Grant Application for Free Competition in the Humanities

1. Project Information

Title:
Mutual intelligibility of closely related languages in Europe: linguistic and non-linguistic determinants

In Europe a large number of languages are spoken. These languages enjoy different statuses, some are officially recognized, while others are spoken by minority populations. Respect for linguistic diversity is a core EU value but the linguistic diversity can lead to communication problems that might only be reconciled with sufficient knowledge about the language situation at hand. In 2007 the High Level Group on Multilingualism (HLGM) therefore published an overview of research topics that should be investigated to improve communication within Europe while still preserving multilingual richness. Two of these topics form the basis for the present investigation. Firstly, the HLGM notes a lack of knowledge about mutual intelligibility between closely related languages in Europe and the lack of knowledge about the possibilities for communicating through receptive multilingualism, i.e., where speakers of closely related languages each speak their own language. Secondly, the HLMG notes a need for an evaluation of the potentials and limitations of the use of English as a *lingua franca* at the European level. More knowledge is needed about how well speakers of various languages in Europe understand each other in English.

We propose a large-scale investigation of the mutual intelligibility of closely related languages within the Germanic, Slavic and Romance language families. The results will be correlated with linguistic factors, such as phonetic and lexical distances, as well as extra-linguistic factors, such as language attitudes towards and familiarity with different languages. Tests will also be carried out with English as a Lingua Franca (ELF) to compare the (mutual) intelligibility of closely related languages with the (mutual) intelligibility of ELF as spoken by the same groups of speakers.

Intelligibility, attitude and familiarity tests will be carried out by means of web-based experiments. The results will be will be made available through the internet. They will provide a basis for developing a model that explains mutual intelligibility between closely related languages. In a more general sense the results will provide a greater understanding of the robustness of the human language processing system. How deviant can a language be before it is no longer intelligible to the listener? The results will also be of great value to European policy makers. A publicly available user-friendly internet application will be developed for use by future target groups of researchers and policy makers. In this way additional languages can be tested later that were not initially included in the project.

2. Details of Applicant

Dr. Charlotte Gooskens
Scandinavian Department
University of Groningen
Center for Language and Cognition Groningen
PO Box 716
9700 AS Groningen
Phone: 050-3635827
E-mail: c.s.gooskens@rug.nl
www.let.rug.nl/gooskens

3. Co-applicant

Prof. dr. Vincent J. van Heuven
Phonetics Laboratory
Leiden University Centre for Linguistics
PO Box 9515
2300 RA Leiden
Phone: +31 71 527 2319
E-mail: v.j.j.p.van.heuven@hum.leidenuniv.nl
http://www.hum.leiden.edu/lucl/organisation/members/heuvenvjjpvan.html

4. Previous and Future Submissions

none

5. Institutional Setting

Center for Language and Cognition Groningen
University of Groningen

6. Period of Funding

1 May 2011 - 1 May 2016

7. Composition of the Research Team

position	name	affiliation
main applicant	dr. Charlotte Gooskens	Scandinavian Department University of Groningen
co-applicant	prof. dr. Vincent J. van Heuven	Phonetics Laboratory Leiden University
postdoc 1	drs. Anja Schüppert	Center for Language and Cognition Groningen University of Groningen
postdoc 2	dr. Wilbert Heeringa	Center for Language and Cognition Groningen University of Groningen
PhD 1	NN	Center for Language and Cognition Groningen University of Groningen Promotor: van Heuven/ Nerbonne
PhD 2	NN	Center for Language and Cognition Groningen University of Groningen Promotor: van Heuven/ Nerbonne
PhD 3	NN	Center for Language and Cognition Groningen University of Groningen Promotor: van Heuven/ Nerbonne
advisers	prof. dr. ir. John Nerbonne	Department of Information Science University of Groningen

	dr. Renée van Bezooijen	Center for Language and Cognition Groningen University of Groningen
	dr. Peter Houtzagers	Department of Slavic Languages and Literature University of Groningen
	dr. Bob de Jonge	Department of Romance Languages and Cultures University of Groningen
	dr. Nanna Hilton	Department of Frisian Language and Literature University of Groningen
	prof. dr. Jiří Nekvapil dr. Marian Sloboda dr. Mira Nábělková	Slovak Akademy of Sciences at the Charles University at Prague
	dr. Simonetta Montemagni	Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR), Italy
	dr. Elisabetta Carpitelli	GIPSA-lab UMR 5216 Université Stendhal-Grenoble3, Bât. E, BP 25 Domaine Universitaire de Saint Martin d'Hères 38040 Grenoble cedex 9 France

8. Structure of the Proposed Research

Subproject I:
Title: A model of mutual intelligibility of closely related languages
Appointment: postdoc
University/faculty: Faculty of Arts, University of Groningen
Supervisor: Gooskens, van Heuven
Name: drs. Anja Schüppert

Subproject II:
Title: Linguistic distance measurements between closely related languages
Appointment: postdoc
University/faculty: Faculty of Arts, University of Groningen
Supervisor: Gooskens, van Heuven, Nerbonne
Name: dr. Wilbert Heeringa

Subproject III:
Title: Mutual intelligibility in the Germanic language area
Appointment: PhD
University/faculty: Faculty of Arts, University of Groningen
Supervisor: Gooskens, van Heuven, Nerbonne
Name: NN

Subproject IV:
Title: Mutual intelligibility in the Slavic language area
Appointment: PhD
University/faculty: Faculty of Arts, University of Groningen
Supervisor: Gooskens, van Heuven, Nerbonne
Name: NN

Subproject V:
Title: Mutual intelligibility in the Romance language area
Appointment: PhD
University/faculty: Faculty of Arts, University of Groningen
Supervisor: Gooskens, van Heuven, Nerbonne
Name: NN

9. Description of the Proposed Research

1. Research topic/background

Within Europe a large number of official languages and an even larger number of minority languages are spoken. Since 2007 multilingualism has been made an EU policy area in its own right with the establishment of a separate portfolio. This is a clear indication of an increased awareness of the multilingual challenge for the EU. As a result of the growth of the EU the challenge has now reached a completely new dimension in terms of size, complexity, and policy relevance. On several occasions, Leonard Orban, the European commissioner responsible for multilingualism, has stated that multilingualism is *the* tool for creating bridges between people and that linguistic diversity will help us develop a European identity.

A high-level group on multilingualism comprising 11 experts from various fields of research was set up in response to the 2005 Commission communication 'A new framework strategy for multilingualism'. In 2007, this High Level Group on Multilingualism (HLMG) delivered its recommendations in a final report (European Commission, 2007). The report states that new knowledge, generated by scientific research, is needed to bring about improvements in the acquisition of multilingual competence and the management of multilingualism. Chapter 7 of the report identifies a number of research topics, among others research on 'Receptive multilingualism' as a communication strategy. This model of receptive multilingualism is based on the fact that some language pairs are so closely related that the speakers are able to communicate each using their own language without prior formal or informal language instruction. This strategy is widely used for communication among speakers of the three mainland Scandinavian languages, Danish, Swedish and Norwegian (Maurud 1976, Bø 1978, and Delsing & Lundin Åkesson 2005), but has received little attention outside of Scandinavia. Moreover, there is a lack of thorough scientific knowledge about the linguistic and non-linguistic factors that determine how well speakers of various closely related languages understand each other.

In her VIDI-project, the applicant has used newly developed methods for quantifying linguistic distance and refined these in order to be able to measure communicatively relevant linguistic distances among the spoken Scandinavian languages. The results show a strong predictive relationship between phonetic and lexical distances on the one hand and intelligibility on the other hand (Gooskens 2007, Gooskens, Beijering & Heeringa 2008, Kürschner, Gooskens & Van Bezooijen 2008). Impe (2010) found similar results for Dutch language varieties. However, it is clear that non-linguistic factors may also contribute to the successful communication between speakers of closely related languages (Maurud 1976, Bø 1978, Börestam 1987, and Wolff 1959). The existence of negative attitudes or social stigmas attached to languages is often seen as a potential obstruction for successful intergroup communication. Ultimately, of course, the level of intelligibility also depends on the amount of experience with the other language(s), including formal instruction.

Communication through receptive multilingualism is not exploited much outside of Scandinavia, but the Scandinavian model has proved its workability and could easily be extended to communication between speakers of other languages in Europe. The EuroCom project (http://www.eurocom-frankfurt.de/) is based on the principle of receptive multilingualism and provides strategies for understanding written Germanic, Slavic or Romance languages when the reader is already familiar with another closely related language. However, the knowledge about the degree of mutual intelligibility in Europe and its linguistic and non-linguistic bases is rather limited. In the present project we will establish the degree of mutual intelligibility of closely related languages in the Germanic, Slavic and Romance language groups in Europe. Both written and spoken intelligibility will be tested, since communication in Europe takes place in both modalities. The intelligibility scores will be explained (statistically predicted) from selected linguistic and non-linguistic factors.

Applying the methodology that has been developed for intelligibility research in Scandinavia within the applicant's VIDI-project to a different and larger data set from other language areas will shed new light on earlier generalizations and methods. So far, only lexical and phonetic distances have been studied as explanatory linguistic factors. The present investigation will be extended to include morphological and syntactic distances as well. This will allow us to apply the explanatory model to groups of closely related languages for which syntactic and morphological levels are expected to be more important for mutual intelligibility than in the Scandinavian situation. Furthermore, non-linguistic factors will be included in the investigation. The results will allow us to determine under what conditions receptive multilingualism works and what its preconditions and its limits are. This will permit us to develop a general, comprehensive and well-founded model of mutual intelligibility between closely related languages that predicts intelligibility on a theoretical basis.

The human language processing mechanism shows a remarkable robustness with respect to incomplete or unfamiliar information. Many possible features are not realized in the signal of a normal linguistic utterance; and on the meaning side too, the interpretation is highly underdetermined by the expression itself. Yet, in the normal case, understanding is not in any way hampered by this. Closely related languages show similarity with different kinds of imperfect language and therefore languages that are intelligible to various degrees form a perfect natural laboratory for the investigation of the robustness of languages. It will give us an answer to the question how corrupt (or deviant) a language can be before it is no longer intelligible.

A second research topic identified by the HLGM is the potential and limitations of the usage of English as a lingua franca (ELF) at the European level. In most cases, ELF is a 'contact language' between persons who share neither a common native tongue nor a common (national) culture, and for whom English is the chosen foreign language of communication. In ELF situations, communication breakdowns can occur when one speaker uses a native speaker idiomatic expression, such as an idiom, phrasal verb, or metaphor, that the interlocutor does not know or when the pronunciation is influenced by the native language of the speaker. Many researcher have been concerned with the description of the characteristics of ELF. However, only few studies actually tested the intelligibility of ELF. Exceptions are Van Heuven & Wang (2007) who tested the mutual intelligibility of American, Chinese and Dutch accented English and found that listeners understand English better when the accent of the speaker is the same as that of the listener (the interlanguage speech intelligibility benefit, cf. Bent & Bradlow 2003). Bent & Bradlow (2003) even claim that the interlanguage intelligibility benefit extends to the situation where the non-native talker and listeners come from different language backgrounds ('mismatched interlanguage speech intelligibility benefit'). Crucially, it is currently unknown whether it is easier or more difficult to understand (non-shared) accented English than the corresponding closely related language. In addition to the mutual intelligibility of closely related languages in Europe, the project will therefore test informants' degree of intelligibility of accented English as well. The results will enable us to compare the mutual intelligibility in English with that of closely related languages. The outcome of this comparison will provide valuable information for policy makers in their decision on how to improve communication between speakers of different languages in Europe.

Previous studies have indicated that mutual intelligibility between language pairs is sometimes asymmetric. Asymmetry has been observed between many language pairs, for example between Spanish and Portuguese (Jensen, 1989) and between Czech and Slovak (Budovičová, 1987). The best-documented case of asymmetric intelligibility is the Danish-Swedish mutual intelligibility. Danes understand Swedish better than Swedes understand Danish (Gooskens, Van Heuven, Van Bezooijen & Pacilly, in press).

These results are usually explained by extra-linguistic factors such as asymmetric attitudes towards the (speakers of the) languages involved and unequal experience with the languages. Danes have a more positive attitude towards Swedes and are more often confronted with Swedish 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 these Swedish loan words than for a Swede to understand the corresponding Danish words.

In addition to these explanations of asymmetry, linguistic differences can also be asymmetric and can also be part of the explanation for asymmetric mutual intelligibility. For example, Danish might have two synonyms for a concept, which has only one equivalent in Swedish. An example is the word for 'room' which is *rom* in Swedish and *rum* or *værelse* in Danish. It will be easy for a Swede to understand the Danish cognate word *rum* but impossible to understand the non-cognate *værelse* unless he or she has somehow learned it. Likewise, phonetic, morphological and syntactic transparency may be asymmetric.

Another explanation for the asymmetric intelligibility might be found in the relationship between the written and the spoken form of the language. For example, 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 (Doetjes & Gooskens 2009).

In the present investigation we will pay special attention to language pairs that show asymmetric intelligibility, since these are languages that also provide crucial information about which factors play an important role for intelligibility.

To summarize, the main aim of our investigation is to develop a model of intelligibility of closely related languages. To achieve this, the following research questions must be answered:

1. What is the mutual intelligibility of closely related languages in the Germanic, Slavic and Romance language groups in Europe?

2. What linguistic distances can be established between the languages in the three language groups at different linguistic levels (phonetics, vocabulary, morphology, syntax, orthography)?
3. What attitudes do various groups of speakers of Germanic, Slavic and Romance languages have towards other languages in the same language family?
4. How much experience do various groups of speakers of Germanic, Slavic and Romance languages have with other languages belonging to the same language family?
5. To what extent are the linguistic and extra-linguistic distances predictors of mutual intelligibility?
6. What explanations can be found for asymmetric intelligibility?

The second aim is to investigate the interlanguage speech intelligibility benefit of English as a lingua franca and compare the mutual intelligibility of accented English to that of closely related languages:

7. How well do various listener groups in the Germanic, Slavic and Romance language areas understand each other in English?
8. How well do speakers of the test languages understand closely related languages compared to how well they understand non-native English?

2. Approach/methodology

The investigation has an experimental set-up. By means of web-based technology, intelligibility measures, attitudes and contact information concerning a large number of European languages will be collected in the countries where these languages are spoken. In addition, linguistic distances at various linguistic levels (phonetic, morphological, lexical, syntactic and orthographic) will be established. By including these data from a large number of languages a test of the relationship between intelligibility (dependent variable) on the one hand and linguistic and non-linguistic predictors (independent variables) on the other hand can be conducted. This data can be analysed statistically and employed to develop a model for predicting (and thereby) explaining mutual intelligibility between closely related languages.

2.1. Languages and informants

The mutual intelligibility of languages within the three major language families in Europe, i.e. Germanic, Slavic and Romance, will be tested. As mentioned above, several investigations of the mutual intelligibility of the three Scandinavian languages (Danish, Norwegian and Swedish) were carried out in the past. A comparison of the results from the proposed investigation with those from previous investigations will be valuable since previous results provide a frame of reference to test the suitability of our test method against (see 2.2). Except for some research on German-Dutch mutual intelligibility (Ház 2005, Kürschner, Gooskens & Van Bezooijen 2008), the mutual intelligibility of the other Germanic languages has hardly been investigated.

Most of the new EU member states are Slavic countries. Still, very little research has been carried out on mutual intelligibility in the Slavic language area. Most research to date is descriptive. We collaborate with researchers from the Slovak Akademy of Sciences at the Charles University at Prague (dr. Marian Sloboda, dr. Mira Nábělková and prof. dr. Jiří Nekvapil) who are currently carrying out an explorative investigation.

Recently, a number of researchers have shown interest in the mutual intelligibility of the Romance languages (http://www.eurocomcenter.eu/), for example in connection with the EuroComRom project which has set up a program supporting receptive multilingual reading competence within the Romance language family. However, hardly any empirical research has been carried out on the mutual intelligibility of Romance languages.

For each language family investigated in the project it must be decided what combinations of test languages and groups of test persons to include. Since the tests will be carried out by means of an internet application, a fairly large number of languages and test persons can be targeted. As a point of departure, all official national languages within the three language families will be included as test languages. The intelligibility of these languages will be tested among speakers of the official standard languages in the same language family. So, for example, the intelligibility of Spanish will be tested with speakers of Portuguese. Similarly, Czech will be tested with speakers of Polish. In addition, the intelligibility of non-native English will be tested in all selected countries. To limit the number of combinations of test languages and groups, a selection will be made of combinations where some threshold level of mutual intelligibility can be expected since there is no point in testing languages that we know beforehand to be unintelligible to the listener group. A well-founded selection will be made on the basis of linguistic distances measured by the Levenshtein algorithm (see Section 2.3) and of literature on mutual intelligibility in the three language families.

2.2 Intelligibility

To be able to compare the level of intelligibility between selected language pairs within the three language families, the same texts will be translated into all test languages. Moreover, the same texts will be used for testing the intelligibility of written and spoken language in order to compare the intelligibility of the two modalities. To limit the length of the testing, each listener will be respond to just six languages. The result of this part of the investigation will be an overview of the intelligibility structure in Europe.

Texts

By using coherent texts in the test situation, we aim to obtain a realistic image of the communicative possibilities in daily life. Each listener will be tested with six closely related languages and the six corresponding non-native English accents. Since each language and accent will be represented by a different text, twelve different texts are needed. These should have identical levels of difficulty and consist of a limited number of short sentences. The same texts will be used for all languages in order to be able to compare the results across languages and language families. Each text will therefore be translated into all languages.

Test

Mutual intelligibility among speakers of the languages within each language family will be tested by means of a cloze test. This test was developed by William Taylor in 1953 in America. Since then it has been a widely-used tool for measuring the intelligibility of written texts. It is generally seen as a reliable and valid measure of reading comprehension. It is suitable for integration in a web-based application and can be corrected automatically (i.e. by a computer), thus ensuring maximum objectivity and efficiency. We expect our listeners to be highly motivated when they know that they will receive individual performance indicators (scores) on completion of the test.

The cloze test exists in various versions. Typically, a number of words are removed from the text and placed in random order above the text. Test subjects are then asked to put the words back in the right place in the text within a certain amount of time. The percentage of words restored correctly is taken as a measure of the intelligibility of the whole text.

Since we are interested in the intelligibility not only of written but also of spoken language, we will develop a variant of the cloze test that can be used in the oral modality as well. In each of the test sentences, one content word is randomly selected. All selected keywords are shown in random order on the computer screen. Keywords will be the same for all languages and will be shown in the native language of the listeners only, thereby avoiding potential confound between imperfect knowledge of the spoken and written forms of the same test language.

When testing the intelligibility of spoken language, subjects will listen to the text sentence by sentence. The keyword in each sentence is replaced by a beep and the listeners' task is to select the alternative from the list of keywords on screen which they think was replaced by the beep.

Since we want to compare the intelligibility of spoken texts to that of written texts, the latter will be presented sentence by sentence as well. The sentences will be presented on the computer screen with the keyword blanked out by a gap of uniform length. Here the reader's task is to identify the correct alternative in a list of options presented on screen. Participants will be allowed the same response time in the spoken and written versions of the cloze test.

Speakers

For testing oral intelligibility, recordings of each text will be made by representative speakers of the target languages. They should speak the standard variety of the language. Pre-university adolescents (e.g. high school children between 17 and 18 years at a level that prepares for a higher education) will serve as speakers. Moreover, several speakers of each language will read each text, so that effects of variability between speakers will average out. In this way we minimize the influence of individual voice characteristics on the formation of attitude judgments. The speakers first read aloud the texts in their own language. Next, they translate the text into English and read the translation aloud. They will make the translation without any help from dictionaries and grammar books and with little preparation time. In this way we make sure that the English is representative of the target group.

Listeners

Listeners from each country should speak the same variety as the speakers of the texts discussed above. In addition the listeners should come from an area which is not close to the border of the countries where the test languages are spoken to avoid a more than minimal contact with the test languages. In some cases it may be desirable to test listeners at more than one location in a country. The listener groups in the various countries should be comparable. The background of the listeners will be similar to that of the speakers (see above). Pre-university adolescents are a well-defined group which can be easily approached through school teachers (Gooskens et al. 2010).

Design

In each country the intelligibility of languages belonging to the same language family will be tested. Listeners will also be tested in their own language to insure that all texts are perfectly intelligible to native speakers of the language. Finally, non-native English will be tested in all of the language families in order to compare the mutual intelligibility of non-native English with the mutual intelligibility of the same persons using semi-communication. This will allow us to answer questions such as 'Is Spanish-accented English more or less intelligible to Portuguese listeners than plain Spanish (and vice versa)?'

Six closely related languages and the six corresponding non-native English accents will be presented to each group of test persons. Half of the test languages will be presented in written form and the other half in spoken form. Furthermore, half of the test persons will listen to one half of the texts and read the other half while for the other half of the test persons it will be the other way round. A (Latin-square) design will be used such that all languages and all texts will be tested equally often in both modalities.

2.4 Non-linguistic factors

After the intelligibility test of each language listeners will be asked to identify the language they just heard or saw. In this way we will know how well listeners recognize the other languages in their language family and whether they base their attitude judgments on the correct language. Next, questions about attitudes towards test languages and their speakers and familiarity with the test languages follow. The results will be used to explain intelligibility results and will also result in an overview of language attitudes and language contact in Europe. Note that attitudes and familiarity may be asymmetric and may be (part of) the explanation for asymmetric intelligibility between some language pairs (see introduction).

The attitudes towards the various languages will be established by means of attitudinal scales concerning the languages (e.g. Zahn and Hopper 1985) and the speakers. Listeners will be asked questions like 'How beautiful does the language of this speaker sound?' and 'How friendly does this speaker sound?'.

The familiarity of the test persons with the languages will be established by means of a questionnaire about how often they have contact with the language in its written and spoken form (via personal contacts or the media). In order to quantify the amount of experience of the listeners with the test language, the listeners will also be asked to translate a number of non-cognate words (i.e. historically non-related words) from the test language. Since non-cognates are per definition unintelligible to listeners with no prior experience with the test language, the number of correctly translated non-cognates is a priori a good measure of previous experience of a language (Gooskens et al. 2010). A number of languages such as English, German, French and Russian have a special status because they are part of the curriculum at school. The test persons will be asked questions about which languages they have learnt at school, for how long and how many hours a week.

2.3 Linguistic distances

To be able to relate the intelligibility scores to linguistic distances, distances will be calculated for all language pairs in the intelligibility tests, i.e. between all pairs of speaker and listener languages. Measurements will be carried out for various linguistic levels separately. The results of the measurements will give an overview of linguistic distances between a large number of language pairs in Europe.

Phonetic and **orthographic** distances will be measured for cognates (i.e. the historically related words) only. For each language pair from the intelligibility test, the test words will be aligned so that it is possible to compute the linguistic distance per word pair. To be able to measure the phonetic distances all texts must be transcribed phonetically. The distances will be calculated by means of the Levenshtein algorithm that has been developed for measuring linguistic distances between dialects (Heeringa 2004) and that has also been used successfully to measure communicatively relevant distances between Scandinavian language varieties (e.g. Gooskens 2007). The overall distance per word pair in a corpus is computed by means of the minimum number of insertions, deletions and substitutions of phonetic segments (or letters in the case of written texts) needed to transform the word in one language into the other, whereby word length is normalised for. More refined measurements will also be carried out by incorporating frequency as well as the phonetic/orthographic nature of the correspondences into the measurements. The distance between two languages is the mean of all word pair distances.

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 from corpora 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.

Morpho-syntactic distances will be assessed by counting the number of morphological and syntactic differences in word order between the test language and the equivalent sentence in the native language of the test person. Also at these levels the effect of frequency (based on corpora) and nature of the phenomena will be assessed. Frequent morphemes and syntactic differences are likely to be easier to understand than infrequent ones. Also, different categories will be analysed separately. The role of syntactic and morphological differences in the intelligibility of a closely related languages has hardly been investigated so far.

As explained in the introduction, linguistic distances between languages may be asymmetric at all linguistic levels.[1] For this reason distances should be calculated in both directions for each pair of languages. In order to account for phonetic and orthographic asymmetry we will also compute Conditional Entropy (Moberg, Gooskens, Nerbonne & Vaillette 2007), which has been shown to successfully model asymmetric intelligibility in Scandinavia.

2.5 Analysis

The percentage of correct responses and the percentage of correct translations will form the dependent (criterion) variables against which the independent (predictor) variables (linguistic distances on different linguistic levels, attitude and familiarity scores) will be tested. Regression analyses will be carried out to determine the relative importance of the various determinants for the (mutual) intelligibility of closely related languages in Europe. On the basis of these results a model of mutual intelligibility between closely related languages will be developed.

In addition to the analysis of the overall intelligibility of closely related languages in Europe, a more in-depth analysis of mutual intelligibility within each of the three language families will be made. Special attention will be paid to language pairs that show an asymmetric intelligibility, since these may provide a deeper understanding of the factors that play a role in mutual intelligibility between closely related languages.

Finally, a separate analysis will be made of the intelligibility of non-native English among test persons from the three language areas. This will give an overview of the English proficiency in Europe. The results will be compared to the results of the tests of intelligibility of closely related languages to be able to draw conclusions about the value of English as a lingua franca as opposed to communication via closely related languages (receptive multilingualism).

2.6 Publicly available web application

A database with the results of the investigation will be made publicly available through the internet for researchers and policy makers to consult. The number of languages spoken in Europe is large and we are forced to make a choice of test

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

languages to be included in the investigation. It might therefore be desirable to extend the investigation to more languages or language varieties after our own data collection has been finished. It will be useful, for instance, to include information about intelligibility or attitudes towards minority languages in the individual countries. For this reason we will develop a simplified version of the web application for public use. Researchers and policy makers will be able to add new test languages and use standardized intelligibility, attitude and familiarity tests in a user friendly way. A tool for measurements of linguistic distances between new pairs of languages will also be made available. This will be made possible by using the ADEPT application which is being developed at the University of Groningen with a CLARIN-grant to John Nerbonne and the principal applicant. This application will facilitate measurement of Levenshtein distances by means of a graphical user interface.

3. Innovation and impact

Very little is known about the mutual intelligibility of closely related languages in Europe. The project will yield a sketch of the intelligibility structure of a large number of languages from the three largest language groups in Europe (Germanic, Slavic and Romance) by means of language tests that makes it possible to compare intelligibility of various languages pairs in their written and spoken form. Also the mutual intelligibility of non-native English will be tested for the first time among a large number of Europeans by means of the same language test, thereby filling a gap in our knowledge identified on a European level. To make this happen, a new version of the cloze test will be developed. In addition, an overview of language attitudes and language contact in Europe will be created for the first time. This is done by means of questionnaires that facilitate comparisons of results across language groups. Finally, an overview of linguistic distances between a large number of European languages will be created.

The databases with the results of the intelligibility tests as well as the inventories of linguistic distances, attitudes and contact patterns will be made publicly available. It will form a valuable source of information for researchers in the area of language variation. In the past, various researchers have tested the mutual intelligibility of closely related languages across the world. Most of these investigations were carried out with the aim to investigate the genealogical relationship between language varieties, for example of Amerindian languages (Voegelin & Harris 1951, Hickerton, Turner & Hickerton 1951, Pierce 1952) and to make an inventory of mutual intelligibility of languages, for example of the Scandinavian languages (Maurud 1976, Bø 1978, and Delsing & Lundin Åkesson 2005). Often, intelligibility testing has been conducted in the context of literacy programs to develop a single orthography to serve multiple closely related language varieties (e.g. Casad 1974, Brye & Brye 2002, Anderson 2005). Only a few experimental investigations on mutual intelligibility of closely related languages focusing on explanatory linguistic and non-linguistic factors have been conducted so far. For example, Tang & Van Heuven (2009) investigated the relationship between phonetic distances and the mutual intelligibility of 15 Chinese dialects. Within Europe, systematic research has only been carried out in the Scandinavian language area and in the Netherlands with Dutch dialects (Van Bezooijen & Van den Berg 1999) recently in an investigation on regional varieties of Standard Dutch (Impe 2010). Since these languages are very closely related and the speakers belong to a historical/cultural entity, the linguistic and social relationship between these languages may be rather different from the relationships between other closely related languages. A comparison with other language areas is therefore crucial for the development of a model of mutual intelligibility. Also the large number of languages in our investigation makes a statistical test of the relationship between intelligibility and the linguistic and non-linguistic factors possible.

The results from this investigation will also be of great value to policy makers. Knowledge about the linguistic determinants of mutual intelligibility is useful for language planning at the national and at the European level. It is important to know under which circumstances linguistic distances can be bridged. If 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. The results will form a basis for the discussion about how large linguistic distances can be before they result in a communication breakdown. Is there a breakdown point or does intelligibility have a gradual relationship with linguistic distances? The results will also provide a necessary basis for the discussion about the use of English as a *lingua franca* in a European context.

The test will be implemented into a web application. The application will be developed in such a way that it will be expandable for future use by for example policy makers or researchers who are interested in testing the intelligibility, attitude and familiarity with a language or language variety not included in the current project. In this public web application we will also use the ADEPT application which has been developed to make it easy to calculate Levenshtein distances by means of a graphical user interface.

References

Anderson, H. (2005). Intelligibility testing (RTT) between Mendankwe and Nkwen. Summer Institute of Linguistics, Dallas, Texas. *Electronic Survey Reports* 205-002: 30.

Bent, T. & A.R. Bradlow (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America* 114, 1600–1610.

Budovičová, V. (1987). Literary languages in contact (A sociolinguistic approach to the relation between Slovak and Czech today). In: J. Chloupek, J. Nekvapil et al. (eds.), *Reader in Czech sociolinguistics*. Amsterdam/Philadelphia: John Benjamins, 156-175.

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.

Brye, E. & E. Brye (2002). *Rapid appraisal and intelligibility testing surveys of the Eastern Beboid group of languages (Northwest Province).* SIL Electronic Survey Reports 2002-019.

Casad, E.H. (1974). *Dialect intelligibility testing*. Summer Institute of Linguistics Publications in Linguistics and Related Fields, 38. Norman: Summer Institute of Linguistics of the University of Oklahoma.

Delsing, L.-O. & K. Lundin Åkesson (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.

Doetjes, G. & C. Gooskens (2009). Skriftsprogets rolle i den dansk-svenske talesprogsforståelse [The role of orthography for Danish-Swedish mutual intelligibility]. *Språk och stil*, 19, 105-123.

European Commission (2007). *Final Report High Level Group on Multilingualism.* Luxembourg: Office for Official Publications of the European Communities

Gooskens, C., S. Kürschner & R. van Bezooijen (submitted). Intelligibility of Low and High German to speakers of Dutch. *Dialectologia.*

Gooskens, C., V.J. van Heuven, R. van Bezooijen & J. Pacilly (2010). Is spoken Danish less intelligible than Swedish? *Speech Communication.* (in press)

Gooskens, C., K. Beijering & W. Heeringa (2008). Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing* 2, 63-81.

Gooskens, C. (2007). The contribution of linguistic factors to the intelligibility of closely related languages. *Journal of Multilingual and Multicultural Development* 28, 445-467.

Ház, E. (2005). *Deutsche und Niederländer. Untersuchungen zur Möglichkeit einer unmittelbaren Verständigung*. Hamburg: Dr. Kovač. (Philologia 68).

Hickerton, H., G.D. Turner & N.P. Hickerton (1952). Testing procedures for estimation transfer of information among Iroquois dialects and languages. *International Journal of American Linguistics* 18, 1-8.

Impe, L. (2010). *Mutual Intelligibility of national and regional varieties of Dutch in the Low Countries*. Dissertation, Katholieke Universiteit Leuven, Leuven.

Kürschner, S., C. Gooskens & R. van Bezooijen (2008). Linguistic determinants of the intelligibility of Swedish words among Danes. *International Journal of Humanities and Arts Computing* 2, 83-100.

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.

Moberg, J., C. Gooskens, J. Nerbonne & N. Vaillette (2007). Conditional Entropy Measures Intelligibility among Related Languages. In: P. Dirix, I. Schuurman, V. Vandeghinste & F. Van Eynde (eds.). *Computational Linguistics in the Netherlands 2006: Selected papers from the 17th CLIN Meeting*. Utrecht: LOT, 51-66.

Pierce, J.E. (1952). Dialect distance testing in Algonquian. *International Journal of American Linguistics* 18, 208-218.

Tang, C. & V.J. van Heuven, (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua* 119, 709-732.

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

Van Heuven, V.J. & H. Wang (2007). Quantifying the interlanguage speech intelligibility benefit. In: W. Barry & J. Trouvain (Eds.) *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken* (pp. 1729-1732). Saarbrücken: Universität des Saarlandes.

Voegelin, C.F. & Z.S. Harris (1951). Methods for determining intelligibility among dialects of natural languages. *Proceedings of the American Philosophical Society* 95, 332-329.

Wolff, H. (1959). Intelligibility and inter-ethnic attitudes. *Anthropological linguistics* 1, 34-41.

Zahn, C.J. & R. Hopper (1985). Measuring Language Attitudes: The Speech Evaluation Instrument. *Journal of Language and Social Psychology* 4, 113-123.

10. Summary in Key Words

intelligibility, language attitudes, linguistic distances, language contact, lingua franca

11. Work Programme

	year 1	year 2	year 3	year 4
Postdoc 1	develop tests for internet experiments	integration of results from all language areas calculate morpho-lexical and syntactic distances	integration of results from all language areas, develop intelligibility model plan conference	
Postdoc 2	develop communicative linguistic distance measurements make internet application for internet experiments	measure linguistic distances develop web application and database for general public use		
PhD 1	background reading, develop internet experiments	recruit test subjects in Germanic language area, phonetic transcriptions, calculate linguistic distances	analysis of results from internet experiment in Germanic language area	write dissertation
PhD 2	background reading, develop internet experiments	recruit test subjects in Slavic language area, phonetic transcriptions, calculate linguistic distances	analysis of results from internet experiment in Slavic language area	write dissertation
PhD 3	background reading, develop internet experiments	recruit test subjects in Romance language area, phonetic transcriptions, calculate linguistic distances	analysis of results from internet experiment in Romance language area	write dissertation

main applicant			plan conference edit a peer-reviewed volume write synthesizing introduction	write monograph

12. Word Count

5364

13. Planned Deliverables and Knowledge Dissemination

Subproject I and II:

- publicly available searchable database with results from intelligibility experiments, attitude and familiarity questionnaires
- publicly available web application for future intelligibility, attitude and familiarity and linguistic distance measurements
- international conference (together with principal applicant) with contributions from all subprojects
- a minimum of eight international publications (together with applicants) on:
  - mutual intelligibility between closely related languages in Europe
  - communicatively relevant linguistic distances measures
  - English as a *lingua franca* in Europe
  - language attitudes in Europe
  - language contact in Europe
  - a model of mutual intelligibility between closely related languages

Subproject III: Mutual intelligibility in the Germanic language area

- three international publications:
  - intelligibility in the Germanic language area
  - language attitudes in the Germanic language area
  - language contact in the Germanic language area
- dissertation

Subproject IV: Mutual intelligibility in the Slavic language area

- three international publications:
  - intelligibility in the Slavic language area
  - language attitudes in the Slavic language area
  - language contact in the Slavic language area
- dissertation

Subproject V: Mutual intelligibility in the Romance language area

- three international publications:
  - intelligibility in the Romance language area
  - language attitudes in the Romance language area
  - language contact in the Romance language area
- dissertation

In addition to the international publications and the international conference, the principal applicant will write a monograph in English about the intelligibility model to be published by an international publisher. She will also edit a peer-reviewed volume featuring contributions from the international conference (including contributions from all subprojects) and a synthesizing introduction.

14. Short Curriculum Vitae Principal Applicant

Dr. Charlotte Gooskens (PhD Nijmegen 1997) is associate professor of Scandinavian Linguistics at the University of Groningen. Her research is concerned with perceptual and communicative effects of language variation, e.g. language attitudes, speaker identity and mutual intelligibility of closely related languages. For her research she uses experimental research methods and exact measurement techniques. In 2005 she was awarded a VIDI-grant for the project *Linguistic determinants of mutual intelligibility in Scandinavia.* She was also co-applicant and coordinator of the CLARIN-project *Assaying differences via edit-distance of pronunciation transcriptions* that is being carried out in 2010 and is a supervisor in the VNC-project *Mutual intelligibility of language varieties in the Low Countries: linguistic and attitudinal determinants* (2005-2011).

Relevant publications:

Dr. Charlotte Gooskens:

Bezooijen, R. van & C. Gooskens (2007). Linguistic and extralinguistic determinants of interlingual text comprehension. In: J. ten Thije & L. Zeevaert (red.). *Receptive Multilingualism and intercultural communication.* Hamburger studies in multilingualism. Amsterdam: Benjamins, pp. 249-265.

Gooskens, C., K. Beijering & W. Heeringa (2008). Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing* 2, 63-81.

Gooskens, C. & S. Kürschner (2009). Cross border intelligibility - on the intelligibility of Low German among speakers of Danish and Dutch. In: A. N. Lenz, C. Gooskens & S. Reker (eds.). *Low Saxon dialects across borders - Niedersächsische Dialecte über Grenzen hinweg*, Zeitschrift fur Dialektologie und Linguistik, Beihefte 138, 273-297.

Gooskens, C. (2007). The contribution of linguistic factors to the intelligibility of closely related languages. *Journal of multilingual and multicultural development,* 28 (6), pp. 445-467.

Gooskens, C. & R. van Bezooijen (2006). Mutual comprehensibility of written Afrikaans and Dutch: symmetrical or asymmetrical? *Literary and linguistic computing,* 21, pp. 543-557.

Prof. dr. Vincent J. van Heuven:

Bezooijen, R. van & V.J. van Heuven (2010). Avant-garde Dutch: a perceptual, acoustic and evaluational study. D. R. Preston. N. Niedzielski (eds.) *Sociophonics*, Cambridge University Press, 357-378.

Gooskens, C., V. J. van Heuven, R. van Bezooijen & J. Pacilly (in press). Is spoken Danish intrinsically less intelligible than Swedish? *Speech communication.*

Heuven, V. J. van (2008). Making sense of strange sounds: (mutual) intelligibility of related language varieties. A review. *International Journal of Humanities and Arts Computing* 2, 39-62.

Heuven, V. J. van & H. Wang (2007). Quantifying the interlanguage speech intelligibility benefit. In W. Barry & J. Trouvain (eds) *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken.* Saarbrücken: Universität des Saarlandes, 1729-1732.

Tang, C. & V. J. van Heuven (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua* 119, 709-732.

15. Summary for Non-specialists

Titel: De onderlinge verstaanbaarheid tussen nauw verwante talen in Europa: linguïstische en niet-linguïstische determinanten

In Europa wordt een groot aantal verschillende talen en dialecten gesproken. De EU heeft veel respect voor deze linguïstische variatie, maar ze ziet ook in dat de diversiteit communicatieproblemen kan opleveren. De problemen kunnen alleen worden opgelost als er voldoende wetenschappelijke kennis beschikbaar is over actuele taalsituaties. In 2007 publiceerde *The High Level Group on Multilingualism* (HLGM*)* daarom een overzicht van onderwerpen die onderzocht zouden moeten worden om de communicatie binnen Europa te verbeteren, met behoud van de meertalige rijkdom. Twee van deze onderwerpen vormen de basis voor het huidige project. Ten eerste signaleert de HLGM een gebrek aan kennis over de onderlinge verstaanbaarheid tussen nauw verwante talen in Europa en de mogelijkheid om te communiceren via receptieve meertaligheid, waarbij sprekers van nauw verwante talen elk hun eigen taal blijven spreken. Ten tweede signaleert de HLMG een behoefte aan een evaluatie van de mogelijkheden en beperkingen van het gebruik van het Engels als *lingua franca* (ELF) op Europees niveau. Het is belangrijk meer kennis te verzamelen over hoe goed sprekers van verschillende talen in Europa elkaar kunnen begrijpen in het Engels.

In haar huidige VIDI-project is de aanvraagster erin geslaagd om bestaande methodes om linguïstische afstanden te kwantificeren zodanig te verfijnen en uit te breiden dat ze communicatief relevante afstanden tussen de gesproken Scandinavische talen kon meten. Haar onderzoeksresultaten tonen aan dat er een sterke relatie bestaat tussen fonetische en lexicale afstanden aan de ene kant en verstaanbaarheid aan de andere kant. Eerder onderzoek naar de Scandinavische talen heeft laten zien dat attitude en vertrouwdheid ook een belangrijke rol spelen. Een negatieve houding tegenover een taal en de sprekers ervan kan een negatieve invloed hebben op de verstaanbaarheid en eerdere opgedane kennis van een taal zal de verstaanbaarheid natuurlijk ten goede komen.

Buiten Scandinavië wordt nog weinig gebruik gemaakt van de mogelijkheid om te communiceren via receptieve meertaligheid en er is weinig onderzoek naar verricht. Kennis over de mate van onderlinge verstaanbaarheid in Europa en de linguïstische en niet-linguïstische basis daarvan is beperkt. De bruikbaarheid van het Scandinavische model is echter duidelijk aangetoond en kan makkelijk worden toegepast op de communicatie tussen sprekers van andere nauw verwante talen in Europa als een alternatief voor een *lingua franca* zoals bijvoorbeeld het Engels.

We stellen daarom een grootschalig experimenteel onderzoek voor naar de onderlinge verstaanbaarheid van nauw verwante talen binnen de Germaanse, Slavische en Romaanse taalfamilies. Omdat communicatie in Europa zowel via schrift als het gesproken woord plaatsvindt, zullen beide vormen van communicatie worden getest. De resultaten zullen worden gecorreleerd met linguïstische factoren, zoals fonetische en lexicale afstanden, en met niet-linguïstische factoren, zoals attitudes tegenover en vertrouwdheid met de testtalen. De verstaanbaarheid van de varianten van het Engels zoals die worden gesproken door de sprekers van de verschillende nauw verwante talen zal ook worden getest. Op deze manier kan de (onderlinge) verstaanbaarheid van nauw verwante talen worden vergeleken met de verstaanbaarheid van ELF.

Verstaanbaarheid, attitude en vertrouwdheid zullen worden onderzocht via web-gebaseerde experimenten. De resultaten zullen beschikbaar worden gesteld via het internet en zullen een waardevolle bron van informatie zijn voor onderzoekers op het gebied van taalvariatie. Zij zullen de basis vormen voor een model dat de onderlinge verstaanbaarheid van nauw verwante talen kan verklaren en voorspellen. De uitkomsten zullen ook van grote waarde zijn voor Europese beleidsmakers.

Er zal een openbaar toegankelijke, gebruikersvriendelijke web-applicatie worden ontwikkeld voor toekomstig gebruik door wetenschappers en beleidsmakers. Zo kan in de toekomst ook de verstaanbaarheid worden getest van taalvariëteiten die niet in het huidige project zijn opgenomen. In deze openbare web-applicatie zal ook de ADEPT-applicatie worden geïntegreerd, die op dit moment in Groningen wordt ontwikkeld om makkelijk fonetische afstanden te kunnen meten via een grafische gebruikersinterface.

Tot nu toe zijn slechts lexicale en fonetische factoren betrokken geweest in het onderzoek van de aanvraagster. In het voorgestelde project zullen in zeer uiteenlopende taalcombinaties ook morfologische en syntactische communicatieve afstanden worden gemeten. Verder zullen ook niet-linguïstische factoren deel uitmaken van het onderzoek. Door deze uitbreidingen zal het mogelijk zijn een generiek goed gefundeerd model van de onderlinge verstaanbaarheid van nauw verwante talen te ontwikkelen dat ons in staat zal stellen verstaanbaarheid op een theoretische basis te voorspellen.

In ongunstige omstandigheden, als informatie ontbreekt of afwijkend is ten opzicht van de taal van de luisteraars, zijn luisteraars opmerkelijk goed in staat te begrijpen wat er wordt bedoeld. De informatie over de verstaanbaarheid van verschillende nauw verwante talen in ons onderzoek vormen een perfect natuurlijk laboratorium voor het bestuderen van de grootte van linguïstische afstanden die mensen kunnen overbruggen. Meer in het algemeen zullen de resultaten van ons onderzoek daarmee inzichten geven in de robuustheid van het menselijke taalvermogen.

16. Research Budget

A. Personnel

Type of appointment	Term	Extent	Salary	Bench fee	
postdoc 1	3 years	0.8 fte	€ 157,308	€ 5,000	€ 162,308
postdoc 2	2 years	0.8 fte	€ 104,872	€ 5,000	€ 109,872
PhD 1	4 years	1.0 fte	€ 200,013	€ 5,000	€ 205,013
PhD 1	4 years	1.0 fte	€ 200,013	€ 5,000	€ 205,013
PhD 1	4 years	1.0 fte	€ 200,013	€ 5,000	€ 205,013
Subtotal A. personnel and bench fee					€ 887,219

B. Other personnel costs

Type of appointment	Term	Extent	Salary	Bench fee	
replacement principal applicant years 3 and 4 for writing monograph and editing peer-reviewed volume	2 years	0.5 fte	€ 50,000		€ 50,000
students assistance for carrying out experiments years 1 and 2	2 years	0.4 fte	€ 30,000		€ 30,000
Subtotal B. other personnel costs					€ 80,000

C. Material

Material	Break down and specify	Year	Amount
Internationalisation activities	1. organisation international conference 2. attendance international conferences project members	year 3 years 2, 3, 4	€ 7,000 € 16,000
Fieldwork/experiments	1. Remuneration for subjects	year 2	€ 9,000
Subtotal C. material			€ 32,000

Overall programme budget:

Subtotal A, personnel	€ 887,219
Subtotal B, replacement	€ 80,000
Subtotal C, material	€ 33,000
Subtotal D, knowledge utilisation	-
Total amount requested	**€ 999,219**