What role does dialect knowledge play in the perception of linguistic distances?

Wilbert Heeringa[1]*, Charlotte Gooskens*[2]*, Koenraad De Smedt*[3]

[1]Meertens Institute, the Netherlands
[2]University of Groningen, the Netherlands
[3]University of Bergen, Norway

Abstract

The present paper reports on an investigation to find an answer to the question to what extent subjects base their judgments of linguistic distances on actual dialect data presented in a listening experiment and to what extent they involve previous knowledge of the dialects when making their judgments. The point of departure for our investigation were distances between 15 Norwegian dialects as perceived by Norwegian listeners. We correlated these perceptual distances with objective phonetic distances measured on the basis of the transcriptions of the recordings used in the perception experiment. In addition, we correlated the perceptual distances with objective distances based on other datasets. On the basis of the correlation results and multiple regression analyses we conclude that the listeners did not base their judgments solely on information that they heard during the experiments but also on their general knowledge of the dialects. This conclusion is confirmed by the fact that the effect is stronger for the group of listeners who recognized the dialects than for listeners who did not recognize the dialects on the tape.

1. Introduction

To what extent do subjects base their judgment of linguistic distances between dialects on what they really hear, i.e. on the linguistic phenomena available in the speech signal, and to what degree do they generalise from the knowledge that they have from previous confrontations with the dialects? This is the central question of the investigation described in this paper. The answer to this question is important to scholars who want to understand how dialect speakers perceive dialect pronunciation differences and may give more insight in the mechanisms behind the way in which linguistic variation is experienced. Our study is of interest to (computational) dialectologists and sociolinguists.

In the spring of 2000, an experiment was performed among Norwegian dialect speakers in order to measure *perceptual linguistic distances* between 15 Norwegian dialects.[1] In each of 15 locations, speakers listened to dialect recordings of the fable ‘The North Wind and the Sun’ in the 15 dialects and were asked to judge the linguistic distance between these dialects and their own dialect. Henceforth we refer to the recordings as the NOS data or just NOS.[2] The experiment is described in Gooskens and Heeringa (2004). The geographical distribution of the 15 locations is shown in Figure 1. The 15 dialects were classified on the basis of the mean judged distances between the dialects. The classification largely agrees with that of traditional dialect maps (e.g. Skjekkeland 1997).

The perceived distances were correlated with *objective linguistic distances*, measured by means of the Levensthein algorithm, with which distances between different pronunciations of the same concept can be measured. The correlation between perceptual and objective distances appeared to be 0.67 which is a significant, but not perfect correlation. There may be different reasons for this. Firstly, the listeners may have been influenced by *non-linguistic* factors such as familiarity and attitude towards the dialects. They may tend to judge familiar dialects as less deviant from their own dialects or they may judge dialects that they seldom hear as more deviant than could be expected from their objective linguistic distance. A negative attitude towards a dialect may cause the listener to judge a dialect as more deviant from their own dialect than expected from the objective linguistic distances and a positive attitude may have the opposite effect.

Secondly, low correlation may be the result of the fact that the objective distances were based on lexical, phonetic and morphological variation only[3], since prosodic and syntactic variation was not measured, while listeners based their judgments on information at all linguistic levels.

In this paper we want to focus on a possible third explanation. When making their judgments listeners may take into account linguistic information that is based on previous contact with the dialect even when that information does not necessarily relate directly to the recordings. When hearing some characteristic linguistic phenomena (shibboleths), they may be able to identify the dialect and judge the distance on the basis of what they *know* about the dialect rather than on what they *hear* when listening to the recording. We want to test the hypothesis that subjects in a perception experiment do not only use the information they hear, but also use *extra linguistic* information not available in the speech signal, thus making the perceptual distances more representative for the dialects than the objective measurements based on the recordings. We focus on the pronunciation level only, i.e. we restrict our analysis to phonetic and phonological variation.[4] Gooskens and Heeringa (2006) have shown that pronunciation is a more important predictor of perceived linguistic distance between the 15 Norwegian dialects than lexical distances.

In order to test our hypothesis, we decided to test the degree to which the perceived distances correlate with objective distances based on other data sets from the same dialects as in the NOS data. We expect that other objective distances correlate significantly with the perceptual distances, despite the fact that only the NOS data is the same as the data used for the perception experiment. Furthermore, if listeners do indeed make judgments based on information that is not present in the recording, we expect that distances measured on other data sets will have a significant contribution in a multiple regression model.

In Gooskens (2005) the listeners in the perception experiment were asked to identify the 15 dialects by putting a cross on a map of Norway to indicate where they thought the speakers came from. When excluding the identifications of the dialects spoken by the listeners themselves, Gooskens found that 25 per cent of the crosses were placed in the correct county. In addition to the perceptual measurements based on all data, we also calculated a) distances as perceived by listeners who identified the dialects correctly; and b) distances as perceived by listeners who did not identify the dialects correctly. We hypothesize that listeners who are able to identify the location of a dialect correctly are likely to use their linguistic knowledge about this dialect when judging the linguistic distance. In this case we would therefore expect a lower correlation between perceived distances and objective linguistic distances, since the judgments are not based solely on the linguistic information present in the recordings. On the other hand, the judgments by listeners who were not able to identify a dialect correctly can be expected to show a higher correlation with the objective linguistic distances since the listeners are more likely to base their judgments solely on the recordings.[5]

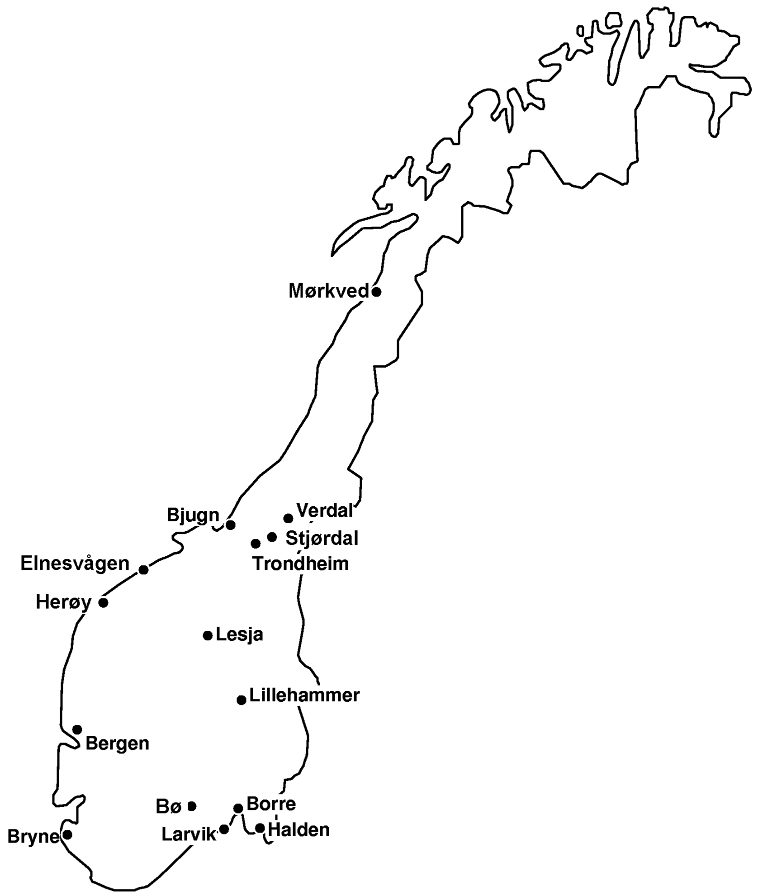


Figure 1. *Map of Norway showing the 15 dialects used in the present investigation.*

We start by describing the four different datasets, the NOS, ALE, ALENOR and NOR data. (Section 2). In Section 3 we describe the perception experiment that was set up in order to measure perceptual distances on the basis of the NOS-text and in Section 4 we present Levenshtein distance, which we use to measure objective phonetic distances. In Section 5 phonetic distances that are obtained on the basis of the NOS, ALE, ALENOR and NOR data are correlated with each other and with the perceptual distance measurements. In Section 6, the results of multiple regression analyses are presented. The main conclusions will be presented in Section 7.

2. Data sets

We used three datasets in addition to the NOS data, namely the ALE, NOR and ALENOR data. The first one contains material from the *Atlas Linguarum Europae* (ALE) transcribed phonetically according to the ALE guidelines. The second data set contains transcriptions of the *Norwegian dialect atlas* (NOR). The pronunciations in this atlas are transcribed according to the phonetic *Norvegia* transcription system. The third source is also based on the ALE data, but the pronunciations are transcribed according to the *Norvegia* transcription system (ALENOR). By comparing ALE to ALENOR, we are able to measure the influence of different transcription systems on the distance measurements.

In order to be able to carry out objective linguistic distance measurements, we needed digital versions of the four datasets. The NOS data were already available in digital form based on the IPA system and X-SAMPA codes.[6] We converted the phonetic transcriptions of NOR, ALE and

ALENOR from the Norvegia and ALE transcription systems into X-SAMPA. Converting the ALE transcriptions to X-SAMPA did not pose a problem, since the phonetic alphabet used in ALE is based on IPA and there is a one-to-one correspondence between IPA symbols and X-SAMPA symbols. The NOR and the ALENOR data are transcribed by means of Norvegia, which is a very detailed phonetic system with more than 100 vowel and consonant symbols. In order to convert the NOR data, we used the information in Nes (1982) where all Norvegia symbols are compared to IPA symbols. Since the NOR transcriptions were very detailed, many symbols representing the diacritics had to be used in the X-SAMPA. In most cases this did not cause any problems. Exceptions are the affricates that we transcribed as two extra short sounds and mediopalatal s, k and g that we transcribed as palatalised s, k and g since there are no symbols for these sounds in X-SAMPA. Norvegia does not make a distinction between tap and alveolar r and between [v] and [ʊ]. We have transcribed these combined sounds with the symbols [r] (X-SAMPA [r]) and [ʊ] (X-SAMPA [P]) respectively.

Most Norwegian dialects distinguish between two tonal patterns on the word level, often referred to as tonemes. Some dialects even have a third toneme, the circumflex. The tonemes are treated differently in the three datasets. In NOS, NOR and ALENOR, the tonemes are marked with [″] for toneme 1, [″″] for toneme 2 and [%%] for the circumflex toneme.[7] In ALE main stress and circumflex toneme are indicated along with the X-SAMPA symbols for rising [_R] and falling [_F] tone respectively. The realization of the tonemes varies considerable across the Norwegian dialects. However, no information was given about the precise realization of the tonemes in the transcriptions.

Ideally, the datasets should be a random sample of the vocabulary in a dialect since this may be expected to be the best representation of perceived distances. In the case of the NOS-text, we expect the choice of a running text to be a sensible approach to selecting a representative sample of a dialect. The three other datasets are word lists, from which we made random selections of 60 words in order to obtain sets of dialect samples that were comparable in size to the NOS dataset.

2.1 The North Wind and the Sun (NOS)

The fable ‘The North Wind and the Sun’ consists of 58 different words. The NOS-database at present contains recordings in more than 50 Norwegian dialects. We included the fifteen dialects which were available when we started our investigation in the spring of 2000 (see Figure 1).[8] The dialects are spread over a large part of the Norwegian language area, and cover most major dialect areas as found on the traditional map of Skjekkeland (1997:276). On this map the Norwegian language area is divided into nine dialect areas. In our set of 15 varieties, six areas are represented.

There were 4 male and 11 female speakers with an average age of 30.5 years, ranging between 22 and 35, except for one speaker who was 66. 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.

The speakers were all given the text in Norwegian beforehand and were allowed time to prepare the recordings in order to be able to read aloud the text in their own dialect. Many speakers had to change some words of the original text in order for the dialect to sound authentic. The word order was changed in three cases.

When reading the text aloud the speakers were asked to imagine that they were reading the text to someone with the same dialectal background as themselves. This was done in order to ensure a reading style which was as natural as possible and to achieve dialectal correctness.

On the basis of the recordings, phonetic transcriptions were made of all 15 dialects in IPA as well as in X-SAMPA. All transcriptions were made by the same trained phonetician, which ensures consistency.

2.2 Atlas Linguarum Europae with ALE transcription (ALE)

The Atlas Linguarum Europae (ALE) was an ambitious cooperation between linguists from 51 European countries. The initiative was taken in 1970 with support from UNESCO, resulting in the publication of maps and commentaries from 1975 onward. Lists of 547 words from different onomasiological categories were collected from 2631 places by means of a questionnaire filled out by hand (Kruijsen 1976). The words were transcribed phonetically with a rather broad phonetic system (Weijnen 1975). In Norway the material was collected from 152 places by linguists in the late 1970s.

For the present investigation it would have been preferable to use ALE material from exactly the same places as those in NOS. However, only two places are covered by both ALE and NOS, so in the other cases we chose material from the neighbouring village instead.[9] The 60 words that we used for the analysis were selected randomly from the ALE word lists. All 15 varieties were transcribed by the same phonetician.

2.3 Atlas Linguarum Europae with Norvegia transcription (ALENOR)

In addition to the phonetic transcriptions with ALE symbols of the words in the ALE corpus, most words were also transcribed in Norvegia (see Section 2). This gives us the opportunity to compare the effect of different phonetic transcription systems. We have therefore made a distinction between the dataset transcribed with ALE symbols (ALE) and a dataset with the same words transcribed with Norvegia (ALENOR). Unfortunately, no Norvegia transcription of the Lillehammer dialect was available. We reconstructed the Norvegia transcription ourselves on the basis of regular correspondences between ALE transcriptions and Norvegia transcriptions in the other 14 varieties.

2.4 Norwegian dialect atlas (NOR)

The Norwegian dialect atlas is based on a collection of 1500 words from most villages and parishes in Norway (see Hoff 1960). Most of the material was collected between 1951 and 1970, but it was supplemented with older material that was collected in the 1940s. The 1500 words are everyday words which cover all grammatical and phonetic details that can be expected to be found in the Norwegian dialects. Most of the words are transcribed phonetically in Norvegia by hand. Several field workers were involved in the collection of the material.

As with the ALE material, it was not always possible to use material from the same places as in the NOS material because the material was not complete. In seven cases, material from a neighbouring village was selected.[10]

The words in the NOR-list are divided into 56 sections according to their grammatical or phonetic characteristics. One word was chosen randomly from each section. In addition, a random word was selected from four random sections. In this way we obtained a random list of 60 phonetically transcribed words from 15 dialects to be used for our analysis.

2.5 Comparison of the four data sets

We have taken care to select four datasets that are as similar as possible in a number of respects. Almost the same number of randomly selected words (58 for NOS, 60 for NOR, ALE and ALENOR) were selected and transcribed phonetically in X-SAMPA in the 15 dialects. We aimed to select the same 15 dialects but in a number of cases we had to chose a neighbouring dialect because some dialects were missing in NOR and ALE/ALENOR. Furthermore, it would have been ideal to work with datasets from the same period. However, the NOS material is more recent than the ALE/ALENOR material and the NOR material is oldest. We have also seen that there is a difference in the phonetic details of the transcriptions. The ALE material is transcribed in a rather broad transcription while the NOR and ALENOR transcriptions are very detailed. Also the NOS transcriptions include more diacritics than ALE. Another difference between the datasets is the way in which the words have been selected. The NOS words come from a running text, the NOR words are a random selection of words covering different phonetic and grammatical categories and the ALE/ALENOR words are a random selection from a list of words representing various onomasiological categories.

3. Perceptual distance measurements

In order to investigate the linguistic distances between 15 Norwegian dialects as perceived by Norwegian listeners, recordings of the fable ‘The North Wind and the Sun’ in the 15 dialects were presented to Norwegian listeners in a listening experiment. The listeners were 15 groups of high school pupils (mean age 17.8 years), 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) who had lived the major part of their lives (on average 16.7 years) in the place where the dialect is spoken. A majority of the 290 listeners said that they spoke the dialect always (60 per cent) or often (21 per cent), the rest spoke it seldom (16 per cent) or never (3 per cent). A large majority of the listeners (83 per cent) had one or two parents who also spoke the dialect.

The 15 dialects were presented to the listeners in a random order preceded by a practice recording. While listening to the dialects, the listeners were asked to judge each dialect on a scale from 1 (similar to one's own dialect) to 10 (not similar to one's own dialect). In addition to the judgment scores, the listeners were presented with a map of Norway with all counties indicated. They were asked to place a cross in the county where they thought the dialect was spoken. This allowed us to make separate analyses of judgments by listeners who recognized the dialects and by listeners who did not recognize them.

Each group of listeners judged the linguistic distances between their own dialect and the 15 dialects, including their own dialect. In this way we obtain a matrix with 15 × 15 distances.[11] There are two mean distances between each pair of dialects, depending on the subject's own dialect. For example, the distance which the listeners from Bergen perceived between their own dialect and the dialect of Trondheim (mean 7.8) is different from the distance as perceived by the listeners from Trondheim (mean 8.6). Different explanations can be given for the fact that different groups perceive the same linguistic distances differently. For example, it is likely that the attitude towards or familiarity with a dialect influence the perception of the linguistic distance. Since in the case of the objective linguistic distance measurements there is only one value per dialect pair, the average of the two mean perceptual distances was calculated, e.g. the average of the distance Bergen-Trondheim and Trondheim-Bergen. This makes it possible to correlate the objective and the perceptual linguistic distances (Section 5).

4. Objective linguistic distance measurements

In this section we describe how pronunciation distances are calculated on the basis of phonetic transcriptions. Since we want to measure pronunciation distances, only cognates (i.e. historically related words) are compared to each other. These cognates may vary both phonetically, phonologically and morphologically. If a particular word corresponds to different lexemes across the 15 Norwegian dialects, we consider variants of the most frequent lexeme only. For example, ‘became’ is translated by *vart* in some dialects and by *blei* in a majority of the dialects. Since we only wanted one cognate lexeme per item, we fill in missing values for dialects that use *vart*.

We use the Levenshtein algorithm to compute the distances between pronunciations in each pair of dialects. This algorithm computes the cost of incrementally changing one pronunciation into the other by inserting, deleting or substituting sounds. In the simplest form of the algorithm, each operation has the same cost, e.g. 1. Assume *gåande* (or the variant *gående*) ‘going’ is pronounced as [^{2}goːɑns] in the dialect of Bø and as [^{2}gɔːnə] in the dialect of Lillehammer. Changing one pronunciation into the other can be performed incrementally as in Table 1 (ignoring suprasegmentals and diacritics for the moment):

Table 1 *Changing one pronunciation into another using a minimal set of operations.*

Bø	goːɑns	subst. o/ɔ	1
	gɔːɑns	delete ɑ	1
	gɔːns	insert ə	1
	gɔːnəs	delete s	1
Lillehammer	gɔːnə		
			4

It is easy to see that there can be different sequences of operations mapping [^{2}goːɑns] to [^{2}gɔːnə], but the power of the Levenshtein algorithm is that it always finds the cost of the cheapest mapping.

In order to achieve distances which are based on linguistically motivated alignments that respect the syllable structure of a word or the structure within a syllable, the algorithm was adapted so that a vowel may only be substituted by a vowel and a consonant only by a consonant. Exceptions are the semi-vowels [j] and [w] and their vowel counterparts [i] and [u], which may correspond to either vowels or consonants. The central vowel schwa [ə] may correspond to any sonorant. In this way, unlikely matches like [o] and [t] or [s] and [e] are prevented.

In our example, the phonetic symbols are aligned as shown in Table 2.

Table 2 *Alignment which gives the minimal cost.*

Bø	g	oː	ɑ	n		s
Lillehammer	g	ɔː		n	ə	
Costs		1	1		1	1

In a previous study, we divided the sum of the operation costs by the length of the alignment. This normalizes scores so that longer words do not count more heavily than shorter ones, reflecting the status of words as linguistic units. However, Heeringa, Kleiweg, Gooskens and Nerbonne (2006) showed that results based on raw Levenshtein distances approximate dialect differences as perceived by the dialect speakers better than results based on normalized Levenshtein distances. Therefore we do not normalize the Levenshtein distances in this paper.

The text 'The North Wind and the Sun' consists of 58 words in most dialects. The distance between two dialects is based on the aggregate distance over at most 58 word pairs. Since we have restricted our analysis to the comparison of cognates only, the number of word pairs per dialect pair will usually be smaller. Therefore the sum of the Levenshtein distances of the available word pairs for a dialect pair is divided by the number of the available word pairs, thus yielding the average Levenshtein distance of the dialect pair.

5. ALE, ALENOR, NOR and NOS compared to each other and to perception

5.1 Consistency

In order to test the consistency of the data sets, Cronbach's alpha was calculated as the average inter-correlation among the words in each data set. A widely accepted threshold in social science is 0.70. All alpha values were higher (see Table 3) and the number of words in the four data sets are therefore proven to be a sufficient basis for a reliable Levenshtein analysis (see Heeringa 2004:170-3).

Table 3. *Cronbach's alpha values on the basis of the four data sets.*

	Cronbach's alpha
ALE	0.89
ALENOR	0.91
NOR	0.89
NOS	0.87

Since our analysis is restricted to the comparison of cognates only, the number of word pairs per dialect pair may vary. Table 4 shows the minimum, average and maximum number of word pairs for each of the four data sources.

Table 4. *Minimum, average and maximum number of word pairs taken into account over all pairwise comparisons.*

	minimum	average	maximum
ALE	37	43.8	50
ALENOR	37	43.8	50
NOR	22	40.3	55
NOS	41	48.7	55

5.2 Correlations

Correlations between different distance measurements

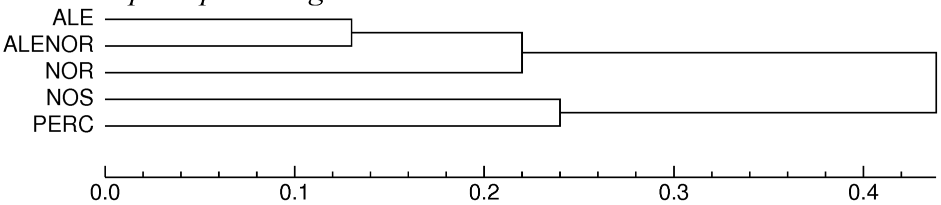
We first calculated the correlations between all objective linguistic distance measurements based on the four datasets and the corresponding perceptual distances. The results are shown in Table 5, which shows, for instance, that the correlation between the perceptual measurements (PERC) and the objective measurements on the basis of the NOS data (NOS) is 0.76. As explained in Section 3, the correlations are based on half matrices. Furthermore, we exclude the distances of dialects with respect to themselves, i.e. the distance of Bergen to Bergen, of Bjugn to Bjugn etc. In computational matrices these values are always 0, whereas in the perceptual matrix they vary, usually being higher than the minimum score. This may be due to the fact that the dialect of the speaker of Bergen, for instance, may differ slightly from that of the listeners from the same location. Since this causes uni-directional distortion of the diagonal distances (they can only be too high, not too low), we exclude them when calculating the correlation coefficient.

Table 5. *Correlations between the different objective linguistic measurements and the perceptual linguistic distances.*

	ALENOR	NOR	NOS	PERC
ALE	0.87	0.75	0.49	0.61
ALENOR		0.81	0.56	0.60
NOR			0.56	0.55
NOS				0.76

From the correlations, distances can be derived by calculating 1-*r*. For example, a correlation of *r*=0.75 can be transformed to a distance of 1-0.75=0.25. Using these distances, we classified the five data sources with cluster analysis (UPGMA). On the basis of this clustering, a dendrogram is constructed (Figure 2), which is a hierarchically structured tree where the varieties are the leaves. A dendrogram provides a good visualization of similarity between (groups of) data sets. The shorter the lines connecting the data sets, the more similar they are. The tree structure shows for example that ALE and ALENOR measurements correlate strongly, while ALE and NOS measurements do not correlate as strongly.

Figure 2. *Cluster analysis based on the correlations between the different objective linguistic measurements and the perceptual linguistic distances*



The highest correlation with the perceptual distances was found for the NOS data (r = .76)[12] and these two distance measurements form one cluster. This is to be expected since the perceptual distances and the NOS distances are based on the same data, viz. the recordings of the ‘North Wind and the Sun’. The objective linguistic distances based on the other data sets form another cluster. The correlation is highest between ALE and ALENOR (r = .87). This does not come as a surprise, since these two distance measurements are based on the same data, but it also shows that these two different transcription systems result in different distance measurements. The NOR distances correlate stronger with the ALE/ALENOR distances (r = .75 and .81) than with the NOS distances (r = .56), which may be explained by the fact that both the ALE/ALENOR and the NOR are based on word lists rather than on running text as in the case of NOS. The NOR in turn correlates more strongly with the ALENOR (r = .81) than with the ALE (r = .75), since the NOR and the ALENOR share the same transcription system. The fact that NOR, ALE and ALENOR correlate stronger internally than with NOS may also be an effect of the tendency of phonemic transcription in these datasets. The NOS dataset is amore detailed phonetic transcription.

Correlations between objective distance measurements and perceptual distance measurements with and without recognition

Subjects could base their judgment of linguistic distances on the purely linguistic characteristics exhibited by the recordings. They could also use additional information about the dialect that is not directly present in the recordings but which was acquired in previous confrontations with the dialect. However, a prerequisite for such a generalization is that listeners were able to recognize the dialect. In Table 6, we repeat the correlations between the perceived distances and the objective linguistic distances in the first column. In the second and third column the correlations are given for the judgments by listeners who recognized the dialects correctly and incorrectly, respectively. We see that the perceived distances including all judgments correlate highly with the judgments by listeners who did not recognize the dialects (r = .99).[13] The correlations with the objective linguistic distances are therefore very similar for these two sets of judgments.

The correlations between perceived distances and the NOS distance measurements are lower when the dialects are recognized correctly (r = .58) than when all judgments are included (r = .76). This is probably due to the fact that the listeners include information in their judgments that are not present in the NOS recordings. On the other hand, the judgments by listeners who did not recognize the dialects are slightly higher (r = .78). These listeners were not distracted by their knowledge about the dialects but based their judgments on what they heard. This interpretation seems to be confirmed by the fact that the correlation with distances as perceived by listeners who recognized the dialects correctly is slightly higher for the ALE data (r = .61) than for the NOS data (r = .58). There are apparently some dialect characteristics present in the ALE data that are not present in the NOS data, and the listeners take these characteristics into account when judging the distances.

Table 6. *Correlations between objective linguistic distances, perceived distances, and perceived distances with and without correct recognition of the dialects*

	PERC	PERC *with* *recognition*	PERC *without* *recognition*
ALE	0.61	0.61	0.57
ALENOR	0.60	0.59	0.57
NOR	0.55	0.56	0.53
NOS	0.76	0.58	0.78
PERC	-	0.84	0.99

6. Regression analyses

In Tables 7, 8 and 9 we present the results of multiple linear regression analyses (stepwise) with perceived distances as the dependent variable and the different objective distances as independent variables. A model including ALE distances in addition to NOS gives the highest correlation (r = . 81). For listeners recognizing the dialects correctly, ALE (and ALENOR) are better predictors of perceived distance than NOS (see Table 8). When the dialects are not recognized correctly, ALE, ALENOR and NOS add little to the model (Table 9), probably due to the fact that the listeners base their judgments solely on what they hear.

Table 7. *Multiple linear regression analysis (stepwise) with perceived distances as dependent variable and different objective linguistic distances as independent variables. All judgments are included.*

Input	Correlation	Significant variables
NOS, ALE, ALENOR, NOR	0.81	NOS, ALE
NOS, ALENOR, NOR	0.79	NOS, ALENOR
NOS, NOR	0.77	NOS, NOR
NOS	0.76	NOS

Table 8. *Multiple linear regression analysis (stepwise) with perceived distances as dependent variable and different objective linguistic distances as independent variable. The perceived distances are based on the judgments made by listeners who recognized the dialects correctly.*

Input	Correlation	Significant variables
NOS, ALE, ALENOR, NOR	0.68	ALE, NOS
NOS, ALENOR, NOR	0.65	ALENOR, NOS
NOS, NOR	0.64	NOS, NOR
NOS	0.58	NOS

Table 9. *Multiple linear regression analysis (stepwise) with perceived distances as dependent variable and different objective linguistic distances as independent variable. The perceived distances are based on the judgments made by listeners who recognized the dialects wrongly.*

Input	Correlation	Significant variables
NOS, ALE, ALENOR, NOR	0.81	NOS, ALE
NOS, ALENOR, NOR	0.79	NOS, ALENOR
NOS, NOR	0.78	NOS
NOS	0.78	NOS

The most important conclusions that can be drawn from the regression analyses are summarized in Table 10. We see that ALE adds most to the model in the case when the dialects are correctly recognized (a difference of 0.1), while ALE adds less to the model when all judgments are included or when only judgments with wrong recognitions are included.

Table 10. *Difference between regression models with NOS data and NOS data combined with ALE data.*

	PERC	PERC *with recognition*	PERC *without recognition*
NOS + ALE	0.81	0.68	0.81
NOS	0.76	0.58	0.78
increase	0.05	0.10	0.03

7. Conclusions

We have shown that it is plausible that listeners do not base their judgments about linguistic distances between dialects solely on what they hear. They seem to generalize their judgments by including information about phonetic dialect characteristics not present in the recordings. This conclusion is based on the fact that in a regression analysis, objective linguistic distances measured on the basis of another data set than the one used for the perception experiment add significantly to the model. This supports the conclusion that listeners take dialect characteristics into account that are not present in the recordings. As could be expected, this effect is even stronger for the group of listeners who recognized the dialects on the tape.

An alternative explanation of the impact of recognition on the judgments made may be that the listeners took geographical distances into consideration when making their judgments. In this case, one would expect geographical distances to show a higher correlation with the judgments with recognition than judgments without recognition. This was, however, not the case: the correlation was .53 in both situations.

To a certain degree, our results explain the rather low correlations that we found between perceived linguistic distance and objective linguistic distances in an earlier investigation. However, the best linguistic model in the present investigation results in a correlation of .81 (66 per cent explained variance) which means that 34 per cent of the variance still needs to be explained. It is likely that the other possible factors mentioned in the introduction (attitude, familiarity and other linguistic factors) also play an important role. In future work, we intend to include attitudes in our analysis. A multiple regression analysis with a combination of attitude scores and objective distance measurements will give us an impression of the relative contribution of these two factors to the perceived distances.

As shown in Section 2.5, we have taken care to select four datasets that are as similar as possible in a number of respects. We suspect that a better matching of the four datasets might have further improved the model. The 15 dialects were not exactly the same, the period of data collection differed and there is a difference in the phonetic details of the transcriptions. Finally, the words have been selected according to different criteria. All these differences between the data sets may have contributed to the rather high amount of variance that remains to be explained.

While we think the model could be further improved, we hope that the present work has already contributed to a better understanding of how people judge phonetic distances between their own dialect and other dialects. Judgements differ depending on whether subjects are able to correctly recognize and place a dialect. We interpret this so that people seem to let their previous knowledge of a dialect contribute to their judgement of how close or far it is from their own dialect.

Acknowledgements

The NOS recordings were made by Jørn Almberg in co-operation with Kristian Skarbø at the Department of Linguistics, University of Trondheim and made available at http://www.ling.hf.ntnu.no/nos/ along with the phonetic transcriptions (in IPA as well as in SAMPA). We are grateful for their permission to use the material. We thank the University of Oslo for access to the handwritten ALE forms and the handwritten NOR volumes (the latter is now available through the Digital Documentation Unit at http://www.edd.uio.no/). We would furthermore like to thank Benedikte Tobiassen for the digitisation of the ALE, ALENOR and NOR data, which are kept at the Department of Linguistic, Literary and Esthetic Studies at the University of Bergen,

References

C. Gooskens (2005). ‘How well can Norwegians identify their dialects?’, *Nordic Journal of Linguistics,* 28 (1), 37-60.

C. Gooskens and W. Heeringa (2004). ‘Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data’, *Language variation and Change,*16 (3), 189-207.

C. Gooskens and W. Heeringa (2006). ‘The Relative Contribution of Pronunciation, Lexical and Prosodic Differences to the Perceived Distances between Norwegian dialects’, *Literary and Linguistic Computing*, *special issue on Progress in Dialectometry: Toward Explanation*, 21 (4), 477-92.

W. Heeringa, P. Kleiweg, C. Gooskens and J. Nerbonne (2006). ‘Evaluation of String Distance Algorithms for Dialectology’, In: J. Nerbonne and E. Hinrichs (eds.) *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006*, 51-62.

W. Heeringa (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* Doctoral dissertation, University of Groningen.

I. Hoff (1960). ‘Norsk dialektatlas. Foredrag i Norsk Forening for Språkvitenskap 3.nov 1959’, *Norsk Tidsskrift for Sprogvidenskap*, 595-622

J. Kruijsen (1976). *Atlas Linguarum Europae (ALE) Premier questionnaire. Onomasiologie, vocabulaire fundamental* (Amsterdam).

O. Nes (1982). *Storms norske lydskriftsystem (med tillegg) definert ved hjelp av IPA’s lydskriftsystem*, Skriftserie for Institutt for fonetikk og lingvistikk, 8 serie B (University of Bergen).

M. Skjekkeland (1997). *Dei norske dialektane: tradisjonelle særdrag i jamføring med skriftmåla* (Kristiansand).

A. Weijnen (1975). *Atlas Linguarum Europae (ALE). Introduction* (Assen).

[1] The term 'dialect' is not used in the strict sense but rather as a variety that is characteristic for some region or place. Variation may be restricted to the phonetic, phonological and morphological level.

[2] In Norwegian 'The North Wind and the Sun' is translated as *Nordavinden og sola*, which we abbreviate to NOS.

[3] Gooskens and Heeringa (2004) also processed lexical differences since pronunciations of translationally corresponding words were compared to each other regardless of whether the words were cognates.

[4] The phonetic and the phonological levels are not distinguished in our analysis. Also morphological variation is included, since we base the measurements on whole words from a running text.

[5] Alternatively the profile of the dialect wrongly identified is activated instead of that of the correct accent. In that case the same goes as for the recognized dialect. However, the question may arise to what extent a listener will be able to activate the profile of a dialect variety wrongly identified, since identifying a wrong variety shows that the listener is unfamiliar with both the correct one and the wrongly identified one.

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

[7] Originally, the circumflex toneme was transcribed as [~} in the X-SAMPA transcriptions of NOS.

[8] At the time the perception experiment was carried out, recordings of only 15 varieties were available. Today more than 50 recordings are available, giving much better possibilities to pick a representative selection of varieties.

[9] We replaced Mørkved with Sørfold, Bjugn with Stjørna, Trondheim with Horg, Stjørdal with Leksvik, Verdal with Skogn, Herøy with Vanylven, Elnesvågen with Bud, Lillehammer with Gausdal, Bergen with Lindås, Larvik with Hedrum, Borre with Nøtterøy, Bryne with Klepp, and Halden with Idd. In some cases there were clear differences between the dialects of the NOS-places and the dialects of the replacements.

[10] We replaced Bergen with Loddefjord, Mørkved with Sørfold, Elnesvågen with Bud, Halden with Enningdal, Larvik with Tjølling, Lillehammer with Follebu, and Stjørdal with Leksvik.

[11] The matrix can be found in Gooskens and Heeringa (2004).

[12] Gooskens and Heeringa (2006) found a lower correlation of 0.68 due to the fact they did not average the perceptual distances, e.g. A-B and B-A (cf. Section 3) but copied *objective* distances (e.g. A-B was copied to B-A). If they had calculated the correlation with the two full matrices in the same way as we did (excluding the diagonal A-A, B-B etc.), the correlation would have been equal to 0.73.

[13] The correlation with the judgments by listeners who recognized the dialects is lower (r = .84), probably due to the fact that the proportion of judgements in this category is much smaller (27.5 per cent) than that of the judgments without recognition (72.5 per cent).