Registration form (basic details)

1a. Details of applicant

-Name, title(s): dr. Charlotte Gooskens
-Male/female: female

-Address for correspondence: Rijksuniversiteit Groningen
Fac. der Letteren/Scandinavistiek
Postbus 716
9700 AS Groningen

-Preference for English correspondence: no
-Telephone: 050 363 58 27
-Fax: 050 363 58 21
-E-mail: c.s.gooskens@let.rug.nl
-Website (optional):

-Doctorate (date): 28 November 1997
-Use of extension clause (see Notes): no

1b. Title of research proposal

Linguistic determinants of mutual intelligibility in Scandinavia

1c. Summary of research proposal

298 words

The three mainland Scandinavian languages, i.e. Danish, Swedish and Norwegian, have a reputation of being mutually intelligible, which means that the speakers are able to communicate each using his or her language. However, in daily practice inter-Scandinavian communication sometimes fails. The results of a number of studies have shown that especially Danes and Swedes have difficulties understanding each other's language. The problems are commonly explained by extra-linguistic factors such as linguistic experience and language attitude. Linguistic explanations have mostly been neglected due to the lack of a suitable method for quantifying linguistic distance. Recently, such methods have been developed. The aim of the present project is to use these newly developed methods and refine them in order to be able to measure communicatively relevant linguistic distances among the spoken Scandinavian languages. On the basis of these measurements, a model will be developed that explains mutual intelligibility in Scandinavia.

First, the model will be developed and tested on adults. However, in order to be able to exclude the influence of extra-linguistic factors such as attitude and experience, the model will be tested on children as well, assuming that most of the extra-linguistic factors have little or no influence on intelligibility among children. In addition, the experiments with children will give new insights into the way children deal with the comprehension of closely related languages, an area that has received little attention so far.

As a point of departure the model will be developed for the standard Scandinavian languages. In a later stage Scandinavian dialects will be included. As the model is expected to be applicable to combinations of closely related languages and languages varieties outside Scandinavia as well, the results will increase our general understanding of the role of linguistic distance in the mutual intelligibility of closely related languages.

Keywords: mutual intelligibility, semi-communication, linguistic distances, Levenshtein distances, speech perception

1d. NWO Council area
GW

1e. Host institution (if known)
RUG

Research proposal

2. Description of the proposed research
3943 words

2a. Research topic

In recent years dialectometric methods have been developed for measuring linguistic distances between dialects and closely related languages. The results of these distance measurements have been used for the classification of dialects within a language area. The aim of the proposed project is to extend the domain of application of these newly developed methods and to refine them in order to develop a model that can predict and explain mutual intelligibility in Scandinavia.

In the literature the three mainland Scandinavian languages, i.e. Danish, Swedish and Norwegian, are often mentioned as examples of so-called 'Ausbausprachen', that is, languages which are separate standard languages only for political and historical reasons. Linguistically they are so closely related that they may be considered as dialects of one language (Kloss 1967). However, the fact that there is no common Scandinavian standard language makes the communication between Scandinavians different from communication between speakers of, for example, Dutch dialects. In the Dutch language area speakers of different dialects in general turn to Standard Dutch when they talk to each other. When speakers of the Scandinavian languages meet, they usually communicate in their own language. Haugen (1966) introduced the term 'semi-communication' for this kind of communication between mutually intelligible languages.

Speakers of the Scandinavian languages are strongly encouraged by the Scandinavian authorities to use their own language rather than a *lingua franca* such as English when communicating with other Scandinavians. 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 the Scandinavian languages (e.g. Maurud 1976, Bø 1978, Börestam 1987). Recently, an investigation supported by the Nordic Cultural Fund was initiated to examine the communicative situation at the beginning of the 21st century (*Internordisk sprogforståelse i en tid med øget internationalisering* - Inter-Nordic communication in an era of increasing internationalisation). The results invariably show that especially the spoken communication between Swedes and Danes is problematic. Moreover, it is a general impression that mutual intelligibility has deteriorated over the last few decades and that English is increasingly adopted as *lingua franca.*

The level of inter-Scandinavian intelligibility may depend upon three factors: (1) attitude towards the languages of the other Scandinavian countries, (2) experience with the other languages (including formal instruction), and (3) linguistic distance between the languages. The studies mentioned above included the first two factors. Furthermore, a number of contrastive descriptions have been made of the linguistic differences between the three Scandinavian languages. On the basis of these descriptions, various isolated linguistic phenomena have been pointed to as part of the reason for problems in semi-communication. However, due to the absence of a suitable method, no substantial

attempt has been made so far to measure the overall linguistic distance between the Scandinavian languages and investigate the role of these distances for inter-Scandinavian intelligibility. The aim of the proposed project is to use newly developed methods for measuring linguistic distances and refine them in order to measure communicatively relevant linguistic distances between the spoken Scandinavian languages at different linguistic levels. These linguistic distance measurements will be used to develop a model for predicting and explaining inter-Scandinavian intelligibility. The model will also be tested on children in order to be able to exclude extra-linguistic factors which might influence intelligibility such as attitude, knowledge of other Scandinavian and non-Scandinavian languages, and knowledge of orthographic correspondences.

The research questions can be formulated as follows:

1. What linguistic distances can be found between the spoken Scandinavian languages at different linguistic levels (phonetics, vocabulary, morphology)?

2. To what extent are the linguistic distances on the different linguistic levels predictors of inter-Scandinavian intelligibility?

So far, not much research has examined which linguistic factors contribute to the mutual intelligibility of closely related languages. In the proposed study, methods for measuring communicatively relevant linguistic distances will be developed. In this way linguistic distance measures, which were originally developed for the classification of dialects, will be used for the first time to systematically predict the relative importance of different linguistic factors for the mutual intelligibility of closely related languages. The investigation will be carried out in the Scandinavian language area but the results will contribute to the general understanding of the relative importance of different linguistic levels for mutual intelligibility between closely related languages and language varieties (for example the Dutch/Frisian/German languages or the Romance languages). Furthermore, insight will be gained into how children cope with the understanding of closely related languages and language varieties.

Knowledge about the linguistic determinants of mutual intelligibility is useful for language planning at the national and European levels. If the smaller languages are to survive in a European context, it is important to gain knowledge about the mechanisms involved in using one's own language for communication with speakers of other, closely related European languages. Knowledge of the role of different linguistic levels for mutual intelligibility is also useful for didactic purposes. It will make it easier to give specific instructions to people trying to gain the necessary passive knowledge needed to understand a language.

2b. Approach

The investigation consists of two projects. In Project A a model of linguistic determinants of mutual intelligibility in Scandinavia will be developed by the applicant in cooperation with a postdoc. In Project B a PhD-student will test the model on children. For this reason Project B will start one year later than Project A.

Project A: developing a model of linguistic determinants of mutual intelligibility

Intelligibility tests
Mutual intelligibility among speakers of the three Scandinavian languages will be tested by means of open questions about a coherent spoken text as well as by translations of isolated spoken words. The texts and the isolated words are read aloud by speakers of the three Scandinavian languages. By using a coherent text a realistic image of the

communicative possibilities in daily life is tested. Isolated words give the possibility to pinpoint exactly which sounds cause the communicative problem and to test the significance of the effect statistically. The percentage of correct answers and the percentage of correct translations will form the dependent variables against which the independent variables (linguistic distances on different linguistic levels) will be tested.

Linguistic distance measurements

Phonetic transcriptions of the texts and the isolated words from the intelligibility tests form the corpus on the basis of which the linguistic distances are calculated. For each language pair, the test words of each language will be aligned so that it will be possible to calculate the linguistic distance per word pair.

The Levenshtein algorithm has been used successfully for measuring distances between dialects (Heeringa 2004). The distance per word pair in a corpus is calculated by means of the minimum number of insertions, deletions and substitutions of phonetic segments needed to transform the word in one language into the other, whereby word length is taken into account. The distance between two languages is the mean of all word pair distances. Application of Levenshtein distance measurements to a corpus will yield a measure of differences between the languages including phonetic, morphological and lexical information. This general distance measure may be correlated with the intelligibility scores. However, results of studies on mutual intelligibility between speakers of Dutch, Frisian and Afrikaans by Gooskens and Van Bezooijen (in progress) show that such a measure is too rough to be a good predictor of intelligibility. It is necessary to calculate linguistic distances at different linguistic levels separately. Below it is shown how linguistic distance can be calculated at the lexical, morphological and phonetic level. At each of these levels both the frequency and the nature of the correspondences will be incorporated into the distance measurements.

At the **lexical** level the percentage of non-cognates expresses the linguistic distance. Not all non-cognates will be equally difficult to understand if the listener has some experience with the language. Frequent words and words which are related to an equivalent in another familiar language (for example French or English) are expected to be easier to understand than infrequent words or words which the listener does not know from another language. For this reason word frequency information will be incorporated into the lexical distance measure as well as information about the nature of the correspondences. Also, different word classes will be analysed separately.

Phonetic distances will be measured for the cognates only. Distances can be measured for whole words but distances will also be measured separately for free and bound morphemes thus creating distance measures at the **morphological** level. Also at the morphological level the effect of frequency and nature of the morpheme will be assessed. Frequent morphemes are likely to be easier to understand than infrequent morphemes. Likewise morphemes which are not related to the same morphemes in the mother tongue are expected to make the meaning less transparent. Also, different categories of morphemes will be analysed separately.

The distances at the **phonetic** level are first measured by means of Levenshtein distances as developed by Heeringa (2004). However, these distances cannot be expected to be perfect determinants of intelligibility, since related languages show patterns of regular phonetic correspondences which are not captured by the Levenshtein measurements but which might facilitate intelligibility. In the same way as for the lexical and the morphological level, frequency as well as the nature of the phonetic correspondences will therefore be incorporated into the phonetic distance measurements. This will be realised as follows.

Related languages show regular correspondences between phonemes. An example is Danish [d], which often corresponds with [t] in medial position in Swedish. Listeners are likely to use such correspondences when listening to a closely related language. Languages with a limited number of regular phonetic correspondences which cover a large part of the distances can be expected to be easier to understand than languages

which need a larger number of correspondences to cover the same percentage of the distances. This might be part of the explanation for the fact that Gooskens and Heeringa (2004) found larger distances between the Scandinavian languages than between Dutch and Frisian while mutual intelligibility seems to be more successful in Scandinavia than in the Dutch-Frisian situation (Van Bezooijen and Gooskens, submitted). Perhaps the Dutch-Frisian correspondences are less regular than the Scandinavian ones.

In addition to the frequency of the correspondences, the nature of the correspondences is also likely to contribute to the degree of intelligibility. Transparent correspondences will increase intelligibility while intransparent correspondences will yield a barrier. It may be expected that the distance between phonemes is decisive for the degree of transparency. For example, it is probably easier for a listener to guess that [p] in another language corresponds with [b] in his own language than with [l], [p] and [b] being phonetically closer than [p] and [l]. The Levenshtein distances have already been refined by calculating the distance on the basis of the number of differences in phonetic features between phonemes (Heeringa 2004). Levenshtein distances calculated on the basis of features might be a better determinant of intelligibility. However, these features are based on articulatory characteristics and their relation to the communicative situation has not been investigated. The communicative reality of phonetic correspondences will therefore be investigated in a number of listening experiments. For example, if a Danish listener often thinks that Swedish [a] corresponds with Danish [e] and not so often with [i], we would conclude that the distance between [e] and [a] is smaller than between [e] and [i]. Probably the weights should be counted differently depending on the phoneme position in the word.

Linguistic transparency between languages may be asymmetric at all linguistic levels. For example, Norwegian might have two synonyms for a concept, which has only one equivalent in Swedish. This will make it difficult for a Swedish listener to understand the non-cognate Norwegian equivalent while the Norwegian listener has no problem understanding the cognate Swedish word. Likewise phonetic or morphological transparency might be asymmetric. For this reason distances should be calculated in both directions for each pair of languages so that the distances reflect transparency as completely as possible.[1]

On the basis of the distance measures it is possible to calculate the relative contribution of the three linguistic levels (lexicon, morphology and phonology) to intelligibility. In the literature, predictions about the relative contribution of different phenomena to intelligibility have often been made. For example, Elert (1981) predicts that content words are more important than function words, that consonants are more important than vowels, and that consonants at the beginning of words are more important than consonants in the middle or at the end of words. Such predictions can be tested experimentally on the basis of our material. On the basis of the distance measures, multiple regression analyses will be carried out with combinations of different linguistic variables such as percentage of non-cognates, percentage of function words and content words, Levensthein distances for consonants and vowels, Levenshtein distances for free and bound morphemes and position in the word as independent variables and intelligibility scores as dependent variable. The outcome will be an evaluation of various factors as predictors of mutual intelligibility among the three standard Scandinavian languages.

Our point of departure for developing the intelligibility model is the Levenshtein distance. A number of researchers have developed algorithms for historical language reconstruction, which also involve determination of phonetic correspondences (see Chapter 2 in Kondrak 2002 for an overview of such algorithms). Even though the purpose of these algorithms was quite different, it might turn out to be fruitful to incorporate

[1] Note that the meaning of 'distance' is broadly defined here, since distances cannot be asymmetric in the strictly mathematical sense of the word.

aspects of such algorithms in our model. Cheng (1996) developed a system for measuring mutual phonological predictability of words of two varieties incorporating regular sound correspondences, but he never tested his findings experimentally. This step will be taken in the proposed project. For the measurements of the lexical distances the method which is developed by and Speelman, Grondelaers and Geraerts (2003) might be an important addition to the model.

Initially, syntactic and prosodic differences will not be included in the investigation since they are likely to play a minor role for the mutual intelligibility of the Scandinavian languages. However, at a later stage it might be interesting to incorporate these linguistic levels into the model as well. This will make it possible to apply the model to other groups of closely related languages, for which the syntactic and prosodic levels are expected to be more important.

Little is known about how well Scandinavians understand different Scandinavian dialects. Especially in the case of Norwegian this is a shortcoming since almost all Norwegians speak their own dialect in both formal and informal situations. In a later stage of the investigation, different Scandinavian dialects will therefore be added. To this end speech material from a number of dialects from different Scandinavian dialect groups will be collected. On the basis of this material an intelligibility test will be developed. One well-defined listeners group (for example speakers of Standard Danish) will participate in this test. This experiment provides the possibility to investigate a larger scale of different linguistic distances at different linguistic levels and their role for intelligibility.

Project B: testing the model on children

A number of previous studies on inter-Scandinavian intelligibility demonstrated that speakers of the three languages do not understand the other Scandinavian languages equally well. These results are usually explained by extra-linguistic factors such as attitude towards the language and its speakers and experience with the language through for example the media and personal contacts. In Scandinavia these factors have also been used as part of the explanation for the fact that intelligibility is sometimes asymmetric. For example, Danes understand Swedish better than Swedes understand Danish. This has been explained by the fact that Danes have a more positive attitude towards Swedes and are more often confronted with their language through the media and on vacation than the other way around. Also, knowledge of other languages might facilitate intelligibility of the other Scandinavian languages. For example, Swedish has many French loan words, which are not found in Danish. Knowledge of French might therefore make it easier for a Dane to understand some Swedish words. Another explanation for the asymmetric intelligibility of Swedish and Danish might be found in the relationship between the written and the spoken form of the language. Spoken Swedish is close to both written Swedish and written Danish, while spoken Danish has developed away from its written form and is therefore rather distant from both Swedish and Danish in their written form. This means that Danes can understand spoken Swedish better because of its close similarity to written Danish while Swedes get less help from written Swedish when understanding spoken Danish. It is, however, important to realise that communicative linguistic distances can be asymmetric (see Project A) and can also be part of the explanation for asymmetric mutual intelligibility.

It is difficult to establish which influence extra-linguistic factors have on mutual intelligibility as compared to linguistic distances. In most Scandinavian studies (see 2a.) there seems to be no direct correlation between attitude and experience on the one hand and intelligibility scores on the other hand. Of course this does not mean that attitude and experience do not have an effect on intelligibility in an actual communicative situation. It is also difficult to establish to what degree listeners can use knowledge of other languages when listening to another Scandinavian language or whether the writing system can indeed help a Danish listener to recognise Swedish words which are pronounced very differently but spelled very similarly. In order to be able to establish

what is the best model of mutual intelligibility based solely on linguistic distances we must find a way of excluding the influence of extra-linguistic factors on intelligibility.

In Project B, which will be carried out in close collaboration with Project A, the intelligibility model will be tested on Scandinavian children around the age of 7. By testing children in addition to adults in Project A it is possible to neutralise the extra-linguistic factors to a great extent, at least if care is taken that the children have had no prior experience with foreign languages and are tested before they learn to read and write. Most children at that age have probably not yet formed strong attitudes towards the other Scandinavian languages. It can therefore be assumed that the intelligibility scores based on experiments with children must be due to linguistic distances. If attitudes, writing systems and prior experience are indeed the most important explanations for the asymmetric intelligibility scores found in previous investigations, asymmetric intelligibility scores are not likely to be found with children. If the intelligibility scores are still asymmetric, it seems reasonable to conclude that linguistic factors determine intelligibility.

The testing method will have to be adapted to the testing of children making use of the experience gained by previous research groups in this area, for example MacWhinney & Bates (see MacWhinney 2005 for an overview). First, vocabulary and sentence constructions should be adapted to the level of young children. Still, the speech material must be representative of the languages in question so that it will be possible to measure linguistic distances in the same way as for the speech material used in Project A. Second, intelligibility should be tested in a setting which is adapted to the level of the children. For example, children can be asked to translate what a 'funny' doll with a built-in audio player says, to answer questions about a story read in the test language, or to play a computer game where a number of tasks as pronounced in the test language must be carried out. The number of tasks which are carried out correctly or the number of correctly translated sentences or words will express the intelligibility level. The linguistic distances and correspondences will be measured and correlated with the intelligibility scores in the same way as in Project A

In addition to greater insight into the purely linguistic determinants of intelligibility, it is also of more general interest to see how well young children are able to understand closely related languages compared to adults. It is possible that children are more flexible than adults when it comes to understanding varieties which are only slightly different from their own language (including dialects). To my knowledge this has never been investigated so far.

2c. Innovation

Predictions about the relationship between linguistic distances at different linguistic levels on the one hand and intelligibility on the other hand will be systematically investigated for the first time by means of experimental research and well-understood techniques for assessing linguistic distance. The Levenshtein distance method which has been developed to measure dialect distances and classify dialects will be used in an innovative attempt to explain language intelligibility. In addition, the method will be refined in such a way that it expresses communicatively relevant distances. New insights will be found into the relative importance of different linguistic levels for the mutual intelligibility of closely related languages and dialects. Furthermore, the ability of children to understand closely related languages will for the first time be tested. In this way intelligibility among children can be compared with adults and purer conclusions can be drawn about the role of linguistic factors devoid of possible interference of extra-linguistic factors. The results will contribute to existing Scandinavian and international research by providing a more solid linguistic basis for the ongoing discussion about mutual intelligibility and language contact.

2e. Literature references

Bezooijen, R. van & C. Gooskens (submitted). Intertalig tekstbegrip. De begrijpelijkheid van Friese en Afrikaanse teksten voor Nederlandse lezers. *Nederlandse Taalkunde.*

Bø, I. (1978). *Ungdom og naboland*. Stavanger: Rogalandsforskning (rapport 4).

Börestam, U. (1987). *Dansk-svensk språkgemenskap på undantag.* Uppsala: Uppsala Universitet.

Cheng C.-C. (1996). Quantifying dialect mutual intelligibility. In: C.-T.J. Huang & Y.-H.A. Li (eds.). New horizons in Chinese linguistics, vol. 36 of *Studies in natural language and linguistics theory*. Dordrecht: Kluwer Academic Publishers, 269-292.

Elert (1981). C.-C. Kvantitativa mått på språklikhet. En jämförelse mellan färöiska, norska och svenska. In: C.-C. Elert (ed.). *Internordisk språkförståelse. Föredrag och diskussioner vid ett symposium på Rungstedgaard utanför Köpenhamn den 24-26 mars 1980.* Umeå: University of Umeå, 85-101.

Gooskens, C. & W. Heeringa (2004). The position of Frisian in the Germanic Language Area. In: D. Gilbers, M. Schreuder og N. van den Berg (red.) *Festschrift for Tjeerd de Graaf* 2004, Groningen: Groningen Universitet, 61-87.

Gooskens, C. & R. van Bezooijen (in progress). How easy is it for speakers of Dutch to understand spoken and written Frisian and Afrikaans, and why? *Linguistics in The Netherlands.*

Haugen, E. (1966). Semicommunication: The language gap in Scandinavia. *Social Inquiry*, 33, 66-81.

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

Internordisk sprogforståelse i en tid med øget internationalisering www.nordkontakt.nu.

Kloss, H. (1967). 'Abstand languages' and 'Ausbau languages'. In: *Anthopological linguistics* 9, 7, 29-41.

Kondrak. G. (2002). *Algorithms for language reconstruction*. Doctoral dissertation,Toronto.

MacWhinney, B. (2005). New directions in the Competition Model. In: M. Tomasello & D.I. Slobin (eds.). *Beyond Nature-nurture. Essays in honor of Elizabeth Bates.*Mahwah, London: Lawrence Erlbaum, 81-164.

Maurud, Ø. (1976). *Nabospråksforståelse i Skandinavia: en undersøkelse om gjensidig forståelse av tale- og skriftspråk i Danmark, Norge og Sverige*. Stockholm: Skandinaviska rådet.

Speelman, D., S. Grondelaers & D. Geeraerts (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37, 317-337.