DiG 13 (2005), 38–62

Traveling time as a predictor of linguistic distance

Charlotte Gooskens

Abstract

The aim of the present investigation[1] was to get an impression of the geographic influences on the dialectal variation in a country. In previous investigations, the correlations between linguistic distances and geographic distances using dialect data from the Netherlands and Norway were calculated (Gooskens and Heeringa 2004, Nerbonne et al. 1996). The results showed a high correlation in the case of Dutch data while the correlation was considerably lower in the case of Norwegian data. This seems to reflect the fact that especially for Norway the direct distance between two settlements does not reflect the difficulty of travel and therefore social contact, which is expected to play a role in keeping linguistic distance within limits. Holland is a country with a flat, regularly populated landscape with few natural obstacles such as mountains and rivers. This is in great contrast with Norway with its high mountains and many *fjords* which made it quite difficult to travel between places, especially in the past. These differences in geographical situations are clearly reflected in the correlations between the linguistic and geographical distances between the dialects of the two countries.

The present investigation is searching for more successful ways of predicting linguistic distances by means of geographic distances in Norway. To this end, old and new traveling data were used providing information about traveling times by road, train, and boat between the places where the different dialects are spoken. The results show that a large part of the linguistic variation can be accounted for by geography in Norway, just as in the Netherlands. However, in the case of a geographically more compli-

---

[1] This article is based on a paper read at the Second International Conference on Language Variation in Europe, Uppsala, 12–14 June 2003. The author wishes to thank the following people for their help at different stages in the investigation: Dag Bjørnland (BI Norwegian School of Management) for providing the data on old travelling times, Femke Jongerius (University of Groningen) for data entering, Wilbert Heeringa (University of Groningen) for help with the Levenshtein distances and statistics and for making the maps, Peter Kleiweg (University of Groningen) for his software for creating the maps and John Nerbonne (University of Groningen) for comments on an earlier version of this paper.

cated country like Norway, traveling times reflect the influence of geography on linguistic variation better than straight-line distances if the historical aspect is also taken into consideration since the old traveling circumstances are reflected in the modern dialects. Traveling times can vary independently from straight-line distances, and this variation was more pronounced in the past, when the present dialect situation crystallized.

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 (r^2=.45). This seems to lend credibility to the continuum view and it suggests that dialect distance reflects mobility and cultural influence. If a place is easily assessable, 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 2004) 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 traveling 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 should be used that takes the ease with which contact between places can take place.

The aim of the present investigation is to see 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 are correlated with travel distances expressed in time. Traveling time can be expected to be a better representation of accessibility between places than straight lines in kilometers on a map in a country like Norway where traveling in straight lines is made difficult because of natural obstacles. First, the linguistic distances were correlated with modern traveling times from the year 2000. However, it can be expected that dialect distances reflect a prior geographical situation. In Norway the modern road system is quite recent 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 from the year 1900. The traveling 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.

This paper is organized as follows. In Section 2 it is shown how the different linguistic and geographical distances were measured. In Section 3 the correlations between the linguistic distances and the geographic distances are presented and the results are discussed and explained by looking at the residuals. In Section 4 there will be a final discussion and suggestions for further steps to be taken in order to get to a better understanding of dialectal variation.

2. Data

2.1. Linguistic distances

There are two kinds of linguistic distance measures involved in the present investigation, the Levenshtein distance measure and the perceptual distance measure. Both of these measures are based on the same material from 15 Norwegian dialects. First this material is described (Section 2.1.1) and next it is 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]

Dialects

The dialects of Norway are in a strong position. In contrast with many European countries people of all ages and social backgrounds use their dialects not only in the private domain but also in official contexts (Omdal 1995). This makes it easy to use recent recordings of young people from all over the country without the risk that some of the speakers might use a standardized variant of their dialect or a variety that is no longer being used in every day life. Another advantage is that it does not feel unnatural for Norwegian people to read aloud a text in their own dialect. This made it possible to use read texts, which was necessary since the same text in different dialects is needed for the calculation of the Levenshtein distances (see 2.1.2). 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.

The speakers all read aloud the same text, namely the fable 'The North Wind and the Sun'.[3] This text has often been used for phonetic investigations, see for example *The International Phonetic Association* (1949 and 1999) where the same text has been transcribed in a large number of different languages. Alternatively a word list might have been used, but the use of a running text ensures that the sample is random. 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.

Speakers

There were 4 male and 11 female speakers. Thirteen of the speakers had filled in a questionnaire about their background. The average age of these speakers was 30.5 years, ranging between 22 and 35, except for one speaker who was 66. All thirteen speakers attended university or already had a university degree.

---

[2] See Gooskens and Heeringa (2004) 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/. I am grateful for their permission to use the material.

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

No formal testing of the degree to which the speakers used their own dialect was carried out. 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.

Recordings

The recordings were made in a soundproof studio in the autumn of 1999 and the spring of 2000. 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 only 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.

The microphone used for the recordings was a MILAB LSR-1000 and the recordings were made in DAT format using a FOSTEX D-10 Digital Master Recorder. They were edited by means of Cool Edit 96 and made available at the World Wide Web (see note 6).

Transcriptions

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. It is known from the literature that the realization of the tonemes can vary considerably 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 realization of the tonemes or intonation in the transcriptions.

2.1.2 Levenshtein distances

A linguistic distance measurement was obtained by means of the Levenshtein distance measurements. With this method, it is possible to measure the phonetic distance between language varieties on the basis of phonetic transcriptions in an objective manner. Using the Levenshtein distance, two dialects are compared by comparing the pronunciation of a word in the

first dialect with the pronunciation of the same word in the second. It is determined how one pronunciation is changed into the other by inserting, deleting or substituting sounds. Weights are assigned to these three operations. In the simplest form of the algorithm, all operations have the same cost. Assume *afternoon* is pronounced as ['æəftəˌnʉ̟ˠn] in the dialect of Savannah, Georgia, and as [ˌæftər'nˠn] in the dialect of Lancaster, Pennsylvania. Changing one pronunciation into the other can be done as in the following way (ignoring suprasegmentals and diacritics for this moment):

æəftənʉ̟n	delete ə	1
æftənʉ̟n	insert r	1
æftərnʉ̟n	subst. ʉ̟/u	1
æftərnun		
		3

In fact many sequence operations map['æəftəˌnʉ̟ˠn] to [ˌæftər'nˠn]. The power of the Levenshtein algorithm is that it always finds the cost of the cheapest mapping. 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 was 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 is 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 (2004).

2.1.3. Perceptual distances

Listeners

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). Each group consisted of 16 to 27 listeners (with a mean of 19). Their mean age was 17.8 years, 52 percent were female and 48 percent male. Only the responses of listeners who had lived the major part of their lives in the place where the dialect is spoken were used for the analysis. On average these listeners had lived in the place in question for 16.7 years. Nine of the 290 listeners (3%) said that they never spoke the dialect, the rest spoke the dialect always (60%), often (21%), or seldom (16%). A large majority of the listeners (83%) had one or two parents who also spoke the dialect.

Procedure

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). They were also asked to judge the dialects and the speakers on a number of attitudinal scales. The experiment was followed by a questionnaire. In this questionnaire the listeners were asked questions about their individual characteristics, such as language background, age and sex. The listeners were paid for their participation.

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 a matrix was achieved 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.

The correlation coefficient between the Levenshtein distances and the perceptual distances is .68 (n=225, p<.000) when excluding the distances as perceived by the listeners to their own dialects and the corresponding Levenshtein distances which are always equal to zero. This shows that the Levenshtein distances are a good representation of the distances between

dialects as perceived by listeners. Furthermore, it shows that listeners base their judgments of dialectal distances on phonetic information to a great extent. For more details about the perceptual distance measurements between Norwegian dialects see Gooskens and Heeringa (2004).

2.2. Traveling time

As mentioned in the introduction, Levenshtein distances and straight geographical lines between dialects correlated highly in the case of the Dutch situation while the correlation was considerably lower for Norwegian dialects. There is reason to expect that the correlation will be higher if the linguistic distances are correlated with distances expressing traveling time since such distances will reflect the ease with which a place can be reached. If two places are separated by for example a mountain, contact between the two places might have been scarce and this is reflected by the long traveling time around or across the mountain. The traveling time between two places which are not separated by a mountain or some other obstacle is shorter.

Van Gemert (2002) used a GIS system to calculate traveling distances between dialects in the Netherlands, incorporating information about roads, rivers and lakes. However, no improvement in correlation with linguistic distances was achieved compared to correlation with straight-line distances. This can probably be explained by the fact that the straight-line distances are a fairly good approximation of the road system in The Netherlands. However, even in present day Norway with its extensive road system, 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 kilometers. When traveling by road, the distance is much longer, 468 kilometers. For this reason the modern traveling time by road is expected to reflect linguistic distance better than straight-line distances in the case of Norway. Still, it can be expected that the present linguistic distances reflect an older geographic situation from the time when little traveling was done by road, but rather by boat or train. In Section 2.2.1 it is shown how the information about modern traveling time between the 15 dialects was gained and in Section 2.2.2 how the traveling times from the year 1900 were calculated.

2.2.1 Modern traveling time

The modern Norwegian road system was constructed quite recently. Until the nineteenth century few roads were suitable for vehicles. During the nineteenth century an increasing number of roads were build, 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.

A number of different software programs exist with which a traveler can plan the route between two places. A rather advanced program, *Oplev Norge 2000* (Discover Norway 2000), was used. This program is developed by Statens Kartverk, the Norwegian geographical institute. With this program it is possible to measure the distance in kilometers by road between two places and furthermore the program can calculate the traveling time by car or by bicycle. The user can define the traveling speed himself. The default values defined for cars were used for the investigation. These 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 traveled by boat. In this way the traveling times by car take into account the size of the roads between two places. The traveling times expressed in minutes were entered in a 15 by 15 matrix with the traveling times between all places. This matrix could be correlated with the linguistic distances (see Section 3). In the appendix on page 61, an overview is given of the modern and old traveling times between all 15 places in the investigation.

2.2.2 Old traveling 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. The traveling times in the year 1900 were chosen. This is a point in history when some parts of the railroad system had already been build while the road system was still rather poor so that a large part of the traveling 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 traveling times with fair precision. Thomas Bennett first published a traveling guide in 1867 which gives detailed information about traveling in Norway. This traveling guide was updated regularly in the years to come. Around the same time the first time schedules appeared for the steamboat along the cost and for the train. Furthermore, at the time there was 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 it is possible to calculate the mean transportation times by horse or carriage and together with the old time tables it is possible to get a reliable picture of traveling circumstances and traveling 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 traveling in the winter was very difficult or even impossible in some parts of Norway. Thus in fact our traveling times 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 and 1989) and Bjørnland/Hajum (1979).

A number of choices had to be made when deciding how to calculate the traveling 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. The fastest route was always chosen 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. But it took much longer – in fact 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 traveling the same distance by car in modern Norway the route between Bergen and Bø goes across the central mountains and the distance can be traveled 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.

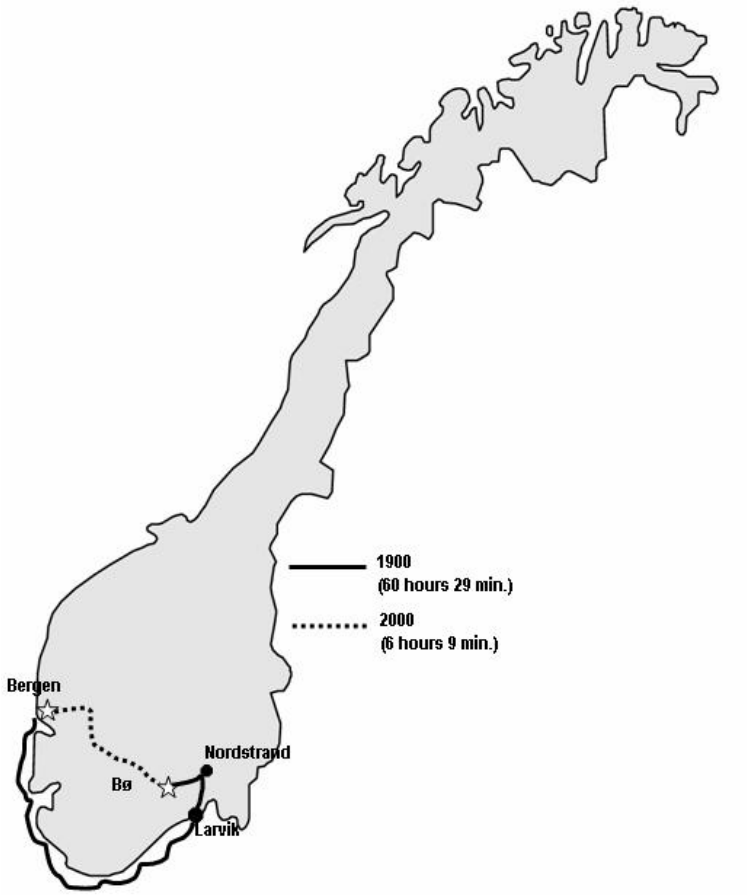


Fig. 2. Map showing the traveling route in 1900 by boat between Bergen and Bø in 1900 (full line) and 2000 (dotted line). In 1900 the journey lasted 60 hours and 29 minutes and went by boat between Bergen and Larvik, by train between Larvik and Nordstrand and by train and horse carriage between Nordstrand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

Just as the modern traveling times, the old traveling times are expressed in minutes and entered in a 15 by 15 matrix, see the appendix on page 61.

3. Results

In Section 3.1 the linguistic distances, Levenshtein and perceptual, are correlated with the modern and old traveling times. In Section 3.2 the residuals are examined in order to explain the results.

3.1 Correlations between linguistic distances and traveling times

As explained in Section 2 the linguistic distance measurements resulted in two matrixes with the distances between all 15 dialects, one for the perceptual distances and one for the Levenshtein distances. For the traveling times there are two different matrixes, one for the modern traveling times and one for the old traveling times. There is also the matrix for the straight-line distances in kilometers between the 15 places. So, in total we had 5 different matrixes. For each pair of matrices the Pearson correlation coefficient was calculated. 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.

	Levenshtein	perceptual	straight lines	modern	old
Levenshtein	-	.68	.29 (.41)	.30 (.41)	.52 (.53)
perceptual		-	.56 (.74)	.54 (.71)	.76 (.86)
straight lines			-	.98	.68
modern				-	.67

Table 1. The correlations between the linguistic and geographical distances between 15 Norwegian dialects. Between brackets logarithmic correlation coefficients are given.

3.1.1 Correlation between linguistic distances and modern traveling times

As expected (see Section 1), the correlations between the linguistic distances between the 15 dialects and straight lines in kilometers are low. It is .29 when correlating with Levenshtein distances and .53 when correlating with perceptual distances. As discussed in Section 2.2 a higher correlation was expected when correlating with traveling 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 when attention was restricted to modern traveling. 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 traveling 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 great extent follows the shortest geographical route. No roads are completely straight, but none of the distances between two dialects have to be traveled via 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 traveling times (r = .98). As became clear in Section 2.2 it also did not improve the correlation coefficient in the case of Dutch dialects to correlate the linguistic distances with modern traveling distances gained from a GIS system.

3.1.2 Correlation between linguistic distances and old traveling 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 traveling times as well. As became clear from 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 traveling times from 1900 are indeed a better reflection of the dialect distances. When correlating the old traveling times with the linguistic distances there is a considerable improvement compared to the correlations with modern traveling times. This goes for the Levenshtein distances (r = .52 versus .30) as well as the perceptual distances (r = .76 versus .54). The logarithmic correlations are even higher (r = .53 versus .41 for the Levenshtein distances and .86 versus .71 for the perceptual distances).

The results suggest that a large part of the Norwegian dialect variation can be accounted for by geography. The degree to which geography predicts dialect variation in Norway is in fact similar to that in The Netherlands. However, in the case of a geographically more complicated country like Norway, traveling times are a better representation of the influence of geography than straight-line distances, in particular when the historical aspect is also taken into consideration. The old traveling circumstances are to a large extent reflected in the modern language.

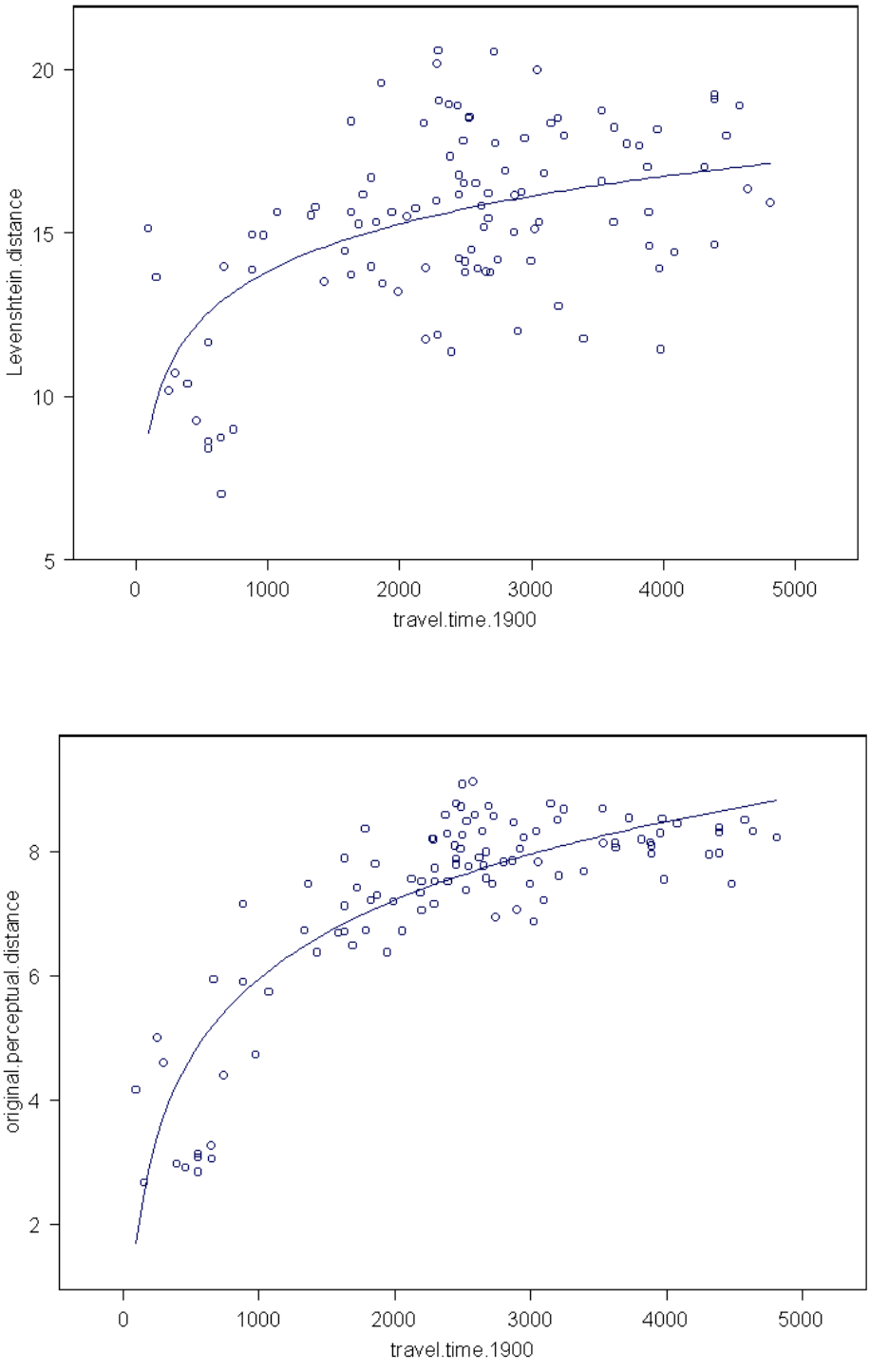
In figure 3a the logarithmic regression line for the old traveling times versus Levenshtein distances is shown and in figure 3b the logarithmic regression line for the old traveling times versus perceptual distances. When comparing the correlations with the two kinds of linguistic distances, perceptual and Levenshtein, it is clear that the old traveling times

are a better representation of the linguistic distances as perceived by listeners than the objective distances expressed in Levenshtein distances.

Perhaps 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 judgments 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 2005) and therefore the Levenshtein distances are a less successful representation of the linguistic distances than the perceptual distances in this respect. In Gooskens and Heeringa (2004) the listeners also judged the linguistic distances between versions of the dialects recordings where intonation had been removed electronically from the signal by means of monotonisation so that they could base their judgments on segmental information only. The correlation between the judgments of this version and the old traveling times is almost identical to the correlations with the Levenshtein distances ($r = .56$ for monotonous version versus .51 for the Levenshtein distances for the linear regression line).

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 judgments on single important characteristics of the dialect, so-called *shibboleths*. One occurrence in a dialect might have a great influence on the judgments while it has only little influence on the Levenshtein distances. It is also possible that the phonetic transcriptions lack information to which the listeners are sensitive.

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 traveling 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.



Figs 3a and 3b. The logarithmic regression lines for the old traveling times versus Levenshtein distances (3a) and the old traveling times versus perceptual distances (3b).

3.2 Residuals

Our results make clear that the ease with which people from different parts of Norway were able to have contact in the past to a large degree reflects the linguistic distances between the dialects as they are spoken today. However, since the correlations are not perfect, it might add to our understanding of the results to look at the residuals of the regression line. By examining the residuals it may be possible to explain them. Only the results from the correlations with old traveling times will be examined, since correlations with modern traveling times did not improve the correlations coefficients compared to the correlations with straight-line distances.

In figure 4a the residuals of the logarithmic regression lines of the straight-line distances versus Levenshtein distances are shown and in 4b the residuals of the old traveling times versus Levenshtein distances. In figures 5a and 5b the residuals are shown for the perceptual distances. Distances larger than predicted by regression on the basis of geographic distances are indicated by red lines, and distances smaller than predicted by regression by blue lines. The intensity of the colour represents the extent of the deviation with respect to the regression value. The intensity of the colour in the maps of the Levenshtein distances (figures 4a and 4b) have been scaled with respect to each other as was the colour of the lines in figures 5a and 5b. This means that the intensity of the colour in the figures 4a and b cannot be compared directly to the intensity in figures 5a and 5b. If the four figures had been scaled in the same way, hardly any lines would be left in figure 5b since there a few large residuals (see figure 3b).

When comparing the residuals of the straight lines (figures 4a and 5a) to the residuals of the old traveling times (figures 4b and 5b) it is clear that in the case of straight lines the linguistic distances between the dialects on each side of the mountains in central Norway are larger than predicted by the geographical distance. However, when the linguistic distances are predicted on the basis of old traveling times, most of the full lines across the mountains have disappeared or have become much lighter. This clearly shows the great impact of the central mountains on the distances between Norwegian dialects.

The colour of many of the lines which remain when correlating with the old traveling times are less intense or no longer visible, but still there are a number of lines present, especially in the north-south direction. As far as the blue lines are concerned there might have been more contact between the places around Oslo in the south-east and Lillehammer than can be predicted by the traveling times (see figure 4b and 5b). Also the contact with other of the more important places Bergen, Trondheim and Bodø might have been more intense than what can be deduced from traveling

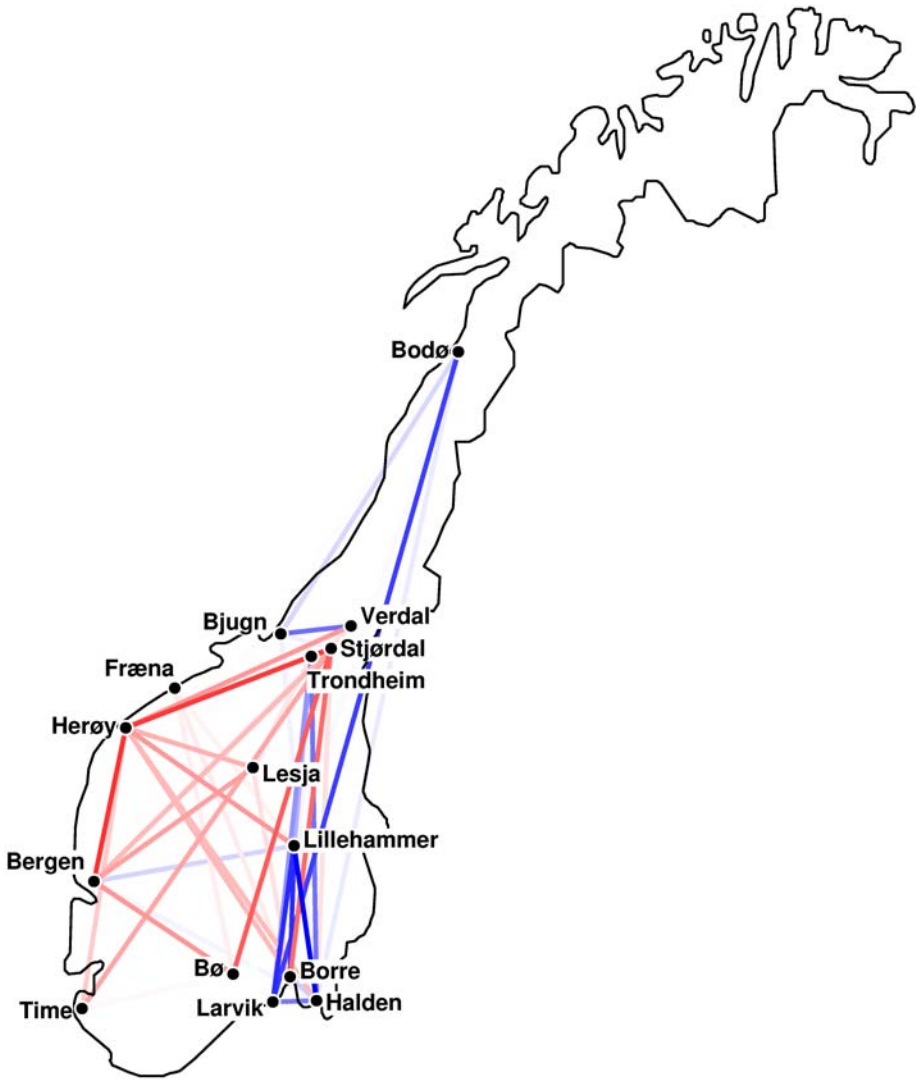
times leading to linguistic distances which are smaller than predicted (see figure 4b).

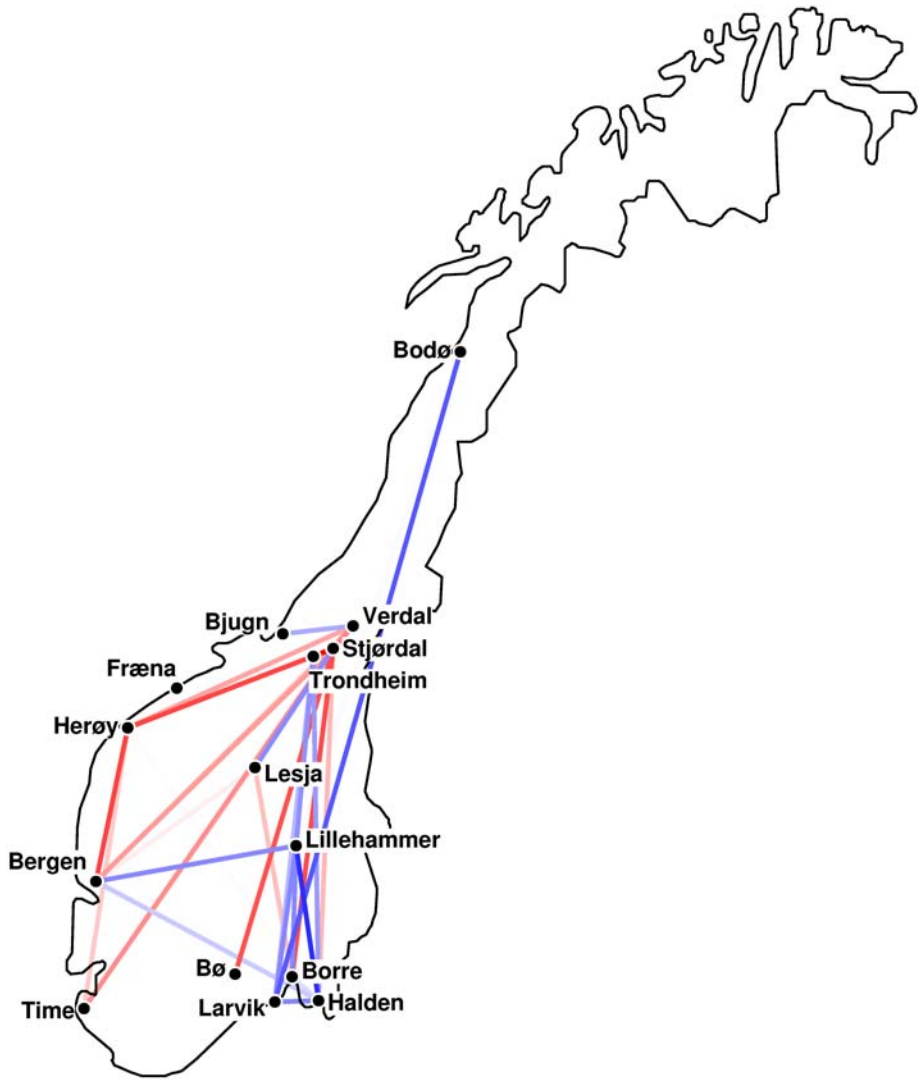
Some residuals in figures 4b and 5b involving smaller places are connected by red lines. Some of these places might have been more isolated than can be deduced by the traveling times, since traveling time does not say anything about the frequency of traveling to a place. It is also possible that the traveling times were in fact longer than the times which were used for the calculations because it was not taken into consideration that the waiting times were sometimes considerable (see 2.2.2). In the case of the Levenshtein distances, especially the dialect of Stjørdal gives rise to many residuals, see figure 4b. If Stjørdal is removed from the data, the correlation between Levenshtein distances and old travel distances is higher (r=.56 for the linear correlation and .64 for the logarithmic correlation) than the coefficients found in Table 1 (r=.52 and .53). This is not the case for correlations with perceptual data where exclusion of Stjørdal results in lower correlation coefficients. It seems that the Levenshtein method is not able to express that this dialect is more deviant than could be expected on the basis of the time it took to travel to Stjørdal 100 years ago. It is possible that some characteristic of this dialect is not captured by the Levenshtein distances or is not given the weight which the listeners in the perception experiment might have given it. One single characteristic of this dialect might have caused the listeners to perceive it as relatively more deviant.

So far the residuals were explained by deficiencies in the calculations of the traveling 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.

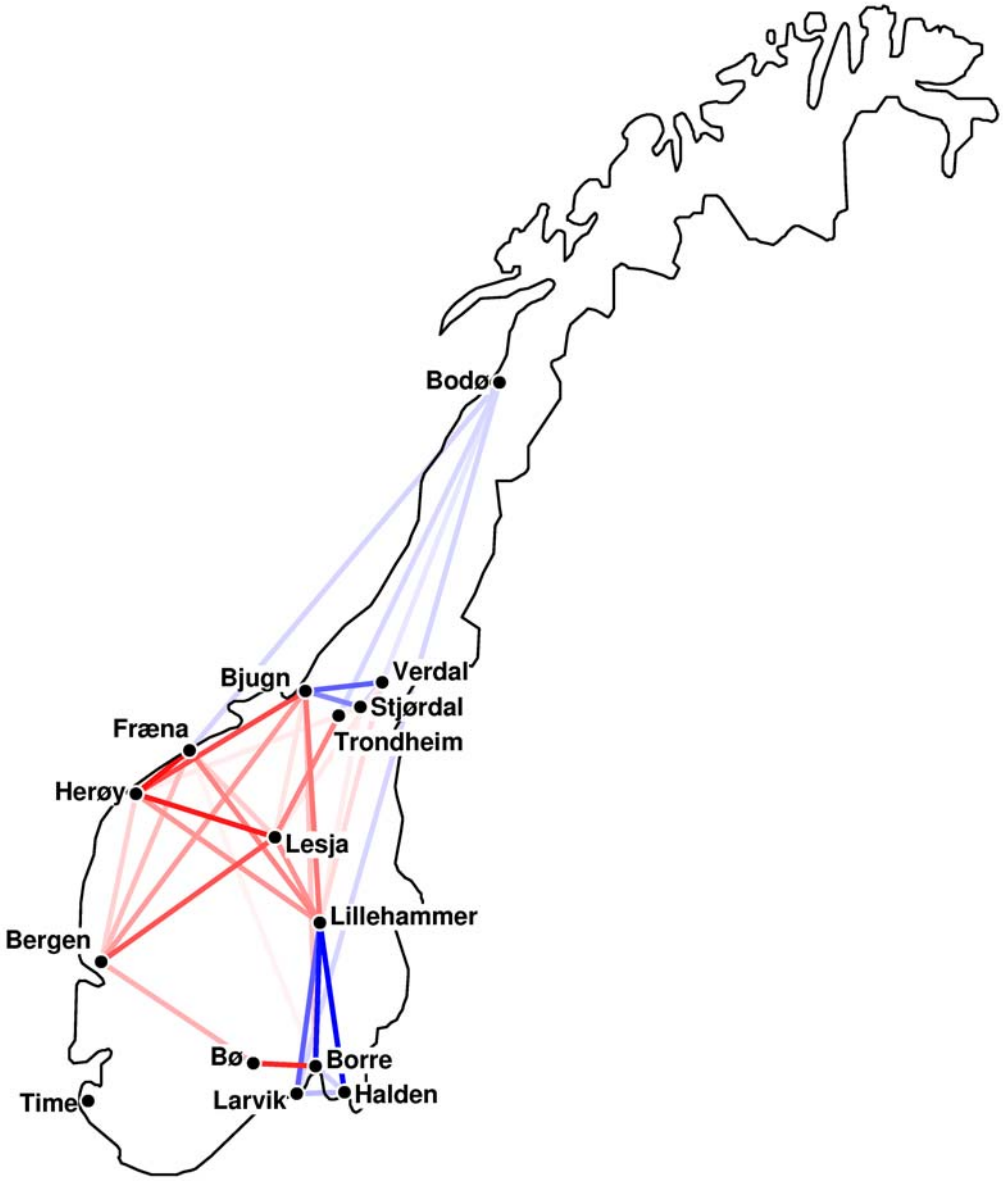
As mentioned in Section 2.1.3, the listeners were asked to judge the different dialects and the speakers on a number of attitudinal scales. These data might give some insight into the influence of attitude on the data. The

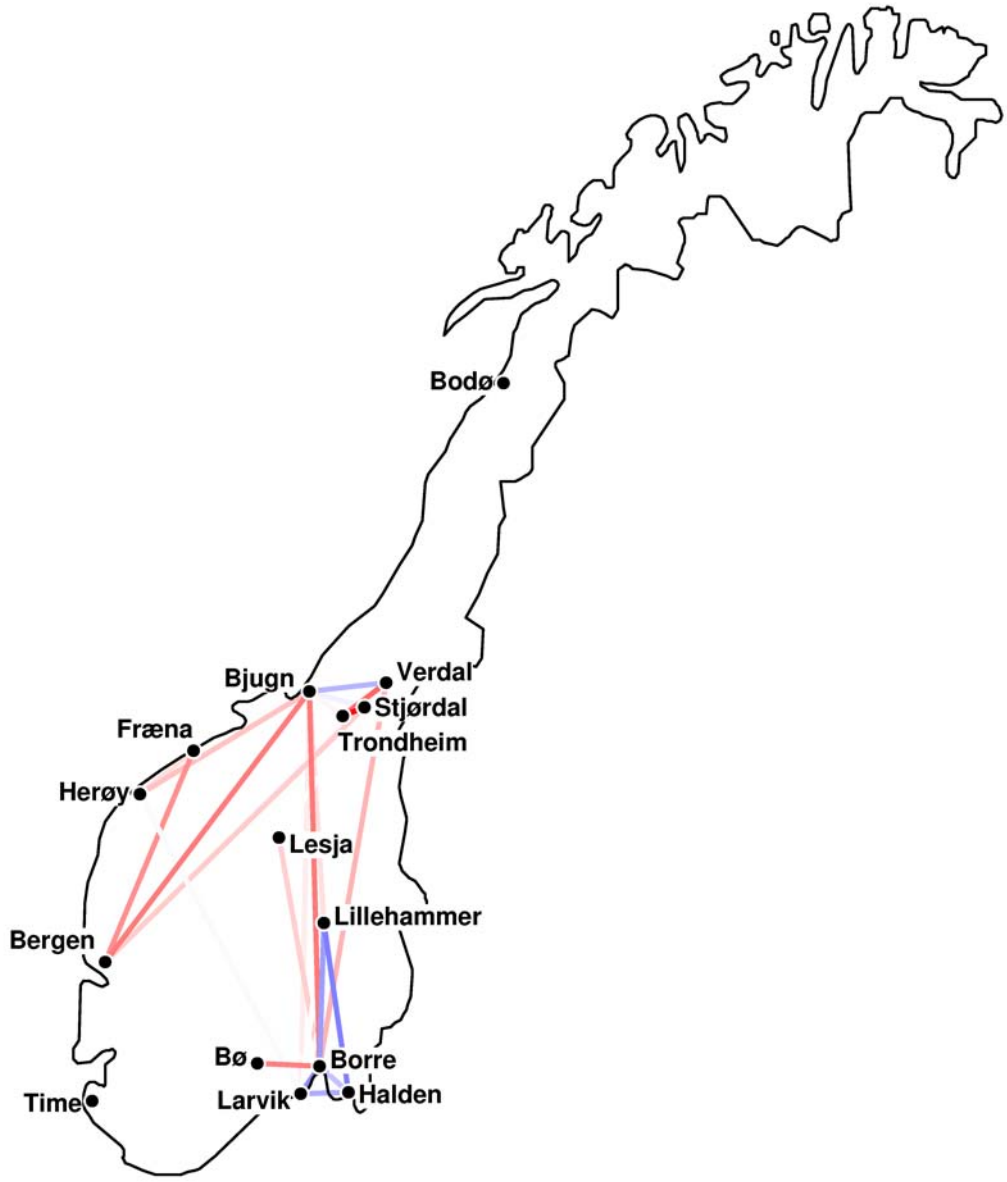
scales concerning the attitudes towards the speakers showed little correlation with the linguistic judgments. However, there turned out to be some degree of correlation between residuals and attitude towards the dialects. When linguistic distance is smaller than expected from old traveling times the attitude is positive and when the distance is larger than expected the attitude is negative. This goes especially for the perceptual distances (r = .48), but only to a limited degree for Levenshtein (r=.26). This means that listeners are inclined to perceive dialects as more deviant when they have a negative attitude. Or the other way round that the negative attitudes have caused the dialects to move apart.





Figs 4a and 4b. The residuals of the logarithmic regression lines for the straight-line distances versus Levenshtein distances (4a) and old traveling times versus Levenshtein distances (4b). Red lines indicate that linguistic distance is greater than geography would predict and blue lines that it is smaller. The intensity of the colour represents the extent of the deviation with respect to the regression value.


Bodø
Bjugn
Verdal
Stjørdal
Trondheim
Fræna
Herøy
Lesja
Lillehammer
Bergen
Bø
Borre
Time
Larvik
Halden



Figures 5a and 5b. The residuals of the logarithmic regression lines for the straight-line distances versus perceptual distances (5a) and the old traveling times versus perceptual distances (5b). Red lines indicate that linguistic distance is greater than geography would predict and blue lines that it is smaller. The intensity of the colour represents the extent of the deviation with respect to the regression value.

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 traveling 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 traveling distances showed that it is not possible to predict linguistic distances entirely on the basis of traveling times from 1900. Correlations might be improved by a more precise calculation of traveling times incorporating waiting time. But even if it was possible to calculate traveling time more precisely, information about the frequency of traveling 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 traveling times from the time before the railway system was constructed might be a better reflection of the present dialect distances.

In order to explain linguistic distances more precisely a number of other geographical and demographic factors should be taken into account. Urban centers 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 modeling 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.

Appendix (following page)**.** *Old travel times (top row) and modern travel times (bottom row) between each of the 15 Norwegian places in the investigation. Numbers of twentyfour hour intervals (only for old travel times), hours and minutes are separated by hyphens.*

Verdal	1-18-13 10-32	0-09-13 2-03	1-08-25 8-52	1-16-54 9-26	2-04-23 14-08	1-23-57 9-58	1-02-28 13-33	1-16-55 11-07	1-16-43 8-14	1-18-25 9-59	2-02-58 4-59	1-20-05 6-55	0-02-37 0-51	0-04-13 1-26
Trond-heim	1-14-00 9-20	0-05-00 2-01	1-04-12 10-06	1-12-41 10-10	2-01-10 14-33	1-19-44 11-03	0-22-15 3-32	1-12-42 8-22	1-12-30 7-17	1-14-12 11-07	1-22-45 3-02	1-15-52 4-53	0-01-36 0-32	
Stjørdal	1-15-36 9-46	0-06-36 2-32	1-05-48 9-43	1-14-17 10-47	2-02-46 13-19	1-21-20 11-46	0-23-51 4-18	1-14-18 8-49	1-14-06 7-50	1-15-48 11-49	2-00-21 3-32	1-17-28 5-16		
Lille-hammer	2-08-37 7-08	1-20-52 6-03	2-20-04 15-43	0-10-51 3-23	1-23-47 9-14	0-17-54 4-14	1-38-07 4-58	0-10-52 4-42	1-52-22 6-33	0-12-22 4-01	1-06-25 3-24			
Lesja	2-04-30 6-45	2-02-00 4-28	3-01-12 13-07	1-03-14 6-34	2-15-40 11-17	0-34-17 8-04	1-07-15 2-30	1-03-15 6-18	1-21-30 4-29	1-04-45 7-24				
Larvik	1-20-15 7-15	1-19-12 10-01	2-18-24 21-30	0-09-11 0-42	0-35-25 5-34	0-16-14 1-13	2-12-27 8-55	0-09-12 2-59	3-02-42 9-40					
Herøy	1-07-00 6-25	1-17-30 7-59	2-16-42 17-35	1-49-11 9-24	1-18-10 11-16	3-08-14 9-51	0-22-45 3-42	3-01-12 10-19						
Halden	2-05-27 9-33	1-17-42 9-34	2-16-54 19-58	0-07-41 2-00	0-44-57 8-26	0-14-44 3-31	2-10-57 8-49							
Fræna	1-05-45 8-13	1-03-15 4-29	2-02-27 13-37	2-10-56 9-09	1-16-55 10-25	2-17-59 10-37								
Bø	2-12-29 6-09	2-00-44 10-25	2-01-56 21-29	0-14-43 1-53	2-03-39 4-28									
Bryne	0-11-10 5-47	2-06-10 14-09	3-05-22 25-00	0-44-36 6-11										
Borre	1-21-46 7-34	1-17-41 9-34	2-16-53 20-29											
Bodø	2-18-12 20-57	1-09-12 10-38												
Bjugn	1-19-00 10-42													
	Bergen	**Bjugn**	**Bodø**	**Borre**	**Bryne**	**Bø**	**Fræna**	**Halden**	**Herøy**	**Larvik**	**Lesja**	**Lille-hammer**	**Stjør-dal**	**Trond-heim**

Literature

Bennett, Thomas. 1867. *Bennett's handbook for Norway for 1867*. Christiania: Bennett's.

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

Bjørnland, Dag. 1989. *Vegen og samfunnet: en oversiktlig fremstilling og analyse i anledning Vegdirektoratets 125-årsjubileum 1864–1989*. Oslo: Vegdirektoratet.

Bjørnland, Dag, and Erik Hajum. 1979. *Jernbanen i samfunnets tjeneste: jernbanens utvikling og betydning frem til 1914*. Oslo: Transportøkonomisk Institutt.

Chambers, John Kenneth and Peter Trudgill. 1989. *Dialectology*. Second edition. Cambridge: Cambridge University Press.

Christiansen, Hallfrid. 1954. "Hovedinndelingen av norske dialekter". *Maal og Minne*: 30–41.

Fintoft, K. and Mjaavatn, P.-E. 1980. "Tonelagskurver som målmerke". *Maal og Minne*: 66–87.

Gooskens, Charlotte. 2005. "How well can Norwegians identify their dialects?" *Nordic Journal of Linguistics* 28 (1): 37–60.

Gooskens, Charlotte, and Wilbert Heeringa. 2004. "Perceptive Evaluation of Levenshtein Dialect Distance Measurements Using Norwegian Dialect Data". *Language Variation and Change* 16 (3): 189–207.

Heeringa, Wilbert. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Dissertation, Groningen: University of Groningen.

Nerbonne, John, Wilbert Heeringa, Eric van den Hout, Peter van de Kooi, Simone Otten and Willem van de Vis. 1996. "Phonetic distance between Dutch Dialects". In: Gert Durieux, Walter Daelemans and Steven Gillis (eds.), *Proceedings of Computer Linguistics in the Netherlands '95*. Antwerpen: Centre for Dutch Language and Speech (UIA): 185–202.

Nerbonne, John and Wilbert Heeringa. 2001. "Computational Comparison and Classification of Dialects". *Dialectologia et geolinguistica. Journal of the International Society for Dialectology and Geolinguistics* 9: 69–83.

Omdal, Helge. 1995. "Attitudes toward spoken and written Norwegian". *International journal of the sociology of language* 115: 85–106.

Sandøy, Helge. 1993. *Talemål*. Oslo: Novus Forlag.

The International Phonetic Association. 1949. *The principles of the International Phonetic Association: being a description of the International Phonetic Alphabet and the manner of using it, illustrated by texts in 51 languages*. London: International Phonetic Association.

The International Phonetic Association. 1999. *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

Van Gemert, Ilse 2002. *Het geografisch verklaren van dialectafstanden met een GIS*. Unpublished MA-thesis, University of Groningen.

Charlotte Gooskens
University of Groningen, the Netherlands
Department of Scanditavian Studies
c.s.gooskens@rug.nl