Measuring Dialect Pronunciation Differences using Levenshtein Distance

Heeringa, Wilbert Jan

Chapter 7

Validating Norwegian dialect distances

From the previous chapters it is clear that a great number of alternative methods is available for comparing dialects. Many of the alternatives are refinements of one another, leading to the question which methods are most suitable in general. In this chapter validation work is reported, which gives an answer to this question.

Section 7.1 starts with an overview of the alternative methods we validate in this chapter. The methods will be validated on the basis of the Norwegian NOS data. This data source is described in Section 7.2. Since measurements are valid only if they are reliable, the reliability of the measurements which are obtained by the word-based methods (frequency method and Levenshtein distance) is checked. The reliability or consistency checking is explained in Section 7.3. Subsequently all methods are validated in Section 7.4. The results of the dialectometric methods are compared to perceptual distances, as found on the basis of a perception experiment. On the basis of reliability checking and validation work we find the optimal comparison method in Section 7.5. We apply this method to the NOS data and show results.

7.1 Overview of methods

In this chapter we validate the different methods with which distances between varieties can be calculated. We examine dialect distance measurements varying several dimensions:

- *Comparison method*
  We examined the corpus frequency method (see Section 2.3.2), the frequency per word method (see Section 2.3.3) and the Levenshtein distance (see Sections 2.3.4, 5.1 and 5.2). The advantage of the frequency per word method compared to the corpus frequency method is that words are re-

garded as linguistic units, and the Levenshtein distance improves on the frequency per word method in that the order of segments in a word is taken into account.

- *Data source*
  All comparison methods are applied to phonetic transcriptions. However, the Levenshtein distance was applied not only to transcriptions (see Section 5.1) but to acoustic word samples as well (see Section 5.2).

- *Transcription segment representation*
  When using transcriptions, speech segments may be represented as phones in the simplest case (see Section 3.1.1). In more refined methods segments are represented by features or acoustically. We examined the feature systems of Hoppenbrouwers & Hoppenbrouwers (H & H, see Section 3.1.2), Vieregge & Cucchiarini (V & C, see Section 3.1.3) and Almeida & Braun (A & B, see Section 3.1.4). These discrete representations are used in combination with the corpus frequency method, the frequency per word method and the Levenshtein distance. To obtain good acoustic representations of canonical segments we examined the Barkfilter (see Section 4.3.1), the cochleagram (see Section 4.3.2) and formant track representations (see Section 4.3.3). Acoustic representations are only used in combination with the Levenshtein distance.

- *Acoustic word representation*
  Above we used acoustic samples of individual segments. In the following section we consider whole word recordings. When using the Levenshtein distance based on acoustic word samples, we experimented with three representations, namely the Barkfilter, the cochleagram and the formant track representation (see Section 5.2.2).

- *Number of length gradations*
  For all variants of comparison methods there is a distinction made between extra-short and non-extra-short sounds which we implemented by changing the transcriptions and weighting non-extra-short sounds at least two times as heavily as extra-short sounds. For the processing of half-long and long we examined two approaches. In the first approach half-long and long are processed by changing feature values or simply ignored when using phone or acoustic segment representations. In this case only *two* degrees of length are represented by weighting segments, namely extra-short and non-extra-short. In the second approach half-long and long are processed by weighting half-long segments three times and long segments four times as heavily as an extra-short sound. In this case *four* degrees of length are represented by weighing segments, namely extra-short, short, half-long and long (see Sections 3.4.2 and 4.6.1).

- *Representation of diphthongs*
  When using transcriptions a diphthong may be processed as the sequence of two segments or as one segment which has a gradual changing color (see Section 3.2 and Section 4.4.

- *Comparison of feature histograms or feature bundles*
  In feature-based measures the distance between feature histograms (corpus frequency method and frequency per word method) or feature bundles (Levenshtein distance) can be determined via Manhattan distance, Euclidean distance, or via a measure based on Pearson's r (see Section 3.6.2.5).

- *Scaling of segment distances*
  Discrete and acoustic segment distances can be used in two different ways in combination with Levenshtein distance. First they can be used unchanged, i.e. linearly. Alternatively, the logarithms of the distances can be used (see Section 3.7).

Although not all of the eight dimensions combine with one another, we nonetheless examine 187 combinations, of which three apply only to acoustic material. The variety reinforces the need for validation techniques.

7.2 Data source

In contrast to many European countries, in Norway dialects are used by people of all ages and social backgrounds both in the private domain and in official contexts (Omdal, 1995). When making recordings the risk is minimal that speakers use a standardized version of their dialect or a variety which is no longer used in every day life. It does not feel unnatural for Norwegian people to read a text aloud in their own dialect.

In the period 1999–2002 Jørn Almberg and Kristian Skarbø (Department of Linguistics, University of Trondheim) compiled a database which consists of recordings of about 50 Norwegian dialects.[1] As a basis the text of the fable 'The North Wind and the Sun' was used. This text was also used in IPA (1949) and IPA (1999) where the text has been transcribed in a large number of different languages. Besides recordings, corresponding transcriptions are also given.

In Gooskens and Heeringa (2003) a perception experiment is described which is based on these recordings (see Section 7.2.1). At the time this experiment was carried out (see Section 7.4.1) recordings of a set of 15 varieties were available. Therefore, the perception experiment was based on 15 varieties. We used the results of this experiment for validation work. In our research the transcriptions of the words in the texts of the same 15 varieties were used as input for a set

[1]The recordings and transcriptions are free available via: `http://www.ling.hf.ntnu.no.nos`.

of 184 transcription-based comparison methods, which are variants of either the corpus frequency method, the frequency per word method and the Levenshtein distance. Samples of the words in the recordings of these 15 varieties were used as input for a set of 3 recording-based comparison methods, which are variants of the Levenshtein distance.

7.2.1 Text recordings

In order to get recordings of translations in different Norwegian dialects of this text, speakers were asked to read the text aloud. The speakers were all given the text in Norwegian beforehand and were allowed time to prepare themselves to be able to read the text aloud in their own dialect. The choice of words and the order of words are sometimes changed to get an authentic rendition. When reading the text aloud speakers were asked to imagine that they were reading the text to someone with the same dialectal background. This was done to ensure a natural reading style and to achieve dialectal correctness.

A set of 15 recordings were used in a perception experiment in order to find perceptual distances among the corresponding 15 varieties. 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. The perception experiment is explained more extensively in Section 7.4.1.

7.2.2 Word transcriptions

On the basis of the recordings transcriptions were made by Jørn Almberg. The transcriptions were made in IPA as well in X-SAMPA (eXtension of Speech Assessment Methods Phonetic Alphabet). In X-SAMPA the IPA symbols are mapped to the ASCII/ANSI characters as found on the keyboard and available in even the most primitive text editors, which makes computational processing of the transcriptions much easier. The big advantage of this data set is that all transcriptions are made by the same person which ensures maximal consistency. The Norwegian translation of the fable 'The North Wind and the Sun' consists of 58 different words. The words are listed in Appendix B Table B.1. The 184 transcription-based computational comparison methods which we validate in this chapter are applied to the transcriptions of translations of these 58 words. For the purpose of validation the same 15 varieties are used as for the perception experiment. Afterwards a larger set of 55 varieties was also used.

Due to the free translation of some phrases a few of the expected words were missing in certain varieties. When two varieties are compared, and when one of the 58 words is missing in a translation of one variety or in both varieties, the word is not taken into account in the calculation of the distance (see Section 5.1.10.1).

Some words occur more than once in the text; e.g., *nordavinden* 'the northwind' normally appears four times in the text. In these cases the mean distance over the variants of one word is used for calculating the distance. (see Section 5.1.10.2). From Section 7.3 it becomes clear that the 58 words are a sufficient basis for reliable dialect comparison.

In Norwegian dialect areas, intonation is one of the most important characteristics. Minimal word pairs can be distinguished by means of tonemes. In the transcriptions three types of tonemes can be found: toneme 1 and toneme 2 (Kristoffersen, 2000) and circumflex (Almberg, 2001). From literature we know that the realization of the same tonemes can vary considerably across the Norwegian dialects. However, no information was given about the precise realization of the tonemes in the transcriptions. We return to this issue below in Section 7.4.1.

7.2.3 Word samples

As mentioned in Section 7.2.2 the Norwegian translation of the fable 'The North Wind and the Sun' consists of 58 different words. For all 15 dialects each of the 58 words were cut from the text, so we usually get 58 word samples per dialect. The 3 recording-based computational comparison methods which we validate in this chapter are applied to the word samples which are selected from the recordings.

The same quantifications we note above in Section 7.2.2 about missing elements in recordings apply here as well. Some words occur more than once in the text. In recording-based comparison only the first occurence is selected since the selection of word samples is rather time-consuming.

The voices of different speakers have different pitches. Most obvious is the difference in pitch between male and female voices. Furthermore, the intonation may vary per speaker. 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 (see Section 5.2.1). In the set of 15 varieties, 4 recordings were recorded by men, and 11 by women (see Section 7.2.4). 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. We are aware of the fact that this choice removes all prosodic information about pitch and intonation contours, and that these 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 are aware of the fact that monotonizing does not remove all gender-dependent information.

7.2.4 Varieties

In Figure 7.1 the geographical distribution of the 15 varieties 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. Figure 7.2 shows which dialect areas the 15 varieties belong to according to the map of Skjekkeland (1997).

For both the perception experiment and the recording-based comparison methods, the distinction between males and females is important. In the set of 15 dialects, the varieties of Bodø, Bø, Herø and Larvik are recorded by male speakers, the other varieties by female speakers.

7.3 Consistency

A measure can only be valid when it is reliable. But it may be reliable without being valid. Since reliability is a necessary condition for validity, we check the reliability of the set of methods which calculate distances as the averages of separate word distances. It concerns a total of 147 methods which are variants of the frequency per word method and the Levenshtein distance. The consistency is measured by calculating Cronbach's α. In Section 7.3.1 an explanation of this measure of reliability is given. In Section 7.3.2 results are discussed.

7.3.1 Cronbach's α

Cronbach's α is a popular method to measure consistency or reliability. Cronbach (1951) proposed the coefficient α as a lower bound to the reliability coefficient in classical test theory. Cronbach's α is not a statistical test, it is a coefficient of consistency.

Using a word-based method, the distances between varieties are obtained per word. When calculating the distances between n_v varieties on the basis of n_w words, n_w matrices are obtained, each containing the distances between the n_v varieties on the basis of the pronunciations of one word. Because the distance of a variety with respect to itself is always 0, these distances from the matrices' diagonals are not considered. Since distances between two word pronunciations are symmetric, only the half of the matrix is used. In a matrix totally $(n_v \times (n_v - 1))/2$ distances are taken into account. For each pair of matrices the correlation coefficient can be calculated. The average inter-correlation $\bar{r}$ among the words is calculated as:

$$\bar{r} = \frac{\sum_{i=2}^{n_w} \sum_{j=1}^{i-1} r(w_i, w_j)}{\frac{n_w \times (n_w - 1)}{2}} \tag{7.1}$$

where $r(w_i, w_j)$ is Pearson's correlation coefficient between the matrices of words w_i and w_j. Cronbach's α can be written as a function of the number of words and the average inter-correlation among the words:

Figure 7.1: The geographic distribution of the 15 Norwegian varieties.

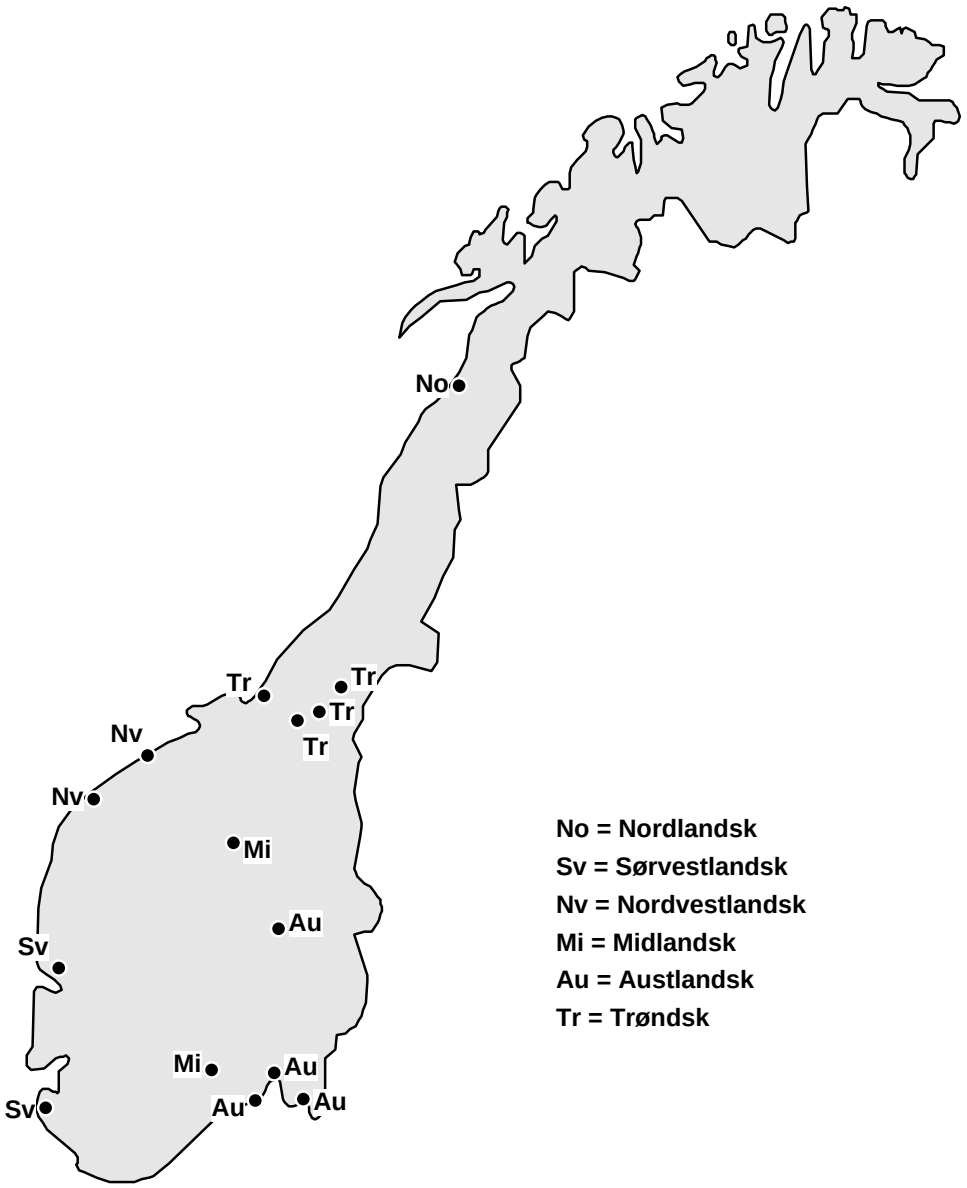

Figure 7.2: According to Skjekkeland (1997) the Norwegian language area can be divided in nine groups. The data points on this map correspond with those in Figure 7.1. In the set of 15 varieties six dialect areas are represented. The same abbreviations are used in the other figures in this chapter.

$$\alpha = \frac{n_w \times \bar{r}}{1 + (n_w - 1) \times \bar{r}} \tag{7.2}$$

As mentioned in Section 7.2.2 the 15 Norwegian varieties are compared on the basis of 58 words. For each matrix corresponding with a word $(58 \times 57)/2 = 1653$ distances are considered. The average inter-correlation is based on $(15 \times 14)/2 = 105$ pairs of matrices.

Usually the Cronbach's α may range between 0 and 1. The higher the α, the more reliable the method. A widely-accepted threshold in social science is that α should be 0.70 or higher for a set of items to be considered a scale (Nunnally, 1978).

7.3.2 Results

In Table 7.1 the results for the 147 word-based methods are summarized. The main division in the table consists of transcription-based methods on the one hand, and recording-based methods on the other hand. For the transcription-based methods different factors are examined. Results are given for different segment representations, for two and four length gradations processed on the basis of changes in the transcription, for different diphthong representations and for different comparison metrics. For the recording-based methods scores are given for the three acoustic representations of the word recordings. When in the table a score is given for a certain combination of factors, the average is taken over the other factors.

Examining the different transcription-based methods used on the basis of different segment representations, the highest scores were found for the frequency per word method using phones, the linear and logarithmic Levenshtein distance using the feature system of H & H, the linear Levenshtein distance using the feature system of A & B and the logarithmic Levenshtein distance using Bark-filters and formant tracks. Further we see that the use of logarithmic segments distances instead of linear ones increases the correlation coefficients of the Levenshtein distances. The high score of the phone-based frequency per word on the one hand, and the improvement which we found when using logarithmic segment distances in the Levenshtein algorithm on the other hand, may indicate that a more reduced number of distance gradations between words will improve the consistency. However, this does not necessarily imply that these most consistent methods will be better methods for validation as well.

From the table it appears that 2 length gradations will in general give higher Cronbach's α values than 4 length gradations. With regard to the diphthong representations, the different representations do not give different scores. Considering the histogram and feature bundle metrics, the Pearson correlation coefficient give the highest scores when using the frequency per word method, and the Euclidean distance gives the best scores when using the Levenshtein distance.

	Freq. word	Lev. lin.	Lev. log.
Transcription-based			
Segment representation discretely			
phones	0.87	0.87	0.87
features H & H	0.84	0.87	0.87
features V & C	0.82	0.85	0.86
features A & B	0.82	0.85	0.87
Segment representation acoustically			
Barkfilter		0.83	0.87
cochleagram		0.82	0.86
formant tracks		0.85	0.87
Number of length gradations			
2 lengths	0.84	0.86	0.87
4 lengths	0.83	0.85	0.86
Diphthongs are represented as			
2 segments	0.83	0.85	0.87
1 segment	0.83	0.85	0.87
Comparison metric			
Manhattan	0.82	0.85	0.87
Euclidean	0.83	0.87	0.88
'Pearson'	0.84	0.85	0.85
Recording-based			
Word representation acoustically			
Barkfilter		0.85	
cochleagram		0.82	
formant tracks		0.77	

Table 7.1: Average Cronbach's α values on the basis of 58 words from 15 Norwegian varieties. The three columns corresponds respectively with the frequency per word method (Freq. word), the linear Levenshtein distance (Lev. lin.) and the logarithmic Levenshtein distance (Lev. log).

Looking at the different word representations which are used in combination with the recording-based Levenshtein distance, the highest score was found when using the Barkfilter representation, the lowest when using the formant track representation. This may be explained by the fact that formant tracks represent only a part of the information in a spectrogram, namely the dominant frequency tracks. This seems to result in less stable results. The formant track-based recording-based Levenshtein distance was at the same time the comparison method with the lowest score compared to all other methods.

In Table 7.1 the lowest Cronbach's α value was equal to 0.77, the highest was equal to 0.87. When examining the Cronbach's α values of all word-based methods separately, it appears that they vary from 0.77 to 0.88. So all methods have a Cronbach's α value which is higher than the threshold of 0.70 (see Section 7.3.1). Our conclusion is that all methods are reliable when using the 58 words of 'the North Wind and the Sun'.

7.3.3 Number of items

With Cronbach's α the number of items can be found which are needed to obtain consistent results. More items result in higher α values. In our data set we used 58 items. Using these items α was 0.77 or higher for all computational methods. When the threshold of α is 0.70 (see Section 7.3.1), we may conclude that with all computational methods reliable results can be obtained on the basis of 58 words.

In this section we investigate the effect of the number of items in more detail. For this purpose we use a variant of the transcription-based Levenshtein distance, where segment distances were found on the basis of the Barkfilter representation, four length gradations are used, diphthongs are represented as a sequence of two segments, and logarithmic segment distances are used. The choice of this method is justified in Section 7.5.1, and the method is applied in Section 7.5.2. With this method, distances between the 15 varieties are calculated for each of the words separately. Subsequently, we calculate α values on the basis of subsets of respectively 2 words, 3 words, and so on, through 58 words. The result is a range of 57 α values. The words in a subset are randomly chosen and each word is unique in a subset.

In Figure 7.3 we find a graph in which the x-axis represents the number of words and the y-axis represents Cronbach's α. For lower number of words the graph fluctuates strongly. When the number of words increases, the graph becomes more stable and a gradual rise can be seen. From 25 words on α is always higher than 0.70. This means that words yield an acceptable degree of consistency, even using only 25 words and with the computational method we mentioned above. The highest α value is equal to 0.86, obtained on the basis of 58 words. To obtain a higher α value, we should use a larger number of words. The α value can be found with the formula (7.2) in Section 7.3.1. From this

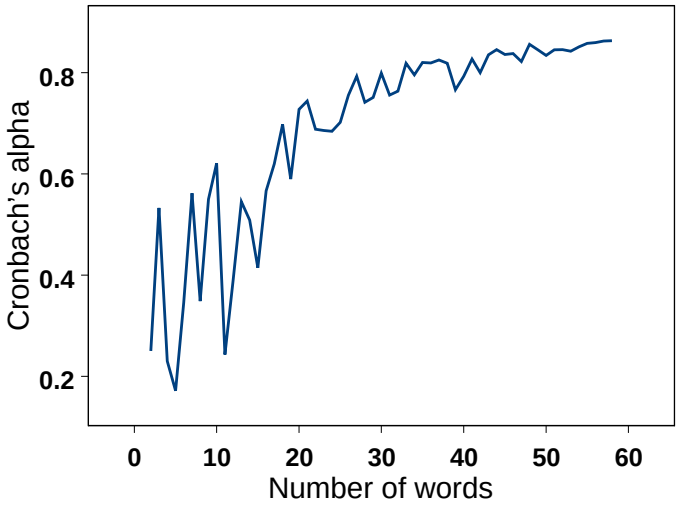

Figure 7.3: Cronbach's α values for 2 through 58 words. For smaller numbers of words the graph strongly fluctuates. When the number of words increases, the graphs becomes more stable and a gradual rise can be seen. From 25 words on α is always higher than 0.70. For 58 words α is equal to 0.86.

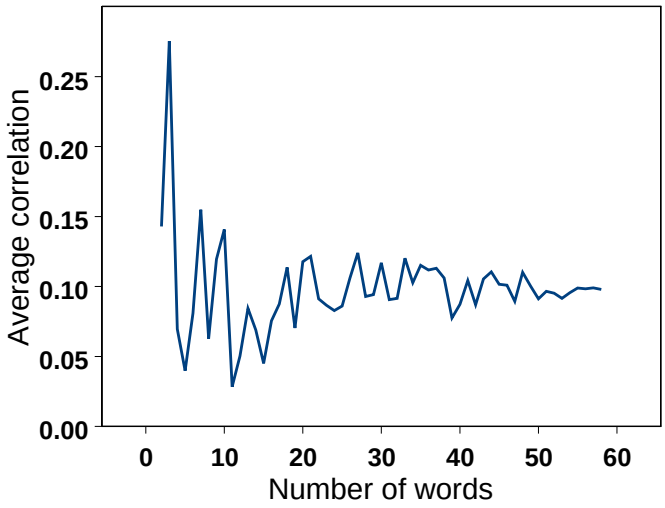

Figure 7.4: Average correlation coefficients for 2 through 58 words. For smaller numbers of words the graph strongly fluctuates. When the number of words increases, the graph becomes more stable. From 55 words on $\bar{r}$ remains stable with a value of about 0.10.

Cronbach's α	Number of words
0.86	55
0.87	60
0.88	66
0.89	73
0.90	81
0.91	91
0.92	104
0.93	120
0.94	141
0.95	171
0.96	216
0.97	291
0.98	441
0.99	891

Table 7.2: Number of words needed to obtain different α values. The numbers are rounded. For $\alpha = 1.00$ the number of words is not defined.

formula, we derived another formula with which the number of words n_w can be found that are required to obtain a certain α value:

$$n_w = \frac{\alpha \times (\bar{r} - 1)}{\bar{r} \times (\alpha - 1)} \tag{7.3}$$

When using this formula, $\bar{r}$ should be known. For each of the subsets containing respectively 2 trough 58 randomly chosen words, we calculated $\bar{r}$. In Figure 7.4, we find a graph where the x-axis represents the number of words and the y-axis the corresponding $\bar{r}$'s. Just as we expected, for lower number of words $\bar{r}$ fluctuates strongly. When the number of words increases, $\bar{r}$ becomes more stable. From 55 words $\bar{r}$ remains stable with a value of about 0.10. Using $\bar{r} = 0.10$, we calculated the number of words that are needed to obtain α values from 0.86 to 0.99. The results are given in Table 7.2. From the formula it appears that the number of words is not defined for $\alpha = 1.00$.

We should emphasize before closing this section that the results depend strongly on the average inter-item correlation $\bar{r}$, which may be expected to vary from one family of dialects to another. The results here apply therefore to determine the sample size needed in other language areas only as a very general indication.

7.4 Validity

Heeringa et al. (2002) validated computational dialect comparison methods by comparing them to a gold standard. The gold standard provides a classification of language varieties with which (nearly) all experts agree. Varieties which experts disagree about are excluded. In this way the gold standard is incomplete, but it represents consensus. Heeringa et al. (2002) based their gold standard on two different Dutch dialect maps.

A gold standard is represented as a partition. The best validation results were obtained when the results of computational comparison methods were also converted to partitions (by clustering) and when these partitions were compared to the gold standard partition. The comparison of partitions was carried out by calculating the Rand index (Rand, 1971) and the Fowlkes and Mallow index (Fowlkes and Mallows, 1983).

We have since recognized disadvantages in this approach. First, dialect maps (and thus the gold standard) show only groups. The maps do not show precisely the proximity of linguistic relationships among the groups and within the groups. Second, varieties that are excluded (mostly borderline cases), play no role in evaluation even when they are classified into linguistically very different groups. Third, this technique cannot take the *degree* of misclassification into account, e.g., when a misclassified variety belongs to a linguistically very close group. Fourth and finally the measurements of the computational comparison methods are converted to partitions by clustering. The consequence is that information about the linguistic relationships within the groups is lost, and that the clustering itself – along with the distance measure – is then subject of validation.

In Gooskens and Heeringa (2003) distances obtained by a variant of the Levenshtein distance are validated by correlating them with perceptual distances. In the variant of the Levenshtein distance segment distances were determined acoustically. Perceptual distances are found in an experiment in which Norwegian listeners judge distances between 15 Norwegian varieties. The validity of the Levenshtein distance is tested by correlating it with the distances obtained by the perception experiment. The advantage compared to the gold standard-based approach is that validation is not based on simplified representations, namely partitions, but on gradual distances found between each possible pair of varieties. Moreover in this approach there is a clear criterion, namely perception, and no dependence on the use of the investigative technique, clustering.

In this section we will validate all 187 methods by correlating the results with perceptual distances. In Section 7.4.1 we describe the perception experiment with which perceptual distances were found. In Section 7.4.2 we discuss the way in which the perceptual distances are correlated with the distances resulting from each of the 187 computational methods. In Section 7.4.3 we discuss the results.

7.4.1 Perception experiment

The perception experiment was carried out by Charlotte Gooskens in the spring of 2000. The experiment is described more detailed in Gooskens and Heeringa (2003). In this Section we give only a brief description.

When the perception experiment was carried out, the recordings of only 15 Norwegian varieties were available. All of them are used. 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 the corresponding recording of the translation of the fable 'The North Wind and the Sun' was presented to Norwegian listeners in a listening experiment. Because the computational comparison method as validated by Gooskens and Heeringa (2003) did not process intonation, both monotonized and original versions of the recordings were used in the perception experiment. The manipulations were carried out with the computer program PRAAT. In order to monotonize the fragments the pitch contours were changed to flat lines. The recordings of the male speakers were monotonized at 134, which is the average pitch of the four male speakers. The recordings of the female speakers were monotonized at 224 Hz. This was the average pitch of the female speakers.

The listeners were 15 groups of high school pupils, one group 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 perception experiment consists of two sessions. In the first session the monotonized texts of the 15 varieties were presented in a randomized order. After a short break, the original texts of the same 15 varieties were presented again in randomized order. Each session was preceded by a practice recording (of a speaker of Stjørdal, but not one of the 15 recordings used in the experiment itself). Between each two recordings there was a pause of 3 seconds. 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 of 15 $\times$ 15 distances for each session.

When comparing the matrices, it appeared that the mean judgments are almost the same (7.19 for the monotonous fragments and 7.25 for the original fragments). However, the standard deviation is smaller in the case of the monotonous fragments (1.38) than in the case of the original fragments (1.68).[2] Two explana-

[2]The variances are not significantly different for α=0.05 since $P(F(\binom{15}{2},\binom{15}{2}) = \frac{1.38^2}{1.68^2}) \Leftrightarrow$ P(F(105,105)=0.6747) $\gg$ 0.05.

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

Table 7.3: Average perceptual distances between all pairs of 15 Norwegian dialects as perceived by 15 groups of listeners judged on a scale from 1 (=similar to own dialect) to 10 (=not similar to own dialect). A column gives the average judgments of different groups of listeners for the same speaker, and a row gives the average judgments for different speakers judged by the same group of listeners.

tions suggest themselves. First the absence of intonation yields unnatural speech. In particular the absence of intonation makes tonemes imperceptible in Norwegian which makes the fragments even more unusual. According to Gooskens and Heeringa (2003) the consequence may be that this makes listeners insecure. This leads to 'safe' judgments, resulting in values which are found closer to the middle of the scale. Second the lower standard deviation for the monotonous distances may have to do with the setup of the experiment. After the first session the listeners know the extremes, i.e., the most similar and most different varieties. This knowledge may be used when judging distances in the second session.

In Gooskens and Heeringa (2003) it is striking that the distances of the comparison method correlate better with the original perceptual distances than with the monotonous perceptual distances, even though the comparison method does not process prosodic information any way. Therefore, we decided only to use the perceptual distances based on the results of the second session, which used the original (non-monotonized) recordings. In this second session the recordings are presented in a natural way. Knowledge about the extremes from the first session is probably used, with the result that the full range of the scale is used. The average judgments as given by the listeners in the second session are given in Table 7.3.

There are two mean distances between each pair of dialects. For example the distance as perceived by the listeners in Bergen with respect to the dialect of Trondheim is different from the distance as perceived by the listeners in Trondheim with respect to the dialect of Bergen. Since both the cluster program and the multidimensional scaling program expect only one value for each pair of different elements, the average of the two mean distances is used when classifying the varieties on the basis of the perceptual distances. In Figure 7.5 a dendrogram is given and in Figure 7.6 a multidimensional scaling plot.

Both the dendrogram and the multidimensional scaling plot accord rather well with the map of Skjekkeland (see Figure 7.2). Sørvestlandsk, Austlandsk and Trøndsk groups can clearly be identified. However, the Midlandsk dialects, Bø and Lesja, do not form a close cluster. Geographically they are rather distant, so they may be rather different although they should be in the same group according to the traditional division. However, in the multidimensional scaling plot the Midlandsk dialect of Lesja is closest to the Midlandsk dialect of Bø. The Nordvestlandsk dialects seem to be very different in both the dendrogram and the multidimensional scaling plot, although they are geographically rather close. Possibly this may be explained by the fact that the map of Skjekkeland is based (partly) on phenomena other than these found in the text 'The North Wind and the Sun'. In our sample the Nordlandsk area is represented by only one variety (Bodø). This variety is grouped with the varieties of the Trøndsk area, which is not unexpected geographically.

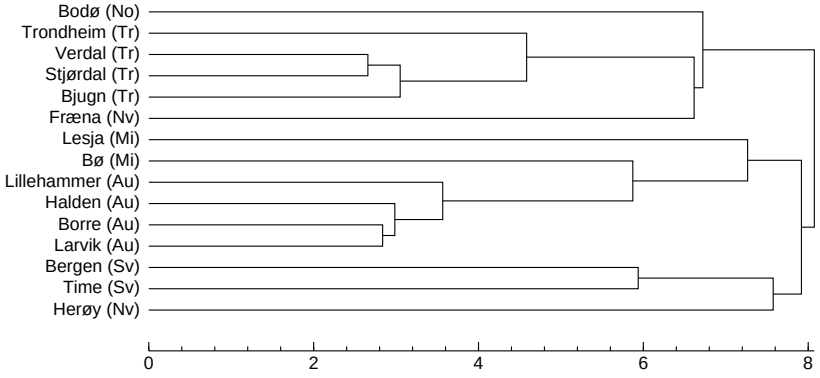

Figure 7.5: Dendrogram derived from the 15 × 15 matrix of perceptual distances showing the clustering of (groups of) Norwegian dialects. UPGMA clustering is used (see Section 6.1.2). On the horizontal scale distances are given in the scale as used by the listeners. The abbreviations between parentheses are explained in Figure 7.2. A Sørvestlandsk, an Austlandsk and a Trøndsk group can clearly be identified. The tree structure explains 91% of the variance.

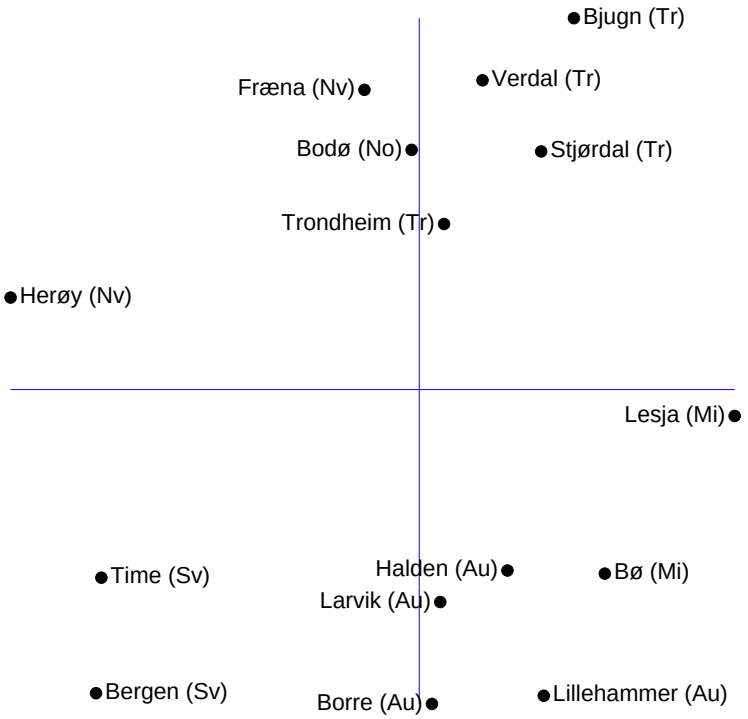

Figure 7.6: Multidimensional scaling of the results derived from the 15 × 15 matrix of perceptual distances. Kruskal's Non-metric MDS is used (see Section 6.2.2). The abbreviations between parentheses are explained in Figure 7.2. The y-axis (first dimension) corresponds with the geographic north-south axis, the x-axis (second dimension) more or less with the west-east axis. Two dimensions explain 67% of the variance.

7.4.2 Correlation

When examining Table 7.3, it is striking that the distances of varieties with respect to themselves are mostly higher than 1 (1 was the lowest possible judgment in the experiment). E.g., the listeners of Trondheim give an average judgment of 3.35 when hearing the recording of a speaker of their own city. This may be explained by the fact that there is variation among the dialect speakers in Trondheim. A slight deviation may reflect different idiolects, but a greater deviation may reflect different varieties spoken in different parts in Trondheim. When the listeners come (partly) from different parts of Trondheim than the speaker of the recording does, their average judgment will never be equal to 1. The fact that the listeners of a given location do not recognize every speaker of the same location as familiar, will cause the distance of a location with respect to itself to be greater than 1. This is different from the use of computational comparison methods where speakers are directly compared to each other without the intervention of listeners' judgments.

When comparing varieties of different locations, the variation within each location may again play a role. Assume we want to find the distance between locations A and B. In A two related varieties A_1 and A_2 are spoken. Assume further that in the experiment the speaker of A spoke variety A_1, and that the listeners' group of A is familiar with variety A_2. In the perception experiment we determine the distance between A_2 and B. However, when using a computational comparison method the distance between speakers is found, i.e. the distance between A_1 and B. Therefore, the perceptual and the computational distance need not to be equal: the one may be higher than the other.

It may be clear that the deviation of perceptual distances between varieties at the *same* location always goes in *one* direction compared to the corresponding computational distances: they will be relatively higher. Therefore, when correlating the matrix of perceptual distances with a matrix of computational distances, these higher perceptual distances may cause distortion, which justifies eliminating them. When calculating the correlation coefficient, the values on the diagonal (from upper left to lower right) are not taken into account. However, the distortion of perceptual distances between *different* locations may go into *two* directions compared to the corresponding computational distances: they can be either relatively higher or relatively lower. Therefore, we regarded these deviations as noise which will cause no significant distortion, when correlating the matrix of perceptual distances with a matrix of computational distances. All distances between different locations are considered when calculating the correlation coefficient.

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 a matrix X and a matrix Y is calculated as:

$$r(X,Y) = \frac{\sum_{i=1}^{n}\sum_{j\neq i}(X_{ij}-\overline{X})(Y_{ij}-\overline{Y})}{\sum_{i=1}^{n}\sum_{j\neq i}(X_{ij}-\overline{X})^2 \sum_{i=1}^{n}\sum_{j\neq i}(Y_{ij}-\overline{X})^2} \tag{7.4}$$

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

For finding the significance of a correlation coefficient we used the Mantel test, just as in Chapters 3 and 4. The Mantel test is explained in Section 3.8.2. 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. In the sections below mostly averaged correlation coefficients are given. Since our implementation of the Mantel test only compares individual correlation coefficients, we were not able to determine whether two averaged correlation coefficients are significantly different.

7.4.3 Results

When examining the correlation coefficients of all 187 methods separately, we found that they vary from 0.33 to 0.67. All correlations were significant. But a perfect correlation was not found. We should be aware of the fact that in the perception experiment the listeners were confronted with recordings of spoken texts. These texts include lexical, phonetic, prosodic, morphological and syntactical information. However, when applying computational comparison methods to the transcriptions of word pronunciations, only lexical, phonetic and morphological information is processed. In the feature-based methods the presence of tonemes is also processed, but only in these methods. However, the exact realization of these tonemes is never processed. They are treated as categorical differences (see Section 3.4.1).

In the Tables 7.4, 7.5, 7.6, 7.7 and 7.8 results for the 184 transcription-based methods are given, and in Table 7.9 results for 3 recording-based methods are given. In the tables for the transcription-based methods results are given for the corpus frequency method, the frequency per word method and the Levenshtein distance (using linear or logarithmic segment distances). In these tables the factors mentioned in Section 7.1 are examined. When particular factors are examined in a table, the average is taken over the factors which are not mentioned in that table.

7.4.3.1 Transcription-based comparison methods

In Table 7.4 correlation coefficients are given for different segment representations for each set of transcription-based methods. Examining the phone-based

methods, we find that the corpus frequency method performs as well as the frequency per word method. The Levenshtein distance in turn performs better than both the corpus frequency method and the frequency per word method. This confirms our conviction that the Levenshtein distance is methodologically better. The linear and the logarithmic Levenshtein distance have the same correlation. When using phones there exist only two differences: 0 (equal) and 1 (different). So the relative distances are not changed when transformed logarithmically.

Examining the feature-based methods, we see that the frequency per word method performs better than the corpus frequency method, the linear Levenshtein distance performs better than the frequency per word method, and the logarithmic Levenshtein distance performs better than the linear Levenshtein distance for each of the three different feature systems. This order reflects the different steps of improvement in this range of methods. When using features the word-based methods are clearly better than the corpus-based methods. When using phones this improvement was not found. As of this writing we have found no explanation for this.

For both the feature-based Levenshtein distances and the acoustic-based Levenshtein distances the use of the logarithmic distances instead of linear distances improves the correlation. It confirms our idea that the use of the logarithm mimics perception better.

It is striking that the phone-based methods and the acoustic-based logarithmic Levenshtein distances perform best. Even the two frequency-based methods using the phone representation perform well. Why do especially the phone-based methods perform better than the more refined feature-based methods? Perhaps this may be explained if, for dialect speakers, all distances between different segments are the same. This means that the distance between [i] and [ɪ] is equal to the distance between [i] and [ɒ]. In fact the first distance is relatively too large (it should be smaller than the distance between [i] and [ɒ]) and the second relatively too small (it should be larger than the distance between [i] and [ɪ]). This may result in a logarithmic effect in which small differences weigh disproportionally, which in turn yields a higher correlation with the perceptual distances.

Although the phone-based methods perform as well (or even any better) than the acoustic-based logarithmic Levenshtein distances, we prefer the methods last mentioned. The results shown here are based on a small set of 15 varieties.[3] Having more varieties results in a network with a higher density, in which small details may become more important. These details are not processed in the phone-based methods, only in the acoustic and feature-based methods. Validation on the basis of a larger set of varieties is advisable in future work.

[3]Having 15 varieties the correlation is still based on $15 \times 14 = 210$ distances!

	Freq. corp.	Freq. word	Lev. lin.	Lev. log.
Segment representation discretely				
phones	0.66	0.66	0.67	0.67
features H & H	0.47	0.61	0.64	0.65
features V & C	0.45	0.59	0.61	0.63
features A & B	0.45	0.58	0.62	0.64
Segment representation acoustically				
Barkfilter			0.65	0.66
cochleagram			0.64	0.66
formant tracks			0.64	0.66

Table 7.4: The effect of methods and representations is shown in the average correlation coefficients of transcription-based distance measures with respect to perceptual distances on the basis of 15 Norwegian varieties. The four columns present respectively the corpus frequency method (Freq. corp.), the frequency per word method (Freq. word), the linear Levenshtein distance (Lev. lin.) and the logarithmic Levenshtein distance (Lev. log). For each method the average scores for different segment or word representations are given.

7.4.3.2 Representation of segments

In Section 7.4.3.1 we found that the phone-based methods and the acoustic-based logarithmic Levenshtein distances perform best. In this section we compare the different segment representations in more detail.

Different feature representations When comparing the different feature representations in Table 7.4, we find that the feature system of H & H gives better results than the other systems. This is especially striking since the V & C and A & B systems were developed especially for measuring transcription differences, while the basis of the H & H system, the features of Chomsky and Halle's *The Sound Pattern of English*, was developed for encoding phonological rules. It was our expectation that the perceptually-motivated system of V & C would give the best results. Even though, the system was originally developed for Dutch. Later on the system was extended so that it contains all vowels and pulmonic consonants of the IPA system (see Section 3.1.3). Although the extensions were made along the lines of the original system, they are not directly based on the perception of listeners. This may possibly explain the lower correlations in Table 7.4.

Feature vs. phone representation Although the use of phones gives better results than the use of features for all methods, we found the greatest difference

for the corpus frequency method in Table 7.4. In a histogram of phones, frequencies are given for the phones individually. In a histogram of features, feature frequencies are given. The frequency per feature gives the number of sounds for which that feature was positive. Different sounds may contribute to the frequency of the same feature. The result is that we lose information. We illustrate this by a hypothetical example. Assume the corpus of a dialect contains a [b] and a [f]. The feature *voice* gets a frequency of 1 (due to the [b]) and the feature *continuant* get a frequency of 1 (due to the [f]). Another dialect contains a [p] and a [v]. The features *voice* and *continuant* get a frequency of 1 (only due to the [v]). The consequence is that the different dialects get the same frequencies and will erroneous appear to be equal when comparing the feature frequencies. This will not happen when using phones. For the first dialect the [b] and the [f] get a frequency of 1, and for the second dialect the [p] and the [v] get a frequency of 1. The dialects are clearly distinguished, although differences may be exaggerated when comparing the phone-based histograms.

Different acoustic representations The acoustic representations are only used in combination with Levenshtein distance. For the Barkfilter we found a higher correlation than for the two other acoustic representations when using the linear Levenshtein distance. Using the Barkfilter representation voiceless and voiced sounds are more sharply distinguished and the vowels and r-like sounds are closer than when using the cochleagram representation. Comparing the Barkfilter with the formant track representation, we get a vowel quadrilateral rather than a triangle. Using the Barkfilter representation plosives and fricatives are clearly distinguished, which is not the case when using the formant track representation (see Sections 4.3.1.2, 4.3.2.2 and 4.3.3.2). When using the logarithmic Levenshtein distance, all acoustic representations have the same correlation. Using logarithmic distances, smaller distances become relatively more important than larger differences. This makes the acoustic-based segment distances more similar.

Acoustic vs. feature representation Since the acoustic representations are only used in combination with Levenshtein distance, the comparison between acoustic and feature representations can only be made on the basis of the results of Levenshtein distance. Among the different feature representations, we found the highest correlation for the feature system of H & H. Equally good or better results are obtained when using acoustic representations. In the feature system of H & H consonants are mainly distinguished by the manner of articulation. This applies for the acoustic representations as well. However, in the systems of V & C and A & B the place of articulation is clearly represented. Our results suggest that the manner of articulation plays a more important role in perception. From

Table 7.4 it appears that the use of acoustic representations gives better results than the use of feature representations in general.

7.4.3.3 Number of length gradations

Table 7.5 presents – for each comparison method and for each segment representation – a comparison of 2 length gradations versus 4 length gradations as processed by changing the transcription. We should be aware of the fact that for phone-based and acoustically-based methods half-long and long are not processed when using 2 length gradations. When using a feature representation, half-long and long are processed by changing a feature value.

In the table in three cases we found that 2 length gradations result in a higher correlation coefficient than 4 length gradations, and in nine cases 4 length gradations is superior.[4] So we conclude that the use of 4 length gradations in general gives better results than the use of 2 length gradations, but the differences are seldom large.

We found no systematic distinction between the phone-based and acoustically-based methods on the one hand, and the feature-based methods on the other hand. For example, the phone-based corpus frequency method performs better when using 4 gradations, but the frequency per word method performs better when using 2 gradations. Looking at the feature-based methods, we found clear improvements for the H & H-based and V & C-based corpus frequency methods when using four gradations. H & H themselves used two gradations while processing half-long and long by changing a feature value.

7.4.3.4 Representation of diphthongs

Table 7.6 compares two diphthong representations for each comparison method and for each segment representation. In the first a diphthong is processed as the sequence of two monophthongs and in the second as one segment. It seems to make no difference whether diphthongs are represented as two segments or as one segment. In three cases we found that the representation as two segments results in a higher correlation than the representation as one segment, and in three cases the representation as one segment results in a higher correlation than the representation as two segments.[5] On the basis of these outcomes a clear conclusion cannot be drawn. This may be explained by the fact that only a small set of diphthongs were defined for the NOS data.

Examining the feature-based corpus frequency methods, we observe that when using the feature system of H & H, the higher correlation coefficient is obtained

[4]Since the phone-based logarithmic Levenshtein distance yields the same results as the linear counterpart, we only counted results of the latter. See Section 7.4.3.1.

[5]Just as in Section 7.4.3.3 we only counted results of the linear Levenshtein distance when considering the phone-based methods.

	Freq. corp.	Freq. word	Lev. lin.	Lev. log.
Phones				
2 lengths	0.66	0.67	0.66	0.66
4 lengths	0.66	0.66	0.67	0.67
Features H & H				
2 lengths	0.43	0.61	0.64	0.65
4 lengths	0.50	0.61	0.64	0.65
Features V & C				
2 lengths	0.43	0.59	0.61	0.62
4 lengths	0.47	0.60	0.62	0.63
Features A & B				
2 lengths	0.45	0.58	0.62	0.64
4 lengths	0.44	0.59	0.62	0.64
Barkfilter				
2 lengths			0.64	0.66
4 lengths			0.65	0.67
cochleagram				
2 lengths			0.64	0.66
4 lengths			0.64	0.66
formant tracks				
2 lengths			0.65	0.66
4 lengths			0.64	0.66

Table 7.5: The effect of *segment length discrimination* is shown in the average correlations of transcription-based methods with perceptual distances on the basis of 15 Norwegian varieties. For each comparison method and for each segment representation the average scores for 2 length gradations versus 4 length gradations can be compared.

	Freq. corp.	Freq. word	Lev. lin.	Lev. log.
Phones				
2 segments	0.66	0.66	0.67	0.67
1 segment	0.66	0.66	0.66	0.66
Features H & H				
2 segments	0.47	0.61	0.64	0.65
1 segment	0.46	0.61	0.64	0.65
Features V & C				
2 segments	0.45	0.59	0.61	0.63
1 segment	0.46	0.59	0.61	0.63
Features A & B				
2 segments	0.44	0.58	0.62	0.64
1 segment	0.45	0.59	0.62	0.64
Barkfilter				
2 segments			0.65	0.66
1 segment			0.65	0.66
cochleagram				
2 segments			0.64	0.66
1 segment			0.64	0.66
formant tracks				
2 segments			0.65	0.66
1 segments			0.64	0.66

Table 7.6: The effect of *diphthong representation* is shown in the average correlations of transcription-based methods with perceptual distances on the basis of 15 Norwegian varieties. For each comparison method and for each segment representation the average scores for two diphthong representations are compared. In the first a diphthong is processed as the sequence of two monophthongs and in the second as one segment. There is very little difference in treating diphthongs as one versus two segments.

when a diphthong is represented as two segments. On the other hand for the feature systems of V & C and A & B the higher correlation coefficients are obtained when a diphthong is represented as one segment. This difference between H & H on the one hand, and V & C and A & B on the other hand, concerns the way in which height is defined for closing diphthongs (in the NOS data there are no centering diphthongs). In H & H the height is equal to the height of the first segment, in the other systems the mean of the heights of the first and the second segment is used.

7.4.3.5 Comparison of feature histograms or feature bundles

In Table 7.7 different metrics for finding distances between histograms and feature bundles are compared for the feature-based comparison methods. For the corpus frequency method, Manhattan gives the best results when using the feature systems of H & H and V & C, and Euclidean gives the best results when using the feature system of A & B. For all systems the use of Pearson's correlation coefficient gives the lowest correlation with the perceptual distances. This is especially striking for the H & H system, since H & H themselves used this metric in all their publications. Looking at the results of the word-based methods, the Euclidean distance unanimously appears to be the best metric.

Why does Manhattan give the better results most of the time when using the corpus frequency method, and does Euclidean appear to be the best metric when using word-based methods? When comparing the Manhattan metric to the Euclidean metric (see the formulas given in respectively (3.1) and (3.2) in Section 3.6.2.5), we see that feature differences are squared when using the Euclidean metric. The result is that larger differences are weighted relatively more heavily than smaller differences. When using the corpus frequency method, dialect distances are measured with the metrics. Using the frequency per word method and Levenshtein distance, respectively word distances and segment distances are calculated with these metrics. This indicates that on the highest level (comparison of dialects) feature differences should be weighted equally, but on the deeper levels (comparison of words or segments) larger differences should be weighted relatively more heavily than smaller ones.

In Table 7.8 in fact the same scores are given as in Table 7.7. However, now for each comparison method and for each histogram or feature bundle comparison metric the average scores for the different feature-based segment representations are given. The table shows that, regardless which comparison method and histogram or feature bundle metric is used, the feature system of H & H gives mostly better results than the other systems. The findings for all three comparison metrics are nearly the same. Using the corpus frequency method the feature system of V & C performs as well or better than the feature system of A & B. For the frequency per word method it turns out that the feature system of V & C gives better results than the system of A & B when using Manhattan or Euclidean,

	Freq. corp.	Freq. word	Lev. lin.	Lev. log.
Features H & H				
Manhattan	0.49	0.61	0.63	0.65
Euclidean	0.48	0.64	0.66	0.66
'Pearson'	0.43	0.58	0.64	0.63
Features V & C				
Manhattan	0.46	0.59	0.61	0.63
Euclidean	0.47	0.61	0.64	0.65
'Pearson'	0.43	0.58	0.59	0.61
Features A & B				
Manhattan	0.46	0.56	0.61	0.64
Euclidean	0.45	0.59	0.65	0.66
'Pearson'	0.42	0.60	0.61	0.63

Table 7.7: The effect of *feature bundle comparison metrics* is shown in the average correlations of transcription-based methods with perceptual distances on the basis of 15 Norwegian varieties. For each comparison method and for each feature-based segment representation the average scores for different feature bundle comparison metrics are given.

but when using Pearson's correlation coefficient it is the opposite. When using the Levenshtein distances the feature system of A & B performs better than that of V & C, regardless the choice of the histogram of feature bundle metric. We conclude that there is no interaction between feature representation and feature metric.

7.4.3.6 Recording-based comparison methods

In Table 7.9 the scores of the recording-based Levenshtein distances are given. As mentioned in the Sections 7.2.3 and 7.2.4 in the set of 15 varieties there were 4 male speakers and 11 female speakers. Correlation coefficients are given for both genders separately, and for all speakers. When using all speakers the use of the formant track representation gives the highest correlation. This correlation is higher than those for the transcription-based corpus frequency methods using features. However, when using Barkfilters or cochleagrams as acoustic word representations, the correlations are lower than those for all transcription-based methods. We explain the low correlations of the recording-based methods by two facts. First the way in which sample sizes are normalized for speech rate is very rough (see Section 5.2.3). This may hamper the finding of correct alignments and corresponding distances by the Levenshtein algorithm. Second voice quality may still play a role, although the samples are monotonized. Differences between

	Freq. corp.	Freq. word	Lev. lin.	Lev. log.
Manhattan				
Features H & H	0.49	0.61	0.63	0.65
Features V & C	0.46	0.59	0.61	0.63
Features A & B	0.46	0.56	0.61	0.64
Euclidean				
Features H & H	0.48	0.64	0.66	0.66
Features V & C	0.47	0.61	0.64	0.65
Features A & B	0.45	0.59	0.65	0.66
'Pearson'				
Features H & H	0.43	0.58	0.64	0.63
Features V & C	0.43	0.58	0.59	0.61
Features A & B	0.42	0.60	0.61	0.63

Table 7.8: The interaction between feature sets and feature bundle comparison metrics is shown in the average correlations of transcription-based methods with perceptual distances on the basis of 15 Norwegian varieties. For each comparison method and for each histogram or feature bundle comparison metric the average scores for the different feature-based segment representations are given. It is clear that Manhattan and Euclidean distance are similar, and that the Pearson measure is inferior.

male and female voices may influence the result (Heeringa and Gooskens, 2003). This appears to be confirmed by the fact that the correlation coefficients on the basis of the female speakers are higher than on the basis of all speakers. The correlation coefficients on the basis of the male speakers are much lower. This may indicate that diversity in voice quality is relatively larger for male speakers than for female speakers. The highest correlation coefficient for male speakers – just as for male and female speakers, taken together – was obtained using formant tracks. Apperently, this representation is less sensitive to voice quality. However, the transcription-based word-based comparison methods perform still much better.

7.5 Choice and results

The findings of the Sections 7.3 and 7.4 enable us to select the optimal method for finding distances between varieties. The choice of the method is made in Section 7.5.1. In Section 7.5.2 we apply this method to Norwegian data and show results.

	Male	Female	All
Barkfilter	0.08	0.44	0.33
cochleagram	0.12	0.55	0.41
formant tracks	0.36	0.55	0.50

Table 7.9: Correlations of recording-based Levenshtein distances with perceptual distances on the basis of 15 Norwegian varieties using three different acoustic word representations. Scores are given for the male and female speakers separately, and for all speakers.

7.5.1 Choice of method

In Section 7.3 we examined the reliability of the word-based methods. We found that all methods are reliable when using the 58 words of 'the North Wind and the Sun'. In Section 7.4 we validated transcription-based and recording-based methods with respect to the results of a perception experiment. Examining the transcription-based methods we found that the phone-based methods and the acoustic-based logarithmic Levenshtein distances give results that correlate strongest with perceptual distances. Among the feature representations, the H & H system yields the best results. Among the acoustic representations we found the Barkfilter representation better than the other two representations, but only when using the linear Levenshtein distance. We found that the use of 4 length gradations will in general give better results than the use of 2 length gradations. For the diphthongs we compared the representation as two segments with the representation as one segment but could not draw a definite conclusion. When using features, the Manhattan metric gives the better results most of the time when using the corpus frequency method. When using word-based methods, the use of the Euclidean metric is preferable. Examining the recording-based methods we found that the three Levenshtein variants gave less satisfying results.

Therefore, we choose one of the transcription-based methods. From this range of methods we have to choose from the phone-based methods and the acoustic-based logarithmic Levenshtein distances. We found the highest average score for the phone-based Levenshtein distances. Nevertheless, we maintained our preference for the acoustic-based logarithmic Levenshtein distances in Section 7.4.3.1. In a small set of 15 varieties, the rougher phone-based methods may perform well, but for a denser sampling, minor differences may play a stronger role. Using the acoustic-based logarithmic Levenshtein distance, these differences are taken into account to a greater deal. Examining Table 7.4, we find that different acoustic representations give the same average scores when using the logarithmic Levenshtein distance. When looking at the results of the linear Levenshtein distance, we find the highest score when the Barkfilter is used. Since we are forced to make a choice, we choose the Barkfilter representation. With regard to the number of length gradations, of course we prefer 4 length gradations. As mentioned

above no obvious conclusion could be drawn about the representation of diphthongs. When examining the scores of the acoustic-based Levenshtein distances in Table 7.4.3.4, we only found different results for the linear Levenshtein distance using the formant track representation. For this method, the two-segmental representation gave better results. Therefore, we choose the two-segmental representation of diphthongs. With that we have made our choice, namely the method with the parameters we chose above. We do not need to choose a feature bundle metric since we use an acoustic representation.

It may be expected that the method which we chose on the basis of different parameters, belongs to the better ones in the set of all of the 187 methods. To examine this, all 187 methods were sorted according to their correlation with the perceptual distances. When examining the sorted list of methods, it appears that our method even has the highest correlation coefficient. The Cronbach's α for this method is 0.86. The correlation coefficient between distances obtained with this method and the perceptual distances is equal to 0.67. Therefore, we applied this method to the NOS data in Section 7.5.2 and Chapter 8, and to the RND data in Chapter 9.

7.5.2 Analysis of Norwegian

In this section first we apply the method chosen in Section 7.5.1, to our set of 15 varieties. In Table 7.10 the distances are given as percentages. The way in which percentages are found is described in the Sections 5.1.8 and 5.1.10. Given the high correlation between the distances obtained with this method and perceptual distances we may expect that the classification results will be similar as those in Section 7.4.1. Since the distance matrix is symmetric, only one half is used, while the zero values on the diagonal from upperleft to lowerright are not used. In Figure 7.7 a dendrogram is given. In the dendrogram, the scale distance is given as a percentage. In Figure 7.8 a multidimensional scaling plot is shown.

Comparing the dendrogram in Figure 7.7 with the dendrogram obtained on the basis of the perceptual distances (Figure 7.5), both show an Austlandsk group which contains the varieties of Larvik, Halden, Lillehammer and Borre, and a Trøndsk group which contains the varieties of Verdal, Bjugn and Stjørdal. Although the two dendrograms do not cluster the Midlandsk varieties as one group, in the perceptual dendrogram they appear to be more related than in the computational dendrogram. In the perceptual dendrogram the Midlandsk dialect of Lesja is clustered with the Austlandsk varieties, although not very close. In the computational dendrogram this dialect belongs to the Trøndsk varieties. Geographically the variety is located about midway between the two areas. In both the perceptual and computational dendrogram Bø is clustered with the Austlandsk varieties, but in the perceptual dendrogram the relation appears to be stronger. The Sørvestlandsk varieties of Bergen and Time form one (rather loose) cluster in the perceptual dendrogram. In the computational dendrogram they do

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

Table 7.10: Average Levenshtein distances between all pairs of 15 Norwegian dialects given as percentages.

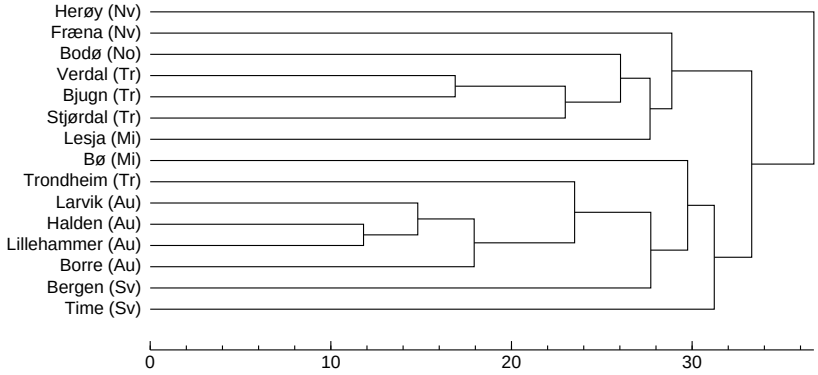

Figure 7.7: Dendrogram derived from the 15 × 15 matrix of Levenshtein distances showing the clustering of (groups of) Norwegian dialects. UPGMA clustering is used (see Section 6.1.2). The scale distance is given as a percentage. The abbreviations between parentheses are explained in Figure 7.2. An Austlandsk and a Trøndsk group can clearly be identified. The tree structure explains 68% of the variance.

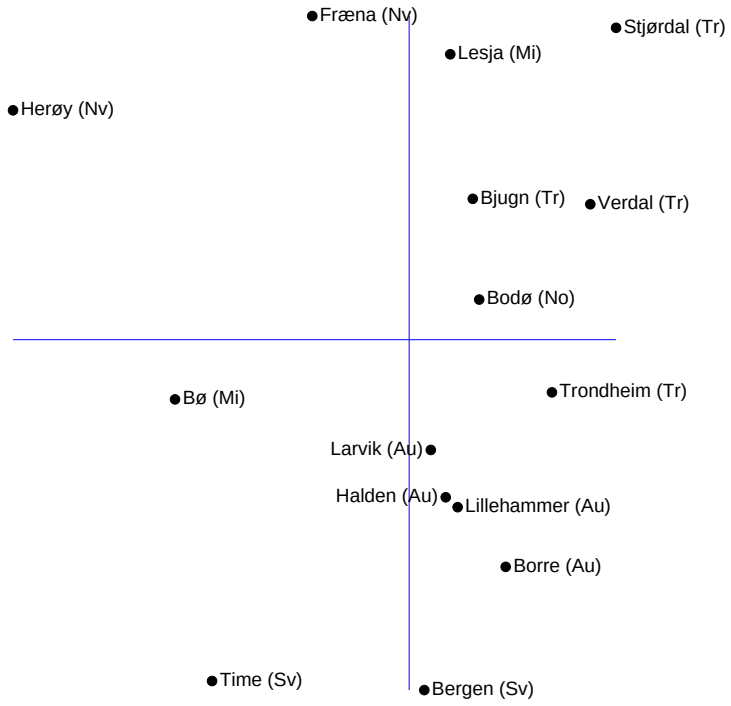

Figure 7.8: Multidimensional scaling of the results derived from the 15 × 15 matrix of Levenshtein distances. Kruskal's Non-metric MDS is used (see Section 6.2.2). The abbreviations between parentheses are explained in Figure 7.2. The y-axis (first dimension) corresponds with the geographic north-south axis, the x-axis (second dimension) more or less with the west-east axis. Two dimensions explain 83% of the variance.

not form one cluster, but appear to be related to some extent (see Section 6.1.4). In both dendrograms the two Nordvestlandsk varieties do not form one cluster. In both Fræna is clustered with the Trøndsk varieties. However, Herøy is clustered with the Sørvestlandsk varieties in the perceptual dendrogram, while in the computational dendrogram it belongs to none of the groups, but appears to be distinct from all the other varieties. In both dendrograms Bodø is clustered with the Trøndsk varieties. However, in the computational dendrogram Bodø looks as if it were closer to the Trøndsk varieties than in the perceptual dendrogram. However, the cluster with Verdal, Bjugn and Stjørdal is geographically not impossible. A striking difference can be found with regard to the dialect of Trondheim, which is clustered with the Trøndsk varieties in the perceptual dendrogram, but in the computational dendrogram it is clustered with Austlandsk varieties. Possibly the listeners recognized the recording of Trondheim as the dialect of Trondheim and let influence their judgments by geography. However, the dialect of larger cities may be in contrast with their surrounding and more related to geographically more distant varieties. We conclude that the two dendrograms are rather similar, due to the fact that especially the closer clusters in the one dendrogram are also found in the other one.

Comparing the multidimensional scaling plot in Figure 7.8 with the multidimensional scaling plot obtained on the basis of the perceptual distances (Figure 7.6), an Austlandsk and a Trøndsk group can be found in both. In the computational plot the Austlandsk varieties are closer than in the perceptual plot. However, in the perceptual plot the Trøndsk varieties are closer than in the computational plot. In the perceptual plot the Trøndsk dialect of Trondheim is most distant from the other Trøndsk varieties. In the computational plot the Trondheim dialect is even more distant to the varieties of the same group. In the perceptual plot the geographically distant Midlandsk varieties of Bø and Lesja are not very close, but in the computational plot they are much more distant. The Nordvestlandsk varieties of Fræna and Herøy are about equally distant in the two plots. In the perceptual plot the Sørvestlandsk varieties of Bergen and Time are closer than in the computational plot. In both plots the Nordlandsk variety of Bodø is found near (perceptually) or among (computationally) the Trøndsk varieties. In the perceptual plot there is a rather sharp division between northern and southern varieties. In the computational plot the northern and southern varieties form a continuum. This may indicate that listeners perceive differences in a more categorical way than the Levenshtein distance suggests. This may also explain the other differences. Since the differences between the plots are relatively small, we conclude that the Levenshtein distances reflect the perceptual distance a great deal.