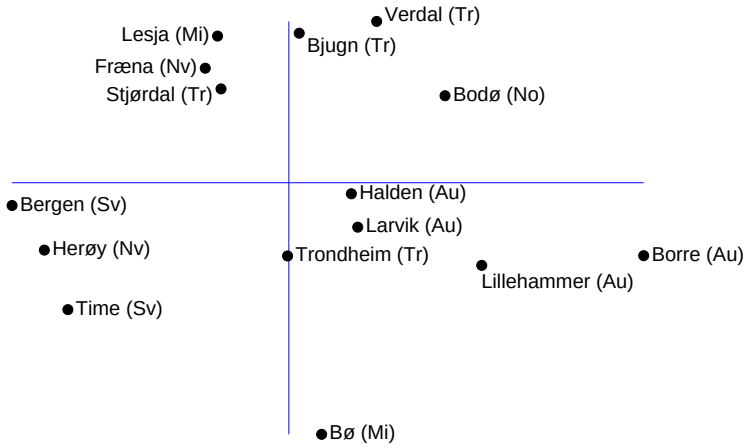


Figure 9: Multidimensional scaling plot obtained on the basis of Levenshtein distances where the combined representation (formant tracks and zero crossing rates) is used. The horizontal axis represents the second dimension, and the vertical axis the third dimension. The two dimensions explain 37.9% of the variance.

6 Conclusion

The aim of this paper was to find a fully acoustic and transcription independent measure for finding dialect distances. Heeringa and Gooskens (2003) presented a semi-acoustic measure. When correlating the distances obtained with that measure with the perceptual distances, they found $r = 0.5$ for all 15 dialects, and $r = 0.55$ for the 11 female dialects (see Heeringa (2004, p. 194)) when using the formant track representation. In this paper we found $r = 0.46$ (15 dialects) and 0.56 (11 dialects) when using a combined representation (formant tracks and zero crossing rates) and normalizing formant frequencies per frame (see Table 3).

When comparing the classiciation results obtained on the basis of perceptual distances with those obtained on the basis of our acoustical measurements, a northern, southeastern and western group can be found for both. However in the acoustical results the influence of gender was found. When using multidimensional scaling, these influence was found in the first dimension. In a plot based on the second and third dimension, the influence of gender is no longer found. However, besides gender-specific variation, the first dimension also represents dialect specific information which is lost when leaving out this dimension. Therefore further research is necessary to filter out the influence of voice quality. Adank et al. (2004) propose different ways of formant frequency normalization which can be examined in future research.

Another issue is speech rate normalization. Our research shows that the results of the semi-acoustic measure are still higher: $r = 0.53$ (15 dialects) and $r =$

0.58 (11 dialects) when using formant tracks and formant frequencies represented in the Bark scale (see Table 2). Therefore it may be useful to seek for a procedure which automatically determines the number of phonetic segments on the basis of the acoustic signal. We found that especially the zero crossing distribution represents the segmental structure to some extent, therefore zero crossings can possibly help in finding the number of segments.

Finally results can be improved by using more data. First, some words occur more than once in the text, for example 'the north wind' occurs four times. We only used the first occurence, but results may be improved when using all occurences of a word and averaging over them. Second, we use one speaker per dialect. Useful future research may be to base results on multiple recordings per dialect.

7 Acknowledgements

We thank 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'. 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 letting us use the programs which he developed for the graphic representation of the maps, dendrograms and multidimensional scaling plots in the present article. This research was carried out within the framework of a talentgrant project, which is supported by a fellowship (number S 30-624) from the Netherlands Organisation of Scientific Research (NWO).

References

Adank, P., Smits, R., and van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, 116:3099–3107.

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

Chambers, J. K. and Trudgill, P. (1998). *Dialectology*. Cambridge University Press, Cambridge, 2nd edition.

Christiansen, H. (1954). Hovedinndelingen av norske dialekter. In Beito, T. B. and Hoff, I., editors, *Frå norsk målføregranskning*, pages 39–48. Universiteitsforlaget, Oslo.

Fintoft, K. and Mjaavatn, P. E. (1980). Tonelagskurver som målmerke. *Maal og Minne*, pages 66–87.

Frankel, J., Richmond, K., King, S., and Taylor, P. (2000). An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces. In *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing.

Goebl, H. (1982). *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*, volume 157 of *Philosophisch-Historische Klasse Denkschriften*. Verlag der Österreichischen Akademie der Wissenschaften, Vienna. With assistance of W.-D. Rase and H. Pudlatz.

Goebl, H. (1993). Probleme und Methoden der Dialektometrie: Geolinguistik in globaler Perspektive. In Viereck, W., editor, *Proceedings of the International Congress of Dialectologists*, volume 1, pages 37–81, Stuttgart. Franz Steiner Verlag.

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

Heeringa, W. (2005). Dialect variation in and around Frisia: classification and relationships. *Us Wurk: Tydskrift foar Frisistyk*, 54(3-4):125–167.

Heeringa, W. and Gooskens, C. (2003). Norwegian dialects examined perceptually and acoustically. *Computers and the Humanities*, 37(3):293–315.

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

Hunt, M. J., Lennig, M., and Mermelstein, P. (1999). Use of dynamic programming in a syllable-based continuous speech recognition system. In Sankoff, D. and Kruskal, J., editors, *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, pages 163–187. CSLI, Stanford, 2nd edition. 1st edition appeared in 1983.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, 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*, pages 60–67, Dublin. EACL.

Kruskal, J. B. (1999). An overview of sequence comparison. In Sankoff, D. and Kruskal, J., editors, *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, pages 1–44. CSLI, Stanford, 2nd edition. 1st edition appeared in 1983.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27:209–220.

Nerbonne, J. (2005). Various Variation Aggregates in the LAMSAS South. In Davis, C. and Picone, M., editors, *Language Variety in the South III*. University of Alabama Press, Tuscaloosa. Accepted to appear.

Nerbonne, J., Heeringa, W., Van den Hout, E., van der Kooi, P., Otten, S., and van de Vis, W. (1996). Phonetic distance between Dutch dialects. In Durieux, G., Daelemans, W., and Gillis, S., editors, *CLIN VI, Papers from the sixth CLIN meeting*, pages 185–202, Antwerp. University of Antwerp, Center for Dutch Language and Speech (UIA).

Nunnally, J. C. (1978). *Psychometric Theory*. McGraw-Hill, New York.

Rietveld, A. C. M. and Van Heuven, V. J. (1997). *Algemene fonetiek*. Coutinho, Bussum.

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

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. 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*, Beijing.

Traunmüller (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America*, 88:97–100.