NORWEGIAN DIALECT DISTANCES GEOGRAPHICALLY EXPLAINED[1]

Charlotte Gooskens

Department of Scandinavian Studies
University of Groningen, the Netherlands

1. Introduction

In traditional dialectology, dialect variation is often represented by areas within which similar dialects are spoken. The dialect areas are found by drawing dividing lines (isoglosses) between areas where different representations are found for selected linguistic variables. However, different isoglosses do not always coincide which makes it difficult to draw borders between the dialect areas. Furthermore, speech variation mostly ranges along a continuum rather than being geographically abrupt. Generally, geographically remote areas are linguistically less similar than geographically close areas so that a high correlation can be expected between linguistic distance and geographic distance.

This does indeed also apply in the Dutch language area. Nerbonne et al. (1996) calculated linguistic distances between 350 Dutch dialects by means of the Levenshtein distance method (see Section 2.1.2). The linguistic distances showed a high correlation with geographic distances (r=.67) which means that a large part of the linguistic variation can be accounted for by geography. This seems to lend credibility to the continuum view and it suggests dialect distance to reflect mobility and cultural influence. If a place is easily accessible, people are more inclined to go to this place and the language varieties of the two places have a greater chance of influencing each other. However, a similar investigation (Gooskens and Heeringa submitted) showed the correlation between linguistic distance and geographic distance to be considerably lower in the case of 52 Norwegian dialects (r=.22).

The difference between correlations in the Dutch and the Norwegian language areas probably reflects the difference in geography. The Netherlands is a flat country with few natural obstacles, which means that it has always been rather easy to travel from place to place. Norway on the other hand has many mountains, which has made it difficult to travel between places. Until recently most of the travelling in Norway has taken place by boat along the coast. Assuming that the degree of accessibility between two places determines the linguistic distances between the two places to a high degree, it does not seem reasonable to correlate linguistic distances with straight geographical lines in the case of Norway since this does not reflect mobility well. Some other measure that takes the ease with which contact between places can take place should be used.

The aim of the present investigation was to investigate how much of the linguistic variation is accounted for by accessibility expressed in terms of the time it takes to travel between two places. To this end the linguistic distances between 15 Norwegian dialects were correlated with travel distances expressed in time. Travelling time can be expected to be a better representation of accessibility between places than straight lines in kilometres on a map in a country like Norway where travelling in straight lines is made difficult because of natural obstacles. The linguistic distances were correlated with travelling time from in year 2000. However, it can be expected that dialect distances reflect a prior geographical situation. In Norway the modern road system has been constructed quite recently and the linguistic distances can be expected to correlate better with historical data. For this reason the linguistic distances were also correlated with travel distances expressed in time in the year 1900. The travelling times were correlated with objective linguistic distances (Levenshtein distances) as well as with the linguistic distances between the dialects as perceived by the language users themselves.

2. Data

2.1 Linguistic distances

Both of the linguistic distance measures, Levenshtein and perceptual, are based on the same material from 15 Norwegian dialects. First this material will be described (Section 2.1.1) and next it will be explained how the Levenshtein distances (Section 2.1.2) and the perceptual distances (Section 2.1.3) between the 15 Norwegian dialects were calculated.

2.1.1 Material[2]

In Figure 1 the fifteen dialects which were used in the investigation are shown. These fifteen dialects represent a large part of the Norwegian language area. Only the dialects spoken in the far north are not represented.

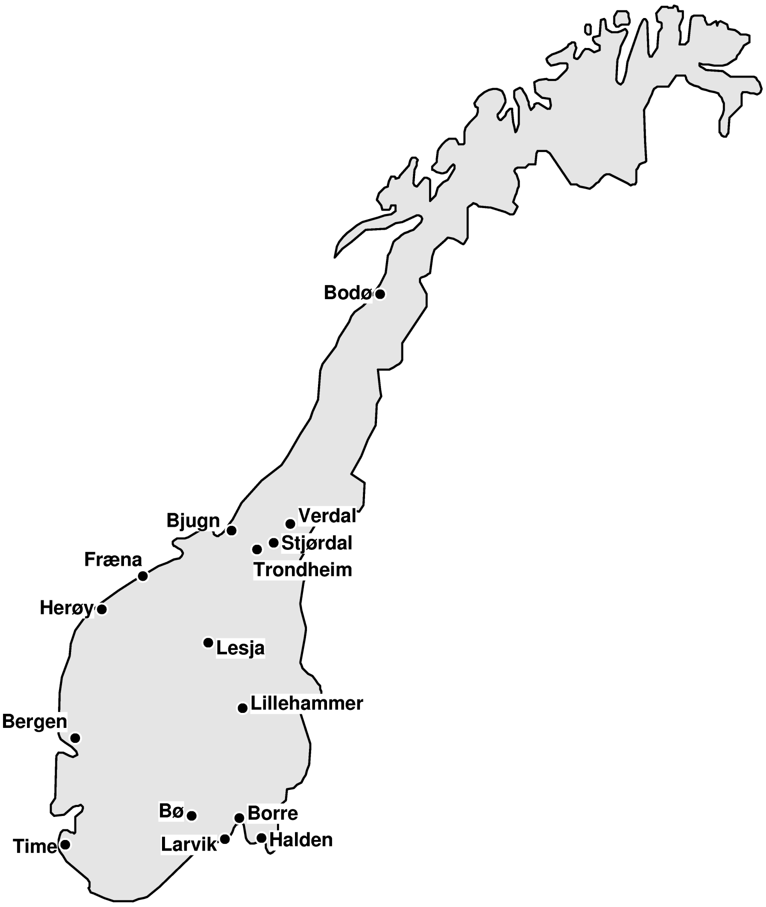

Figure 1. *Map of Norway showing the geographical distribution of the 15 Norwegian dialects used in the present investigation.*

The speakers all read aloud the same text, namely the fable 'The North Wind and the Sun'.[3] The Norwegian text consists of 58 different words which were used to calculate the Levenshtein distances. The recordings of the whole texts were used for the listening experiments which resulted in the perceptual distance measurements.

There were 4 male and 11 female speakers with an average age of 30.5 years. No formal testing of the degree to which the speakers used their own dialect was done. However, they had lived at the place where the dialect is spoken until the mean age of 20 (with a minimum of 18) and they all regarded themselves as representative speakers of the dialects in question. All speakers except one had at least one parent speaking the dialect.

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

On the basis of the recordings, phonetic transcriptions were made of all 15 dialects. These transcriptions were used to calculate the Levenshtein distances. The transcriptions were made in IPA as well as in X-SAMPA (eXtended Speech Assessment Methods Phonetic Alphabet). This is a machine-readable phonetic alphabet which is still human readable. Basically, it maps IPA-symbols to the 7 bit printable ASCII/ANSI characters. All transcriptions were made by the same person which ensures consistency. Most Norwegian dialects distinguish between two tonal patterns on the word level, often referred to as tonemes. We know from the literature that the realisation of the tonemes can vary considerable across the Norwegian dialects. Intonation is considered to be one of the most important characteristics of the different Norwegian dialect areas by Norwegian scholars (e.g. Christiansen 1954, Fintoft and Mjaavatn 1980, Sandøy 1993). However, no information was given about the precise realisation of the tonemes or intonation in the transcriptions.

2.1.2 Levenshtein distances

A linguistic distance measurement was gained by means of the Levenshtein distance measurements. With this method it is possible to measure objectively the phonetic distance between language varieties on the basis of phonetic transcriptions. The Levenshtein distance may be understood as the cost of (the least costly set of) operations mapping one string to another. The basic costs are those of (single-phone) insertions, deletions and substitutions. Insertions and deletions cost half that of substitutions. The simplest versions of this method are based on calculation of phonetic distance in which phonetic overlap is binary: non-identical phones contribute to phonetic distance, identical ones do not. Thus the pair [a,p] counts as different to the same degree as [b,p]. A more sensitive version is one in which phones are compared on the basis of their feature value, so the pair [a,p] counts as more different than [b,p]. However, it is not always clear which weight should be attributed to the different features. For this reason a version is used which compares spectrograms of the sounds.

It is a disadvantage of the method that it only takes segmental phenomena into consideration and leaves little room for the role which, for example, syntax and supra-segmental features such as intonation might play. In our case, morphology is included in the distance measurements, since words from a running text with different morphological forms are compared. For further detail about the Levenshtein distances see Nerbonne and Heeringa (2001) and Heeringa (2004).

For calculating the distance between two dialects, a large number of Levenshtein distances are determined – one difference per word, and the mean difference over all words is calculated. The Norwegian text consists of 58 different words which proved to be a sufficient basis for a reliable Levenshtein analysis (Cronbach's alpha was as high as 0.82). Some words occur more than once in the text. In these cases the mean distance over the variants of one word is used for calculating the Levenshtein distances. The distances between all pairs of dialects were put in a 15 by 15 matrix. Only half of the matrix was filled since the lower half is the mirror image of the upper half. The diagonal is always zero and is left out of consideration in our analysis. The results of the Levenshtein distance measurements can be found in Gooskens and Heeringa (submitted).

2.1.3 Perceptual distances

The listeners were 15 groups of high school pupils, one group from each of the places where the 15 dialects are spoken (see Figure 1). The listeners listened to the complete fable about the North Wind and the Sun in all 15 dialects. While listening to the dialects the listeners were asked to judge each dialect on a scale from 1 (similar to own dialect) to 10 (not similar to own dialect). Each group of listeners judged the linguistic distances between their own dialect and the 15 dialects, including their own dialect. Accordingly, there are two distances between each pair of dialects. In this way we get a matrix with 15 by 15 distances. However, in order to be able to correlate the distances with the Levenshtein distances and the geographical distances, the mean values of the upper and the lower half of the matrix were calculated. Furthermore, the diagonal was excluded, as in the case of the Levenshtein distances. For more details about the perceptual distance measurements between Norwegian dialects see Gooskens and Heeringa (submitted).

2.2 Travelling time

2.2.1 Modern travelling time

The modern Norwegian road system has been constructed quite recently. Until the nineteenth century few roads were suitable for vehicles. During the nineteenth century an increasing number of roads were built, first of all in order to improve the administration of the country. Today an extensive road system exists which makes it possible to travel by car to all places in the country. However, the detour which has to be made to travel between two places can be considerable. For example, the straight-line distance between Bergen and Oslo is 305 kilometres. When travelling by road, the distance is much longer, 468 kilometres. For this reason we expected the modern travelling time by road to reflect linguistic distance better than straight line distances in the case of Norway.

In order to calculate the travelling times, we used the program *Oplev Norge 2000* (Discover Norway 2000), which was developed by Statens Kartverk, the Norwegian geographical institute. With this program it is possible to measure the distance in kilometres by road between two places and furthermore the program can calculate the travelling time by car or by bicycle. The user can define the travelling speed himself. We used the default values defined for cars, which are 70 km per hour on national roads, 50 km per hour on county roads, 30 km per hour on smaller roads and private roads and 10 km per hour at the stretches which have to be travelled by boat. So, in fact the travelling times by car take into account the size of the roads between two places. The travelling times expressed in minutes were entered in a 15 by 15 matrix with the travelling times between all places. This matrix could be correlated with the linguistic distances (see Section 3).

2.2.2 Old travelling time

Dialects change constantly across time under influence from among others the contact with other language varieties. Circumstances in the past still have an effect on modern dialects. When investigating the role of accessibility on the linguistic distance between dialects it is therefore obvious that one should look at accessibility in the past. However, it is difficult to decide which time in the past has had the strongest influence on dialects as they are spoken today. We chose to look at travelling times in the year 1900. This is a point in history when some parts of the railroad system had already been built while the road system was still rather poor, so that a large part of the travelling had to be done by boat along the coast. From the end of the nineteenth century, regular services were established within public transportation. The year 1900 is well-documented so that it is possible to retrieve data about travelling times with fair precision. From 1868 a travelling guide which gives detailed information about travelling in Norway was published and around the same time, the first time-tables appeared for the steamboat along the cost and for the train. Furthermore, there was also an extensive system of conveyance by horse which was regulated by law. This system included permanent posting stations at the main roads. From information about this system we are able to calculate the mean transportation times by horse or carriage and together with the old time tables it is possible to get at reliable picture of travelling circumstances and travelling times in the year 1900 on all routes connecting the 15 places in our investigation. However, it was not possible to take into account the waiting time when changing from one mean of transportation to another. This waiting time could sometimes amount to several hours or even days. Furthermore one should bear in mind that travelling in the winter was very difficult or even impossible in some parts of Norway. Thus, the travelling times used in the present investigation are based on the ideal situation without waiting time and bad weather. For more information about transportation in Norway in the past, see Bjørnland (1977).

A number of choices had to be made when deciding how to calculate the travelling times. Sometimes there were two routes leading from one place to another. For example, it was often possible to go by horse carriage as well as by train. We always chose the fastest route even though this might not have been the best choice in all cases. Even though the route by train or boat was sometimes twice as long as by road, it often turned out to be the quickest way between two places. For example the distance between Bergen and Bø is calculated as follows (see also Figure 2):

Bergen-Larvik by steamship:	37 hours 15 minutes
Larvik-Nordstrand by train:	5 hours 21 minutes
Nordstrand-Bø by train and horse carriage:	10 hours 53 minutes
Total:	60 hours 29 minutes

As becomes clear from Figure 2 a large detour had to be made in order to travel around the mountains of central Norway. Still it took much longer or it was almost impossible to travel across the mountains by horse. Also it took a longer time to travel through the mountains from the coast to Bø than by train and horse via Nordstrand. When travelling the same distance by car in modern Norway the route between Bergen and Bø goes across the central mountains and the distance can be travelled ten times as fast as in 1900 (6 hours and nine minutes, see Figure 2). In reality the difference was probably even larger since it was hardly possible to travel non-stop for 60 hours and 29 minutes.

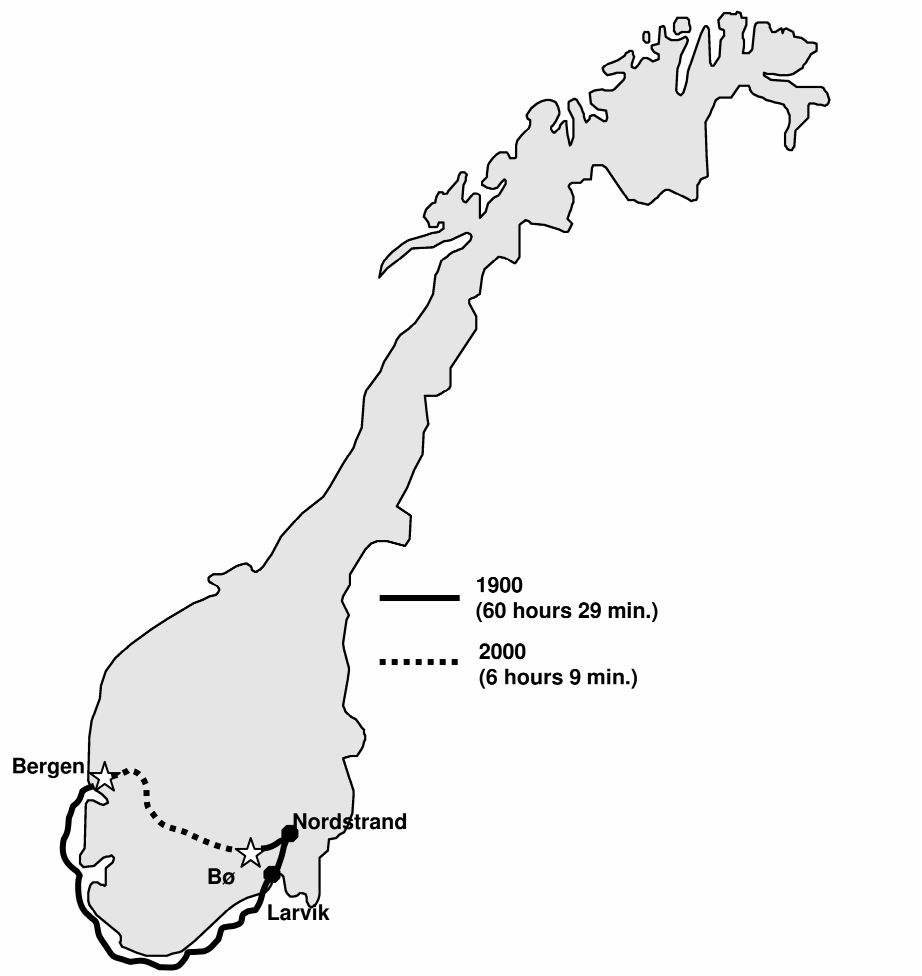

Figure 2. *Map showing the travelling route in 1900 between Bergen and Bø in 1900 (full line) and 2000 (dotted line). In 1900 the journey lasted 60 hours and 29 minutes and went by boat between Bergen and Larvik, by train between Larvik and Nordstrand and by train and horse carriage between Nordstrand and Bø. In 2000 the journey took 6 hours and 9 minutes by car.*

3. Results

3.1 Correlations between linguistic distances and travelling times

As explained in Section 2, the linguistic distance measurements resulted in two matrices with the distances between all 15 dialects, one for the perceptual distances and one for the Levenshtein distances. For the travelling times we have two different matrixes, one for the modern travelling times and one for the old travelling times. We also have the matrix for the straight-line distances in kilometres between the 15 places. So, in total we correlate 5 different matrixes. For each pair of matrices we calculated the Pearson correlation coefficient. The results are shown in Table 1. In addition to the linear correlations, the logarithmic correlations are given for the correlations between linguistic distances and geographical distances. The logarithmic correlation coefficients are higher in these cases because dialect distance increases when geographical distance increases, but only to a certain extent.

Table 1 *The linear correlations and the logarithmic correlations (between brackets) between the linguistic and geographical distances between 15 Norwegian dialects.*

	Levenshtein	perceptual	straight lines	modern	old
Levenshtein	-	.68	.27 (.41)	.30 (.41)	.51 (.54)
perceptual		-	.54 (.68)	.54 (.71)	.76 (.86)
straight lines			-	.98	.68
modern				-	.67

3.1.1 Correlation between linguistic distances and modern travelling times

As expected (see Section 1), the correlations between the linguistic distances between the 15 dialects and straight lines in kilometres are low. It is .27 when correlating with Levenshtein distances and .54 when correlating with perceptual distances. As discussed in Section 2.2, a higher correlation is expected when correlating with travelling times because it takes into account the detour which has to be made around a mountain or a lake, or the time delay when a river has to be crossed by boat. However, this did not turn out to be the case. The correlation was the same in the case of the perceptual distances (.54) and only slightly higher in the case of the Levenshtein distances (r = .30). Also the logarithmic coefficients hardly differ. Apparently the modern travelling times are not a better representation of the amount of contact between the Norwegian dialects than the straight-line representation. This can probably be explained by the well-developed modern road system which to a high extent follows the shortest geographical route. No roads are completely straight, but none of the distances between two dialects have to be travelled by a very large detour or at least the detour is similar for all travel distances which results in little difference in correlation with linguistic distances. This is also reflected by a high correlation between the straight lines and the modern travelling times (r = .98).

3.1.2 Correlation between linguistic distances and old travelling times

The dialect situation can be expected to be a reflection of the amount of contact between dialects in the past. This was the reason to look at old travelling times as well. As we saw in Figure 2, the routes which had to be followed between two places were sometimes very different in the years 1900 and 2000. The question is now whether the old travelling times from 1900 are indeed a better reflection of the dialect distances. When correlating the old travelling times with the linguistic distances we see a considerable improvement compared to the correlations with modern travelling times. This goes for the Levenshtein distances (r = .51 versus .30) as well as the perceptual distances (r = .76 versus .54). The logarithmic correlations are even higher (r = .54 versus .41 for the Levenshtein distances and .86 versus .71 for the perceptual distances).

The results clearly show that, as with Dutch dialects, a large part of the linguistic variation can be accounted for by geography. However, in the case of a geographically more complicated country like Norway, travelling times are a better representation of the geographical situation than straight line distances at least if the historical aspect is also taken into consideration since the old travelling circumstances are to a large extent reflected in the modern language. When examining the residuals resulting from the correlations between linguistic distances and the old travelling times as compared to correlations with the straight line distances, we see that the correlation with the old travelling times is higher because the old travelling times between places at both sides of the central mountain are better predictions of the linguistic distances than straight lines or modern travelling times. The residuals still remaining when correlating with old travelling times, concern the distances between the places around Oslo in the south-east and Lillehammer. The contact between these places has probably been more intensive than can be deduced from the travelling times in 1900 and therefore the linguistic distances are smaller than what could be expected from travelling times. On the other hand linguistic distances to some smaller places are larger than predicted from the old travelling times. Some of these places might have been more isolated than can be deduced by the travelling times, since travelling time does not say anything about the frequency of travelling to a place. It is also possible that the travelling times were in fact longer than the times which we used for our calculations because we did not take into consideration that the waiting times were sometimes considerable.

When comparing the correlations with the two kinds of linguistic distances, perceptual and Levenshtein, it is clear that the old travelling times are a better representation of the linguistic distances as perceived by listeners than the objective distances expressed in Levenshtein distances. Probably the explanation for this difference can be found in the different amount of linguistic data on which the two kinds of linguistic distances are based. The listeners based their judgements on complete texts. This means that the perceptual distances are based on all linguistic information including prosody. The Levenshtein distances were based on phonetic transcriptions of isolated words, which means that intonation and tonemes are not taken into consideration when calculating the distances (see Section 2.1.2). Intonation and tonemes are important characteristics for the perception of Norwegian dialects (see Gooskens submitted) and therefore the Levenshtein distances are a less good representation of the linguistic distances than the perceptual distances in this respect.

Furthermore, the difference between the Levenshtein distances and the perceptual distances might be explained by the fact that all segments are given the same weight when calculating the Levenshtein distances while listeners might base their judgements on single important characteristics of the dialect, so-called shibboleths. One occurrence in a dialect might have a great influence on the judgements while it has only little influence on the Levenshtein distances.

Finally, part of the explanation for the difference in correlation with the perceptual distances and Levenshtein distances might be that the listeners were able to use their knowledge about geographical distances and travelling time when making their judgments. If a listener for example knows that a dialect is spoken far away, he might be influenced when making his judgment and judge the dialect to be very deviant from his own dialect, not basing his judgments entirely on linguistic information. If it is indeed the case that listeners use geographical knowledge when making their judgments, this would mean that they also take natural obstacles such as mountains into account.

4. Conclusion and discussion

The results of the present study clearly show that accessibility in the past still has influence on the dialects spoken in the Norwegian language area today. By correlating old travelling times with linguistic distances, it became clear that places which are easily reached are more likely to show a greater linguistic similarity with other dialects than more isolated places.

However, the correlations between linguistic distances and travelling times showed that it is not possible to predict linguistic distances entirely on the basis of travelling times from 1900. Correlations might be improved by a more precise calculation of travelling times incorporating waiting time. But even if we were able to calculate travelling time more precisely, information about the frequency of travelling would also be an important addition since this would give a more accurate picture of the amount of contact between speakers of different dialects. It is also possible that even older travelling times from the time before the railway system was constructed might be a better reflection of the present dialect distances.

We explained the residuals by deficiencies in the calculations of the travelling times or the linguistic distances. There are, however, other factors that might explain some of the residuals. One of them is the attitude towards the different dialects. It is known from the literature that different groups of the population have different attitudes towards different dialects. It is possible that such attitudes influence the perceived distance between dialects. For example, listeners might judge dialects which they have a negative attitude towards as being more deviant from their own dialect than expected from pure linguistic characteristics of the dialect. Or the other way round, if for some reason they are very positive about a dialect, they might judge it to be very similar to their own dialect. It is also possible that attitudes have influenced the real linguistic distances. If a group of dialect speakers have a negative attitude towards another group of dialect speakers they will not want their own dialect to sound similar and there is not likely to be much contact between the speakers. The result might be that the dialects grow apart.

In order to explain linguistic distances more precisely a number of geographical and demographic factors should also be taken into account. Urban centres are important in the spreading of linguistic innovations and might therefore cause dialects to converge to dialects spoken in economically, politically and culturally dominant places (Chambers and Trudgill p. 172). This effect is reinforced in modern time under the strong influence of the spoken mass media. This means that the size of the place where the dialect is spoken should be taken into account when modelling linguistic distances. Migration and immigration might also result in the spreading of linguistic variables. Furthermore, population density might play an important role. In densely populated areas there is more contact between dialect speakers which might cause the dialects to converge. Political and historical borders on the other hand might have the opposite effect of divergence. It would be instructive to incorporate the above-mentioned geographic, demographic and attitudinal factors into a model for predicting linguistic distances between dialects. This might lead to a greater understanding of mechanisms involved in dialectal variation.

Literature

Bjørnland, D. (1977). *Innenlands samferdsel i Norge siden 1800: del 1: Demring (1800-1850-tallet)*, Transportøkonomisk institutt, Oslo.

Chambers, J. K. and P. Trudgill (1989). *Dialectology.* Second edition. Cambridge: Cambridge University Press.

Christiansen, H. (1954). ‘Hovedinndelingen av norske dialekter’. *Maal og Minne*: 30-41.

Fintoft, K. and Mjaavatn, P.-E. (1980). ‘Tonelagskurver som målmerke’. *Maal og Minne* 1980: 66-87.

Gemert, I. van (2002). ‘Het geografisch verklaren van dialectafstanden met een GIS’. Unpublished MA-thesis, University of Groningen.

Gooskens, C. (submitted). ‘How well can Norwegians identify their dialects?’ *Nordic Journal of Linguistics.*

Gooskens, C. and W. Heeringa (submitted). ‘Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data‘. (submitted to *Language Variation and Change*).

Heeringa, W. (2004). ‘Measuring dialect pronunciation differences using Levenshtein distance’. Dissertation, Groningen: University of Groningen.

Nerbonne, J.; Heeringa, W.; Hout, E. van den; Kooi, P. van der; Otten, S.; Vis, W. van de. (1996). ‘Phonetic distance between Dutch Dialects’. In: G. Durieux, W. Daelemans and S. Gillis (eds*.). Proceedings of Computer Linguistics in the Netherlands ’95*, Antwerpen, 185-202.

Nerbonne, J. and Heeringa, W. (2001). ‘Computational comparison and classification of dialects’. *Dialectologia et geolinguistica. Journal of the International Society for Dialectology and Geolinguistics,* 9, 69-83.

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

Notes

[1] The authors wish to thank the following people for their help at different stages in the investigation: Dag Bjørnland for providing the historical data, Femke Jongerius for data entry, Wilbert Heeringa for help with the Levenshtein distances and Peter Kleiweg for digitalizing and visualizing the maps.

[2] See Gooskens and Heeringa (submitted) for more details about the material.

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