Intuitions on linguistic distance: geographically or linguistically based?*

Renée van Bezooijen, Dept. of Linguistics, Radboud University Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, r.v.bezooijen@let.ru.nl
Wilbert Heeringa, Humanities Computing, University of Groningen, P.O. Box 716, 9700 AS Groningen, w.j.heeringa@rug.nl

1. Introduction

The present study explores the basis of non-linguists' intuitions about the linguistic distances among language varieties in the Netherlands and Flanders, focussing on geographic and linguistic determinants. The idea that 'ordinary', non-linguistically trained people have some notion of the regional variation present in the language area where they live has been recognized for a long time. This knowledge was first exploited in a dialect survey carried out in the Netherlands in 1939. This survey included the following question: *In which place(s) in your area does one speak the same or about the same dialect as you do?* Locations indicated to be linguistically similar were connected by means of arrows on the map. Concentrations of arrows were interpreted as dialect groups and arrowless areas as dialect boundaries. Several maps were constructed using this so-called 'pijltjesmethode' (e.g. Weijnen 1946; Daan 1969). However, the validity of the method was never systematically tested.

In a study by Van Hout & Münstermann (1981) subjects heard samples of nine dialects from different areas in the Netherlands. For each of these they rated a 7-point scale expressing the linguistic distance from Standard Dutch. In addition, they indicated on a map where they thought the target variety was spoken. Van Hout & Münstermann measured on the map the distance between the centre of the Randstad, i.e. the general area where Standard Dutch is commonly located, and the places where the target varieties were *actually* spoken. They also measured the distance between the centre of the Randstad and the places where the subjects *thought* that the varieties were spoken. All three measures, i.e. perceived linguistic distance, objective geographic distance and subjective geographic distance, yielded the same order of the nine varieties. Van Hout & Münstermann did not assess the *objective* linguistic distance of the fragments.

In our study of laymen's intuitions on the linguistic distances in the Netherlands and Flanders we took both geographic and linguistic data into consideration, using a quantitative approach. We focussed on the distances between regional varieties and Standard Dutch. In contrast to Van Hout & Münstermann (1981) and in accordance with the 'pijltjesmethode', we did not present auditory stimuli but examined people's *preconceived* ideas. In Section 2 we describe the way in which the subjective estimates of linguistic distance were obtained. The subjective linguistic distances are compared with objective geographic distances in Section 3, with objective linguistic distances in Section 4, and with these two measures combined in Section 5. Finally, in Section 6, our findings are summarized.

2. Subjective estimates of linguistic distance

2.1 Method

140 subjects, 69 males and 71 females from different parts in the Netherlands, were used to assess the intuitions on linguistic distance. Their mean age was 28.1 years, ranging between 15 and 63. Most had a high level of education. The subjects were presented with a map with the twelve provinces of the Netherlands and the five Dutch-speaking provinces of Belgium. They were asked to write a number between 0 and 100 for the Frisian language as spoken in Friesland and for the varieties of Dutch (Gronings, Drenths, etc.) as spoken in the other provinces, 0 expressing no linguistic distance and 100 expressing the largest linguistic distance to Standard Dutch (not indicated on the map).[1] The subjects were free to assign the same value to different language varieties if they thought their distance to Standard Dutch to be equal.

2.2 Results

Cronbach's alpha was .93, which indicates that the subjects agreed to a high extent on the values for the different language varieties. In Figure 1 the mean estimates of the linguistic distance between the varieties spoken in the various provinces and Standard Dutch (with a possible range of between 0 and 100) are indicated on the map. There are clear gaps between the values, suggesting four groups of varieties.

The group with the smallest distance from Standard Dutch, with values between 17 and 22, comprehends Noord-Holland, Zuid-Holland, Utrecht, and Flevoland. The first three provinces constitute the Randstad, the culturally, politically and economically dominant western part of the Netherlands. The Randstad is commonly seen as the region where Standard Dutch was formed and from where it expanded to the other regions. The fourth variety in this group is Flevoland, which consists of two polders that were reclaimed from the former Zuiderzee. People from all over the Netherlands settled in these polders, bringing with them a great variety of dialects. From this a fairly neutral language variety emerged, very similar in grammar and vocabulary (but not always in pronunciation) to the standard language (Scholtmeijer 1992).

The second group of varieties, with somewhat higher distance values of between 34 and 41, includes four provinces in the Netherlands, i.e. Overijssel, Gelderland, Noord-Brabant, and Zeeland. They form a semi circle around the provinces of the first group. Apparently, this area is considered to be a zone of transition.

The third group, with fairly high distance values of between 50 and 58, comprises all five provinces in Flanders, as well as two provinces in the northeast of the Netherlands, i.e. Groningen and Drenthe. Except for Antwerpen, all these provinces are in direct contact with areas where other languages are spoken, namely (Walloon) French in the south and German in the east. From the viewpoint of people living in the Randstad the area included in this group is peripheral.

Fourthly, the group with the largest distance from Standard Dutch comprises only Friesland, with a value of 72. Note that Friesland received a higher linguistic distance value than Groningen, despite the fact that it is geographically closer to the Randstad. It could be that the special status of Frisian, presently spoken by about half the inhabitants of Friesland, has played a role in the estimation of

linguistic distance. Frisian is the second national language in the Netherlands. This is common knowledge in the Netherlands and it was also explicitly mentioned in the instructions.

Figure 1. Mean estimates of the linguistic distance (minimum 0, maximum 100) between regional varieties of Dutch / Frisian and Standard Dutch. The four groups distinguished (see text) are indicated with different shadings of grey, darker grey indicating greater estimated distance.

3. Subjective estimates of linguistic distance compared with geographic distance

The results from the rating task suggest that estimates of linguistic distance are closely related to geographic distance in the sense that varieties are generally (but not always) estimated to be more distant linguistically from Standard Dutch as they are more distant geographically from the Randstad. To test the relationship in a more formal way, we calculated the geographic distance between each province and Standard Dutch. We located Standard Dutch at the position of Haarlem, which is typically seen as the place where Standard Dutch is spoken in its purest form (see Smakman & Van Bezooijen 1997). The distances from Haarlem in Noord-Holland to the centres of the other provinces were measured in mm in a straight line on the map and then rescaled to values between 0 and 100. The results are given in Figure 2. To facilitate the comparison with the estimates of linguistic distance in Figure 1, the geographic distances were divided into four groups as well. As there were no clear gaps, we divided the geographic continuum into four groups of (about) equal width, namely 14-34, 35-56, 57-78, and 79-100.

The correlation between estimated linguistic distance and geographic distance is .87 ($p<.01$).[2] The high correlation confirms the impression that one gets from comparing Figures 1 and 2, namely that there is a substantial though not complete overlap between subjective linguistic reality and geographic reality. To examine the differences in more detail, we used regression analysis with geographical distances as the independent variable and linguistic distance ratings as the dependent

variable.

Figure 2. Geographic distance (originally in mm, rescaled from 0 to 100) between regional varieties of Dutch and Frisian to Standard Dutch. The four groups distinguished (see text) are indicated with different shadings of grey, darker grey indicating a greater geographic distance.

The province with the greatest negative residual value is Flevoland, which means that the subjects estimated its linguistic distance to standard Dutch to be lower than one would expect on the basis of the geographical distance to Haarlem. As explained in 2.3, the linguistic situation in Flevoland is exceptional, because it is a new area with no history of dialect like the other provinces. The province with the greatest positive residual value is Friesland, which means that the estimated linguistic distance is larger than one would expect on the basis of geography. As stated in 2.3, the symbolic meaning of the special status of Frisian may have increased the linguistic distance to Standard Dutch in the subjects' minds.

4. Subjective estimates of linguistic distance compared with objective distance

Next to geographical distances, we considered the role of objective linguistic distances in the perception of linguistic variation. As we were not sure whether non-linguists base their ideas on a perhaps somewhat romanticized linguistic situation in the past or whether they keep track of the rapid changes affecting the use and structure of dialects in the Dutch language area, we decided to use two dialect samples of different dates, the Old Dialect Sample and the New Dialect Sample. More information on these samples is given in 4.1. Linguistic distance to Standard Dutch was calculated using the Levenshtein algorithm, which expresses the degree of dissimilarity between (groups of) words at the phonetic-phonological level. This measure is explained in 4.2. In 4.3 the Levenshtein distances and their relationship with the subjective distances are presented.

4.1 Dialect samples

For the Old Dialect Sample, we made use of the *Reeks Nederlandse Dialectatlassen* (*RND*) (Blancquaert & Pée 1925-1982). The *RND* contains the transcribed translations of 141 sentences into 1956 different dialects in the Dutch and Frisian language area. The data are between 30 and 80 years old. The Old Dialect Sample contains a subset of 125 words for 324 dialects, regularly distributed across the area under consideration, as shown in Figure 3. In this sample the province of Flevoland is not represented. The only dialect available for this province in the *RND* is Urk. The inhabitants of Urk traditionally speak a Low-Saxon dialect, whereas the remaining part of Flevoland is characterized by an approximation of Standard Dutch. As the variety of Urk would give a distorted picture of the language spoken in the province as a whole, we deemed it unfit to be included. The *RND* does not contain a sample of Standard Dutch. For the purpose of the present study, we derived a transcription of an older form of Standard Dutch from the *Tekstboekje* of Blancquaert (1939).

The New Dialect Sample was compiled by the first author in 2001. One hundred nouns referring to objects and concepts of every day life were translated and spoken onto tape by present-day speakers of 26 Dutch varieties from different parts of the Netherlands and Flanders, at least one for each province (see Figure 3). The dialect sample was supplemented with recordings of speakers of Standard Dutch and Standard Frisian. Just like in the Old Dialect Sample, the province of Flevoland is not represented in the New Dialect sample. Transcriptions were made by the first author.

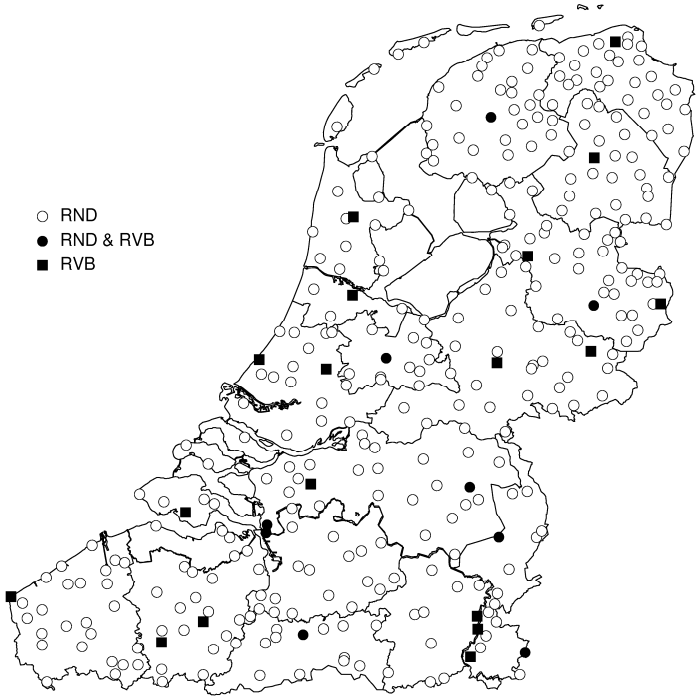

Figure 3. Location of the 324 dialects in the Old Dialect Sample (RND) and the 26 dialects plus Standard Frisian (located in Grou) in the New Dialect Sample (RVB). Some dialects coincide.

4.2 Levensthein distance

The Levenshtein distance was introduced by Kessler (1995) as a tool for measuring linguistic distances among Irish dialects. Later the technique was applied to other dialect areas, such as Dutch (Heeringa, 2004) and Norwegian (Gooskens & Heeringa 2004). The Levenshtein distance compares the pronunciation of words in one dialect with the pronunciation of semantically corresponding words in another dialect, assessing how the one pronunciation can be efficiently converted to the other by inserting, deleting or substituting sounds. The fewer operations are needed, the smaller the linguistic distance between the words is assumed to be. An example may illustrate the basic procedure. In our New Dialect Sample, Standard Dutch *konijn* [koːnɛin] 'rabbit' is pronounced as [kninə] in the dialect of Zwollekerspel (province of Overijssel). The first pronunciation can be converted to the second in the following way:

Word string	Operation	Cost
koːnɛin	delete oː	1
knɛin	substitute ɛi by i	1
knin	insert ə	1
kninə		
		3

In this particular example, all three possible operations are applied once. In fact, there are many alternative string edit operations that map [koːnɛin] to [kninə]. However, the Levenshtein algorithm always selects the mapping with the lowest cost. A representation of the segments in terms of linguistic units, as in the example given, is not the only option when calculating the Levenshtein distance. In the present study pronunciations were compared on the basis of spectrograms. A spectrogram is the visual representation of the acoustical signal and the visual differences between the spectrograms reflect the acoustical differences. 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. The different sounds were isolated from the recordings and monotonized at the mean pitch of each of the two speakers with the program PRAAT. Next, PRAAT was used to make a spectrogram for each sound using the so-called Bark filter, which is a perceptually oriented model. On the basis of the Bark filter representations, segment distances were calculated.

Using acoustic segment distances, it is easy to find gradual weights for insertion and deletion operations. We calculated the distance from the sound to be inserted or deleted to 'silence' as a sound for which the intensities of all frequencies are equal to zero at each point in time. In our study we cut a small silent part on the IPA tape and used this as a 'silence' sample. This assures that silence has approximately the same background noise as the other sounds. Comparing segments to 'silence' we found the [a] to be the most distant, and the [ʔ] to be the closest.

We used logarithmic segment distances, which means that small distances, i.e. small differences in pronunciation, are weighed relatively more heavily than large distances, i.e. large differences in pronunciation. Since the logarithm of 0 is not defined, and the logarithm of 1 is 0, distances were increased by 1 before the logarithm was computed. To obtain percentages, we divided the logarithmic segment distances by the highest possible segment distance and multiplied the value by 100.

To reckon with syllabification, the Levenshtein algorithm was adapted so that only a vowel may match with a vowel, a consonant with a consonant, [j] or [w] with a vowel (or vice versa), [i] or [u] with a consonant (or vice versa), and a central vowel (in our research only the schwa) with a sonorant (or vice versa). So [i], [u], [j] and [w] align with anything, but otherwise vowels align with vowels and consonants with consonants. In this way unlikely matches, e.g. [p] with [a], are prevented.

4.3 Results

The Old Dialect Sample contained data for 324 dialects. The Levenshtein values for the dialects within the same province were averaged to calculate the objective linguistic distance between each province (except Flevoland) and Standard Dutch. The results are shown in Figure 4. It can be seen that there are no clear gaps in the continuum. To allow comparison with the previous classifications, here also four groups were distinguished, using (approximately) equal intervals (22-26, 27-32, 33-38, 39-43). Most distant from Standard Dutch is the language spoken in Friesland. It is quite more deviant than the other provinces with relatively large values, such as West-Vlaanderen and Groningen. The province that is closest to Standard Dutch is Zuid-Holland, directly followed by Noord-Holland and Utrecht.

Figure 4. Objective linguistic distances to Standard Dutch based on the Old Dialect Sample (1921/1922-1975).

Figure 5. Objective linguistic distances to Standard Dutch based on the New Dialect Sample (2001).

The New Dialect Sample contained data for 27 varieties (26 dialects plus Frisian) in all provinces

except Flevoland. The results of the Levenshtein measurements are shown in Figure 5, again with a division into four groups (12-18, 19-26, 27-34, 35-41). The correlation between the old distances and the new distances is .80 ($p<.01$). We think that the degree of similarity between the rank orders of the provinces in the Old and New Dialect Sample is striking, considering the time lapse and the difference in size.

The correlation between subjects' estimates of linguistic distances and the linguistic distances based on the Old Dialect Sample is .93 ($p<.01$). The provinces of Belgisch Limburg, Noord-Brabant, Friesland and Antwerpen are overrated most strongly. Undervalued are the provinces of Overijssel and Gelderland. The correlation between the subjective estimates of linguistic distances and the linguistic distances based on the New Dialect Sample is .80 ($p<.01$). Most strongly overrated are the provinces of Drenthe, Friesland, Belgisch Limburg, Groningen and Vlaams Brabant. Most undervalued are the provinces of Utrecht and Gelderland. So, it appears that the estimated distances correlate more strongly with the old objective distances than with the new objective distances.

5. Subjective estimates of linguistic distance compared with objective linguistic distance and geographical distance combined

In previous sections we looked at the relationship between the two determinants of popular ideas of linguistic distance, i.e. geography and linguistic distance, separately. One may wonder to what extent a better prediction is obtained when the two determinants are combined. We therefore performed a number of multiple regression analyses. In view of the special position of Friesland, specifically the fact that the subjects estimated its linguistic distance to Standard Dutch to be larger than one would expect on the basis of its geographical location, the analyses were run both with and without this province. The results are presented in Table 1.

Table 1. Results of four multiple regression analyses with geographical distance and objective linguistic distance as independent variables and estimated linguistic distance as dependent variable

	Old Dialect Sample		New Dialect Sample	
	Friesland included			
	r	r^2	r	r^2
Geographical distance	.87	75.1	.87	75.1
Objective linguistic distance	.93	86.2	.80	64.7
Geographical and objective linguistic distance	.94	89.3	.90	80.3
	Friesland excluded			
Geographical distance	.98	96.8	.98	96.8
Objective linguistic distance	.91	83.0	.78	61.2
Geographical and objective linguistic distance	.99	97.6	.98	96.9

The data in Table 1 show again that the old dialect data are a better predictor of the subjects' intuitions than the new dialect data, not only when Friesland is included (86.2% vs. 64.7% explained variance) but also when Friesland is excluded (83.0% vs. 61.2%). Furthermore, what we already suspected with respect to the special position of Friesland is now confirmed, geographical distance is

a much better predictor of estimated linguistic distance if Friesland is left out of consideration than when it is included in the analysis. The percentage of explained variance is raised from 75.1 to 96.8. In fact, if only the varieties of Dutch are considered, adding linguistic information to the equation has (virtually) no effect on the total percentage of explained variance. Geographical distance alone is an almost perfect predictor.

6. Conclusion

The aim of the present study was to gain insight into the basis of people's intuitions about the linguistic situation in the Netherlands and Flanders. More specifically we were interested in the basis of non-linguists' ideas about the linguistic distances between Standard Dutch on the one hand and Dutch dialects and Frisian on the other. We wanted to know whether these ideas are (mainly) geographically or linguistically based.

As for the linguistic determinant, it appears that the distances calculated on the basis of the data gathered for 324 dialects between 1921/1922 and 1975 are a better predictor of present-day views held by non-linguists than the data gathered for 27 dialects in 2001. This could mean that subjects' intuitions are outdated and that they are based on a situation that no longer exists. It could also mean that the old objective distances are more representative of regional varieties because they are based on a larger number of dialects per province.

In the results for the Older Dialect Sample, the inclusion or exclusion of the province of Friesland was found to make a substantial difference. When Friesland is included, objective linguistic distance is a better predictor of estimated linguistic distance; adding geographic distance hardly makes a difference. When Friesland is excluded, the opposite is the case. Then geographical distance is such a good predictor that adding objective linguistic distance has a minimal effect.

We do in fact think that the subjects have based their estimates of linguistic distance largely on geographical factors. We know for certain that non-linguists believe that Standard Dutch is spoken in its purest form in de Randstad (Smakman & Van Bezooijen 1997) and it is very likely that they assume that linguistic distance will increase with increasing geographic distance. As geographical distance does in fact correlate with objective linguistic distance, the subjects' estimates could in principle also be based on knowledge of dialectal differences. However, we think that linguistic knowledge has at the most played a supportive role. Except for a few regions such as Limburg, most dialects are used almost exclusively in the informal domains of communication, for the interaction with friends and family. When confronted with outsiders, most dialect speakers will switch to Standard Dutch, preserving at the most some regional pronunciation characteristics in their speech. In Flanders, accommodation will be less complete, but then, all subjects in our study were from the Netherlands, so they probably had little experience with the varieties of Dutch as spoken at the other side of the border. Also, in most provinces only a minority speaks dialect nowadays.

There is one province though that forms an exception, and that is Friesland. In Friesland more than half of the population speaks Frisian, which is officially recognized to be a separate language from Dutch. This special position has played an important role in the subjects' estimates of linguistic distance. For although Friesland is geographically closer to the Randstad than for example Groningen, it has nevertheless received a much larger linguistic distance rating. The correlation with the objective linguistic measurements suggests that the subjects may have based their deviant rating

on linguistic knowledge. However, again this is probably not the case. In analogy with the other provinces, all Frisians who speak Frisian also speak Dutch and with non-Frisians around they will automatically use the national language. Tourists visiting Friesland may not hear one word of Frisian. We think that the subjects simply assumed that a variety that is officially recognized to be a separate language must have many linguistic characteristics of its own and that it is this assumption that has led them to give high distance ratings.

In short, in our view the popular views on linguistic deviancy in the Netherlands and the northern part of Belgium are largely based on three widely shared beliefs, namely that (1) the varieties of Dutch spoken in the west of the Netherlands resemble Standard Dutch most closely, (2) linguistic distance from Standard Dutch increases with increasing geographical distance, and (3) the linguistic distance of Frisian to Dutch is larger than one would expect on the basis of its geographical distance from the west because it is a separate language.

Notes

* We thank Peter Kleiweg for letting us use the programs that he developed for constructing and visualizing maps.

1. In addition to the varieties spoken in the Netherlands and Flanders, the subjects were also asked to estimate the linguistic distance of French, German, and English to Standard Dutch. These data are not presented here.
2. Since for the objective linguistic analysis the province of Flevoland was excluded (see Section 4.1), we also calculated the correlation without this province. The exclusion of Flevoland did not affect the coefficient.

References

Blancquaert, E. (1939, second edition), *Tekstboekje.* Nederlandse Fonoplaten van Blancquaert en van de Plaetse, Eerste reeks [Text book. Dutch phonograms of Blanquaert and Van de Plaetse. First series], Antwerpen: De Sikkel.

Blancquaert, E. & W. Pée, Editors (1925-1982), *Reeks Nederlandse Dialectatlassen,* Antwerpen: De Sikkel.

Daan, J. (1969). Dialecten [Dialects], in J. Daan & D.P. Blok (Eds.), *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde* [From border city to country border; clarification of the map: Dialects and Onomastics] *Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen*, 37:7-43, Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.

Gooskens, C. & W. Heeringa (2004), Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data, *Language Variation and Change* 16: 189-207.

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

Kessler, B. (1995), Computational dialectology in Irish Gaelic. *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics,* EACL, Dublin: 60-67.

Scholtmeijer, H. (1992), Het Nederlands van de Ijsselmeerpolders [The Dutch spoken in the IJsselmeerpolders]. Dissertation Universtiy of Leiden.

Smakman, D. & R. van Bezooijen (1997), Een verkenning van populaire ideeën over de standaardtaal in Nederland [An exploration of popular ideas on the standard language in the Netherlands], *Taal en Tongval*, special issue 10: 126-139.

Van Hout, R. & H. Münstermann (1981), Linguistische afstand, dialekt en attitude [Linguistic distance, dialect and attitude]. *Gramma*, 5: 101-123.

Weijnen, A.A. (1946), De grenzen tussen de Oost-Noordbrabantse dialecten onderling [The borders between the dialects of eastern North Brabant], in A. Weijnen, J.M. Renders, & J. van Ginneken (Eds.), *Oost-Noordbrabantse dialectproblemen* [Eastern North Brabant dialect problems]:.1-15. Bijdragen en Mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam, 8.