Measuring Dialect Pronunciation Differences using Levenshtein Distance
Heeringa, Wilbert Jan

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2004

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Chapter 8

Measuring Norwegian dialect distances

In Chapter 7 a range of computational comparison methods was validated. The method with the highest score is a variant of the Levenshtein distance, where (i) segment distances are found on the basis of the Barkfilter representation, (ii) four length gradations are used, (iii) diphthongs are represented as a sequence of two segments, and (iv) logarithmic segment distances are used (Section 7.5.1). This method was applied to a small set of 15 Norwegian dialects (Section 7.5.2). In this chapter we apply the same method to a larger set of 55 Norwegian varieties. Results will be compared to the dialect map of Skjekkeland (1997).

In Section 8.1 the set of 55 varieties will be discussed. On the basis of the Levenshtein distances we will perform cluster analysis and multidimensional scaling. In Section 8.2 results of cluster analysis are presented, and in Section 8.3 the results of multidimensional scaling. The discussion of the results should be considered as an initial impetus. Further analysis of the results may be useful future work. In Section 8.4 we draw some conclusions.

8.1 Data source

In Section 7.2 we described a database which contains recordings of different Norwegian varieties. The database was compiled by Jørn Almberg and Kristian Skarbo. For each variety a recording and a transcription is given of the fable ‘The North Wind and the Sun’. The text consists of 58 words which are given in Appendix B Table B.1.

When the perception experiment was carried out (see Section 7.4.1), recordings of only 15 varieties were available. Later on this database was extended. In this chapter results are presented which are obtained on the basis of a set of 55 varieties. Figure 8.1 shows the geographical distribution of the dialects. The set of 55 varieties covers all nine dialect areas as found on the map of Skjekkeland.
Figure 8.2 shows the distribution of the varieties over the dialect areas as given by Skjøkkeland. For some locations more than one recording and transcription was available. Therefore, these locations are numbered in the figures in this chapter.\(^1\)

- **Alstahaug**
  The two versions are based on different recordings of different informants, the first from Sandnessjøen (Alstahaug 1) and the second from Tjøtta (Alstahaug 2). The first version is most representative for the area of Alstahaug.

- **Bergen**
  The two versions are based on different recordings of the same informant. The older version (Bergen 1) is no longer available on the web, but was used in validation work (see Section 7.2). The newer version (Bergen 2) is the better one according to the speaker.

- **Bodø**
  The two versions are based on different recordings of the same informants. The older version (Bodø 1) is no longer available on the web, but was used in validation work. The newer version (Bodø 2) is the better one according to the speaker.

- **Rana**
  The two versions are based on different recordings by different informants, both from Rana (Rana 1 and Rana 2). The second version is more representative for the area around Rana.

- **Stavanger**
  The two versions are based on different recordings by different informants, the first from Hafrsfjord (Stavanger 1) and the second from Hundvåg (Stavanger 2). Both are equally representative for the surrounding of Stavanger, but when we are forced to make a choice, we select the second version.

- **Stjørdal**
  The three versions are based on different recordings of different informants, the first and the second from Stjørdal (Stjørdal 1, Stjørdal 2), and the third from Stjørdalshalsen (Stjørdal 3). The first version is most representative. In validation work the second version is used, which was available earlier.

---

\(^1\)We are grateful to Jørn Almberg (personal communication) for advice at several points below, e.g., the question as to which of two versions is the more typical for a given site.
8.2 Classification

8.2.1 Cluster analysis

Using the Levenshtein variant we mentioned at the beginning of this chapter, we calculated the distances between the 55 varieties. On the basis of these distances we applied cluster analysis (see Section 6.1). In Figure 8.3 a dendrogram is given, showing the classification of 55 Norwegian varieties. In the dendrogram, the scale distance shows percentages.

Examining the nine most significant groups we find from upper to lower the dialects of Heroy and Fræna, a central group, the dialect of Bø, an eastern group, a southeastern group, a northern group, a western group and a southwestern group. The same groups are geographically visualized in Figure 8.4.

When regarding only the 5 most significant groups, Heroy and Fræna appear to be one cluster. Both varieties belong to the Nordvestlandsk varieties. However they are not clustered with the other Nordvestlandsk varieties, which are found in the western group. When considering the 9 most significant groups, each of these two varieties appears to be a separate dialect, not clustered with any of the other groups. This indicates that the two varieties are very marked dialects among the other Nordvestlandsk varieties.

The central group contains the Trøndsk varieties of Sunndal and Oppdal and the geographically rather close Midlandsk variety of Lesja. It is striking that Sunndal and Oppdal are not clustered with the other Trøndsk varieties, which are for the greater part found in the northern group. We expected that Lesja would be clustered with the other Midlandsk variety of Bø. However, just as for the set of 15 varieties, this is not the case. Geographically the two varieties are distant. The variety of Bø appears to be a separate variety which does not belong to any of the other varieties. It is striking that Bø is suggested to be closest to the eastern varieties, and not with the geographically closer varieties of the southeastern group.

The Austlandsk varieties are divided into an eastern group and a southeastern group. In the southeastern group the geographically adjacent Sørlandsk varieties are found as well. More striking is the presence of the Sørvestlandsk variety of Bergen and the Trøndsk variety of Trondheim in this group. We cannot explain this. However, it is not uncommon that varieties of larger cities are dialect islands.
Figure 8.1: The geographic distribution of the 55 Norwegian varieties.
Figure 8.2: 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 8.1. In the set of 55 varieties all dialect areas are represented. The same abbreviations are used in the other figures in this chapter.
which are related to geographically more remote dialects.\(^2\) Rather unexpected is that the varieties of Tynset and Lillehammer are in the southeastern group, and not in the eastern group. Lillehammer is closest to Halden. Both Halden and especially Lillehammer are nearly standard (i.e., close to bokmål), and therefore not very typical dialect versions from their respective geographic regions. The reading of Tynset is also quite standard, which may be the reason why it is judged to be closer to Oslo. In the southeastern group two versions of Bergen, which are recorded by the same informant, do not form one cluster, but are rather close.

The largest group in the dendrogram is the northern group. It contains Nordlandsk, Helgelandsk, Troms-Finnmarks-mål and Trøndsk. The group may be divided in a Trøndsk group on the one hand, and a group containing the other varieties on the other hand. In the latter group, no systematic division between Nordlandsk, Helgelandsk and Troms-Finnmarks-mål varieties can be found. Perhaps the division in these three areas has become blurred over time. The two varieties of Rana are rather close, although they do not form one cluster. The varieties of Alstahaug are obviously more distant, indicating dialect diversity in a small area. Stjørdal 1 and 2 are rather close. Compared to these two varieties Stjørdal 3 is relatively distant, indicating again strong variation in a small area. The two versions of Bodøy are recorded by the same person. They neatly form one cluster.

In the western group Nordvestlandsk varieties are mainly found. The adjoining Sørvestlandsk varieties of Vaksdal and Voss are in this group as well. More surprisingly is that the Sørlandsk dialect of Fyresdal is also in this group. It would be more fitting if this dialect were clustered with other Sørlandsk varieties. We cannot explain this. In the dendrogram the Sørlandsk varieties cannot be found as a group. The southwestern group mainly contains Sørvestlandsk varieties. The geographically adjacent Sørlandsk variety of Mandal is also in this group. The two varieties of Time neatly cluster together, just as the two varieties of Stavanger.

### 8.2.2 Area map

In the map in Figure 8.4 we treated some varieties as dialect islands, i.e. their color was not expanded to their surrounding by triangulation. The varieties of Herøy, Fræna and Bo do not belong to groups, instead they are treated as dialect islands. The varieties of Trondheim, Tynset, Lillehammer and Bergen were clustered with the southeastern varieties. As stated above, this is unexpected since they are geographically rather distant from the other varieties in the southeastern group and found among varieties of other groups. Therefore, we dealt with them as language islands. Finally the classification of Fyresdal with

\(^2\)Compare, e.g., the town Frisian dialect islands in the Frisian dialect area on the map of Daan and Blok (1969).
8.2. CLASSIFICATION

Figure 8.3: Dendrogram derived from the $55 \times 55$ matrix of Levenshtein distances showing the clustering of (groups of) Norwegian dialects. UPGMA clustering is used (see Section 6.1.2). The scale distance shows percentages. The abbreviations between parentheses are explained in the caption to Figure 8.2. The nine most significant groups are labeled and geographically visualized in Figure 8.4. The tree structure explains 48% of the variance.
Figure 8.4: The nine most significant groups as derived from the dendrogram in Figure 8.3. UPGMA clustering is used (see Section 6.1.2). Varieties treated as dialect islands are marked with a diamond. The dialects of Herøy, Bø and Fræna are marked with a white diamond. Each of them is a separate group.
the western group is unexpected as stated above. The variety is treated as a language island. In this way we get a clear division into six areas. For those locations for which more than one transcription was available, we selected only the most reliable transcription, or the transcription which is most representative for the surroundings of the location (see Section 8.1 for more details).

When comparing our map in Figure 8.4 with the map of Skjekkeland (1997) in Figure 8.2 it is most striking that the Midlands group on the map of Skjekkeland is not found on our map. As explained above, the fact that Lesja and Bo do not form one group may be explained by the fact that they are geographically rather distant. The northern part of the Austlands group corresponds with our East group. Our Southeast group covers the southern part of the Austlands varieties and the eastern part of the Sørlands varieties. Our Southwest group corresponds with the southern part of the Sørvestlands varieties, but also the Sørlands variety of Mandal is in this group. Our West group includes the Nordvestlands varieties. However on our map the southern border is shifted to the South. Our central group covers the southern part of the Trøndsk group. The northern Trøndsk varieties, the Helgeland variety, the Nordlands varieties and the Troms-Finnmark-mål are found as a North group on our map, covering a large geographic area.

8.3 Continuum

8.3.1 Multidimensional scaling

We also applied multidimensional scaling to the Levenshtein distances between the 55 Norwegian varieties (see Section 6.2). When applying this classification technique we found that one dimension explains 55% of the variance, two dimensions 79%, three dimensions 89%, four dimensions 91%, five dimensions 93%, six dimensions 94% and seven dimensions 95%. Using more than three dimensions only a small improvement of the explained variance is obtained. Therefore, we regard the three-dimensional solution which is shown in Figure 8.5.

The y-axis represents inversely the first dimension. At the top the central Trøndsk varieties are found, and at the bottom the southern Sørlands varieties. This accords with geography. However it is striking that the northern Helgeland, Nordlands and Troms-Finnmark-mål varieties and some Austlands varieties are found about intermediate between the Trøndsk and Sørlands varieties, which does not agree with geography. This suggests that the northern varieties are more related to the southern varieties than might be expected on the basis of simple geographical distance. The x-axis represents the second dimension. For the southern varieties a division in western and eastern varieties can be found. On the left the western Nordvestlands and Sørvestlands varieties are found, and on the right the eastern Austlands and most Sørlands varieties
are found. For the central and northern varieties we found no clear division in West and East. The grey tones represent the third dimension. The Austlandsk varieties of Fyresdal, Bø, Nordre Land, Stange, Trysil and Rendalen are represented by black dots (low values), the other Austlandsk varieties by darker grey dots, most remaining varieties by lighter grey dots, and the Nordlandsk variety of Sortland by a white dot (high value).

To get insight into the relation between the variation per dimension on the one hand, and variation in word pronunciations on the other hand, first we calculated distances between varieties per dimension. When two varieties have respectively the values $x$ and $x'$ in a dimension, the distance is equal to $|x - x'|$. In this way a distance matrix of $(55 \times 54)/2$ distances is obtained per dimension. Subsequently Levenshtein distances are calculated on the basis of the pronunciations of a single word. In this way we get 58 matrices for 58 words, each containing $(55 \times 54)/2$ distances. The distances which we calculated per dimension are correlated with the distances of each of the 58 matrices. In this way we found the strongest correlating word per dimension.

It appears that the distances in the first dimension correlate most strongly with distances obtained on the basis of pronunciations of the word *mann* ‘man’ ($r = 0.55$). In the northern and central varieties this word is mostly pronounced as [ˈmʌn] while it is usually pronounced as [ˈman:] in the southern varieties. They differ by the last segment: [n] versus [n]. Distances in the second dimension correlate most strongly with distances obtained on the basis of pronunciations of the word *enige* or *samde* ‘agreed’ ($r = 0.52$). In the western varieties forms like [tʰe:nɪ], [tʰemɪ], and [tʰemɪ] are used. In the eastern varieties mostly forms like [tʰemɪ], [tʰemɪ], and [tʰemɪ] are found. So in the western varieties a [g] is pronounced, but in the eastern varieties the [g] is elided. Furthermore, only in Herøy did we find the form [tʰsamɪ]. Distances in the third dimension correlate most strongly with distances obtained on the basis of pronunciations of the word *kven* or *hven* ‘who’ ($r = 0.44$). In Rendalen and Stange the word is pronounced as [tʰokːaːn], in Nordre Land and Bø as [tʰokːaːn], and in Trysil as [tʰokːaːn]. In the other varieties forms like [tʰkʰemɪ], [tʰkʰemɪ] and [tʰemɪ] are used. Lexical differences are represented by this word. Per dimension we also examined other strongly correlating words. However we found no system in the phenomena which causes differences between the pronunciations.

### 8.3.2 Continuum map

In Section 6.2.4 we explained that on the basis of three dimensions of the three-dimensional solution each variety can be represented by a color. The three dimensions are mapped to the intensities of red, green and blue. We used this approach to create a map in which each variety gets its own unique color. We assigned the colors to the three dimensions so that the different areas can be recognized rather clearly. The first dimension represents inversely the intensity of green, the second
Figure 8.5: Multidimensional scaling of the results derived from the $15 \times 55$ matrix of Levenshtein distances. Kruskal’s Non-metric MDS is used (see Section 6.2.2). The abbreviations between parentheses are explained in Figure 8.2. The y-axis represents inversely the first dimension, the x-axis represents the second dimension and grey tones the third dimension. Three dimensions explain 89% of the variance.
Figure 8.6: Dialect variation represented by color variation. The first MDS dimension is mapped inversely to green, the second is mapped to blue and the third to red. Kruskal’s Non-metric MDS is used (see Section 6.2.2). The color of intermediate points is determined by interpolation using Inverse Distance Weighting. Dialect islands are marked with a diamond. They are not involved in the interpolation process.
dimension represents the intensity of blue, and the third dimension the intensity of red.

In Figure 8.6 a color map based on three MDS dimensions is shown. In the map the space between the points is colored on the basis of MDS values which are found by interpolation using Inverse Distance Weighting (see Section 6.2.4). In this way the dialect landscape is represented as a continuum. Dialect islands are excluded from the interpolation process. The same varieties are treated as dialect islands as in the map in Figure 8.4. Where two or more transcriptions were available for the same location, just as in Figure 8.4 we selected only the most reliable one or the one which is most representative for its surroundings (see Section 8.1 for more details).

Comparing this map with the map of Skjekkeland (see Figure 8.2) we see that the eastern blue area corresponds with the Austlandsk varieties. Examining the blue shades in more detail, we find three areas. In and around Rendalen we find greenish blue. A more pure blue area is represented by Trysil, Nordre Land and Stange. Darker blue is found in and around the varieties of Oslo, Borre, Halden and Larvik. The purple area in the furthest south corresponds with the Sørlandsk varieties. In the West we find an area varying from red in the North to red-purple more south. The area corresponds with the Sørvestlandsk varieties. In the Northwest different colors are found, illustrating that the Nordvestlandsk varieties do not form a homogeneous group. The southern Trøndsk varieties of Oppdal and Sunndal and the Midlandsk variety of Lesja represent a central green area. This area is not found on the map of Skjekkeland (1997). In and around most northern Trøndsk varieties we find a blue-green color. The Hellelandsk, Nordlandsk and Troms-Finnmarks-mål varieties are represented by different shades of purple. However the three different groups are not clearly distinguished. The purple shades suggest a strong relation with the Sørlandsk varieties in the furthest south. In our map we found no Midlandsk area.

8.4 Conclusions

Examining the dendrogram in Figure 8.3, the map in Figure 8.4, the multidimensional scaling plot in Figure 8.5 and the map in Figure 8.6 we found some minor and some major differences compared to the map of Skjekkeland (1997). We explain this by two factors. First the text ‘the North Wind and the Sun’ is a rather short text. We are not sure of the extent to which the translations of this text are representative pictures of the varieties. In Bolognesi and Heeringa (2002) a word list is used where the words are randomly chosen from a corpus. In that way the data will be more representative. The classification as given in this section may not interpreted as the classification of Norwegian dialects, but as one classification which only reflects the variation in the translations of the fable ‘the North Wind and the Sun’. Second the map of Skjekkeland (1997) is based on
a restricted number of phenomena. Possibly the map may reflect the historical situation to some extent. To get more clarity about this, it would be interesting if a new map were created on the basis of the arrow method, just as was done by Daan and Blok (1969) for the Netherlandic part of the Dutch language area.

When comparing the map in Figure 8.4 with the map of Skjekkeland, we found some similarities, but also a lot of differences. In our opinion the map in Figure 8.5 is much more like the map of Skjekkeland. Figure 8.4 is based on the dendrogram in Figure 8.3. This dendrogram explains 48% of the variance. Figure 8.6 is based on the multidimensional scaling plot in Figure 8.5. This plot explains 89% of the variance. Therefore, we judge the map in Figure 8.6 to be more reliable than the map in Figure 8.4.