Dialectologia 9 *(2012), 27-51.*
ISSN: 2013-2247

Received 2 May 2011.
Accepted 15 June 2011.

NON-LINGUISTS' JUDGMENTS OF LINGUISTIC DISTANCES BETWEEN DIALECTS

Charlotte GOOSKENS
University of Groningen
c.s.gooskens@rug.nl

Abstract

This study explores the relative contribution of geographic and objective linguistic distances to the perceived and estimated linguistic distances between Norwegian dialects as judged by non-linguists. The perceived linguistic distances were quantified by playing recordings of fifteen Norwegian dialects to groups of subjects from the same fifteen places and having them judge the linguistic distance of each dialect to their own dialect. The estimated linguistic distances were collected by asking the subjects to judge the distances on the basis of the place names only. Geographic distances were quantified as straight line distances and as traveling times from the year 1900. The objective linguistic distances were computed by means of the Levenshtein algorithm. The results show that non-linguists' preconceived ideas about linguistic distances are based mainly on geographic information while both linguistic and geographic information play a role when they judge distances on the basis of dialect samples.

Key words
perceptual dialectology, linguistic distances, Norwegian dialects

CONSIDERACIONES NO LINGÜÍSTICAS SOBRE LAS DISTANCIAS LINGÜÍSTICAS ENTRE DIALECTOS

Resumen

Este estudio explora la contribución relativa de las distancias geográficas objetivas en la percepción y estimación que tienen no lingüistas sobre las distancias lingüísticas que existen entre los

dialectos noruegos. Las distancias lingüísticas percibidas fueron cuantificadas a través de la escucha de grabaciones de quince dialectos noruegos que hicieron grupos de sujetos que pertenecían a los mismos lugares y que emitían juicios sobre la distancia lingüística que existía entre cada dialecto y su propia variedad. Las distancias lingüísticas estimadas se han obtenido pidiendo a los sujetos que valoraran las distancias únicamente a partir de los nombres de las localidades. Las distancias geográficas se cuantificaron como una línea recta y aplicando los tiempos de viaje desde el año 1900. Las distancias lingüísticas objetivas se calcularon mediante el algoritmo de Levenshtein. Los resultados muestran que las ideas preconcebidas de los no lingüistas sobre las distancias lingüísticas se basan principalmente en la información geográfica mientras que las informaciones lingüísticas y geográficas tienen un papel relevante cuando las distancias se valoran a partir de muestras dialectales.

Palabras clave
dialectologia perceptiva, distancias lingüísticas, dialectos noruegos

1. Introduction

Traditional dialectology is almost solely based on production data gathered by means of dialect surveys and on linguists' view of the geographic distribution of dialect areas and borders. More recently dialectometrical methods have made it possible to measure objective distances between dialects. The present investigation can be situated in the field of perceptual dialectology (Preston 1989; 1999; Long & Preston 2002) which can be seen as a complement of traditional dialectology and dialectometry. It is concerned with 'the ordinary speaker's perception of language variation' (Preston 1989: 2). Many studies in the area of perceptual dialectology have been concerned with the construction of dialect maps. Non-linguists have for example been asked to draw dialect borders on geographic maps. The present study is concerned with non-linguists' judgments of distances between dialects. Previous research has provided evidence that non-linguists' perception of language variation may be different from that of linguists due to the fact that factors other than linguistic differences play a role in their mental representation of dialect variation. An example of such a factor is geographic distances. The aim of the present study is to investigate to which extent non-linguists base their judgments of linguistic distances between dialects on objective linguistic distances and to which extent on geographic distances.

In previous investigations on distances between dialects as judged by non-linguists, a distinction can be made between two kinds of judged distances, namely perceived linguistic distances and estimated linguistic distances. *Perceived linguistic distances* are gained by playing recordings of dialects to subjects and having them judge the distances to some other variety, for example their own variety or the standard variety of the language. In order to collect *estimated linguistic distances*, subjects are asked to judge the distances without auditory input but purely on the basis of geographic place names.

Subjects may base their perceived and estimated judgments on *objective linguistic distances* as well as on *geographic distances*, but the two factors can be expected to play different roles for the two kinds of judgments. When subjects base their judgments on auditory information they can only use geographic information if they recognize the dialect and have an idea about the geographic distance to the dialect. On the other hand, the estimated linguistic distances can only be based on linguistic information if the subjects know how the dialect sounds.

	Dependent variables		**Explaining factors**	
	Estimated linguistic distances	**Perceived linguistic distances**	**Geographic distances**	**Objective linguistic distances**
Kuiper (1999)	X		X[1]	
Van Hout & Münsterman (1981)		X	X[1]	
Van Bezooijen & Heeringa (2006)	X		X[1]	X
Present study	X	X	X[1,2]	X

[1]Straight line distances

[2]Traveling times (see Section 2.2.4)

Table 1. Overview of dependent and independent variables in previous investigations and in the present study.

In Table 1, a schematic overview is given of three previous investigations of judged distances between dialects compared to the present study. None of the previous studies included both dependent variables (perceived and estimated linguistic distances)

and only one investigation included both explaining factors (geographic and objective linguistic distances).

Kuiper (1999) asked 76 arbitrarily chosen Parisian men and women of all ages and socio-economic classes to rate the degree of difference between their own speech and French as spoken in 24 regions in France, Francophone Belgium and Switzerland on a four-point scale (1 for speech exactly like the respondent's, 4 for incomprehensible speech). These estimated linguistic distances correlate significantly with straight line distances.[1]

In a study by Van Hout & Münsterman (1981), subjects were asked to rate the linguistic distance between nine dialects from different areas in the Netherlands and Standard Dutch on a 7-point scale, 0 indicating that the dialect was not deviant from Standard Dutch and 7 that is was very deviant. The listeners based their judgments on recordings of the dialects rather than on preconceived ideas as was the case in Kuiper (1999). The geographic and linguistic distances yielded the same order of the nine dialects which shows that there is also a relationship between perceived linguistic distances and geographic distances.

The two studies mentioned above did not assess the objective linguistic distances and therefore we do not know to which degree the subjects' judgments may also have a linguistic basis. Van Bezooijen & Heeringa (2006) measured three distances between Standard Dutch and dialects spoken in the twelve provinces of the Netherlands and the five Dutch-speaking provinces of Belgium: objective linguistic distances based on an old language sample, objective linguistic distances based on a new language sample, and geographic distances. They correlated these objective distances with estimates of linguistic distances by subjects from all Dutch provinces on a scale from 0 (no linguistic distance) to 100 (largest linguistic distance to Standard Dutch). They found high correlations with both objective linguistic distances and geographic distances. A multiple regression analysis showed that a combination of the two factors had no effect

[1] Kuiper did not look for explaining factors himself in the original study. However, I measured the straight line distances between Paris and the centers of the regions (excluding Belgium and Switzerland) and correlated these geographic distances with the mean difference rates per region as listed by Kuiper. The correlation was .66. Especially the two regions in the north-east, Lorraine and Alsace, were judged to be linguistically more deviant than expected from the geographic distances. Without these two regions the correlation is .79.

on the total percentage explained variance.[2] The authors assume that the subjects based their estimates of linguistic distance largely on geographic factors. However, since geographic distance shows high correlations with objective linguistic distance, the subjects' estimates could also be based on their knowledge of dialectal differences.[3]

From previous investigations it can thus be concluded that there seems to be a clear relationship between geographic distances and linguistic distances as judged by non-linguists, both on the basis of auditory input and on the basis of preconceived ideas. However, it is still uncertain to what extent non-linguists actually take objective linguistic distances into account when judging linguistic distances. If objective linguistic distances are involved at all in the judgments, they are likely to play a more important role for perceived linguistic distances than for estimated linguistic distances. On the other hand, it is reasonable to expect estimated linguistic distances to be based on geographic distances to a higher degree than perceived linguistic distances. However, to make such conclusions both kinds of judgments have to be compared within the same investigation.

1.1 The present investigation

In the present investigation the basis of non-linguists' judged distances between fifteen Norwegian dialects is explored. In contrast with previous investigations, both perceived and estimated linguistic distances are judged by the same subjects (see Table 1). This makes it possible to assess to which degree distances based on dialect samples presented auditorily correspond to distances judged on the basis of knowledge and preconceived ideas. Furthermore, it is possible to compare the relative importance of the two independent factors (geographic and objective linguistic distance), for the two dependent factors (perceived and estimated linguistic distances).

By investigating a geographically complex area like Norway with its many fjords and mountains, it may be possible to draw stronger conclusions about the relative

[2] The linguistic distances based on the old language sample were a better predictor of the estimated linguistic distances than the new sample. When Frisian, which is in fact a separate language, was included in the analysis, objective linguistic distance was a better predictor of estimated linguistic distance (r = .93) than geographic distance (r = . 87). When Frisian was excluded, geographic distance was a better predictor (r = .98) than objective linguistic distance (r = .91).

[3] No correlation coefficient is mentioned in the paper, but high correlations have been found in previous investigations (see Heeringa & Nerbonne 2001).

importance of geographic and objective linguistic distances than Van Bezooijen & Heeringa (2006) did for the Dutch language area. Heeringa & Nerbonne (2001) calculated linguistic distances between 350 Dutch dialects by means of the Levenshtein algorithm (see Section 2.2.3). The linguistic distances showed a high correlation with geographic distances (r = .66) which means that a large part of the linguistic variation can be accounted for by geography (r^2 = .45). However, a similar investigation (Gooskens & Heeringa, 2004) showed the correlation between linguistic distance and geographic distance to be considerably lower in the case of 52 Norwegian dialects (r = .22). In contrast with Dutch dialects, the two factors can therefore be expected to show low covariation in the case of Norwegian dialects. This makes it possible to separate the role of the two independent factors for the perceived and estimated linguistic distances and make stronger conclusions about the relative contribution of objective linguistic distances and geographic distances than in Van Bezooijen & Heeringa (2006).

The geographic distances will be measured in two ways. First, they will be measured by means of straight lines ('as the crow flies'). Next, geographic distances will be expressed by means of old traveling times since these have proven to be a more successful predictor of linguistic distance between Norwegian dialects (Gooskens, 2005a).

In contrast with many European countries, the position of the dialects in Norway is strong. Officially there is no standard variety in Norway. In fact the Norwegian Parliament decided in 1878 that no spoken standard should be taught in elementary and secondary schools. The variety spoken in and around the capital seems to some extent to be perceived as and have some functions of a spoken standard, but it does not have a very strong position as compared to spoken standards in many other 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) and people are often exposed to dialects spoken in different parts of Norway, for example via the media. Accordingly, both the exposure to and familiarity with dialects is exceptionally high among Norwegians. As a result, estimated linguistic distances are more likely to be based at least partly on linguistic characteristics than in the Netherlands where dialects are only used locally in unofficial contexts.

The research questions can be formulated as follows:

1. Do Norwegians judge linguistic distances differently when they hear the dialects (perceived linguistic distances) than when they have no auditory input (estimated linguistic distances)?

2. What is the relative contribution of geographic and linguistic distances to perceived linguistic distances?

3. What is the relative contribution of geographic and linguistic distances to estimated linguistic distances?

2. Method

There are four kinds of distance measures involved in the present investigation. Three of these distances are linguistic and one is geographic. The perceived linguistic distances and the objective linguistic distances are based on the same material from fifteen Norwegian dialects. First this material is described (Section 2.1) and next it is explained how the linguistic and geographic distances were calculated (Section 2.2).

2.1 Material

As mentioned in the introduction, Norwegian dialects are widely used by all age groups in different contexts. This makes it possible to use recent recordings of young people from all over the country with a minimal 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. Furthermore, it is possible for Norwegian people to read aloud a text in their own dialect. This was necessary since the same text in different dialects is needed for the calculation of the objective linguistic distances (see Section 2.2.3). Figure 1 shows the places where the fifteen dialects in the investigation are spoken. These fifteen dialects represent a large part of the Norwegian language area.

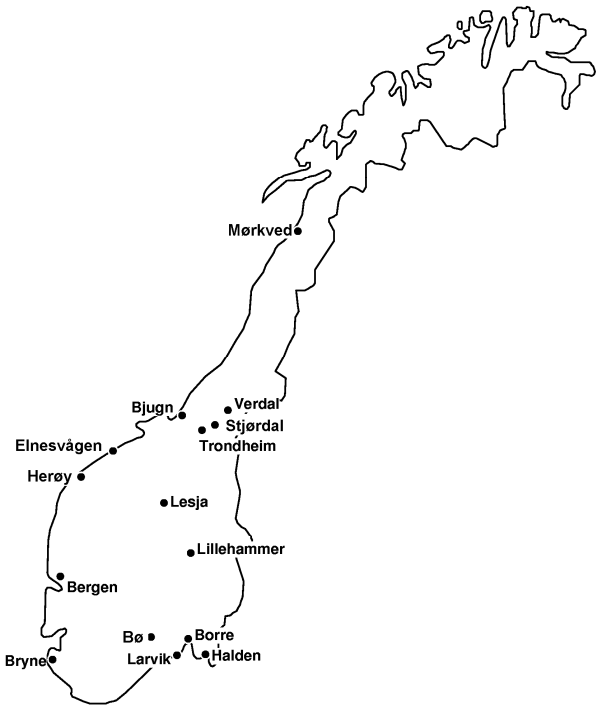


Figure 1. Map of Norway showing the geographic distribution of the fifteen Norwegian dialects used in the present investigation.

The speakers all read aloud the same text, namely the fable 'The North Wind and the Sun'.[4] The recordings of the whole texts were used for the listening experiments which resulted in the perceived linguistic distance measurements (see Section 2.2.1). The text consists of 58 different words which were used to calculate the objective linguistic distances.

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 the speakers attended university or already had a university degree.

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.

[4] 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.

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

On the basis of the recordings, phonetic transcriptions were made of all fifteen dialects. These transcriptions were used to calculate the objective linguistic distances (see Section 2.2.3). The transcriptions were made in IPA as well as in X-SAMPA (eXtended Speech Assessment Methods Phonetic Alphabet).[5] 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.

2.2 Measuring linguistic and geographic distances

Four kinds of distances were measured: perceived and estimated linguistic distances (the dependent variables) and objective distances and geographic distances (the independent variables). In the following four sections, the measurements will be explained.

2.2.1 Perceived linguistic distances

Fifteen groups of high school pupils, one group from each of the places where the fifteen dialects are spoken (see Figure 1), participated in the investigation (in total 285 pupils). Each group consisted of 16 to 27 subjects. Their mean age was 17.8 years, ranging between 17 and 20 years. 52 percent were female and 48 percent male. Only responses of subjects who had lived the major part of their lives in the place where the

[5] See http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm.

dialect is spoken were used for the analysis. On average these subjects had lived in the place in question for 16.7 years. Nine of the 290 subjects (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 subjects (83%) had one or two parents who also spoke the dialect.

In order to assess perceived linguistic distances the subjects listened to the complete fable about 'The North Wind and the Sun' in all fifteen dialects. While listening to the dialects the subjects were asked to judge each dialect on a scale from 1 (similar to own dialect) to 10 (not similar to own dialect). They were not told which dialects they heard. They were given a practice recording first (of a speaker of Stjørdal, but not one of the 15 recordings used in the experiment itself). In this way the listeners could get used to the task.

For each pair of dialects the mean perceived linguistic distance was calculated. Each group of listeners judged the linguistic distances between their own dialect and each of the fifteen dialects, including their own dialect. Accordingly, there are two mean distances between each pair of dialects, from dialect A to dialect B and from dialect B to dialect A. For example the distance which the listeners from Bergen perceived between their own dialect and the dialect of Trondheim (mean judgment is 7.8) is different from the distance as perceived by the listeners from Trondheim (mean judgment is 8.6). Different explanations can be given for the fact that different groups perceive the same linguistic distances differently. For example, it is possible that the attitude towards a dialect influence the perception of the linguistic distance (Van Bezooijen, 1994). In this way a matrix was achieved with 15 by 15 distances. In the case of the geographic and the objective linguistic distances, there is only one distance per dialect pair (see Sections 2.2.3 and 2.2.4) and accordingly only half of the matrix was filled for these distances. Therefore, in order to be able to correlate the perceived linguistic distances with the objective distances and the geographic distances the matrix with the perceived linguistic distances was made symmetrical by averaging corresponding cells above and below the diagonal, i.e. the cell contents of contra-diagonal cells *i, j* and *j, i* were averaged. The diagonal (for example the distance between Bergen and Bergen) was excluded just like in the case of the objective and geographic distances where the distances are always zero.

In addition to the judgment task, it was tested whether the subjects could identify the dialects correctly by having them place a cross on a map of Norway in the province where they thought that the dialect was spoken. In this way we got an indication of whether the listeners recognized the approximate place where the dialect is spoken. The purpose of this identification task will be explained in Section 3.

2.2.2 Estimated linguistic distances

The same fifteen groups of subjects as described in Section 2.2.1 were given a randomized list of place names, one from each province (*fylke*) in Norway. The names of the places were identical to the names of the places where the fifteen recordings of the 'North Wind and the Sun' were made. In addition, six places were added in order for all provinces to be represented. Four provinces were represented twice because two dialects were spoken in the same province. The names of the provinces were placed between brackets after the place names since it was possible that the subjects did not know all place names on the list.

The subjects were asked to estimate the linguistic distance from their own dialect to the dialect spoken in each of the places on a scale from 1 (similar to own dialect) to 10 (not similar to own dialect). They judged the estimated linguistic distances after the perceived linguistic distances (Section 2.2.1). This means that they were familiar with the idea of judging linguistic distances, when asked to estimate the distances on the basis of the place names only. They did not know that they were asked to estimate the distances to the dialects that they had already heard in the first part of the experiment.

For each pair of dialects the mean estimated linguistic distance was calculated in the same way as for the perceived linguistic distances (see Section 2.2.1).

2.2.3 Objective linguistic distances

A linguistic distance measurement was obtained by means of the Levenshtein distance measurements. The same method was used by Van Bezooijen & Heeringa (2006) in their investigation (see Section 1.1). With this method, it is possible to measure the linguistic distance between language varieties on the basis of phonetic transcriptions in an objective manner. Using the Levenshtein distance, the distance

between two dialects is measured 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.

For illustration, let us consider a simplified example of the calculation of the difference between two words. Assume *gåande* or *gående* 'going' is pronounced as [goːɑns] in the dialect of Bø and as [gɔːnə] in the dialect of Lillehammer. Changing one pronunciation into the other can be done as in Table 2 (ignoring suprasegmentals and diacritics for the moment). In fact, many sequence operations map [goːɑns] to [gɔːnə]. The power of the Levenshtein algorithm is that it always finds the cost of the cheapest mapping. Comparing pronunciations in this way, the distance between longer words will generally be greater than the distance between shorter words. The longer the words, the greater the chance for differences with respect to the corresponding word in another dialect. Because this does not accord with the idea that words are linguistic units, the sum of the operations is divided by the length of the longest alignment which gives the minimum cost. The longest alignment has the greatest number of matches. The alignment of our example is shown in Table 2. The total cost of 4 (1+1+1+1) is now divided by the length of 6. This gives a word distance of 0.67 or 67%.

Alignments	**1**	**2**	**3**	**4**	**5**	**6**
Bø	g	oː	ɑ	n		s
Lillehammer	g	ɔː		n	ə	
Costs		1	1		1	1

Table 2. Alignment which gives the minimal cost

The simplest versions of this method are based on a notion of phonetic distance in which phonetic overlap is binary: non-identical phones contribute to phonetic distance, identical ones do not. Thus the pair [i, ɒ] counts as different to the same degree as [i, ɪ]. In more sensitive versions, phones are compared on the basis of their feature values, so the pair [i, ɒ] counts as more different than [i, ɪ]. 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. A spectrogram is the visual representation of the acoustical signal and the visual differences between the spectrograms are reflections of the acoustical differences. When using spectrograms it is not necessary to make decisions about the weight of the different features. The spectrograms were made on the basis of recordings of the sounds of the International Phonetic Alphabet as pronounced by John Wells and Jill House on the cassette The Sounds of the International Phonetic Alphabet from 1995.[6] The different sounds were isolated from the recordings and monotonized at the mean pitch of each of the two speakers with the program PRAAT.[7] Next, with PRAAT a spectrogram was made for each sound using the so-called Barkfilter which is a perceptually oriented model. Segment distances were calculated by using the Barkfilter distances as operation weights. In this way the fact that for example [i] and [e] are phonetically closer to each other than [i] and [a] is taken into account. Gradual weights for insertions and deletions are obtained by measuring distances between the IPA sounds and silence. Differences in length are formalised as insertions or deletions (indels), for example [a] versus [aː] is represented as a versus aa, which results in two indels. More information about the Levenshtein distances on the basis of spectrograms can be found in Heeringa (2004: 79-119).

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. 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. Gooskens, 2005b; Hognestad, 1999; Skjekkeland, 1997; Sandøy, 1991; Fintoft & Mjaavatn 1980). However, since the transcriptions gave no information about the precise realization of the tonemes or intonation, we were not able to include this linguistic level in the analysis. On the other hand, morphology and lexicon are included

[6] See http://www.phon.ucl.ac.uk/home/wells/cassette.htm.
[7] The program PRAAT is a free public-domain program developed by Paul Boersma and David Weenink at the Institute of Phonetic Sciences of the University of Amsterdam and available at http://www.fon.hum.uva.nl/praat.

in the distance measurements since words from a running text with different morphological and lexical forms are compared. For further details about the Levenshtein distances see Nerbonne & Heeringa (2001) and Heeringa (2004).

In order to calculate the distance between two dialects, a large number of Levenshtein distances are determined – one distance per word. Next, 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 & Heeringa (2004).

2.2.4 Geographic distances

In previous investigations, the geographic distances between dialect data in the Netherlands and Norway were calculated using straight line distances. Heeringa & Nerbonne (2001) measured linguistic and geographic distances between 350 Dutch dialects and found a correlation of .66 between these two distances. The correlation was considerably lower in the case of Norwegian data (r = .22, Gooskens & Heeringa 2004). 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. In Gooskens (2005a) geographic distances were measured on the basis of information about traveling times in the year 1900 by road, train, and boat between the places where the different dialects are spoken. In addition to old traveling times, new (year 2000) traveling times were calculated. The results show that in the case of a geographically

complicated country like Norway, old traveling times reflect the influence of geography on linguistic variation better than modern traveling times and straight-line distances. Modern traveling times and straight line distances correlate highly (.98) and for this reason only the old traveling times and the straight line distances will be included in the present analysis.

The old traveling times were calculated on the basis of time schedules for the steamboat along the cost and for the train in the year 1900. Furthermore, there was an extensive system of conveyance by horse which was regulated by law. This system included permanent posting stations at the main roads, see Bjørnland (1977 and 1989) and Bjørnland & Hajum (1979). 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 was possible to get a reliable picture of traveling times in the year 1900 on all routes connecting the fifteen places in our investigation. More details and the results of the calculations can be found in Gooskens (2005a).

3. Results

In Section 3.1, the results showing the role of linguistic and geographic distances for the perceived linguistic distances will be presented. The subjects had linguistic information on which they could base their judgments. They were not told which dialects they heard, but in those cases where listeners could identify the dialects, they could base their judgments on geographic information as well. We can get an impression of the relative contribution of geographic and linguistic distances to the perceived linguistic distances by correlating the perceived linguistic distances with the geographic and objective linguistic distances and by performing a multiple regression analysis.

We furthermore made a separate analysis of the judgments by listeners who identified the dialects correctly (see Section 2.2.1), since for these judgments we can be sure that the subjects had linguistic as well as geographic information to base their

judgments on.[8] This makes it possible to draw stronger conclusions about the relative contribution of geographic and objective linguistic distances for the perceived linguistic distances than could be done in previous research.

It is also informative to analyze the judgments by listeners who were not able to identify the dialects correctly. Geographic information cannot have played an important role for these judgments since the subjects did not know where the dialects were spoken. This will give an impression of how well non-linguists are actually able to judge objective linguistic distances on a purely linguistic basis without information about the geographic distances.

In Section 3.2, corresponding analyses will be presented with estimated linguistic distances as the dependent variable. The correlation between perceived and estimated linguistic distances is rather high (r = .75), but still there are differences between the two measures. The relative contribution of geographic and objective linguistic factors to the perceived and estimated linguistic distances is therefore likely to be different. When the subjects judged the perceived linguistic distances they had linguistic input on which they could base their judgments while this was not the case when they judged the estimated linguistic distances. Since the subjects only had information about the place names and provinces when estimating the linguistic distances, they had to base their judgments purely on their knowledge and intuitions about the dialects.

We also made a selection of the estimated linguistic distances made by listeners who were not able to identify the dialects correctly when listening to them. These estimated linguistic judgments cannot have been based on the correct linguistic characteristics, since the subjects obviously did not know what the dialects sounded like. By isolating these results we get a better idea of how well the estimated linguistic judgments correspond with the geographic distances.

Finally we also analyzed the estimated linguistic distances with correct identifications. For these judgments we know for sure that the subjects knew the linguistic characteristics of the dialects.

[8] A dialect was considered correctly identified if the cross was placed in the correct province. In total 4350 identifications were made by the 285 subjects, and 28.0% of these identifications were correct.

3.1 Correlations with perceived linguistic distances

In Table 3, the correlations of perceived linguistic distances with geographic and objective linguistic distances are shown. Also the results of a multiple regression analysis with objective linguistic distances and old traveling times are shown. The logarithmic correlations are presented since these are higher than the linear correlations in all cases, probably due to the fact that in perception, small differences in pronunciation may play a relatively strong role in comparison to larger differences. The effect of using logarithmic distances is that small distances are weighed relatively more heavily than large distances.

As explained in Section 2.2.1, the subjects were asked to place a cross on a map of Norway in the province where they thought that the dialect was spoken. The correlations with the judgments by subjects who identified the dialects correctly are presented in the middle column and the judgments by subjects who identified the dialects incorrectly are presented in the right column.

	Perceived linguistic distances		
	all identifications	**correct identifications**	**wrong identifications**
	r	***r***	***r***
Objective linguistic distances	.76	.59	.82
Geographic distances			
straight line	.71	.68	.68
old traveling times	.85	.79	.81
Objective linguistic distances and old traveling times (regression analysis)	.90	.80	.90

All correlations are significant at the .01 level.

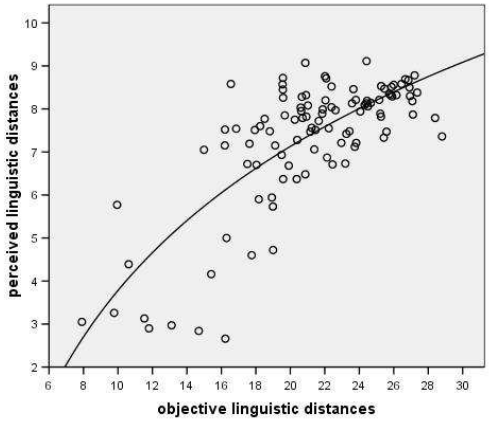
Table 3. The logarithmic correlations of perceived linguistic distances with objective and geographic linguistic distances between fifteen Norwegian dialects and the results of a regression analysis including objective linguistic distances and old traveling times (bottom row).

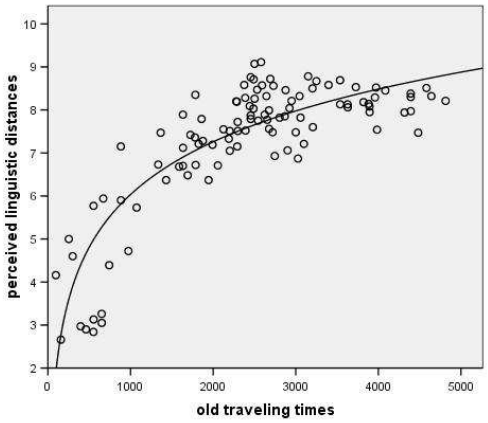
When we look at the correlations with the perceived linguistic distances in Table 3, we see that the geographic distances expressed as old traveling times correlate strongly with perceived linguistic distances (r = .85). The correlation is higher than with the straight line distances (r = .71). This means that the old traveling times are to a higher extent reflected in the mental maps of the listeners than the straight line distances even though the traveling circumstances have changed dramatically over the last century. We will therefore exclude the straight line distances from further analysis.

The old traveling times correlate stronger than the objective linguistic distances with the perceived linguistic distances (.85 versus .76). This shows that the subjects base their judgments on geographic distances to a larger extent than on objective linguistic distances. However, the results of a multiple regression analysis show that a better prediction of the perceived linguistic distances is obtained when the two determinants are combined (r = .90, p = .000 for both determinants). This means that the subjects base their judgments of linguistic distances on both geographic and objective linguistic information.

The correlations with objective linguistic distances and old traveling times are visualized in the scatterplots in Figures 2a and 2b. By comparing these two figures it again becomes clear that the old traveling times are a better predictor of the perceived linguistic distances than the objective linguistic distances. A closer look at the residuals in Figure 2a showed that the subjects tended to underestimate the linguistic distance to the dialects spoken close to the place where they lived, probably due to the fact that they often know these dialects well and therefore perceive them as less deviant than they in fact are. On the other hand they often overestimated the linguistic distance to dialects spoken further away. This tendency is confirmed by a significant correlation between the deviance of residuals from the regression line in Figure 2a and geographic distances (r = .48 with straight line distances and .54 with old traveling times, $p < .01$).

Dialectologia 9 *(2012), 27-51.*
ISSN: 2013-2247





Figures 2a and 2b. Scatterplots showing perceived linguistic distances versus objective linguistic distances (r = .76, p = .000) and perceived linguistic distances versus old traveling times (r = .85, p = .000).

If we only look at the judgments by the subjects who identified the dialects correctly, the correlation with objective linguistic distance becomes lower (r = .59). It looks as if the geographic knowledge has distracted the subjects from basing their judgments on the linguistic characteristics of the dialects. The difference between the correlations with old traveling times and objective linguistic distances is even larger than when all subjects are involved (r = .76 versus .85 for all identifications and .59 versus .79 for correct identifications only). A linear regression analysis shows that the objective linguistic distances result in a slightly better prediction (r = .80) than traveling time alone, but the contribution of the objective linguistic distances are only significant at the .05 level (p = .029), while the contribution of the old traveling times is significant (p = .000).

When looking at the selection of judgments by listeners who were not able to identify the dialects and thus can be assumed to have based their judgments mainly on the linguistic characteristics, the correlation with objective linguistic distances gets higher (r = .82) and lower with old traveling times (r = .81) than when all judgments are included (r = .76 and .85). This shows that the listeners are to a high degree able to judge linguistic distances on the basis of objective linguistic distances only. In a regression analysis, however, both distances contribute significantly (p = .00) and the predictive value is .90.

3.2 Correlations with estimated linguistic distances

In Table 4, the logarithmic correlations of estimated linguistic distances with geographic and objective linguistic distances are shown as well as the results of a multiple regression analysis with objective linguistic distances and old traveling times. Like for the perceived linguistic distances, the correlation with the old traveling times is high (r = .78 when including all data) and lower when correlated with the straight line distances (r = .72) and again we will therefore exclude the straight line distances from further analysis.

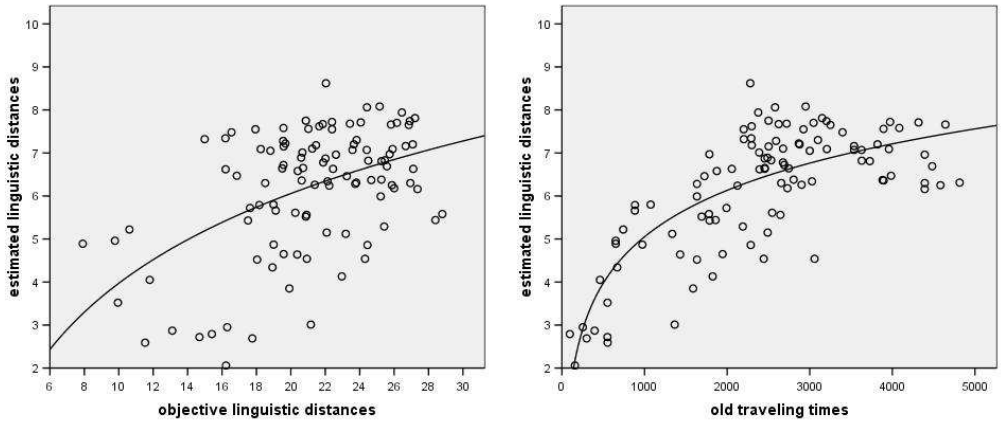
The correlation with the objective linguistic distances is lower than was the case for the perceived linguistic distances (r = .50 versus .76) and also much lower than with the old traveling times (r = .78). The subjects clearly base most of their judgments on the geographic distance between the dialects. A multiple regression analysis shows that we do not obtain a better prediction of the estimated linguistic distances when the objective linguistic distances and the old traveling times are combined (r = .78). Therefore the objective linguistic distances are excluded by the procedure (p = .917). Only the old traveling times are included (p = .000).

	Estimated linguistic distances		
	all identifications	**correct identifications**	**wrong identifications**
	r	***r***	***r***
Objective linguistic distances	.50	.49	.48
Geographic distances			
straight line	.72	.60	.76
old traveling times	.78	.68	.77
Objective linguistic distances and old traveling times (regression analysis)	.78	.69	.77

All correlations are significant at the .01 level.

Table 4. The logarithmic correlations of estimated linguistic distances with objective and geographic linguistic distances between fifteen Norwegian dialects and the results of a regression analysis including objective linguistic distances and old traveling times (bottom row).

In Figures 3a and 3b, the scatterplots are shown between the estimated linguistic distances and the objective linguistic distances (3a) and the old traveling times (3b). We see a larger dispersion than in the case of the perceived linguistic distances, but again it becomes clear that the old traveling times are a better predictor than objective linguistic distances. The residuals from the correlation between estimated linguistic distances and objective distances (Figure 3a) showed the same trend as for the perceived linguistic distances. Dialects spoken geographically close to the dialects of the subjects are underestimated and dialects spoken further away are overestimated (r = .56 for straight line distances and .51 for old traveling times). As far as the residuals from the old traveling times are concerned (Figure 3b) no clear trend could be found.



Figures 3a and 3b. Scatterplots showing estimated linguistic distances versus objective linguistic distances (r = .50, p = .000) and estimated linguistic distances versus old traveling times (r = .78, p = .000).

The conclusion that the subjects base their judgments almost solely on geographic distances is confirmed by the results of the judgments of the subjects who did not identify the dialects correctly in the right column of Table 4. As explained above, we expect these judgments to have been based on geographic distances only. When the subjects do not hear the dialects and do not know how the dialects sound, they are obviously not able to involve objective linguistic distances in their judgments. It hardly makes a difference whether all estimated linguistic distances are included (left column) or whether the analysis is based on the distances with wrong identifications only (right column). The subjects base their judgments on the geographic distances to the same

extent (r = .78 and .77). The correlation with estimated linguistic distances may mostly be explained by covariance between geographic and objective linguistic distances. This is confirmed by a multiple regression analysis which shows that a combination of old traveling times and objective linguistic distances do not result in a better prediction of the estimated linguistic distances (r = .77). Only the old traveling times are included by the procedure (p = .000) while the objective linguistic distances are excluded (p = .887).

Also when we analyze the results of the judgments by subjects who identified the dialects correctly we see no improvement (r = .69). These subjects seem to know what the dialects sound like but still do not use this knowledge of the dialects when they do not hear them. Again only the old traveling times are included by the procedure (p = .000) while the objective linguistic distances are excluded (p = .259).

4. Conclusions and discussion

In the present investigation, the role of geographic and objective linguistic distances for the perceived and estimated linguistic distances has for the first time been tested with the same group of non-linguists. This provided the opportunity to investigate the basis of non-linguists' preconceived and perceptual ideas of dialectal variation and compare the role of two explaining factors, geography and linguistic distances. The results show that perceived and estimated distances only correlate to a certain extent. This makes clear that listeners form their ideas of the linguistic distances in different ways when they hear the dialects than when they have no auditory input.

The estimated linguistic distances in the present investigation are mainly based on geographic information. This result confirms the expectation by Van Bezooijen & Heeringa (2006) that their subjects had based their estimates of linguistic distances largely on geographical factors. An advantage of an investigation with Norwegian dialects is that the correlation between objective linguistic distances and geographic distances is rather low. Accordingly, covariation is low and this allows us to separate the role of the two factors for the judged distances. In the investigation by Van Bezooijen & Heeringa (2006) it was harder to draw strong conclusions because the correlation between objective linguistic distances and geographic distances was high

and both objective linguistic distance and geographic distance correlated highly with estimated linguistic distances.

Both the results of the estimated linguistic distances in the present investigation and in the investigation by Van Bezooijen & Heeringa raise the question whether non-linguists are at all able to judge linguistic distances on the basis of objective linguistic distances. The results of the correlations with perceived linguistic distances showed that objective linguistic distances play an important role for the judgments of linguistic distances when the subjects hear recordings of the dialects on which they can base their judgments. However, when the listeners know where the dialect is spoken, they base their judgments almost exclusively on geographic information. When the subjects did not know where the dialects were spoken and thus had to base their perceived judgments on linguistic information, the correlation between perceived linguistic distances and objective linguistic distances was higher.

This means that non-linguists are indeed well capable of using linguistic information when judging linguistic distances, but only when auditory dialect samples are presented as a basis for the judgments. When no auditory samples are presented, listeners base their judgments mainly on geographic distances. Even though Norwegian listeners have more experience with dialectal variation than listeners from most European countries, they are apparently still not well capable of using this knowledge. When investigating non-linguists' ideas of language variation in future investigations it is therefore important to consider whether dialect samples should be played to the listeners or not.

When comparing the results of the estimated and the perceived distances it should be kept in mind that the two distance measures are not completely comparable. When estimating the linguistic distances, the subjects were told the place and the province of the dialect but still we cannot be sure that they did indeed know the exact location of the place. We also do not know exactly which geographic information the subjects had when judging the perceived distances. Even when analyzing the correct identifications only, we only know that they recognized the correct province. For the sake of comparability it may have been an advantage to inform the subjects which dialects they heard on the tape. Then we would have been sure that they had the same geographic information both when judging the perceived and the estimated distances. A

disadvantage of this would have been that we would not know whether subjects would actually be able to judge linguistic distances on the basis of linguistic information only.

It should also be kept in mind that Norwegian subjects may not be representative for non-linguists in general. As explained in the introduction, the language consciousness of Norwegians may be higher than that of subjects from other language areas due to the strong position of Norwegian dialects. It is possible that Norwegians are better at identifying the dialects and judging differences. However, the fact that they almost only use geographic information when estimating the distances shows that their language consciousness is limited. Norwegians only take linguistic distance as their point of reference for judging dialectal difference if they have no other clues and linguistic distance plays only a minor role for the perception of distance between dialects if they have clues about geographic distance.

A further difference with previous investigations is that our subjects judged the distances to their own dialect while in most investigations they are asked to judge the distance to the standard language. It is uncertain which effect this has had on the results, but in the light of the strong position of the dialects in Norway (see Section 1.2) it can be expected to be easier for Norwegian subjects to use their own dialect as a point of reference than the standard language when judging deviance.

References

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 & Erik HAJUM (1979) *Jernbanen i samfunnets tjeneste: jernbanens utvikling og betydning frem til 1914*, Oslo: Transportøkonomisk Institutt.

FINTOFT, Knut & Per Egil MJAAVATN (1980) "Tonelagskurver som målmerke", *Maal og minne*, 66-87.

GOOSKENS, Charlotte (2005a) "Travel time as a predictor of linguistic distance", *Dialectologia et Geolinguistica*, 13, 38-62.

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

GOOSKENS, Charlotte & 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*. Doctoral dissertation, University of Groningen.

HEERINGA, Wilbert & John NERBONNE (2001) "Dialect Areas and Dialect Continua", *Language Variation and Change* 13. 375-400.

HOGNESTAD, Jan K (1999) "Nye toner? Språklig tonalitet i talemålsundervisningen", *Norsklæreren*, 1, 49-53.

KUIPER, Lawrence (1999) "Variation and the norm. Parisian perceptions of regional French", in Dennis R. PRESTON (ed.), *Handbook of Perceptual Dialectology*, Vol. 1, Amsterdam/Philadelphia: John Benjamins, 243-262.

LONG, Daniel & Dennis R. PRESTON (eds.) (2002) *Handbook of Perceptual Dialectology*, Vol. 2, Amsterdam: Benjamins.

NERBONNE, John & Wilbert HEERINGA (2001) "Computational comparison and classification of dialects", *Dialectologia et Geolinguistica*, 9, 69-83.

OMDAL, Helge (1995) "Attitudes toward spoken and written Norwegian", *International Journal of the Sociology of Language*, 115, 85-106.

PRESTON, Dennis (1989) *Perceptual Dialectology: Nonlinguists' Views of Areal Linguistics*, Dordrecht: Foris.

PRESTON, Dennis (ed.) (1999) *Handbook of Perceptual Dialectology*, Vol. 1, Amsterdam: Benjamins.

SANDØY, Helge (1991) *Norsk dialektkunnskap*, Oslo: Novus Forlag.

SKJEKKELAND, Martin (1997) *Dei norske dialektane. Tradisjonelle særdrag i jamføring med skriftmåla*, Kristiansand: Høyskoleforlaget.

VAN BEZOOIJEN, Renée (1994) "Aesthetic evaluation of Dutch language varieties", *Language and Communication*, 14, 253-263.

VAN BEZOOIJEN, Renée & Wilbert HEERINGA (2006) "Intuitions on linguistic distance: geographically or linguistically based?", in Tom KOOLE, Jacomine NORTIER & Bert TAHITU (eds.), *Artikelen van de Vijfde Sociolinguïstische Conferentie*, Eburon, Delft, 77-87.

VAN HOUT, Roeland & Henk MÜNSTERMAN (1981) "Linguistische afstand, dialect en attitude [Linguistic distance, dialect and attitude]", *Gramma*, 5, 101-123.