INTRODUCING MICRELA: PREDICTING MUTUAL INTELLIGIBILITY BETWEEN CLOSELY RELATED LANGUAGES IN EUROPE[1]

Vincent J. van Heuven[i, ii, iii], **Charlotte S. Gooskens**[ii], **Renée van Bezooijen**[ii]

[i] Pannon Egyetem, Veszprém, Hungary
[ii] Groningen University, The Netherlands
[iii] Leiden University, The Netherlands

v.j.j.p.van.heuven@hum.leidenuniv.nl

1. Introduction

The Wikipedia lists 91 indigenous languages that are spoken in Europe today.[2] Forty-two are currently being used by more than one million speakers. Only 24 are recognized as official and working languages of the European Union. Most of these belong to one of three major families within the Indo-European phylum, i.e. Germanic, Romance and Slavic languages. Languages and language varieties (dialects) within one family have descended from a common ancestor language which has become more diversified over the past centuries through innovations. Generally, the greater the geographic distance and historical depth (how long ago did language A undergo an innovation that language B was not part of), the less the two languages resemble one another, and – it is commonly held – the more difficult it will be for speakers of language A to be understood by listeners of language B and *vice versa*. A working hypothesis would then be that the longer ago two related languages split apart, the less they resemble one another and the smaller their mutual intelligibility.

We are involved in a fairly large research project that was set up to test these hypotheses. The basic idea was to first measure the level of mutual intelligibility between all pairs of languages within a family, them compute the degree of structural linguistic similarity between the members, and try to predict the observed level of intelligibility from the linguistic distances measured. The historic component is seen as a subsidiary issue: we will not try to predict mutual intelligibility between languages from the distance between them in the traditional linguistic family tree (cladistic distance) but simply check to what extent the linguistic relatedness between two languages corresponds with their mutual intelligibility and/or measured linguistic distance.

It should be noted right from the start that predicting mutual intelligibility between two related languages from structural linguistic differences can only be done as long as the interlocutors have no prior knowledge of, or experience with, the other language. It is often difficult, if not impossible, to find a sufficient number of speakers/listeners who have never heard the other language before, and are *tabula rasa*. Nevertheless, if we want to determine the degree of inherent mutual intelligibility between two languages, the *tabula rasa* requirement is imperative. In actual practice, speakers of related languages will have some degree of

[1] Research sponsored by a M€ 1 grant from the Netherlands Organisation for Research (principal investigators: Charlotte Gooskens and Vincent van Heuven). See the Micrela (**M**utual **i**ntelligibility between **c**losely **r**elated **la**nguages) project homepage at http://www.let.rug.nl/~gooskens/project/.
[2] http://en.wikipedia.org/wiki/Languages_of_Europe

prior exposure to other languages in the same family, either because the languages are spoken in neighbouring countries or have been learnt through explicit instruction in a school setting. Such forms of language contact will then lead to a variable degree of *acquired* mutual intelligibility – depending on the intensity of the contact, the quality of the educational system and the personal language-aptitude of the learner. Obviously, the effects of non-linguistic factors such as prior exposure and language education can be far more influential in determining the level of intelligibility than the *inherent* linguistic similarity (see e.g. Bahtina and Ten Thije, 2013). In fact, even if two languages are totally unrelated, such as English and Mandarin Chinese, native speakers of Chinese have often learned to understand English, and there is also a (relatively small) group of English speakers who are able to understand Mandarin – through years of exposure and training. In our project we make a strict division between inherent and acquired intelligibility between languages. In order to achieve this goal we collected data on the individual familiarity of our participants with other languages within their language family on the one hand (extra-linguistic factors), and the inherent degree of similarity between the pairs of languages involved (linguistic factors) on the other. The ultimate goal would then be to map out, for the major European languages, the separate contribution of these linguistic versus extra-linguistic factors to their mutual intelligibility.

The results of this study should be interesting enough from a theoretical linguistic perspective. We would argue that an adequate theory of language should allow us to predict with great accuracy how well a listener with a native language A will understand a related language B – assuming no prior exposure to the related language. However, the enterprise may also have a practical spin-off for language policy makers. When educated Europeans of different nationalities meet, the default strategy would seem to be to use English as a *lingua franca*. It is an open question when the same or even a better level of mutual understanding can be achieved if the speakers would avail themselves of the advantages of what is often referred to as *receptive multilingualism* (Hockett, 1958; Zeevaert, 2004; Braunmüller, 2007; Ten Thije and Zeevaert 2007), *plurilingual communication* (Lüdi 2007), *semi-communication* (Haugen, 1966), *intercompréhension* (Grin 2008) and *lingua receptiva* (Rehbein, Ten Thije and Verschik 2012). In this type of communication, each interactant continues to speak his/her native language, which the other person then understands sufficiently well to sustain a meaningful exchange of information. Our experiments would allow policy makers to ascertain the feasibility of this mode of communication between speakers with different native languages.

In the remainder of this article we will describe the overall set-up of the research project, with emphasis on methodological choices made. Then we will give a rough overview of the results of the intelligibility tests and of the linguistic distances measured between the pairs of languages studied. In the final section we will determine how well the degree of mutual intelligibility between two languages can be predicted. Within the space limitations of this article, we cannot give a comprehensive overview of our findings. Rather we will present a selection of results just to demonstrate how the research is done and what kind of answers can be given to the research questions. For details we refer the reader to three dissertations that will be defended later this year at Groningen University: Swarte (2015) on Germanic languages, Golubovic (2015) on Slavic languages and Voigt (2015) on Romance languages.

2. Data collection

We selected five Germanic languages, five Romance languages and six Slavic languages, as listed in table 1. These are the three largest language families in Europe (in terms of numbers of speakers). The languages chosen are the official main languages within their respective countries. This criterion excludes regional languages such as Basque, Catalan, Romansh, Frisian, Luxembourgish, and many others.

Germanic		Romance		Slavic	
Danish	Da	French	Fr	Bulgarian	Bu
Dutch	Du	Italian	It	Croatian	Cr
English	En	Portuguese	Pt	Czech	Cz
German	Ge	Romanian	Ro	Polish	Pl
Swedish	Sw	Spanish	Sp	Slovak	Sk
				Slovene	Sn

Table 1. Official EU languages selected for the Micrela project. Languages are listed in alphabetical order within families. Abbreviations will be used in Figure 2.

2.1. Intelligibility test battery

Intelligibility testing can be done in two ways, called opinion testing and functional testing (Van Bezooijen and van Heuven, 1997). Opinion testing was applied by American structuralists in the 1950s when examining the mutual intelligibility among Amerindian languages. The method was called 'ask the informant'. In opinion testing the researcher asks how much of a spoken or written text the informant thinks s/he understood. In functional testing, the informant reads, or listens to, a text sample and has to demonstrate to the experimenter whether the contents were understood, for instance by noting a spoken text down on paper, by answering content questions or by providing a summary or translation. This method was called 'test the informant' (Voegelin and Harris, 1951; Hickerton, Turner and Hickerton, 1952; Pierce, 1952; Wolff, 1959). Generally, the functional method is more time consuming than opinion testing. The two methods tend to be strongly correlated but the correlation is often imperfect (Tang and van Heuven 2009, 2015a) so that it seems wise to employ functional testing rather than approximate the level of intelligibility through the short cut of opinion testing. In the present case we tested intelligibility between language pairs through the administration of six functional tests, which covered spoken and written communication at the level of (i) single words, (ii) detailed sentence intelligibility and (iii) global message understanding at the text level. We will now briefly sketch the six tests we developed.

Word-level test. A list was compiled of the 100 most frequently used nouns in the British National Corpus (BNC Consortium, 2007). This list was translated into the 16 target languages and recorded by four native speakers (two female) for each of the 16 target languages. A subset of 50 words was selected from the larger set, avoiding polysemous words but otherwise a random choice. Two versions of the word test were prepared, one for visual presentation and one for oral presentation. In the written version the stimulus word remained on screen until the participant finished typing the response (by pressing the return key) with a timing out of 10 seconds. In the oral mode the stimulus words was made audible twice with one second between tokens with a maximum time lapse of 10 seconds (trial to trial onset). Each recorded speaker contributed one quarter of the stimulus words per listener. Speakers and words were rotated through a Latin square design. The stimulus word had to be translated

from the stimulus language into the participant' s native language by typing the translation on the keyboard. Responses were considered correct if they were (nearly) identical to the target word or one of its synonyms. A maximum of two misspelled letters was allowed as long as the resulting form was not identical to another word in the target language. Any other response was counted as an error.

Sentence intelligibility test. Four English texts at the B1/B2 level of difficulty, as defined by Common European Framework of Reference for languages (Council of Europe, 2001), were selected and adjusted to a length of 200 words each. These texts were translated into each of the 16 target languages and recorded by the same four speakers as above. Sentence intelligibility was tested through a cloze procedure.[3] In our version of the cloze test, twelve words were deleted from each text at roughly equidistant intervals and replaced by a gap of uniform length. In the written version of the cloze test, the passage was presented on screen in its entirety. A grid with response alternatives was continually shown at the top of the screen, listing four nouns, four adjectives and four verbs in three columns. The respondent's task was to drag and drop each of the twelve response alternatives in the appropriate slot in the text. The task had to be completed within ten minutes. In the oral version of the test the same twelve target words were replaced by a beep (1000-Hz sine wave) of uniform length (1 second). Each sentence was made audible twice in a row, with 1 second between first and second presentation. Each sentence contained just a single beep. Within 30 seconds the participant had to click one of the twelve response alternatives in the grid presentation on screen such that it best fitted the missing word (beep) in the utterance just heard. Once clicked the response alternative was greyed out. Responses were automatically scored as correct or wrong.

Text comprehension test. The same four 200-word texts were also used to establish the participant's global comprehension of the text, i.e. to determine whether the participant got the gist of what the text is about. The passage was shown or played to the participant in its entirety after which the participant had to click one of four pictures that were shown on the screen such that the picture chosen optimally matched the contents of the passage presented. The four pictures were constructed such that they embodied the correct or wrong representation of two essential ingredients in the passage. For instance, if the passage was about driving a car in winter, one picture showed a car driving in a wintery landscape, another picture showed a car driving in summer (with a sunny landscape and trees and flowers in full bloom), a third picture would show a plane flying over a wintery landscape and a last picture contained a plane in a summer setting. When both content features were correctly identified the participant got full marks, when both aspects were wrongly identified no mark was given, when one feature was correct the participant was given half marks.

The intelligibility tests as well as a questionnaire (see next section) were presented to participants over the internet. Potential participants were alerted to the existence of the test through social media (especially Facebook). In exchange for their service participants

[3] A cloze test (also cloze deletion test) consists of a portion of text with words removed at regular intervals (e.g. every sixth word), where the participant is asked to replace the missing words. Cloze tests require the ability to understand context and vocabulary in order to identify the correct words or type of words that belong in the deleted passages of a text. This exercise is commonly administered for the assessment of native and second language learning and instruction. The word cloze is derived from closure in Gestalt theory. The exercise was first described by Taylor (1953). Our version of the cloze test varies from the classical edition by presenting a list of printed alternatives to choose from.

received a stake in a lottery with attractive prizes to be won, such as a tablet. It is important to note here that any single participant took just one of the six tests we developed. This had to be done in order to keep the total test time within acceptable bounds: filling in the rather lengthy questionnaire left just enough time for one intelligibility test. Moreover, a participant did the test in just one non-native language, and never did participants take a test in their native language (we simply assumed near-ceiling performance in the native language). Tests, languages and speakers were assigned to participants (blocking and rotated by a Latin square design) by the internet application which controlled the assignment of conditions and on-line data collection from a central server at Groningen University.[4]

2.2. Individual data (extra-linguistic variables)

In order to collect information on extra-linguistic variables that might influence the intelligibility of a stimulus language, each participant was asked to answer a fairly comprehensive questionnaire on language attitudes towards, prior exposure to, and familiarity with, a number of European languages. Participants were asked to specify their age, sex, country they had grown up in, country they had spent most of their lives in (and how many years), whether they lived within 50 kilometres from an area where another language than their own is spoken, which language they normally heard at home, and how many years they had spent learning English.[5] The next stage in the questionnaire asked the participants to specify their self-estimated level of understanding for each of the related languages within the family of their native languages. For instance, participants with a Germanic native language were asked to rate their understanding (on a five-point scale between 1 ‘ none at all’ to 5 ‘ very good’) of Afrikaans, Danish, Dutch, Faroese, Frisian, German, Icelandic, Luxembourgish, Norwegian, and Swedish. Next, they were asked to indicate how ugly or beautiful they found each of the five or six official languages within their family, again on a five-point scale (1 ‘very ugly’ to 5 ‘very beautiful’). Participants gave their ratings once without having heard a spoken sample of the languages. They then listened to each of the five or six stimulus languages within their language family and rated the samples for perceived beauty. For each language, they then heard the first paragraph of the declaration of human rights spoken by one of the four speakers (random choice) who had also recorded the stimulus materials for the intelligibility tests. Participants now also indicated whether they knew what language they had just listened to. Participants were then informed which language they would next hear (or see) as the stimulus language in a functional intelligibility test. They were asked to indicate whether they had learned this language, and to estimate in a rather detailed manner how much exposure to the stimulus language they had had by listening to live speakers, watching speakers through the media (television, DVD, movies), playing computer games, talking to people (live, or through chatting/skyping over the internet) and by reading books, newspapers or text on the internet. The intensity of the non-native language exposure was rated on a five-point scale between 1 ‘never’ and 5 ‘daily’.

This extra-linguistic information was collected immediately before the participant did the functional intelligibility test (one of six tests). Once participants had completed their test, they were asked again to rate the five or six languages within their language family in terms of beauty. This was done to check *post hoc* whether having had a positive or negative

[4] The whole experiment (questionnaire and intelligibility testing) was done online through a custom-made web application available at www.micrela.nl/app.

[5] One part of the Micrela project is about the pros and cons of using English as a Lingua Franca (ELF) as opposed to using receptive multilingualism. The present paper does not deal with the ELF aspect of the project.

experience with the functional test might have an influence on the perceived beauty of the stimulus language.[6]

3. Linguistic distances

Within each of the three language families, linguistic distances were measured between all pairs of the five (Germanic, Romance) or six (Slavic) stimulus languages used in the project. Languages do not differ from one another along just a single dimension. Therefore we computed distances at various linguistic levels, so that we would be able to correlate distances at each of these levels with each other (intercorrelations) to see to what extent the distances at the various levels express the same type of information, so that we may reduce the dimensionality to one or two underlying distance parameters through factor analysis. Ultimately our goal is to predict the functional intelligibility scores emerging from the experiment from the linguistic distances found between the stimulus language and the listener's native language (taking account at the same time of the listener's past exposure to the stimulus language and his/her attitude towards it).

Distances were computed at five linguistic levels, viz. lexical (whole words, affixes only), phonetic, orthographic, and syntactic. We will briefly characterize the measures below. It should be noted that we have chosen to compute the distance measures for just the materials that were used in the experiment. A different choice could have been made here, for instance by adopting distances computed over large language resources such as complete dictionaries or large (spoken or written) text corpora. We know that distance measures become more stable and correlate better with mutual intelligibility scores (whether functional or opinion) as the materials the distances are computed on get larger (Tang and van Heuven, 2015a) but this relationship may well be different if the distance measures are specifically based on the stimulus materials used in the intelligibility tests, as shown by Tang and van Heuven (2015b).

Lexical distance. In our project lexical distance is defined as the percentage of cognates in the native language of the listener given the stimulus language. Lexical distance may be asymmetrical. A word in language A may have a cognate in language B but a synonym in B need not have a cognate synonym in A. For instance, the Dutch word *plek* 'place, spot, location' has no cognate in German. The nearest equivalent for *plek* in German would be *Ort*, which is cognate to Dutch *oord*. Generally, of course, the larger the percentage of cognates shared between two languages, the better the mutual intelligibility between them.

Orthographic distance. Orthographic distance is computed only for cognate word pairs in two languages. Cognate words have descended from the same word in the ancestor language from which languages A and B have developed historically. As a result, cognates bear a certain degree of cross-linguistic similarity, such that a speaker of language A may still recognize the cognate form in language B. The degree of similarity between cognates is computed by the Levenshtein algorithm, which computes the smallest number of string edit operations needed to convert the orthographic string in language A to the cognate string in B. Possible string operations are deletions, insertions and substitutions of symbols. Since a substitution is a logical combination of a deletion and an insertion, the latter two incur half a penalty point, whereas the former counts as a full penalty point. Operations merely involving

[6] In the Micrela project it is assumed that the perceived beauty (or ugliness) of a language also expresses the participant's positive (or negative) attitude towards the language concerned and its speakers. Hence the ugly-beautiful scale serves as an overall language attitude scale.

a diacritic (e.g. umlaut, accent) incur half a penalty point. The string operations are applied to aligned cognate pairs. The total number of penalty points is then divided by the length of the alignment (number of alignment slots) to yield a length-normalised Levenshtein distance for a cognate word pair. The overall orthographic distance is the arithmetic mean of the normalised distances for all cognate word pairs in a corpus. The measure is symmetrical between language A and B by definition (for more explanation and background, see Heeringa 2004). As an example, the orthographic distance between Dutch *oord* and German *Ort* would be established as follows:

Dutch	o	o	r	d	
German		O	r	t	
	.5			1	= 1.5 / 4 = 0.375

Obviously, the larger the distance between the cognates, the more difficult it will be for the reader to realize that A is the counterpart of B.

Affix distance. The above orthographic distance is computed on whole words. The affix distance does the same thing, in principle, but the aligned strings contain the affix portions of the cognate words only. This measure serves as an estimate of the degree of morphological similarity between two related languages (and excludes the similarity between the word stems). Also, the more the affixes are alike in two related languages the easier it is to isolate the word stems.

Phonological distance. This distance measure is computed in the same way as is done for the orthographic distance but now the string operations are applied to (either broad or narrow) phonetic transcriptions. A stress or tone mark is considered a base symbol rather than a diacritic. A stress difference between cognates counts as one penalty point (it yields a transposition, i.e. a deletion of a symbol in one alignment and an insertion of the same symbol at another).

Syntactic distance. Although syntactic differences between related languages are generally small, hearing (or seeing) words in unusual positions in a sentence may compromise intelligibility. We defined a syntactic distance measure by computing the correlation between syntactic trigram frequencies between two languages (see Nerbonne and Wiersma, 2006).[7] For a given language family some ten lexical categories are defined, e.g. noun, adjective, adverb, modal, verb, determiner, pronoun, quantifier, …, sentence boundary, other. All trigrams (different sequences of three lexical category labels) are then inventoried and counted in a text corpus in language A and (preferably a maximally literal translation of the texts) in language B (as long as the translation is not ungrammatical in B). This yields different frequencies of the same trigrams in language A than B. Syntactic distance is then defined as 1 minus the Pearson correlation coefficient (r) found between the trigram frequencies.[8]

[7] Several other syntactic distance measure were piloted in the course of the Micrela project, such as the mean distance over which displacements of word is observed between literal translations of the same sentence in languages A and B. We finally adopted the trigram measure as this distance measure correlated strongly with the other methods but was much more cost-efficient (computationally easy, no human intelligence needed once parts of speech are assigned).

[8] Theoretically the correlation between the trigram frequencies in two languages can be negative, i.e. when trigrams that occur quite frequently in A never occur in B and vice versa. However, if we subtract −1 from 1, the

Before looking at the results of the intelligibility tests, we will now present some findings obtained in terms of linguistic distances. Here we will limit the presentation to just one language family, viz. Romance.

Table 2 list for each of ten Romance language pairs the distance scores obtained for each of the five distance types. In this table we specify asymmetrical as well as symmetrical distances. The latter type is computed by averaging over the AB and BA distances within a pair.

Language pair		Asymmetrical (N = 20)					Symmetrical (N = 10)				
Stimulus	Listener	Lex	Orth	Suffix	Phon	Trigr	Lex	Orth	Suffix	Phon	Trigr
French	Italian	21.6	47.3	34.4	47.2	21.0	23.0	48.0	34.4	47.2	21.0
French	Portuguese	17.8	44.7	38.7	43.3	23.0	19.4	44.4	38.7	43.3	23.0
French	Romanian	49.0	51.1	45.0	44.5	31.0	53.4	51.8	45.0	44.5	31.0
French	Spanish	20.0	38.8	28.2	51.3	16.0	22.2	38.1	28.2	51.3	16.0
Italian	French	24.4	48.8	34.4	47.2	21.0					
Italian	Portuguese	11.5	43.6	38.6	39.4	13.0	16.4	42.2	38.6	39.4	13.0
Italian	Romanian	54.0	50.0	39.0	32.8	26.0	53.1	49.6	39.0	32.8	26.0
Italian	Spanish	14.4	40.2	32.6	27.7	11.0	12.4	40.5	32.6	27.7	11.0
Portuguese	French	21.0	44.1	38.7	43.3	23.0					
Portuguese	Italian	21.3	40.8	38.6	39.4	13.0					
Portuguese	Romanian	52.0	51.7	48.7	39.0	20.0	53.0	45.4	48.7	39.0	20.0
Portuguese	Spanish	6.0	26.4	23.9	37.9	14.0	3.9	27.2	23.9	37.9	14.0
Romanian	French	57.8	52.5	45.0	44.5	31.0					
Romanian	Italian	52.2	49.1	39.0	32.8	26.0					
Romanian	Portuguese	53.9	39.0	48.7	39.0	20.0					
Romanian	Spanish	54.0	49.2	41.5	37.1	26.0	50.0	49.7	41.5	37.1	26.0
Spanish	French	24.3	37.4	28.2	51.3	16.0					
Spanish	Italian	10.4	40.7	32.6	27.7	11.0					
Spanish	Portuguese	1.8	28.0	23.9	37.9	14.0					
Spanish	Romanian	45.9	50.2	41.5	37.1	26.0					

Table 2. Lexical, orthographic, affix, phonological and syntactic distances computed for 20 Romance language pairs. In the right-hand part of the table the same distances are specified but after averaging over AB and BA orders within the same language pair.

	Linguistic level									
	orthography		suffix		phonic		lexical		trigrams	
orthography			.761	< .001	.045	.425	.731	< .001	−.733	< .001
suffix	.796	.003			−.075	.377	.799	< .001	−.616	.002
phonic	.047	.449	−.075	.418			−.030	.450	−.218	.378
lexical	.784	.004	.808	.002	−.030	.467			−.793	< .001
trigrams	−.767	.005	−.616	.029	−.218	.273	−.802	.003		

.

Table 3. Upper triangle of correlation matrix computed for five linguistic distance variables. Pearson correlations were computed for 20 pairs of Romance languages, where reversal of directionality between stimulus language and receiver language yielded a different pair. In the lower triangle the same correlations were computed after averaging over the AB and BA order of the languages within each pair (ten pairs). Each cell contains the r-value and the associated p-value. Probabilities were established by 1-tailed testing.

---

syntactic distance would be +2. Strictly speaking, then, the syntactic trigram distance should be rescaled between 0 and 1 as follows: distance = (1 − *r*) / 2. In our project we chose not to do this.

Generally there is not much difference in the strength of the correlation coefficients that were computed for the 20 AB and BA language pairs separately and those that were obtained for the ten symmetric pairs that remain after averaging over the AB and BA orders within language pairs. This means that the distance measures established for the five Romance languages in our sample are more or less symmetrical. We will therefore limit the remainder of this section to only the ten language pairs that remain after averaging. Closer inspection of table 3 then reveals that distances at all linguistic levels are strongly intercorrelated (be it negatively when syntactic trigram distance is involved) with the exception of the phonic level, which behaves rather independently of the other levels. This is confirmed by a factor analysis which yields two principal components with an eigenvalue > 1. The first factor (PC1) accounts for 65% of the variance and has high loadings on all non-phonic distances (after Varimax rotation with a Kaiser window); the second factor (PC2) accounts for another 21% of the variance and loads exclusively on the phonic distance. We decided to compute a simple non-phonic index by taking the (unweighted) geometric mean of the non-phonic distances. Figure 1 shows the result of multi-dimensional scaling (MDS) of the five Romance languages using the phonic and non-phonic distances as input to the proximity scaling procedure. Note that the arrangement of the five languages in the output graph closely resembles the geographic location of the five languages. In order to see the resemblance Figure 1 has to be rotated counter-clockwise by some 45 degrees and then horizontally flipped – which transformations do not affect the distances between the languages in any way. One possible interpretation of this finding is that in the case of the five Romance languages geographic distance lead to greater linguistic differences between languages with a common origin. It is beyond the scope of this article to speculate on the mechanisms of language change that are responsible for this isomorphism between geographic and linguistic distance.[9]

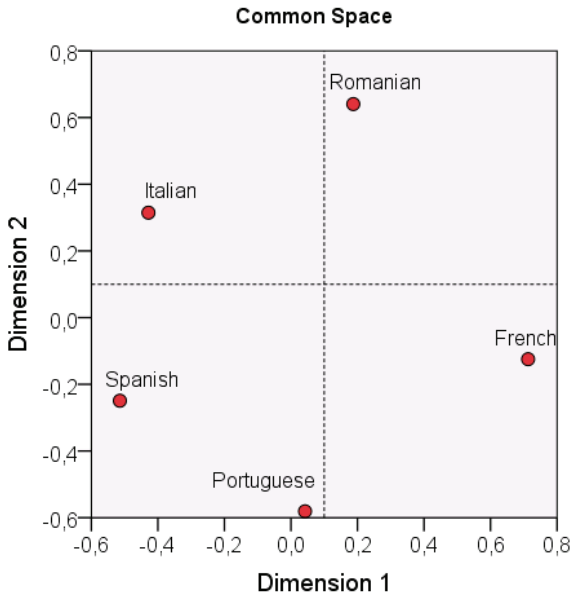

Figure 1. MDS of five Romance languages by ProxScal procedure with Phonic versus Non-phonic distance as input variables.

[9] The strong correlation between geographic and linguistic distance in the Romance language sample has been computed (without taking syntactic trigram distance into account – which measure was not available at the time) by Heeringa et al. (2013). The MDS mapping is a novelty of the present paper.

4. Results

Rather than present a comprehensive overview of the results obtained by our participants on the six intelligibility tests, we will limit the presentation in this article to just one test, viz. the results obtained for the spoken cloze test. For a complete overview of the results we again refer the reader to the individual doctoral theses mentioned earlier.

In all, at the moment of writing this article, more than 40,000 persons contributed their data through the internet application. These were male or female participants of all ages between 18 and 80, with diverse levels of education and with varying degrees of familiarity with non-native languages, and hailing from a great many European countries. From this pool of participants we made a first selection based on the following inclusion criteria. Participants had to be between 18 and 34 years of age, should have at least a secondary-school education, be born and bred in one of the 16 countries that are the native soil of our target languages, i.e., Denmark, Germany, Netherlands, Sweden, United Kingdom (for the Germanic family), France, Italy, Portugal, Romania, Spain (for the Romance group), and Bulgaria, Croatia, the Czech republic, Poland, Slovakia and Slovenia (for the Slavic family). This left roughly 13,000 participants, distributed over 70 stimulus language-native language combinations. Each of the six tests for each of the 70 language combinations was filled with between 12 and 65 participants (one participant would participate in just one test – see above).

Figure 2A-B-C plots the percentage of correctly restored words in the spoken cloze test for each of the 20 Germanic, 20 Romance and 30 Slavic language combinations. Each language pair occurs twice, viz. as an AB and a BA pair. We know that some of these pairs must yield asymmetrical results. For instance, it has been documented profusely that Swedish is easier to understand for Danish listeners than vice versa (e.g. Gooskens et al., 2010 and references therein). Similarly, Brazilian-Portuguese listeners obtained better intelligibility scores with Spanish than Argentinian and Mexican Spanish listeners did with (Brazilian) Portuguese (Jensen, 1989).

The overall level of mutual intelligibility does not differ very much from one language family to the next. Germanic languages seemingly do better than Romance and Slavic languages but it should be borne in mind that English is a compulsory subject at secondary schools in the four Germanic countries where English is not the native language. The four language combinations with English as the target language yield the highest scores within the Germanic group (between 86 and 94 percent correct, with little variation between native languages). The figure also shows that within each language family the intelligibility scores vary enormously, between some 10 percent intelligibility to near-ceiling performance, depending on the specific language combination. The highest degree of mutual intelligibility is found for the Czech-Slovak (93%) and Slovak-Czech (95%) combinations.

Our results do corroborate the classical asymmetry reported between Spanish and Portuguese. In our study Portuguese listeners obtain better scores with Spanish (77%) than the other way around (37%). This asymmetry is, in fact, substantially greater than what was reported by Jensen (1989). The Danish-Swedish asymmetry is not reproduced in our data; the reversed difference, however, is not statistically significant (52% for Swedes responding to Danish versus 57% for Danes listening to Swedish).

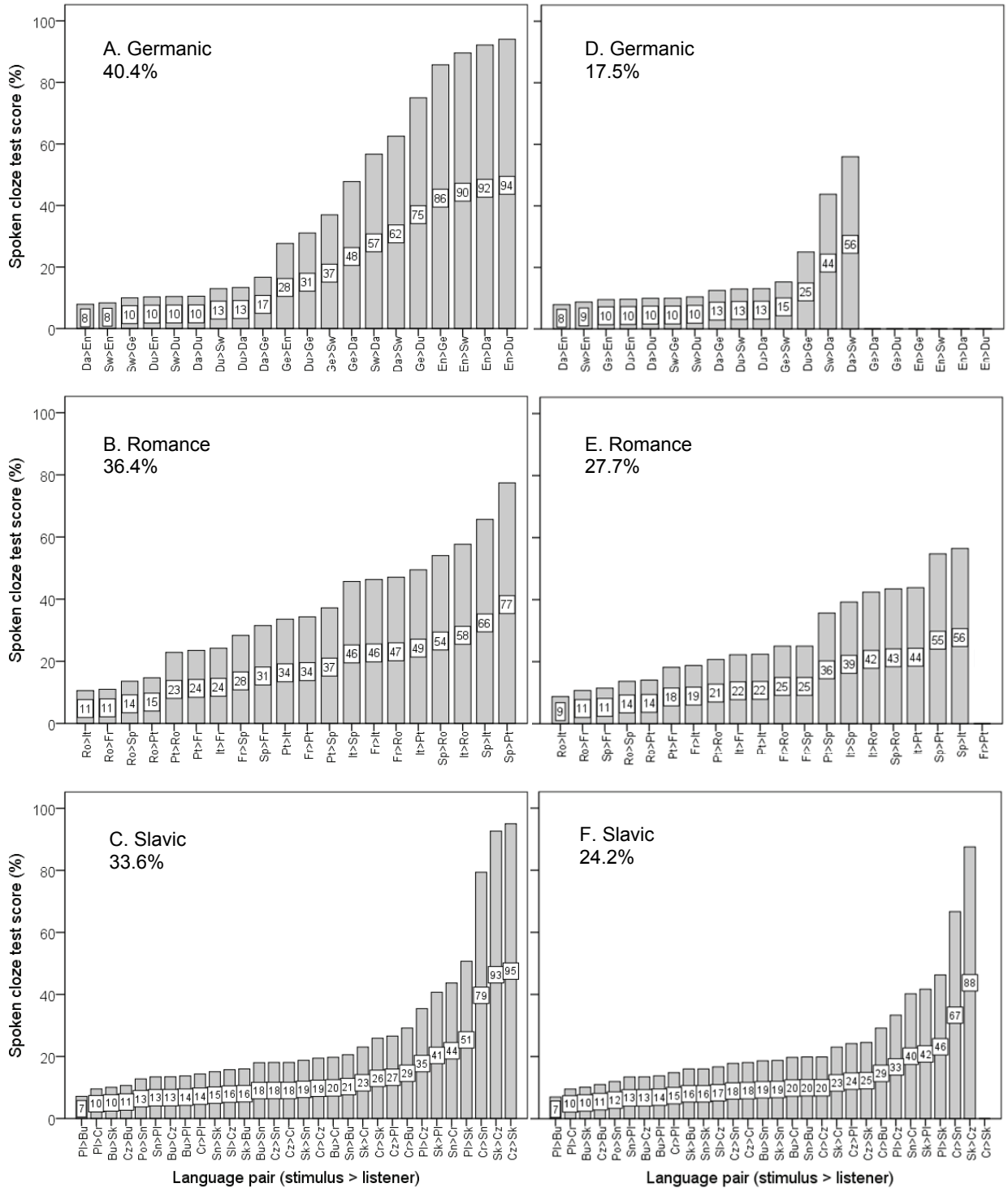

Figure 2. Intelligibility scores (percentage of correctly restored items in a spoken cloze test) for 20 Germanic, 20 Romance and 30 Slavic speaker-listener language combinations. Panels A-B-C show results for all participants who met the original inclusion criteria (inherent + acquired intelligibility). Panels D-E-F contain the results after further selection leaving only participants with little or no exposure to the stimulus language (inherent intelligibility only). Overall means across all language combinations in a family are indicated.

We would now like to know how well the intelligibility scores for each language combination can be predicted from the linguistic and non-linguistic variables we collected in the project.

We regressed the intelligibility scores against all linguistic and non-linguistic predictors in a step-wise procedure. Non-linguistic predictors were the mean exposure the participant reported (averaged over five categories of exposure, see section 2), and the number of years participants had learnt the target language at school). Linguistic predictors were lexical distance, phonic distance and syntactic trigram distance. The linguistic distances did not vary on a subject-individual basis but were constant per language combination. Other predictors were much weaker and were eliminated in an early stage of the analysis.

Table 4 specifies the results of the regression analysis. The variability in the intelligibility scores obtained by each of the 70 language combinations (i.e. across Germanic, Romance and Slavic languages) can be predicted quite well with just three predictors.[10]

Model	All languages (N = 70)	R^2
1	Exposure	.82
2	Exposure + Lexical	.88
3	Exposure + Lexical + Phonic	.89
Model	Germanic (N = 20)	R^2
1	Exposure	.86
2	Exposure + Lexical	.93
Model	Romance (N = 20)	R^2
1	Exposure	.74
2	Exposure + Phonic	.80
Model	Slavic (N = 30)	R^2
1	Exposure	.85
2	Exposure + Lexical	.90

Table 4. Stepwise regression models optimally accounting for cross-language intelligibility scores (spoken cloze test) for all language pairs together, and for Germanic, Romance and Slavic language pairs separately. R^2 values are cumulative.

The most powerful predictor, obviously, is the amount of exposure to the target language reported by the individual participant. This predictor by itself accounts for 82% of the variance. Number of years studying the target language at school does not make a contribution. The intercorrelation between exposure and years of study is so high that years of study cannot make an independent contribution. Also, attitudes, whether positive or negative, on the part of the listener towards the non-native stimulus language do not influence the intelligibility scores. When it comes to linguistic predictors, lexical distance is the only predictor that makes a significant contribution to predicting mutual intelligibility. Including this predictor in the regression model increments the R^2 value by another six points. Adding phonic distance adds one more point bringing the total explained variance to 89%.

If we run the regression analysis separately for each language family, exposure remains the most influential predictor with R^2 values between 74% (Romance languages) and 86% (Germanic languages). Here, too, secondary contributions are made by linguistic predictors adding some 5 (Slavic) to 9 (Germanic) points. In the Romance group phonic distance is the only linguistic predictor that makes a significant contribution.

[10] The statistics we report here are based on group means only (where a group comprises all selected participants in a specific combination of stimulus language and listener language). Predictions are substantially noisier when the extra-linguistic variables are entered on a participant-individual basis.

It is obvious that exposure to the target language is a factor of overriding importance when it comes to cross-language intelligibility. Listeners who are familiar with the non-native language will understand it better than those who have had no prior exposure to it. The intelligibility scores presented above would therefore seem to be largely a matter of acquired bilingualism (see introduction). We will now attempt a second analysis, in which we intend to zoom in on intrinsic bilingualism. In the ideal case we would need listeners with no prior exposure to the target language whatsoever, the so-called *tabula rasa* condition (see introduction). This condition can be approximated by making a further selection from our set of participants such that those who remain included have (virtually) no experience with the stimulus language. We therefore selected a subset of listeners who had indicated that their exposure with the target language was either 1 ('none at all') or 2 ('little') and with less than one year of learning the target language at secondary school. This reduced not only the number of participants to 1,255 but also the number of language combinations represented by at least one listener. For instance, there are no Dutch participants older than 18 years who have not learnt English in school for less than one year. Eight out of the original 70 language combinations are no longer represented, six of which are in the Germanic group.

The results of this exercise are shown in Figure 2D-E-F, which has been drawn analogously to Figure 2A-B-C. It is seen in panels D-E-F that the overall intelligibility has dropped as a results of the more stringent selection criteria. Inherent cross-linguistic intelligibility within the Germanic group has decreased to 17.5%. In the Romance and Slavic groups, however, considerable mutual intelligibility persists, with means of 27.7 and 24.2%, respectively. Be this as may, there are some language combinations in each of the three languages groups that afford a reasonable cross-linguistic intelligibility (> 50%).

The only source of information from which we may predict the differences in intelligibility is in the linguistic distance between the languages. We tested the predictive strength of the linguistic distance measures in a stepwise multiple regression analysis. The results are summarized in Table 5.

Model	All languages (N = 62)	R^2
1	Lexical	.51
2	Lexical + Phonic	.63
3	Lexical + Phonic + Syntactic	.68
Model	Germanic (N = 14)	R^2
1	Lexical	.91
Model	Romance (N = 19)	R^2
1	Syntactic	.31
Model	Slavic (N = 29)	R^2
1	Lexical	.60
2	Lexical + Syntactic	.66

Table 5. Stepwise regression models optimally accounting for cross-language intelligibility scores (spoken cloze test) for all language pairs together, and for Germanic, Romance and Slavic language pairs separately. R^2 values are cumulative. Only listeners with little or no exposure to the stimulus language were included. No listener had learnt the stimulus language for more than twelve months in the school curriculum.

Prediction of inherited mutual intelligibility between European languages is possible with considerable success: nearly 70 percent of the variance in the spoken cloze test scores can be

predicted by a combination of three linguistic predictors, i.e. lexical, phonic and syntactic distance between the listener's native language and the non-native related stimulus language. When we run the analyses separately for each language group, mutual intelligibility between Germanic languages can be predicted at 91 percent, where the prediction is done solely on the basis of the percentage of cognates shared between the native language and the related language. The poorest predictions are afforded in the Romance family. Only one linguistic predictor makes a significant contribution, i.e. syntactic distance. Syntactic distance here is strongly intercorrelated with lexical distance (at r = .91) but the latter does not make an independent contribution once the intercorrelation is factored out. Mutual intelligibility between our Slavic languages can be accounted fairly well from a combination of two linguistic predictors, i.e. lexical and syntactic distance. Phonic distance is strongly inter-correlated with lexical distance (r = .79) but does not contribute independently.

5. Conclusions and discussion

The first and most important conclusion that can be drawn from the results presented is that mutual intelligibility between related languages within the three major language families within Europe, is relatively poor, with an average within language families of some 35% intelligibility – as measured by the spoken cloze test (better results may be afforded by written tests and/or tests targeting global text comprehension). Moreover, a large part of the intelligibility appears to be due to non-linguistic factors, of which prior exposure to the related language is the most important one. Interestingly, positive or negative attitudes towards the stimulus language, do not make a significant contribution to the prediction of mutual intelligibility – counter to what is often claimed in the literature.

If we practically eliminate the influence of non-linguistic factors, our results show that mutual intelligibility suffers considerably. Scores go down from 35% to a mere 25% on average. Still, there is mutual intelligibility between the related languages, which may be quite good in some cases, and quite poor in others. Obviously, advocating receptive bilingualism as the preferred mode of communication is ill advised. In each specific pairing of a native language and a related target language, receptive bilingualism should be offset against using English as a *lingua franca*. Which of the two will be the preferred mode of communication, will depend on the specific language combination in each of the three major European language families.

When acquired bilingualism is abstracted away from, by selecting non-native listeners with little or no prior exposure to the related stimulus language, only the inherent similarity between the related languages remains as a way to crack the code of the non-native language. Within the Germanic language group the lexicon is the key to cracking the code. The results show that shared vocabulary is what makes the difference; whatever differences may exist in the phonetics or phonology between the related words can be overcome. This finding is in line with Van Heuven (2008), who suggests that inherent mutual intelligibility between related languages does not rely on special mechanisms but falls out as a by-product from human speech perception, which is extremely robust and relies on multiple mechanisms to deal with noisy and suboptimal input speech.

In the Slavic language group the predictive power of linguistic factors is somewhat smaller than what was found for the Germanic languages but still appreciable. Nevertheless, for this family, too, lexical similarity is the strongest predictor of inherited mutual intelligibility.

References

Bahtina, D. and Thije, J. D. ten (2013). Receptive Multilingualism. In: Chapelle, C. A. (ed) *The Encyclopedia of Applied Linguistics*. West Sussex: John Wiley online. 1-6.

Bezooijen, R. van and Heuven, V.J. van (1997). Assessment of speech synthesis. In: D. Gibbon, Moore, R., and Winski, R. (eds) *Handbook of standards and resources for spoken language systems*, Berlin/New York: Mouton de Gruyter. 481-653.

British National Corpus, version 3 (BNC XML Edition), 2007, distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Braunmüller, K. (2007). Receptive multilingualism in Northern Europe in the Middle Ages: A description of a scenario. In: Thije J. D. ten and Zeevaert, L. (eds) *Receptive multilingualism.* Amsterdam: John Benjamins. 25-47.

Council of Europe (2011). *Common European Framework of Reference for: Learning, Teaching, Assessment*. Strasbourg.

Golubovic, J. (2015). *Predicting mutual intelligibility of Slavic languages from linguistic and extra-linguistic factors*. LOT dissertation series. Utrecht: LOT (in preparation).

Gooskens, C., Heuven, V. J. van, Bezooijen, R. van, and Pacilly, J. (2010). Is spoken Danish less intelligible than Swedish? *Speech Communication*, 52, 1022-1037.

Grin, F. (2008). L'intercompréhension, efficience et équité. In: Conti, V. and Grin, F. (eds) *S'entendre entre langues voisines: vers l'intercompréhension*. Chêne-Bourg: Georg. 79-109.

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

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

Heeringa, W, Golubovic, J., Gooskens, C., Schüppert, A., Swarte, F., and Voigt, S. (2013). Lexical and orthographic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance. In: Gooskens, C. and Bezooijen, R. van (eds) *Phonetics in Europe: Perception and Production*. Frankfurt a. M.: Peter Lang. 99-137.

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.

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

Hockett, C. F. (1958). *A course in modern linguistics*. New York: Macmillan.

Jensen, J. B. (1989). On the mutual intelligibility of Spanish and Portuguese. *Hispania*, 72, 849-852.

Lüdi, G. (2007). The Swiss model of plurilingual communication. In: Thije, J. D. ten and Zeevaert. L. (eds) *Receptive Multilingualism: Linguistic analyses, language policies and didactic concepts*. Amsterdam: John Benjamins. 159-178.

Nerbonne, J. and Wiersma, W. (2006). A measure of aggregate syntactic distance. In: Nerbonne, J. and Hinrichs, E. (eds) *Linguistic Distances*. Sydney: International Committee on Computational Linguistics and the Association for Computational Linguistics. 82-90.

Pierce, J. E. (1952). Dialect distance testing in Algonquian. *International Journal of American Linguistics*, 18, 203-210.

Rehbein, J., Thije, J. D. ten, and Verschik, A. (2012). Lingua receptiva (LaRa) – remarks on the quintessence of receptive multilingualism. *International Journal of Bilingualism*, 16, 248-264.

Swarte, F. (2015). *Predicting mutual intelligibility of Germanic languages from linguistic and extra-linguistic factors*. LOT dissertation series. Utrecht: LOT (in preparation).

Tang, C. and Heuven, V. J. van (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua*, 19, 709-732.

Tang, C., Heuven, V. J. van (2015a). Predicting mutual intelligibility of Chinese dialects from multiple objective linguistic distance measures. *Linguistics*, 52, 285-311.

Tang, C., Heuven, V. J. van (2015b). Mutual intelligibility of Chinese dialects: Predicting cross-dialect word intelligibility from lexical and phonological similarity. *Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow*.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30, 415-433.

Thije, J. D. ten and Zeevaert, L. (2007). *Receptive Multilingualism.* Amsterdam: John Benjamins.

Voegelin, C. F. and Harris, Z. A. (1951). Methods for determining intelligibility among dialects of national languages. *Proceedings of the American Philosophical Society*, 95, 322-329.

Voigt, S. (2015). *Predicting mutual intelligibility of Romance languages from linguistic and extra-linguistic factors*. LOT dissertation series. Utrecht: LOT (in preparation).

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

Zeevaert, L. (2004). *Interskandinavische Kommunikation. Strategien zur Etablierung von Verständigung zwischen Skandinaviern im Diskurs.* Hamburg: Verlag Dr. Kovač.