The Contribution of Linguistic Factors to the Intelligibility of Closely Related Languages

Charlotte Gooskens
Department of Scandinavian Studies, University of Groningen, Groningen, The Netherlands

The three mainland Scandinavian languages (Danish, Swedish and Norwegian) are so closely related that the speakers mostly communicate in their own languages (semicommunication). Even though the three West Germanic languages Dutch, Frisian and Afrikaans are also closely related, semicommunication is not usual between these languages. In the present investigation, results from intelligibility tests measuring the mutual intelligibility of Danish, Norwegian and Swedish were compared with results of similar tests of mutual intelligibility between speakers of Dutch, Frisian and Afrikaans. The results show that there are large differences in the level of intelligibility depending on test group and test language. Correlations between the intelligibility scores and linguistic distance scores showed that intelligibility can to a large extent be predicted by phonetic distances, while intelligibility is less predictable on the basis of lexical distances.

doi: 10.2167/jmmd511.0

Keywords: linguistic distances, mutual intelligibility, semicommunication, receptive bilingualism

Introduction

Most individuals have to invest considerable time and effort in order to master a language other than their mother tongue. However, some genetically related languages are so similar to each other in terms of grammar, vocabulary and pronunciation that speakers of one language can understand the other language without prior instructions. Speakers of such languages are able to communicate with each other without a lingua franca or without one speaker using the language of the other. This type of interaction, which is referred to with terms such as 'semicommunication' (Haugen, 1966) or 'receptive multilingualism' (Braunmüller & Zeevaert, 2001), has many advantages, in any case on the production side.

The Scandinavian languages, Danish, Norwegian and Swedish, are an example of languages which are so closely related that they are mutually intelligible. In the past, a number of studies were carried out in order to get a precise picture of the actual level of understanding between speakers of these languages (e.g. Bø, 1978; Börestam, 1987; Maurud, 1976). Recently, an investigation supported by the Nordic Cultural Fund was carried out to examine the communicative situation at the beginning of the 21st century (see

0143-4632/07/06 445-23 $20.00/0
J. OF MULTILINGUAL AND MULTICULTURAL DEVELOPMENT

Vol. 28, No. 6, 2007

Delsing & Lundin Åkesson, 2005). In the present paper this investigation will be referred to as the 'INS-investigation'.[1]

The three West Germanic languages, Dutch, Frisian and Afrikaans, form another group of languages that are so closely related that a high level of mutual intelligibility can be expected. However, in contrast with the Scandinavian languages, semicommunication is not the usual manner of communication between the speakers of these languages. Speakers of Dutch are generally not interested in the Frisian vernacular whereas all Frisians are bilinguals. Afrikaans is understood on the basis of Dutch, at least to a certain extent. Little research has been conducted in order to investigate how well the speakers can understand the other two languages. Exceptions are a number of experiments testing the intelligibility of Frisian among Dutchmen (Van Bezooijen & Van den Berg 1999a, 1999b, 1999c, 2000) and an investigation by Van Bezooijen and Gooskens (2005), who replicated the spoken intelligibility tests from the INS-investigation (see second section).

The intelligibility of a closely related language mainly depends upon three factors:

(1) the listener's attitude towards the language,
(2) the listener's contact with the language and other language experience, and
(3) linguistic distance to the listener's language.

The Scandinavian studies mentioned above included questions about attitude towards and contact with the test language. The authors found some degree of relationship between the non-linguistic factors (contact, language instruction and attitude) and the intelligibility scores, but correlations are low and the direct relationship is difficult to prove (see Gooskens, 2006). Gooskens and Van Bezooijen (2006), too, found only a weak relationship between attitude and mutual intelligibility of written texts in a study of Afrikaans and Dutch. The third factor, linguistic distance, has been largely neglected so far, first of all due to the absence of a suitable method to measure linguistic distances between languages. In recent years, new methods have been developed for measuring linguistic distances in the area of dialectometry (see Heeringa, 2004: 14–24 for an overview). In the present investigation, the so-called Levenshtein distances will be used. This method has proved a useful way of measuring distances between dialects and closely related languages (Gooskens & Heeringa, 2004a, 2004b). The method is explained in the third section. Van Bezooijen and Gooskens (2005) used Levenshtein distances for the first time to explain mutual intelligibility between spoken Dutch, Frisian and Afrikaans. However, as too few language varieties were included in their investigation, the relation between Levenshtein distances and intelligibility scores could not be tested statistically. Gooskens and Heeringa (2004a) found Levenshtein distances between Dutch and Frisian varieties to be smaller than distances between the three Scandinavian Languages (a mean difference of 6% between the two language groups). These results raise the question whether mutual intelligibility between speakers of Dutch, Frisian and Afrikaans is just as high as mutual intelligibility in Scandinavia and it asks for a more detailed

comparison of the linguistic distances within the two language groups. It is possible, for example, that the relationship between intelligibility and linguistic distances is different at each linguistic level.

According to most overviews of linguistic differences between the Scandinavian languages, the morphological and syntactic differences between the Scandinavian languages can be assumed to be of hardly any importance for the mutual intelligibility. The phonetic and lexical differences between the three languages are often sketched as in Figure 1 (see for example Delsing & Lundin Åkesson, 2005; Torp, 1998). The Norwegian–Danish communication is impeded by phonetic differences but facilitated by lexical similarities. On the other hand, the most hindering factor in the Swedish–Norwegian communication is the differences in vocabularies, while the phonetic similarities are an advantage for the mutual intelligibility. In the communication between Danes and Swedes both the lexicon and the phonological system form a hindrance.

The linguistic differences between Dutch, Frisian and Afrikaans are a result of the historical relationships between the languages. Like the Scandinavian languages, Dutch and Afrikaans originate from the same language, but in the course of history the two languages have diverged. Originally, Afrikaans was a dialect that developed among a small group of Dutch colonists who settled in South Africa at the beginning of the 17th century. In the course of time its nature changed, among others because it was largely used by non-native speakers with an insufficient command of Dutch and because it was influenced by other local languages. Frisian, on the other hand, is historically related to English but has become increasingly similar to Dutch. Due to the dominance of Dutch in the media, education and administration, Frisian loses more and more of its typical characteristics. The historical relationships between Dutch, Afrikaans and Frisian have influenced the different linguistic

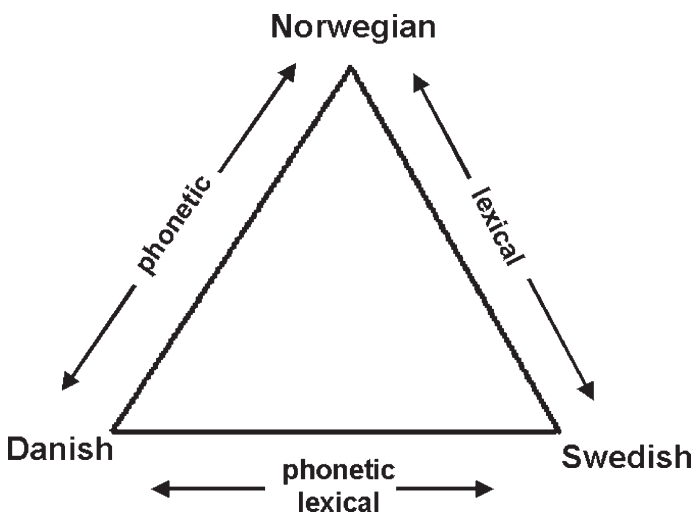

Figure 1 Schematic overview of the linguistic differences that form the largest obstacle for the mutual intelligibility between the Scandinavian languages

levels in different ways. Van Bezooijen and Gooskens (2005) found Dutch listeners to be better able to understand Afrikaans than Frisian. They attributed this to a smaller proportion of non-cognates and a smaller phonetic distance of the cognates between Dutch and Afrikaans than between Dutch and Frisian. In the rest of this paper the Dutch/Frisian/Afrikaans language group will be referred to as the West Germanic group.

The aim of the present investigation is to explain part of the results of the intelligibility tests from the INS-investigation by means of linguistic distances. Linguistic distances are measured at two linguistic levels, the phonetic level and the lexical level. Further, the results from the investigation on the mutual intelligibility of Dutch, Frisian and Afrikaans by Van Bezooijen and Gooskens mentioned above are included in the analysis. In this way the role of lexical and phonetic distances in the two language groups can be compared

In the second section, it will first be shown how mutual intelligibility was tested in the two language groups and the results of the tests will be presented. Next, phonetic distance scores (the third section) and lexical distance scores (fourth section) will be calculated. In the fifth section, the phonetic and the lexical distance scores will be related to the intelligibility scores and finally some conclusions will be drawn in the sixth section.

Intelligibility

The Scandinavian data used for the present investigation are a limited set of data from the INS-investigation mentioned in the first section. The West Germanic data (Dutch, Frisian and Afrikaans) are from the project carried out by Van Bezooijen and Gooskens (2005). For the sake of comparability, the West Germanic data were as similar as possible to the Scandinavian data. The listeners were matched as well as possible and so were the tasks. The recordings and the tests were administered in the mother tongue of the subjects in all cases. In the first part of this section it is shown how intelligibility was measured and in the second part the results are presented.

Method

Listeners

Scandinavian listeners. For the present investigation a selection of listeners was made from the INS-investigation.[2] In this section only the listeners who were selected for the present investigation are described. First of all only the results from the Scandinavian listeners from the INS-investigation were used. These listeners came from nine different towns in Denmark, Norway, Sweden and the Swedish-speaking part of Finland (see Figure 2). In earlier investigations on inter-Scandinavian intelligibility, mutual intelligibility has often been tested in the capitals only (e.g. Maurud, 1976). However, as the capitals of Norway and Denmark are close to Sweden while the capitals of Sweden and Finland are distant from the neighbouring countries, the amount of contact with the neighbouring languages[3] may differ considerably. Therefore, groups

Figure 2 The four countries and nine cities (indicated by stars) included in the Scandinavian investigation

of listeners in one or two additional towns were tested in each of the four countries.

In Table 1, an overview is given of the 14 groups of listeners that were included in the present investigation. The shaded cells in the table show which languages and which groups of listeners were tested. Except for Vaasa and Malmö, where no group of listeners was included for Danish,[4] two groups were tested in each town, one for each neighbouring language. This means that 16 groups were tested. In total 488 secondary school pupils between the age of 16 and 19 years with a mean age of 17.0 years participated. In general, more girls (55.1%) than boys (40.4%) participated (4.5% of the listeners did not give their sex).

Only the results by listeners who attended pre-university education were analysed. Furthermore, only those listeners who reported speaking the official Scandinavian language of the country of residence (Danish, Norwegian or Swedish) at home were included. Listeners who spoke more than one language at home were excluded to make sure that all listeners had a high native competence in the Scandinavian language of the relevant country. If, for example, a listener claimed to speak Danish and Turkish at home, this listener

Table 1 Number of Scandinavian listeners, per town and intelligibility test (shaded) and total, percentage of boys and girls, and mean age per town

Town	*Intelligibility test*			*Total number of listeners*	*% boys**	*% girls**	*Mean age*
	Danish	*Norwegian*	*Swedish*				
Denmark							
Århus	–	30	42	72	34.7	63.9	17.8
Copenhagen	–	39	26	65	35.4	56.9	17.2
Norway							
Bergen	22	–	19	41	56.1	41.7	17.1
Oslo	54	–	77	131	36.6	58.0	16.8
Sweden							
Malmö	**	43	–	43	51.2	41.9	16.6
Stockholm	28	19	–	47	34.0	63.8	16.9
Finland							
Mariehamn	22	25	–	47	40.4	59.6	17.2
Vaasa	**	12	–	12	33.3	66.7	16.4
Helsinki	9	21	–	30	56.7	43.3	16.8
Total	135	189	164	488	40.4	55.1	17.0

*The percentages of boys and girls do not always add up to 100, as not all listeners answered the question about their sex.
**Danish was not tested in Malmö and Vaasa.

was excluded from the investigation, as it is possible that he had been raised primarily in Turkish by Turkish-speaking parents. In the case of Finland, however, listeners who spoke Swedish as well as Finnish at home were included as it could be assumed that in those cases Swedish was the mother tongue of at least one of the parents, Finnish being the majority language of the country. In Helsinki, a majority of the listeners (70%) were bilinguals, while in Mariehamn and Vaasa this percentage was much lower (15% and 17%). Furthermore, in Finland only pupils from schools where Swedish is the language of instruction were included. In this way all Finnish listeners could be assumed to have native competence in the Swedish language.

West Germanic listeners. The West Germanic material was collected by Van Bezooijen and Gooskens (2005). A total of 81 West Germanic listeners participated (see Figure 3 and Table 2). Half of them listened to one related language and the other half to the other language. However, no Frisian group of listeners listened to Dutch, as all Frisian children learn Dutch as well as Frisian. Accordingly, five groups of listeners were tested in the West Germanic language group. They meet the same criteria as the Scandinavian listeners. They were between 16 and 17 years with a mean age of 16.6 years and all

Figure 3 Maps of the Netherlands and South Africa showing where the Frisian, Dutch and South African listeners came from (indicated by stars)

Table 2 Number of West Germanic listeners, per region and intelligibility test (shaded cells) and total, percentage of boys and girls, and mean age per place

Region	*Intelligibility test*			*Total number of listeners*	*% boys**	*% girls**	*Mean age*
	Dutch	*Frisian*	*Afrikaans*				
Netherlands	–	16	16	32	25.0	75.0	16.3
Friesland	**	–	17	17	52.9	47.1	16.3
South Africa	15	17	–	32	34.4	65.6	17.1
Total	15	33	33	81	34.1	64.6	16.6

*The percentages of boys and girls do not always add up to 100, as not all listeners answered the question about their sex.
**Dutch was not tested in Friesland.

attended pre-university education. All listeners were native speakers of Dutch, Frisian or Afrikaans. Listeners who spoke more than one language at home were excluded. An exception was made for seven Frisian listeners, who spoke

Dutch in addition to Frisian at home. In the Scandinavian investigation, listeners were tested in two or three towns for each country as listeners from different parts of the country may have different linguistic and extra-linguistic backgrounds that may influence their understanding of the neighbouring languages. In the West Germanic investigation, geography was expected to play a minor role, as South Africa is isolated from the other two language areas and the Frisians were not tested for Dutch. Geography may play a role only in the case of Dutch listeners listening to Frisian, as listeners living close to Friesland may have had more contact with the Frisian language than listeners living further away. In order to minimise the influence of geographical proximity, only Dutch listeners from provinces that are not adjacent to Friesland were included. About half originated from Zwolle and surroundings in the north-eastern part of the country; the other half came from The Hague in the west. The South African subjects originated from Hennenman and surroundings in the province of Vrystaat. The Frisian listeners came from different villages in the surroundings of Leeuwarden.

Task

To assess the intelligibility of a running spoken text, the listeners listened to a news item about a run-away kangaroo being chased in the streets of Copenhagen.[5] The text was translated from the original Norwegian text into Danish, Swedish and Dutch, and from Dutch into Frisian and Afrikaans and read aloud by native speakers of the standard languages. The speakers all had professional experience with reading aloud texts. The number of words varied between 256 and 262 for the Scandinavian texts and between 272 and 290 for the West Germanic texts. Each group of listeners listened to the recording in one of the two related languages (see Tables 1 and 2). There were five open questions about the text. The listeners wrote down their answers while listening to the recordings. Three degrees of correctness were distinguished: completely correct (2 points), partly correct (1 point) and incorrect (0 points). The maximum score was therefore 10 points per listener. The percentage of correct answers formed the intelligibility score.

The intelligibility test was preceded by personal questions about the language situation at home, sex and age.[6] The answers to these questions were used for the selection of the listeners (see section on 'Listeners').

Results

In Table 3, the mean intelligibility scores are given for each group of listeners.[7] In general, the Scandinavian results are similar to results found in previous investigations (see first section). Mutual intelligibility is highest between Norwegians and Swedes (82.6% and 83.7% correct answers for the Swedes and a little higher, 88.9% and 88.3%, for the Norwegians); Danish is hard to understand, especially for Swedish-speaking listeners (24.3% correct for Stockholm, 21.8% for Mariehamn and 6.7% for Helsinki). We see that there are large differences in intelligibility depending on test language and the places where the listeners live. The percentages of correct answers may differ considerably within one country. For example, the Danish listeners in Copenhagen answered 58.1% of the questions about the Swedish recording

Table 3 Percentages of correct answers in the intelligibility test, broken down for group of listeners and test language

Listeners	***Intelligibility test***		
	Danish	***Norwegian***	***Swedish***
Denmark			
Århus (år)	–	58.0	48.3
Copenhagen (co)	–	55.9	58.1
Norway			
Bergen (be)	80.9	–	88.9
Oslo (os)	69.3	–	88.3
Sweden			
Malmö (mö)	–	82.6	–
Stockholm (st)	24.3	83.7	–
Finland			
Mariehamn (ma)	21.8	82.0	–
Vaasa (va)	–	86.7	–
Helsinki (he)	6.7	57.1	–
Mean	40.6	72.3	70.9
	Dutch	***Frisian***	***Afrikaans***
Netherlands (nl)	–	55.6	62.4
Friesland (fr)	–	–	66.6
South Africa (af)	44.0	25.0	–
Mean	44.0	40.3	64.5

correctly, while only 48.3% of the questions were answered correctly in Århus. This shows how important it is to pay attention to the geographical background of the listeners. This means that there is no general answer to the question how well Scandinavians understand each other's languages.

Intelligibility is not always symmetrical. For example, the two groups of Danes understand Swedish better (48.3% and 58.1% correct answers) than the three groups of Swedish-speaking listeners understand Danish (24.3%, 21.8% and 6.7%).[8] Also the Norwegian listeners understand Swedish and Danish better than the Swedes and the Danes understand Norwegian. These asymmetries have also been found in previous investigations.

In contrast with Scandinavia, there is not a tradition for semicommunication between speakers of the West Germanic languages under investigation. The mean score of the West Germanic listeners is significantly lower (a mean score of 50.8%) than for the Scandinavian listeners (62.0%), $p = 0.000$, $t = 5.952$. The

lower mean score is primarily caused by low scores by the South African listeners. As in the Scandinavian language group, asymmetrical intelligibility scores are found: the South African listeners understand less Frisian and Dutch than the Dutch and Frisian listeners understand Afrikaans. The Dutch and the Frisian listeners perform rather well when confronted with the other languages, better than Swedes who are confronted with Danish, for example.

Phonetic Distances

Method

In order to investigate the importance of phonetic differences for intelligibility, a phonetic distance score had to be calculated for each of the 21 intelligibility scores in Table 3. Therefore, the distance between the language variety of each group of listeners and the test language (standard Danish, Norwegian or Swedish) was measured (a total of 21 phonetic distances). For example, the phonetic difficulties that had to be overcome by the listeners from Stockholm in Sweden when listening to the news item in Danish had to be examined. To this end the phonetic distance had to be measured between the Swedish Stockholm variety and Danish as pronounced by the Danish newsreader on the tape used for the intelligibility test.[9] This means that recordings had to be made in each of the nine Scandinavian towns (see Figure 2). Together with the three recordings that were used for the intelligibility tests, this forms the material that was used for the phonetic distance measurements between the Scandinavian language varieties. In the case of the West Germanic language group, geographical distances are likely to play a smaller role than in Scandinavia and therefore no distinction was made between groups of listeners from different towns. For this language group the distances were measured on the basis of the original recordings of the three standard languages, for example, the distance between standard Dutch and standard Afrikaans. For this reason, no extra recordings had to be made.

The nine extra Scandinavian recordings were made at schools that participated in the intelligibility test in each of the nine towns (see Figure 2). A pupil from each town was instructed to read the news item aloud in the language variety that he or she used for daily communication with his or her classmates. The language variety of these pupils was regarded as representative for the language variety of their classmates by their teacher and their classmates. The language of the pupils can in all cases be characterised as a locally coloured accent (regiolect) rather than a dialect.

The recordings used for the listening experiments as well as the new recordings of the pupils from the nine Scandinavian towns were all transcribed phonetically using X-SAMPA.[10] This is a machine-readable phonetic alphabet, which maps IPA-symbols to the seven-bit printable ASCII/ANSI characters.

To measure the distances for each of the 21 combinations of language varieties (see Table 3) the texts were aligned, i.e. the phonetic transcriptions of the corresponding words were placed next to each other. As an example of an

aligned word string, the first words of the kangaroo text in Standard Swedish and the Danish Århus variety are presented in Table 4 in the orthographic as well as the phonetic form.

The phonetic distance between word pairs was assessed by means of the so-called Levenshtein distance (see Heeringa, 2004). This is an objective measure which can be calculated automatically by computer. The Levenshtein distances are based on phonetic transcriptions of the aligned texts as described above. The distances were calculated on the basis of the cognates only, as it makes no sense to calculate phonetic distances between historically non-related words. This means that distances were measured between cognate word pairs like Swedish *gatorna* and Danish *gaderne* 'the streets', but not between non-cognate word pairs like Swedish *skuttar* and Danish *hopper* 'jump' (non-cognates).

The Levenshtein distance between corresponding words is based upon the minimum number of symbols that need to be inserted, deleted or substituted in order to transform the word in one language into the corresponding word in another language. The more operations are needed, the larger the distance. In the present study, costs were assigned in the following way:

- insertions and deletions 1 point,
- identical symbols 0 points,
- substitutions of a vowel by a vowel or of a consonant by a consonant 0.5 point,
- substitutions of a vowel by a consonant or of a consonant by a vowel 1 point,
- diacritics were joined with the preceding symbol, adding an extra 0.25 point.

Table 4 Example of a text alignment showing the orthographic version, the phonetic transcription of Standard Swedish and the Århus variety of Danish, and the English translation

Orthographic version		***Phonetic transcription***		***English translation***
Standard Swedish	***Århus variety***	***Standard Swedish***	***Århus variety***	
Kängurur	Kænguruer	çɛŋgɵɾʉɾ	kɛŋguːʁuʔɔ	Kangaroos
som	som	sɔm	sɔm	which
skuttar	hopper	skɵtːɑɾ	hɔbɔ	jump
runt	rundt	rɵnt	ʁɔnd	around
på	på	pɔ	pɔ	on
gatorna	gaderne	gɑ ː tənə	gɛ ː ðɔnə	the streets
är	er	ə	aʁ	are
inte	ikke	ɪntə	egə	not
…	…	…	…	…

So, for example the distance between [a] and [a:] was 0.25, that between [a] and [o] 0.5, and that between [o] and [a:] 0.75. The unwanted effect of word length was compensated for by dividing the total sum of costs by the number of symbols aligned. In Heeringa (2004) a more extensive explanation of the procedure is given. As an example the calculation of the distance between the word *gaderne*, 'the streets', in the pronunciation of the Århus variety of Danish and the corresponding standard Swedish word *gatorna* is presented.

Alignments	1	2	3	4	5	6
Århus variety of Danish	g	ɛː	ð	ɔ	n	ə
Standard Swedish	g	ɑː	t	ə	n	ə
Costs	0	0.5	0.5	0.5	0	0

It can be seen that the transformation involved one substitution of a consonant by another consonant (ð by t) and two substitutions of a vowel by a vowel (ɛː by ɑː and ɔ by ə). The sum of costs ($0.5 + 0.5 + 0.5 = 1.5$) is divided by the number of alignments (6). The result is a distance of 25%. The total distance between two languages is the mean distance over all word pairs. The maximum distance score is 100%.

Results

Table 5 shows the phonetic distance between the language varieties of the listeners and the test languages, measured on the basis of the cognates. As far as the distances between the Scandinavian varieties are concerned, Norwegian is clearly the language in the middle. It is most similar to both the Swedish and the Danish language varieties and the mean distance between Norwegian and all Swedish and Danish varieties is smallest (21.1%). The largest distances are found between standard Danish and the Swedish varieties (distances between 29.7% and 31.1%). The phonetic distances are not the same in two directions. For example, Standard Swedish is phonetically closer to the Danish varieties spoken in Århus (28.5%) and Copenhagen (28.2%) than Standard Danish to the Swedish varieties in Stockholm (30.8), Mariehamn (31.1) and Helsinki (29.7). The differences within one country are not large. This is what could be expected, as the subjects all spoke a regiolect rather than the local dialect.

Within the West Germanic language group, the smallest distances are found between Afrikaans and Dutch (18.5% and 18.2%), and the largest distances are found between Dutch and Frisian (26.2%) and between Afrikaans and Frisian (25.0% and 25.5%).

When comparing the linguistic distances within the two language groups, we see that the smallest distances are in fact found between Dutch and Afrikaans, and that the largest distances are found between the Swedish and Danish varieties. The mean distances within the two language groups are

Table 5 Phonetic distances between the varieties spoken in nine Scandinavian towns and the three Scandinavian standard languages (top) and between Dutch, Frisian and Afrikaans (bottom)

Varieties	*Standard varieties*		
	Danish	*Norwegian*	*Swedish*
Denmark			
Århus	–	21.6	28.5
Copenhagen	–	20.3	28.2
Norway			
Bergen	23.8	–	23.4
Oslo	23.1	–	22.0
Sweden			
Malmö	–	22.5	–
Stockholm	30.8	21.2	–
Finland			
Mariehamn	31.1	19.9	–
Vaasa	–	21.2	–
Helsinki	29.7	20.7	–
Mean	27.7	21.1	25.5
	Dutch	*Frisian*	*Afrikaans*
Netherlands	–	26.2	18.5
Friesland	–	–	25.0
South Africa	18.2	25.5	–
Mean	18.2	25.9	21.8

rather similar, 24.2% for Scandinavia and 22.7% for the West Germanic language group.

Lexical Distances

Method

Lexical distances between two language varieties were expressed as the percentage of non-cognates, i.e. historically non-related words, which the listeners heard during the test. Non-cognates should be unintelligible to listeners with no prior knowledge of the test language and a large proportion of these words will therefore impede comprehension. In addition to phonetic

distances, the percentage of non-cognates is therefore an obvious candidate for predicting intelligibility.

In contrast with the phonetic distances, it is not necessary to measure the lexical distances from the variety of each town to the test language, as there is hardly any variation at the lexical level between the varieties spoken by the groups of listeners within one country. For example, Danish listeners from Århus are likely to be confronted with the same number of non-cognates as listeners from Copenhagen when listening to Swedish. Even between Swedish as spoken in Sweden and Swedish in Finland, the lexical differences are small. Andersson and Reuter (1997): 81) estimate the lexical differences between the two language varieties to be less than 1%. For this reason the distances were calculated between each pair of languages, for example between standard Swedish and standard Danish.

To measure the lexical distances, the word pairs of the aligned texts were given points. A non-cognate was given one point, a compound that is partly cognate was given half a point, and a cognate was given zero points. In some cases a word pair consisted of non-cognates, but still a common synonym cognate existed in the native language of the listeners which would make it possible for them to understand the word in the other language. In such cases the word pair was also given zero points, as what matters is how well the listeners would be able to understand the word.

Distances were calculated in two directions, for example from Swedish to Danish and also from Danish to Swedish. This results in two lexical distances between each language pair. These two distances can be different, as two languages do not always have the same synonyms. For example, in the original Swedish text, the word *förvånade* 'surprised' corresponded to the Danish non-cognate *forbløffede.* However, in Swedish also the cognate word *förbluffade* exists and therefore the Danish word is likely to be intelligible to Swedish listeners. This word pair was therefore given zero points when measuring the distance from Swedish to Danish. The Swedish word *förvånade*, on the other hand, does not have a cognate synonym in Danish and therefore Danish listeners cannot be expected to understand the Swedish word. When measuring the distance from Danish to Swedish, the word pair was therefore given one point.

Results

In Table 6 the lexical distances between each language pair are presented. As far as the Scandinavian language area is concerned, we see that the Norwegians were confronted with no non-cognates when listening to the Danish text and the Danes encountered only very few non-cognates (1.2%) when listening to the Norwegian text. The highest percentage is found for the Swedes listening to Danish (3.6%).

In all cases, the percentage of non-cognates is much higher for the West Germanic languages than for the Scandinavian languages. As might be expected, the South Africans and the Frisians have to deal with many non-cognates when listening to each other's language (16.8% and 12.0%).

Table 6 Percentage of non-cognates between the Scandinavian languages and between Dutch, Frisian and Afrikaans

Listeners	***Danish***	***Norwegian***	***Swedish***
Danish	–	1.2	2.6
Norwegian	0.0	–	1.4
Swedish	3.6	3.4	–
Varieties	***Dutch***	***Frisian***	***Afrikaans***
Dutch	–	9.4	6.6
Frisian		–	12.0
Afrikaans	8.9	16.8	–

The Relationship Between Intelligibility and Linguistic Distances

Correlation between intelligibility scores and phonetic distances

In order to investigate the relationship between intelligibility and phonetic distance scores, the results of the intelligibility tests as found in Table 3, i.e. the mean intelligibility results per group of listeners, were correlated with the phonetic distance scores (Table 5). There was a negative correlation of -0.64 ($r^2 = 0.41$, $p = 0.002$) when all data were included. This correlation was stronger ($r = -0.80$, $r^2 = 0.64$, $p = 0.000$) when only the Scandinavian data were included.

In Figure 4, a plot is shown of the correlation between the intelligibility scores and the phonetic distance scores. In this plot, different symbols (triangles, squares and circles) correspond to the different test languages and the different groups of listeners are indicated by abbreviations. In general, intelligibility can be well predicted from the phonetic distance scores. We see that the asymmetry in mutual intelligibility between Swedes and Danes at least to some extent seems to be due to different phonetic distance scores. The Swedish varieties from Stockholm (st), Helsinki (he) and Mariehamn (ma) are less similar to standard Danish (symbolised by ∇) than the Danish varieties from Copenhagen (co) and Århus (år) are to standard Swedish (□), corresponding with lower intelligibility scores among Swedish-speaking subjects than among Danish subjects. These different phonetic distances may also explain the asymmetrical intelligibility scores that have been found in earlier investigations of mutual intelligibility between Swedes and Danes (see first section). On the other hand, the asymmetry in mutual intelligibility between Norwegians (be∇ and os∇) and Danes (co□ and år□), between South Africans (af▼) and Dutchmen (nl■) and between South Africans (af●) and Frisians (fr■) cannot be explained by differences in phonetic distance scores. In these cases lower intelligibility scores do not correspond with larger phonetic distances.

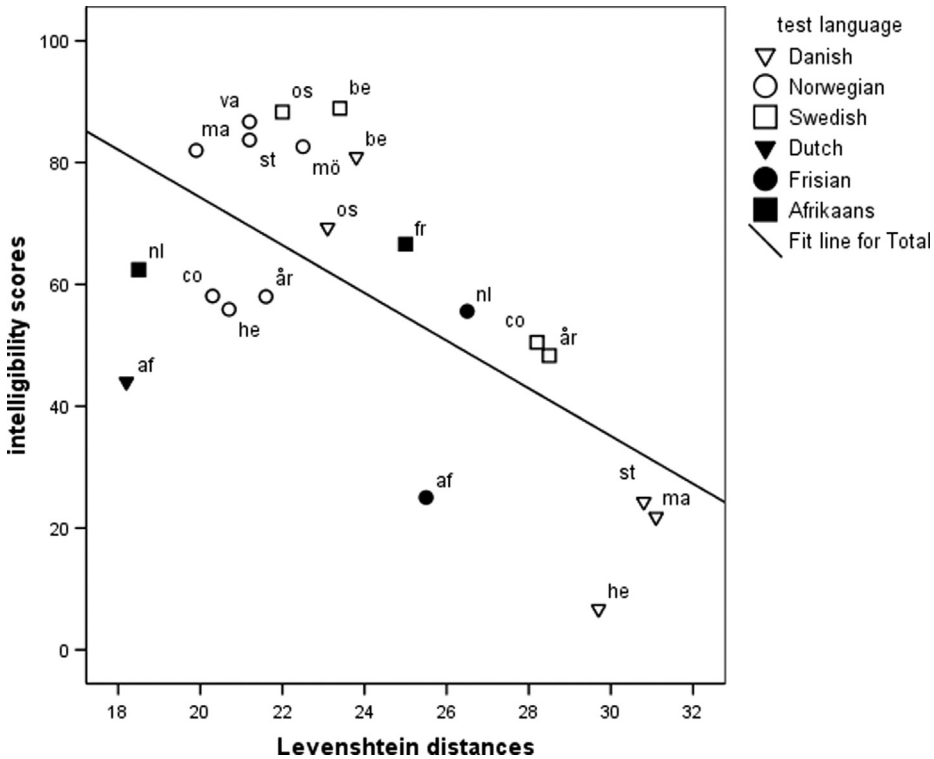

Figure 4 Scatterplot showing the relationship between the intelligibility scores in Table 3 and the phonetic distances in Table 5 ($r = -0.64$). The meaning of the abbreviations can be found in Table 3

Correlation between intelligibility scores and lexical distances

In addition to phonetic differences, lexical differences can also be expected to contribute to intelligibility. In Figure 5, a scattergram is presented which shows the relationship between intelligibility scores and lexical distance expressed as the percentage of non-cognates. The correlation between intelligibility scores and percentage of non-cognates is low and not significant ($r = -0.36$, $p = 0.11$) when all groups are included and a little higher but still not significant ($r = -0.42$, $p = 0.11$) when only Scandinavian groups are included. The low correlation is especially due to the small variation in lexical distances within the Scandinavian language area (values between zero and 3.6, see Table 6).

As far as the West Germanic languages are concerned there is a larger variation in lexical distances. For this group, there seems to be a relationship between lexical distance and intelligibility (see also Van Bezooijen & Gooskens, 2005). Dutch listeners are confronted with fewer non-cognates (6.6%, see Table 6) when listening to Afrikaans than the other way round (16.8%). This difference corresponds with asymmetrical intelligibility scores that are higher for the Dutchmen (62.4%) than for the South Africans (44.0%). The Frisians are confronted with a large number of non-cognates when

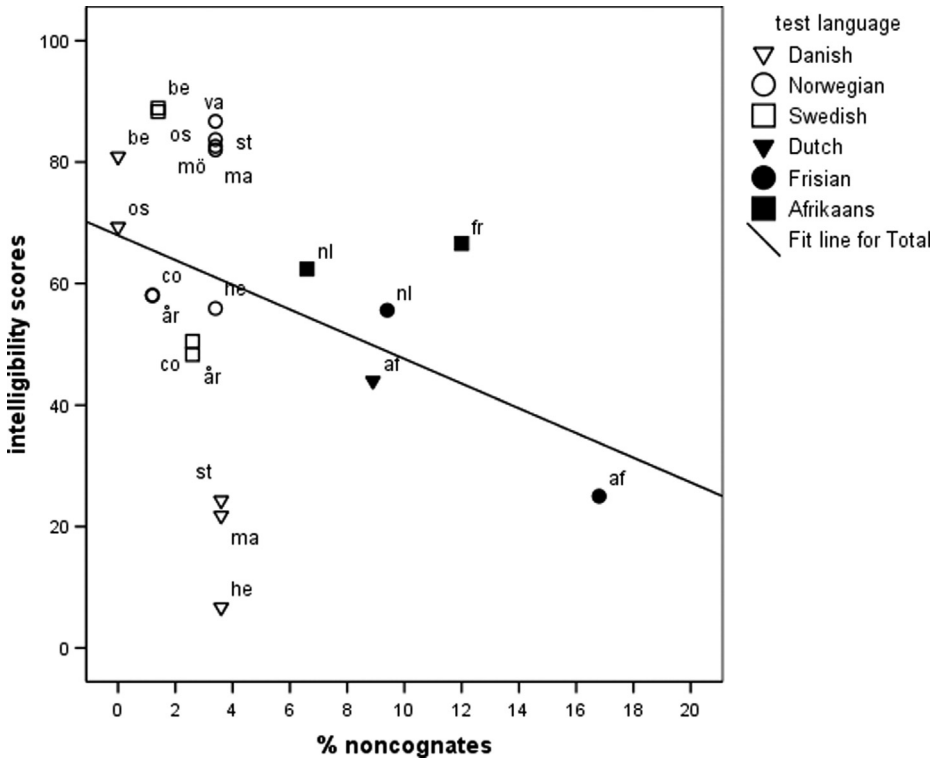

Figure 5 Scatterplot showing the relationship between the intelligibility scores in Table 3 and the percentage of non-cognates in Table 6 ($r = -0.36$). The meaning of the abbreviations can be found in Table 3

hearing Afrikaans (12.0%), but still perform better than the South Africans who are confronted with an even larger number of non-cognates (16.8%).

Explanations of the results

The correlation between intelligibility scores and lexical scores was low and not significant (see second part of fifth section). This is probably due to the fact that the effect of lexical differences is difficult to predict. One single non-cognate word in a sentence or even a larger part of a text can lower intelligibility considerably if the non-cognate word is a central concept. On the other hand, if the non-cognate words in a text have little semantic content, intelligibility is less heavily influenced. Furthermore, not all non-cognates are necessarily unintelligible. Foreign words (from for example English or Latin decent) may help to facilitate mutual understanding. The first part of the fifth section showed that there is a significant relationship between intelligibility and phonetic distances. However, as only 41% of the variance is explained, explanations should be considered for the residuals found in Figure 4.

Two groups of listeners performed considerably less well on the intelligibility test than expected on the basis of phonetic distance, namely the South Africans listening to Dutch and to Frisian and the listeners from Helsinki in

Finland listening to Danish and to a smaller extent to Norwegian. It seems reasonable to conclude that part of the explanation for the fact that the South Africans understand Frisian and Dutch less well than expected from the phonetic distances may be the large percentage of non-cognates that the South Africans are confronted with when listening to these two languages.[11] However, as far as the Helsinki results are concerned, the explanation cannot be found in the lexical distances, as the lexical distances are not larger than for the other Swedish-speaking listeners. The most likely explanation for the low performance of the Helsinki group is the fact that they are often bilingual (see part on 'Listeners' in the second section) and that their Swedish language competence is low due to the strong position of the Finish language in Helsinki. A study by Leinonen and Tandefelt (2007) showed that Finland–Swedish high school students from southern Finland had lower Swedish language proficiency than students from the western part of Finland. Especially their knowledge of idiomatic expressions and stylistic variation was poorer. In Southern Finland, Swedish is only spoken by a small minority and many Finland–Swedes are bilingual, while in the western part, represented by Vaasa in the present investigation, the language situation is more strongly dominated by Swedish. In Mariehamn most inhabitants speak only Swedish in everyday life.[12]

In contrast with the listeners from Helsinki, bilingualism seems to be an advantage for the Frisians. They can make use of their knowledge of both Dutch and Frisian when trying to understand Afrikaans while the listeners from Helsinki have no help from their knowledge of the genetically unrelated Finnish. This is probably an explanation for the fact that the Frisians have better test results for Afrikaans than could be expected from both the phonetic and the lexical distances. The percentage of correct answers is even higher than for the Dutch listeners.

The Norwegian listeners from Bergen and Oslo listening to Swedish and listeners from Bergen listening to Danish perform considerably better than could be expected from the phonetic distance score. As became clear in the second part of the fifth section, the lexical distances within the Scandinavian group are small and show little variation within the Scandinavian area, and the lexical distances do not seem to explain the differences within the Scandinavian language area. In the literature, the good results of Norwegians are often explained by another kind of bilingualism than that found in Finland and Friesland. Norwegian dialects hold a strong position and are widely used in both formal and informal situations. For this reason Norwegians are used to listening to speakers with different dialectal backgrounds. This may make it easier for them to understand yet another Scandinavian variety even though it is not Norwegian.

So far linguistic distances and different kinds of bilingualism have been mentioned as predictors of intelligibility. Other factors that are often mentioned as important predictors of intelligibility are contact and attitude. As mentioned in the introduction, a direct relationship between these factors and intelligibility has not been found for the data presented in this paper (see Gooskens, 2006; Gooskens & Van Bezooijen, 2006). Still, contact and attitude

are likely to be important factors in addition to linguistic distances in real-life settings and more research is needed in this area.

Conclusions and Discussion

The results of the present investigation show that intelligibility can vary considerable both between and within language areas. The results from the intelligibility tests in Scandinavia confirm the results of previous investigations. Swedish-speaking listeners have the greatest difficulties understanding Danish. Danish listeners have fewer difficulties with Swedish, though the percentages of correct answers are still low for this group. Swedish–Norwegian mutual intelligibility is high. Even though semicommunication is more widely used in the Scandinavian language area than between speakers of Dutch, Afrikaans and Frisian, the results show that the mutual intelligibility between speakers of these West Germanic languages is higher than between Danish and Swedish. In the West Germanic group, the largest problems are found when South African listeners are confronted with Frisian and Dutch but they still have better test results than when Danes and Swedes are confronted with each other's languages.

In order to look for linguistic explanations for the intelligibility scores, phonetic and lexical distances were measured. Distances expressed in percentages can now be specified in a schematic presentation of the linguistic distances between the Scandinavian languages shown in Figure 1. In Figure 6a, the mean distances between the Scandinavian languages (leaving out Finland Swedish for the sake of comparability with Figure 1) at the lexical and the phonetic levels are added. From this figure it becomes clear that in contrast to the impression given in Figure 1, the phonetic distance between Norwegian and Danish is not larger than between Norwegian and Swedish. As expected, there are hardly any lexical differences between Norwegian and Danish while some differences are found between Norwegian and Swedish. This means that neither the phonetic distances nor the lexical distances can explain why Norwegians understand Swedish better than Danish (see Table 3). Future research will hopefully result in refinements of the distance measurements and

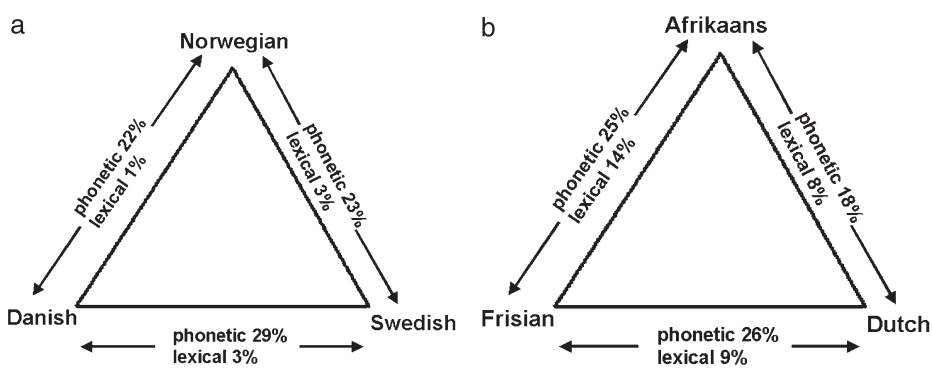

Figure 6 Mean phonetic and lexical distances between each language pair within the Scandinavian (a) and West Germanic language group (b)

the development of measurements at other linguistic levels that can explain such unexpected results. The influence of extra-linguistic factors such as contact, language instruction and attitude should also not be neglected. The distance between Danish and Swedish is largest at the phonetic level. A comparison with the distances found for Dutch, Frisian and Afrikaans (Figure 6b) makes clear that the lexical distances are relatively small in Scandinavia. We find large lexical distances between all three languages in the West Germanic language group, especially between Frisian and Afrikaans, while the phonetic distances are lower than between Danish and Swedish or, in the case of Afrikaans and Dutch, are even lower than between any Scandinavian language pair.

The main purpose of this investigation was to explore the relationship between linguistic distances and intelligibility. Correlations between the intelligibility scores and linguistic distance scores showed that intelligibility to a large extent can be predicted by phonetic distances. Even the asymmetrical intelligibility between Danish and Swedish which has also been observed in previous investigations seems, at least to some extent, to be due to different phonetic distances between the test language and the language of the listeners. Lexical distances between the Scandinavian languages are small and show little variation. Therefore, mutual intelligibility in this language area can hardly be predicted from the lexical distances. On the other hand, the lexical distances between Dutch, Frisian and Afrikaans are larger and the asymmetrical Dutch–Afrikaans intelligibility should probably to a large extent be attributed to differences in lexical distances. The large lexical distances between the languages may be part of the explanation for the fact that semicommunication is not widely used between speakers from this language group in spite of the fact that the phonetic distances are not larger than in the Scandinavian area. Another explanation might be that semicommunication is not institutionalised and promoted in the same way as in Scandinavia.

The present investigation has shown that it is important to include different varieties when investigating the importance of linguistic distances for the intelligibility. However, on the basis of the present data it is only possible to get an overall impression of the role of phonetic and lexical distances for the intelligibility. In future research, experiments are planned which will give a more precise picture of the relative contribution of different linguistic levels, including the prosodic, morphological and syntactical levels, to the intelligibility of closely related language varieties. Furthermore the intelligibility of more different varieties of the Scandinavian languages will be tested and a more detailed analysis of the relationship between intelligibility and linguistic distances will be carried out with the aim to pinpoint the linguistic factors that form the largest obstacle in the communication between different groups of Scandinavians. As far as the phonetic distances are concerned, more sophisticated measures will be developed that are able to express the fact that for example consonants are more important for decoding cognates than vowels and that not all phonotactic positions are of equal importance for understanding. The onset is clearly the most important position at least within the Germanic language family. The lexical measurements could also be improved. For example, the knowledge of other languages, for example

English, should be taken into account. Similarly, false friends, i.e. pairs of words in two languages that sound similar, but differ in meaning, should be incorporated into the lexical distance measures. The text used for the present investigation consisted of only 256–290 words. More reliable lexical distance measurements will be achieved by including a larger number of words. Also, at other linguistic levels more advanced measurements will be developed in the future.

Acknowledgements

I am grateful to the Nordic Cultural Fund for their permission to use the results from the INS-investigation and in particular to Lars-Olof Delsing from the University of Lund who has been very helpful in providing me with parts of the database. I furthermore thank the Gratama-fonds for funding the collection of additional material and the phonetic transcriptions and Andreas Vikran and Jørn Almberg for making the phonetic transcriptions of the texts. Finally, I thank three anonymous reviewers for their useful comments on an earlier version of this paper.

Correspondence

Any correspondence should be directed to Charlotte Gooskens, Department of Scandinavian Studies, University of Groningen, Postbus 716, 9700 AS Groningen, The Netherlands (c.s.gooskens@rug.nl).

Notes

1. INS is short for *Internordisk sprogforståelse i en tid med øget internationalisering* – 'Inter-Nordic communication in an era of increasing internationalisation'.
2. The INS-investigation was more extensive than previous investigations on mutual intelligibility in Scandinavia. In contrast with earlier investigations it included the testing of the intelligibility of the three Scandinavian languages in all Nordic countries. It tested reading and listening comprehension of both adolescents and adults with different levels of education and with different language backgrounds.
3. 'Neighbouring language' is the translation of the Scandinavian *nabosprog/grann(e)språk* and refers to the two Scandinavian languages spoken in the other Scandinavian countries. For example, the neighbouring languages of a Norwegian person are Swedish and Danish. Note that only Swedish-speaking subjects were tested in Finland.
4. In the INS-investigation, listeners from two different levels of education were tested (practical and theoretical educations). Listeners from the practical educations were excluded from the present investigation in order to match the background of the listeners from the two language groups as well as possible. Danish was in fact tested in Malmö in the INS-investigation, but only data from listeners attending practical educations were available when the present investigation was carried out. In Gooskens (submitted), the intelligibility of different Nordic dialects including Southern Swedish among Danish listeners was investigated. Southern Swedish was more difficult to understand than Standard Swedish, probably due to the deviant vocabulary.
5. A second text about counting frogs to determine the quality of the environment was used as well in the INS-investigation. However, this text turned out to be very difficult, probably due to the fact that the word for frog is different in each language and due to the abstract nature of the text. Even when tested for their own native language, the listeners had low scores on this test.

6. The questionnaire also included questions about attitude towards and contact with the other languages. However, as no clear effect of attitude and contact on the intelligibility scores was found (see Gooskens, 2006; Van Bezooijen & Gooskens, 2005), the results of these questions will not be dealt with here. Another listening comprehension test preceded the test and it was succeeded by a reading test. However, these tests were not included in the present investigation.
7. The mean degree of understanding is much lower in the INS-investigation due to the fact that the results from listeners attending practical educations were included. These listeners performed less well than the listeners attending pre-university education. Furthermore the results in the INS-investigation are based on two texts which also resulted in lower results (see note 6).
8. In Delsing and Lundin Åkesson (2005) this asymmetry was not found due to the bad performance of the listeners attending the practical education in Copenhagen. They attribute this to a lack of interest in the investigation among these listeners (p. 146). It should also be remembered that no listeners from the geographically close Malmö were tested for Danish. The difference would probably have been smaller if such a group had been included. For a further discussion of the results, see Delsing and Lundin Åkesson (2005).
9. In the present investigation it is not taken into account that some listeners may in fact be able to use phonetic information from more than one language variety (for example, their own variety as well as the standard language) when listening to the neighbour language.
10. See http://www.phon.ucl.ac.uk/home/sampa/.
11. The South Africans also perform less well than the Dutch and the Frisian subjects when tested in their own language. This shows that part of the low scores for the South Africans may be explained by non-linguistic factors such as a lack of interest or experience. However, Gooskens and Van Bezooijen (2006) showed that there was no relationship between the results for their native language and for the test language. Some listeners performing well for their native language had low intelligibility scores for Dutch or Frisian, while listeners who had low scores for their own language performed well on the second-language test. For this reason extra-linguistic factors do not seem to be the only explanations for the low performance of the South Africans.
12. The Swedish proficiency of the listeners from Mariehamn was just as high as that of the Swedish listeners. Unfortunately the Swedish proficiency of the listeners from Helsinki and Vaasa was not tested in the present part of the investigation. However, in another part of the test, where the listeners had to answer questions about a video recording, the listeners from Helsinki had considerably lower scores than the Swedes. Also the listeners from Vaasa had lower scores while the Mariehamn listeners had just as many correct answers as the Swedes (see Delsing & Lundin Åkesson, 2005 for the precise results).

References

Andersson, E. and Reuter, M. (1997) Finlandssvensk språkvård som minoritetsstrategi [Fenno-Swedish language preservation as a minority strategy]. In S. Løland, A.M. Gustafsson, K. Arnason and B. Lindgren. (eds) *Språk i Norden* (pp. 78–92). Oslo: Novus.

Bø, I. (1978) *Ungdom og naboland* [*Youth and Neighboring Country*]. Stavanger: Rogalandsforskning (rapport 4).

Börestam, U. (1987) *Dansk-svensk språkgemenskap på undantag* [*Danish–Swedish Language Community as a Special Case*]. Uppsala: Uppsala Universitet.

Braunmüller, K. and Zeevaert, L. (2001) *Semikommunikation, rezeptive Mehrsprachigkeit und verwandte Phänomene. Eine bibliographische Bestandsaufnahme* [*Semicommunication, receptive multilingualism and related phenomena. A bibliographical overview*]. Arbeiten zur Mehrsprachigkeit – Folge B, nr. 19. Universität Hamburg.

Delsing, L-O. and Lundin Åkesson, K. (2005) *Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska* [*Does the Language Keep the Nordic Countries Together? A Research Report on How Well Young People Understand Danish, Swedish and Norwegian*]. Copenhagen: Nordiska ministerrådet.

Gooskens, C. (2006) Linguistic and extra-linguistic predictors of inter-Scandinavian communication. In J. van de Weijer and B. Los (eds) *Linguistics in the Netherlands* 23 (pp. 101/113). Amsterdam: John Benjamins.

Gooskens, C. (submitted) Internordisk sprogforståelse i et dialektperspektiv [Internordic intelligibility in a dialect perspective]. In P. Widell and U. Dalvad Berthelsen (eds) *11. Møde om Udforskningen af Dansk Sprog Århus 2006.*

Gooskens, C. and van Bezooijen, R. (2006) Mutual comprehensibility of written Afrikaans and Dutch: symmetrical or asymmetrical? *Literary and Linguistic Computing* 23, 543–557.

Gooskens, C. and Heeringa, W. (2004a) The position of Frisian in the Germanic language area. In D. Gilbers, M. Schreuder and N. Knevel (eds) *On the Boundaries of Phonology and Phonetics* (pp. 61–87). Groningen: University of Groningen.

Gooskens, C. and Heeringa, W. (2004b) Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16 (3), 189–207.

Haugen, E. (1966) Semicommunication: The language gap in Scandinavia. *Sociological Inquiry* 36, 280–297.

Heeringa, W. (2004) Measuring dialect pronunciation differences using Levenshtein distances. Groningen: Groningen dissertations in linguistics (Grodil).

Leinonen, T. and Tandefelt, M. (2007) Evidence of language loss in progress? Mother tongue proficiency among students in Finland and Sweden. *International Journal of the Sociology of Language* 187, 185–204.

Maurud, Ø. (1976) *Nabospråksforståelse i Skandinavia: en undersøkelse om gjensidig forståelse av tale- og skriftspråk i Danmark, Norge og Sverige* [*Mutual Intelligibility of Neighbouring Languages in Scandinavia. A Study of the Mutual Understanding of Written and Spoken Language in Denmark, Norway and Sweden*]. Stockholm: Nordiska rådet.

Torp, A. (1998) *Nordiske språk i nordisk og germansk perspektiv* [*Nordic Languages in a Nordic and Germanic Perspective*]. Oslo: Novus forlag.

Van Bezooijen, R. and van den Berg, R. (1999a) Verstaanbaarheid van het Gronings, Fries, Limburgs en West-Vlaams: Waar zitten de problemen? [Intelligibility of the Groningen, Frisian, Limburg and West-Flemish language varieties: Where are the problems?]. *Artikelen van de Derde Sociolinguïstische Conferentie* (pp. 49–60). Lunteren. Delft: Eburon.

Van Bezooijen, R. and van den Berg, R. (1999b) Taalvariëteiten in Nederland en Vlaanderen: hoe staat het met hun verstaanbaarheid? [Language varieties in the Netherlands and Flanders: How well are they understood?]. *Taal en Tongval* 51, 15–33.

Van Bezooijen, R. and van den Berg, R. (1999c) Word intelligibility of language varieties in the Netherlands and Flanders under minimal conditions. In R. van Bezooijen and R. Kager (eds) *Linguistics in the Netherlands* (Vol. 16, pp. 1–12). Amsterdam: John Benjamins.

Van Bezooijen, R. and van den Berg, R. (2000) Hoe verstaanbaar is het Fries voor niet-Friestaligen? [How well can non-Frisians understand Frisian?]. *Filologia Frisica* (pp. 9–26). Ljouwert: Fryske Akademy.

Van Bezooijen, R. and Gooskens, C. (2005) How easy is it for speakers of Dutch to understand spoken and written Frisian and Afrikaans, and why? In J. Doetjes and J. van de Weijer (eds) *Linguistics in the Netherlands* (Vol. 22, pp. 13–24). Amsterdam: John Benjamins.