Speech Communication 89 (2017) 25–36

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Measuring cross-linguistic intelligibility in the Germanic, Romance and Slavic language groups

Charlotte Gooskens[a,*], Vincent J. van Heuven[a,b]

[a] University of Groningen, Department of Applied Linguistics, Postbus 716, 9700 AS Groningen. The Netherlands
[b] Pannon Egyetem, Department of Hungarian and Applied Linguistics, Egyetem Ut. 10. 8200 Veszprém, Hungary

ARTICLE INFO

Article history:
Received 13 August 2016
Revised 23 February 2017
Accepted 27 February 2017
Available online 2 March 2017

Keywords:
Intelligibility testing
Functional language tests
Cloze test
Judged intelligibility
Estimated intelligibility
Perceived intelligibility

ABSTRACT

We administered six functional intelligibility tests, i.e., spoken and written versions of (i) an isolated word recognition test, (ii) a cloze test at the sentence level and (iii) a picture-to-text matching task at the paragraph level. The scores on these functional tests were compared with each other and with intersubjective measures obtained for the same materials through opinion testing, i.e., estimated and perceived intelligibility. The native language of the speakers and listeners belonged to one of three groups of European language families, i.e., Germanic (Danish, Dutch, English, German, Swedish, yielding 20 within-family pairs of different speaker and listener languages), Romance (French, Italian, Portuguese, Romanian, Spanish, yielding 20 language pairs) and Slavic (Bulgarian, Croatian, Czech, Polish, Slovak, Slovene, i.e., 30 pairs). Results from 13,566 participants were analyzed for the 70 within-family combinations of speaker and listener languages. The word recognition test and the cloze test revealed similar patterns of intelligibility but correlated poorly with the picture-to-text matching scores. Both measures of judged intelligibility (estimated and perceived) correlated highly with one another and with the functional test scores, especially those of the cloze test. We conclude that lay listeners are able to judge the intelligibility of a non-native test language from within their own language family. Moreover, participants understood written language better than the spoken forms. Advantages and disadvantages of the various intelligibility measures we used are discussed. We conclude that the written cloze procedure which we developed is the optimal cross-language intelligibility test in the European language area.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

A considerable research effort has been made during the past decades to establish the extent to which speakers of one language variety may understand the speech of another, related variety. The degree of mutual intelligibility between language varieties has been used as a criterion to decide whether the varieties were dialects of the same language or varieties of different languages (e.g., Voegelin and Harris, 1951; Hickerton et al., 1952; Pierce, 1952; Wolff, 1959). Moreover, it has been argued that an explanatory theory of language should be able to predict cross-linguistic intelligibility between languages from systematic comparison of their formal, structural properties (Van Heuven, 2008). Such an undertaking presupposes the availability of valid cross-linguistic intelligibility data.

There may also be more practical reasons to study mutual intelligibility between language varieties. When two related languages are mutually intelligible even at first confrontation, as in the Scandinavian language group (Haugen, 1966; Maurud, 1976a, b; Bø, 1978; Börestam Uhlmann, 1991; Delsing and Lundin Åkesson, 2005; Gooskens et al., 2010), there would be no need to teach the related language in the school curriculum. If related languages have a moderate degree of mutual intelligibility, educational policy makers may want to know if it would be more efficient to design a curriculum to bridge the intelligibility gap between the related languages (Klein and Stegmann, 2000; Hufeisen and Marx, 2014) than to teach both language groups to use a third language as a lingua franca (e.g., English).

A wide variety of methods have been developed to capture the degree of cross-language intelligibility in a single number. The methods differ depending on the modality (speech versus written language), age of the participants (young, adult), precision and coverage aimed at, as well as time and means available for administering the test(s). Our aim is to present, compare and evaluate results of various measures of cross-language intelligibility between closely related languages. Specifically, we are interested in

* Corresponding author.
E-mail addresses: c.s.gooskens@rug.nl (C. Gooskens), v.j.j.p.van.heuven@hum.leidenuniv.nl (V.J. van Heuven).

http://dx.doi.org/10.1016/j.specom.2017.02.008
0167-6393/© 2017 Elsevier B.V. All rights reserved.

the question of whether opinion testing can be used as a valid substitute for the more laborious functional intelligibility tests.

Methods of intelligibility testing can be categorized into two types, which we will refer to as *judged intelligibility* ('ask the informant') and *functional intelligibility* ('test the informant'). When using judged intelligibility (also called 'opinion testing') subjects indicate how well they think they understand a language or some sample produced in it. This is a quick and easy way to obtain an estimate of the intelligibility of a language.

The simplest kind of judged intelligibility testing involves no written or spoken language samples at all. Participants are simply asked how well they think they would understand language X. We refer to this method as *estimated intelligibility*. An example of such an investigation is Haugen (1966), who asked Danish, Norwegian and Swedish participants how well they believed they would understand speakers of the other two Scandinavian languages (i) if they had never heard such a speaker before, (ii) given their present experience with the other language, and (iii) how well they would themselves be understood by the other person. Responses could be indicated on a scale such as: 'not at all' — 'with great difficulty' — 'if I listen attentively' — 'all but a few words' — 'I would understand everything'. An advantage of this written method is that no stimulus materials have to be presented, which eliminates the problem of having to find speakers with equally intelligible voices in each of the languages tested. It is unclear, however, to what extent the respondents' opinions are influenced by (positive or negative) experiences with, or attitudes towards, the other language, stereotyping (e.g., 'Danish is not a language, it is a disease of the throat') and political correctness.

Alternatively, we may ask language users to judge the intelligibility of actually presented samples of spoken or written language, in which case we speak of *perceived intelligibility*. For instance, Tang and Van Heuven (2007) had native listeners of 15 different Chinese languages ('dialects') rate the intelligibility of the fable "The North Wind and the Sun", spoken in each of these 15 Chinese dialects. Participants indicated how well they believed monolingual speakers of their own dialect would understand the fable if they had never heard it before, using an 11-point scale ranging between 0 ("They will not understand a word") to 10 ("They will understand everything"). There is no guarantee, of course, that these judgments are a valid representation of the actual ('functional') intelligibility.

Most researchers, therefore, prefer to measure actual speech comprehension, i.e., by *functional intelligibility testing*. Here, listeners have to demonstrate that they effectively understood the input, e.g., by carrying out instructions given in the non-native variety or by deciding whether a sentence is true or false (for a survey of techniques see Van Bezooijen and Van Heuven, 1997; Gooskens, 2013).

Not many studies have systematically compared results of different methods of intelligibility testing. Doetjes (2007) measured the intelligibility of Swedish for Danes (and *vice versa*) using the same materials in six different tests: true/false questions, multiple-choice questions, open questions, word translation, summary, and short summary. Scores decreased from 93% correct for true/false questions to 66% for short summaries, showing that it is difficult to compare intelligibility scores across tasks. Nevertheless, Danes, for example, obtain higher scores on Swedish tests than *vice versa*, not only in Doetjes (2007) but in the majority of studies on mutual intelligibility of spoken Scandinavian languages (for a survey see Schüppert, 2011). Test scores obtained from different intelligibility tests involving the same languages and listeners groups should therefore be strongly correlated. Indeed, Tang and Van Heuven (2009,2015) report high – but imperfect – correlation coefficients for scores obtained by judged intelligibility and by functional intelligibility tests at the word and sentence levels: correlations ranged between $r = .77$ and .82 (with 225 combinations of test dialects and listener dialects). This still leaves 33 to 41 percent of the variance unexplained, and therefore the authors prefer functional testing to opinion testing, and recommend the latter method only if functional methods are either too costly or too time-consuming. Tang and Van Heuven, however, used different materials when collecting judgment data (readings of the North Wind and the Sun fable) than in their functional intelligibility tests (isolated words and a list of short everyday sentences). The correlation between judged and functional intelligibility might have been stronger if the same materials had been used for both types of test.

The present paper reports on the results of a comprehensive project on cross-language intelligibility between closely related languages within the Germanic, Slavic and Romance language groups in Europe (for background and details see Van Heuven et al., 2015; Golubović, 2016; Swarte, 2016). We used three different functional tests, i.e., a word translation task, a cloze test and a picture-to-text matching task, and applied these tests in both the written and spoken modality. The tests were carried out by means of an Internet application, so that a large number of languages and participants could be targeted. In total, 16 languages in 70 combinations of participant language and test language were functionally tested. We also collected judged intelligibility scores by having the same participants answer questions about their estimated and perceived understanding of the test language both in the spoken and written modalities. This body of data allows us to systematically compare the results of three different functional tests and to compare the results of functional tests with the opinion scores, and to establish the extent to which the results may differ across modalities.

2. Method

2.1. Functional intelligibility test battery

We tested the intelligibility of 16 languages (Section 2.2) functionally by administering six tests, which covered spoken and written communication at the level of (i) intelligibility of single words (word recognition test), (ii) intelligibility at the sentence level (cloze test), and (iii) global message understanding at the paragraph level (picture-to-text matching task). We will now briefly describe these six functional tests.

2.1.1. Cloze test

A cloze test (also called 'cloze deletion test') consists of a text with words removed at regular intervals (e.g., every sixth word), creating gaps where the participant is asked to restore 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 gaps. The term 'cloze' is derived from the notion of 'closure' in Gestalt theory. The exercise was first described by Taylor (1953). Cloze tests are often used to assess (progress in) proficiency in native and second language (e.g., Abraham and Chapelle, 1992; Keshavarz and Salimi, 2007), but it has also been used to measure intelligibility, e.g., by Scharpff and Van Heuven (1988) and Van Bezooijen and Gooskens (2005).

Our cloze test differs from the classical version in that it presents a list of printed alternatives to choose from. This obviates the problem of having to decide whether the words filled in by the subjects are acceptable alternatives. Responses can now be evaluated automatically. Four English texts at the B1 level of difficulty, as defined by the Common European Framework of Reference for languages (Council of Europe, 2001), were adjusted to a length of approximately 200 words each (16 or 17 sentences). The texts were translated into each of the 16 test languages and recorded by four native female speakers aged between 20 and 40

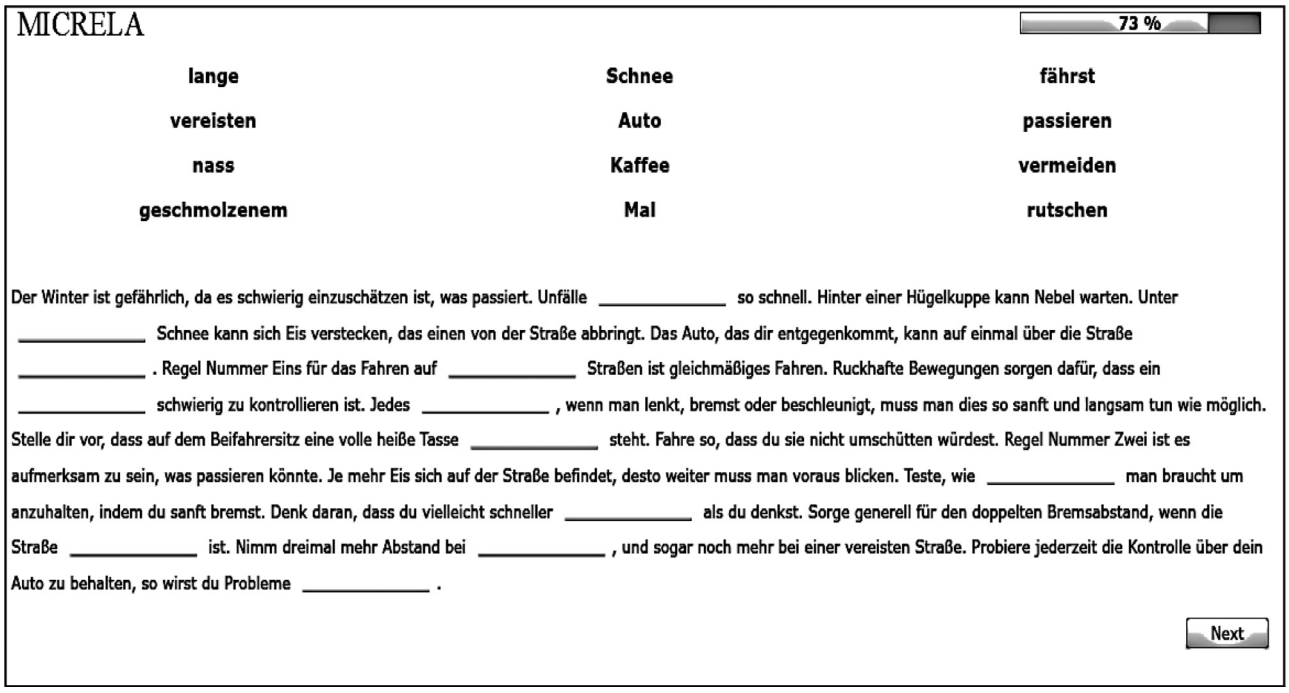

Fig. 1. Screen shot of one version of the German written cloze test.

years for each of the 16 test languages.[1] By using four different speakers for each test language we aimed to neutralize the potential influence of voice quality differences on the results. The designated speakers had been carefully selected from a larger group of speakers in online surveys. Native listeners (between 21 and 197 per language) rated the speakers by answering the question "How suitable is this speaker for presenting the news on national television?" on a 5-point scale ranging from "not at all suitable" to "very suitable". The voices of the four best-rated speakers per language were used in the experiment. The same 16×4 speakers also recorded the listening materials for the picture-to-text matching task and the word recognition test (see below).

The recording of one text from each speaker was randomly chosen to be used in the experiment. We divided each text into twelve units, corresponding to sentences or clauses. From each fragment we removed one word and replaced it by a gap of uniform length. In the written version, the passage was presented on the screen in its entirety. Twelve response alternatives were continually shown at the top of the screen, in three columns: four nouns, four adjectives and four verbs. When moving the mouse over a word a translation of the word into the native language of the participant was revealed. This was done because we wanted to test the intelligibility of whole texts. If some of the response alternatives were unknown to the participants they would not be able to place them in the right gaps, even if they understood the fragments per se. The respondent's task was to drag and drop each of the twelve response alternatives into the appropriate slot in the text. Inserted words were greyed out in the selection area, in order to help the participants keep track of their choices. In case they wanted to change an answer, they could simply drag and drop a different word into the same gap. Their original word of choice would then re-appear in black in the selection area above the text. The entire task had to be completed within ten minutes. An example of a written cloze test is given in Fig. 1.

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 s). Each fragment contained one beep. Participants listened to each fragment twice in a row, with 1 s between presentations. The participant had 30 s to select 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 selected, the response alternative was greyed out; however, it could be reused if needed in the same manner as described above for the written cloze test.

2.1.2. Picture-to-text matching task

We anticipated that the cloze test might be too difficult for some participant groups that were tested in a language which they were not familiar with or that was too distant from their native language to provide them with sufficient clues for understanding the text. Therefore, we developed a test to establish the participant's global comprehension of the same four 200-word texts that were used for the cloze test. The text 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. The four pictures 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 contained a car driving in a wintery landscape, another picture a car driving in summer (with a sunny landscape and trees and flowers in full bloom), a third picture contained a plane flying over a wintery landscape and the final picture contained a plane in a summer setting. Only when both content features were correctly identified, did the participant get full marks (picture c in Fig. 2), in all other responses no mark was given.

2.1.3. Word recognition test

The cloze test and the picture matching task test language understanding at the higher levels of linguistic organization and the

[1] All participants in this study, i.e., both the speakers and translators, as well as the listeners and readers in the on-line tests, gave their informed consent prior to their involvement. The project was approved in its entirety by the ethics committee of the Humanities Faculty at Groningen University.

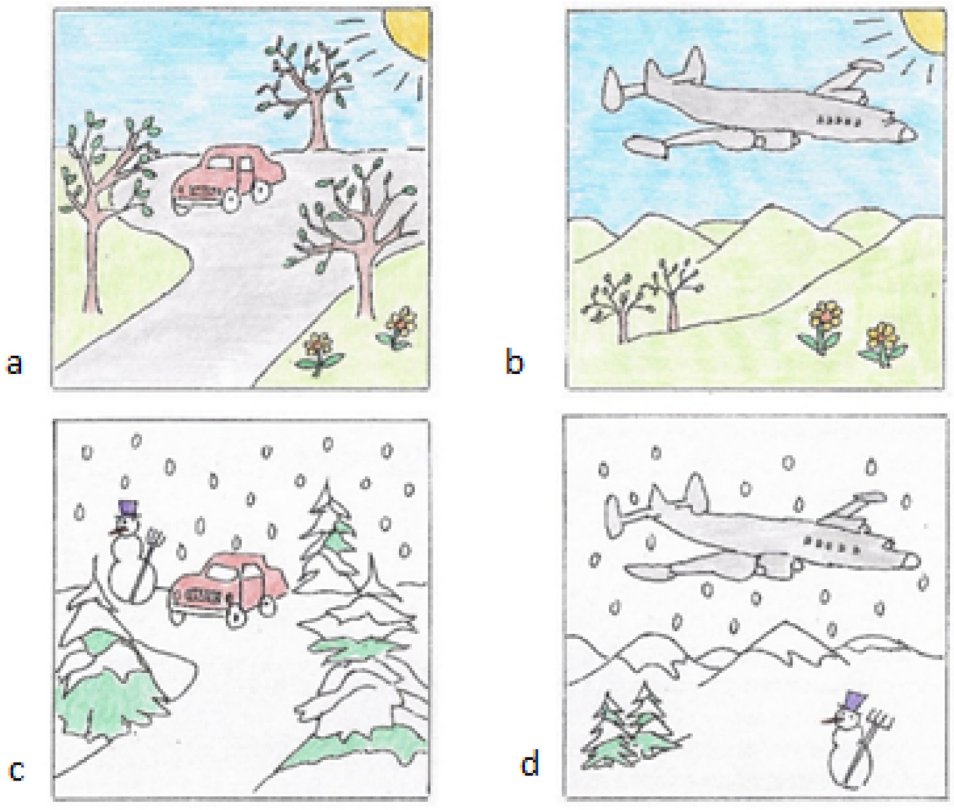

Fig. 2. Set of pictures to be matched with the text 'Driving in winter' used in the picture task. Frame (c) is the correct alternative.

exploitation of contextual redundancies. It is often expedient to test language understanding at a lower level as well, specifically at the level of word recognition, which is seen as the key process in language understanding, since the word (or lexeme) is the smallest linguistic unit in which the sensory input matches with representations stored in long-term memory (e.g., Marslen-Wilson and Welsh, 1978; Van Heuven, 2008; Cutler, 2012).

To test word recognition, a list was compiled of the 100 most frequently used nouns in the British National Corpus (BNC Consortium, 2007). The list was slightly adapted to exclude pairs of words with similar meanings (for instance, in most Slavic languages the most commonly used translation of the words 'work' and 'job' is the same word). If we excluded a word, we added the next one on the frequency list. The 100 words selected are listed in Appendix 1. The target words were translated into the other test languages and recorded by the same 16×4 speakers as used for the cloze test. Each speaker contributed a different quarter (i.e., 25 words) of the stimulus words.

Each listener was presented with a random subset of 50 words from the larger set of 100 words, to keep the duration of the test within limits. Two versions of the word test were prepared, one for visual presentation and one for oral presentation. In the written version each stimulus word was presented on the computer screen and remained visible until the participant finished typing the response (by pressing the return key) with a time out after 10 s. In the spoken modality, stimulus words were played twice with one second between tokens, again with a maximum time lapse of 10 s (trial-to-trial onset).

Participants were instructed to translate each stimulus word into their native language using the computer keyboard. While the responses to the cloze test and the picture task could be evaluated automatically, this was only partially feasible for the word test, since typing errors or synonyms might occur that were not recognised by the software. Therefore we manually checked any response that the software could not recognize. Responses were considered correct only if they were (nearly) identical to the target word or one of its synonyms. A single deletion, addition or substitution of a letter, as well as switching two letters, were considered acceptable errors, as long as the string of letters was not identical to another existing word in the participant's native language. For example, English *eye* is *oog in Dutch;* the response *oor* 'ear' is incorrect since it is an existing Dutch word while the non-word *oof* is considered correct.

2.2. Test languages

We limited our investigation to the three largest language families in the EU member states in terms of numbers of speakers, i.e., Germanic, Slavic and Romance. All 16 official national EU languages within these three language families were included. If a language is an official language in more than one country we only included the standard variety from the country with the largest number of speakers. For example, Dutch is an official language of both Belgium and the Netherlands, but since the larger number of speakers live in the Netherlands, Netherlandic Dutch was included as the test language and Flemish Dutch was not. Fig. 3 shows a map of Europe with the 16 languages/countries included in the investigation.

2.3. Participants

Since the participants were tested online, no restrictions concerning their background were set beforehand. We selected participants for further analysis by matching groups according to specific criteria. Since most participants were young adults, we focused on this group and excluded participants younger than 18 years or older than 33 years. Intelligibility was tested only among

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 (PO)
Swedish (SW)	Spanish (SP)	Slovak (SK)
		Slovene (SL)

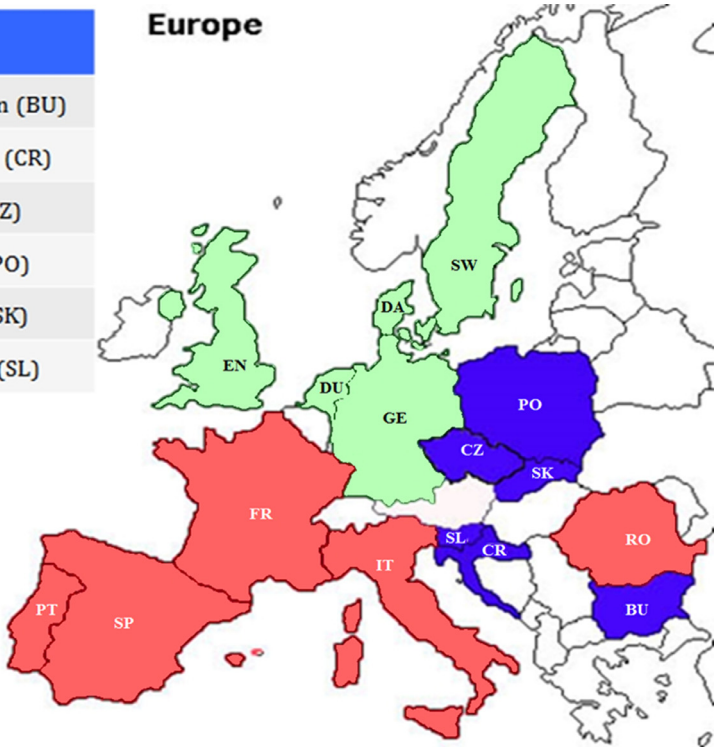

Fig. 3. Map of Europe (excluding Iceland) indicating the 16 countries/languages in the investigation (alphabetically listed within language families).

speakers of languages within the same language family. For example, the intelligibility of Danish was tested with Swedish listeners, i.e., both Germanic languages, but not with Spanish (Romance) listeners. This resulted in 70 combinations of participant and test languages. Participants had all grown up and lived most of their lives in the countries included in the project and spoke the language of the country as their native language. We excluded participants who spoke another language at home. Some of the test languages are also school languages. So, many participants had learned the test language at school. We included these participants because they are part of the representative sample of young educated Europeans we targeted. However, we excluded participants who had learned the test language for longer than the maximum period offered in the secondary-school curriculum.

The criteria described above resulted in a selection of 13,566 participants (22% from the Germanic, 31% from the Romance and 47% from the Slavic language area). A majority of the participants (65%) were female and the mean age across all participants was 23.7 years. Most participants (78%) were (or had been) university students. Given 70 combinations of speaker and participant group and six different tests (written and spoken word tests, cloze tests and picture tasks), a total of 420 different tests were conducted. The number of participants ranged between 14 and 58 per test, with a mean of 25.1.

2.4. Procedure

The intelligibility tests as well as a questionnaire (see below) were presented through an online application.[2] Potential participants were alerted to the existence of the test through social media (predominantly Facebook). In exchange for their service participants received an entry into a lottery with prizes, such as a tablet. It is important to note that each participant completed only one of the six functional tests. This was to keep the total test time within time bounds: filling in the lengthy questionnaire (see below) left enough time for only one intelligibility test. Moreover, a participant completed the test in only one randomly assigned non-native language. Participants never took a test in their native language (we assumed near-ceiling performance in the native language).

To collect information on extra-linguistic variables that might influence the test performance, each participant filled in a 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 where they had grown up, country they had spent most of their lives in (and how many years) and which language they normally heard at home. Participants were then informed which language they would next hear (or read) as the test language in a functional intelligibility test. They were asked to indicate whether they had learned this language, and to estimate in a detailed manner how much exposure to the test language they had by listening to live speakers, watching speakers through the media (television, DVDs, 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 frequency of the non-native language exposure was rated on five-point scales between 1 'never' and 5 'daily'.

Crucially, the participants indicated how well they thought they would be able to understand the test language (i.e., estimated intelligibility), prior to being exposed to any stimuli, on a 5-point scale from 1 = 'not at all' to 5 = 'very well'. It was not made explicit whether this question referred to the understanding of spoken or written language. The participants filled in the questionnaire immediately before they did the functional intelligibility test (one of six tests). Once they had completed the test, they were asked to indicate how well they thought they had understood the test language (i.e., perceived intelligibility) on the same scale as before the stimulus presentation.

[2] http://www.micrela.nl/app

3. Results

3.1. Preliminaries

We computed eight mean intelligibility scores for each of the 70 language combinations, i.e., six functional intelligibility scores (spoken and written word translation tests, cloze tests, and picture matching tasks) and two judged intelligibility scores (estimated and perceived intelligibility). The scores on the functional tests were converted to percentages by dividing the number of correct responses by the number of items in the test (and multiplying the result by 100). Judged intelligibility scores on the 5-point scales were converted by the formula (score – 1) × 25 so that 1 ('do not understand at all') and 5 ('understand very well') correspond with 0 and 100%, respectively.

The results will be analysed in two sections. In Section 3.2 we will examine the relative difficulty of the tests we developed and discuss how well each test differentiates between language combinations with high and low cross-language intelligibility, as a function of the factors Test type (word test, cloze test, picture-matching test), Modality (written vs. spoken language) and Language family (Germanic, Romance, Slavic). The analytic tool we used is a Repeated Measures Analysis of Variance with the Test type and Modality as factors within language pairs, and with Language family as a between-pairs factor. Degrees of freedom and *p*-values were Huyhn-Feldt corrected when the assumption of sphericity was violated. Partial eta squared ($p\eta^2$) is used as the metric for effect size. The discriminatory power of the tests will be quantified in terms of the standard deviation in the test scores obtained for the 20 or 30 language pairs tested within each family as well as for all 70 language pairs together.

In Section 3.3 we will examine the relationships between the eight tests we administered in more detail in order to establish how well the scores on one test can be predicted from those of another test. Specifically, we will try to determine how well the scores on the spoken functional tests discriminate between high and low cross-language intelligibility, that is, how well the cloze test results can be predicted from the perceived intelligibility scores and the opinion score obtained for the same language combinations. This part of the study will be based on an analysis of the correlation structure of the eight tests we administered.

3.2. Task difficulty and discriminatory power

Fig. 4 is a summary of the test results, averaged over the 20 or 30 language pairs in each family.

The ANOVA proves that the main effect of Language family (Germanic 60%, Romance 54%, Slavic 53%) is insignificant, $F(2, 67) < 1$. The main effect of Test type, however, is highly significant, $F(1.6, 110.4) = 184.2$ ($p < .001$, $p\eta^2 = .733$). All three test types differed from each other (Bonferroni test, $p < .05$), i.e., the Word translation test was more difficult (41%) than the Cloze test (54%), while the Picture test was the easiest (71%). The written tests (59% correct) were easier overall than the spoken versions (52%), $F(1, 67) = 76.7$ ($p < .001$, $p\eta^2 = .534$). Written and spoken tests were almost equally difficult in the Slavic language group but not in the other language groups, so that the Family × Modality interaction reached significance, $F(2, 67) = 7.0$ ($p = .002$, $p\eta^2 = .174$). Also the interaction between Test type and Modality proved significant, $F(1.6, 105.8) = 27.4$ ($p < .001$, $p\eta^2 = .290$), since written versions were easier than spoken versions for the Word test and the Cloze test but not for the Picture test. The third-order interaction fell just short of significance, $F(3.2, 105.8) = 2.4$ ($p = .073$, $p\eta^2 = .066$).

Fig. 5 shows the spread of the scores expressed as the standard deviation for each of the eight intelligibility tests. The larger the standard deviation, all else being equal, the better the discriminatory power of the test at issue.

Fig. 5 shows that there is a general tendency for the cloze tests to differentiate better between language pairs than any of the other types of test. Moreover, the written versions of the cloze tests differentiate slightly better than the spoken versions, which is especially true for the Romance language group. The two judgment tests discriminate almost as well as the cloze tests in the Germanic and Slavic group, but not in the Romance group. The discriminatory power of the Word translation and Picture matching tests, irrespective of their modality, is clearly poorer than that of the other test types.

In Fig. 6 the judged intelligibility scores (estimated and perceived) and the results of the six functional tests are presented per language family in more detail. Within each panel, the intelligibility scores (in %) are given for the 20 (Germanic and Romance families) or 30 (Slavic) different language combinations, with separate lines for each of the eight tests. Language combinations are listed along the horizontal axis in descending order of the perceived intelligibility score. The eight different intelligibility measurements generally follow the same pattern. This visual impression is confirmed by a high Cronbach's alpha (.923) computed over all 70 language combinations tested.

The cloze test is the most difficult of the three functional tests. It also has the best differentiating potential of the tests we administered (see above). Some participant groups scored almost at ceiling, e.g., the Germanic groups when tested in English, Portuguese participants tested in written Spanish, and Slovaks tested in Czech. Other language combinations yielded very low scores, e.g., English participants tested in the other Germanic languages, Romance groups tested in Romanian, and many Slavic participants, especially when tested in the spoken modality. The fact that the results cover the entire range from almost zero to almost 100 percent (SD = 25.4% for the spoken and 27.8% for the written cloze test) proves that this test differentiates very well between different levels of intelligibility.

The word translation scores were in general in the middle of the range but cover a smaller range (SD = 17.4% and 16.6% for the spoken and written task), especially in the Romance language area. Only in the Romance language combinations were the results of the word translation tasks often lower than those of the cloze test, especially in the spoken version. Romance participants appeared to find it difficult to recognize isolated spoken words, but when they were provided with context information the task seemed easier. The participants in the other language groups found it more difficult to understand running text than isolated words.

The scores on the picture task were highest overall: Many participant groups succeeded in at least extracting the essence of a text. Still, there were participant groups with high scores and groups with rather low scores (SDs of 16.1% and 18.5% for the spoken and written task, respectively), so this test, too, differentiated reasonably well between high and low intelligibility.

Importantly, we see in Fig. 6 that the judged intelligibility scores followed the scores on the cloze test (rather than those on the other tasks) quite closely, not only in a relative sense but also absolutely, and like the cloze test they covered the entire scale (SD = 23.5 for both estimated and perceived intelligibility). We hypothesized that participants might change their intelligibility judgments when they found the actual test more (or less) difficult than they expected. The judgments obtained before (estimated intelligibility) and after (perceived intelligibility) the functional test, however, were almost identical, suggesting that the participants were slightly more optimistic before ($M = 35.3\%$) than after ($M = 33.6\%$), $t(69) = 4.3$ ($p < .001$).

Let us now compare the performance on the spoken (solid lines) and written (dotted lines) tests. In general, participants un-

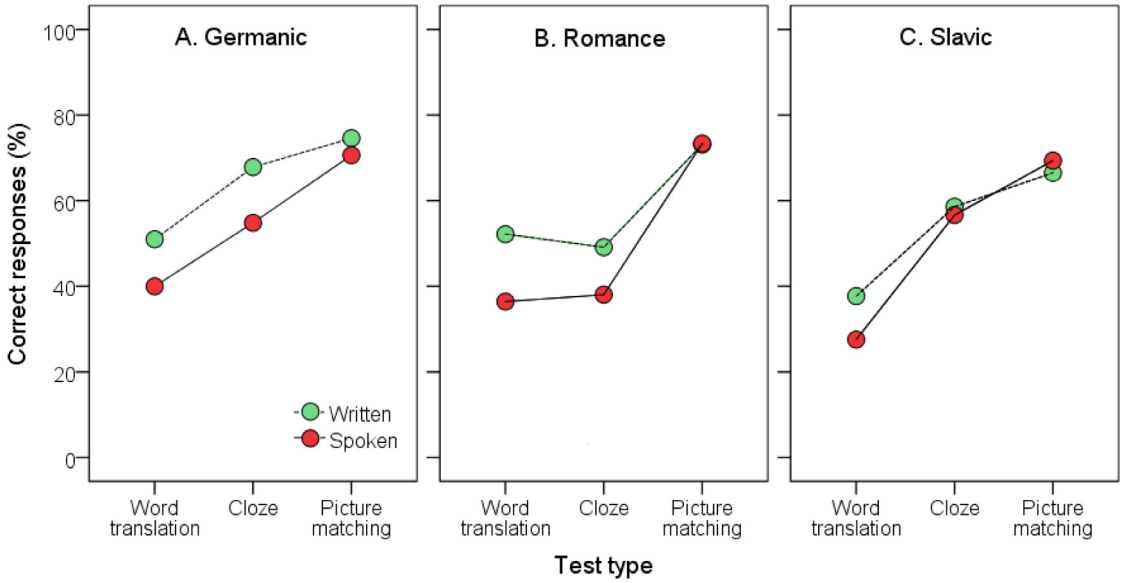

Fig. 4. Percentage of correct responses to intelligibility tests as a function of Test type and Modality for 20 Germanic language pairs (panel A), 20 Romance language pairs and 30 Slavic language pairs.

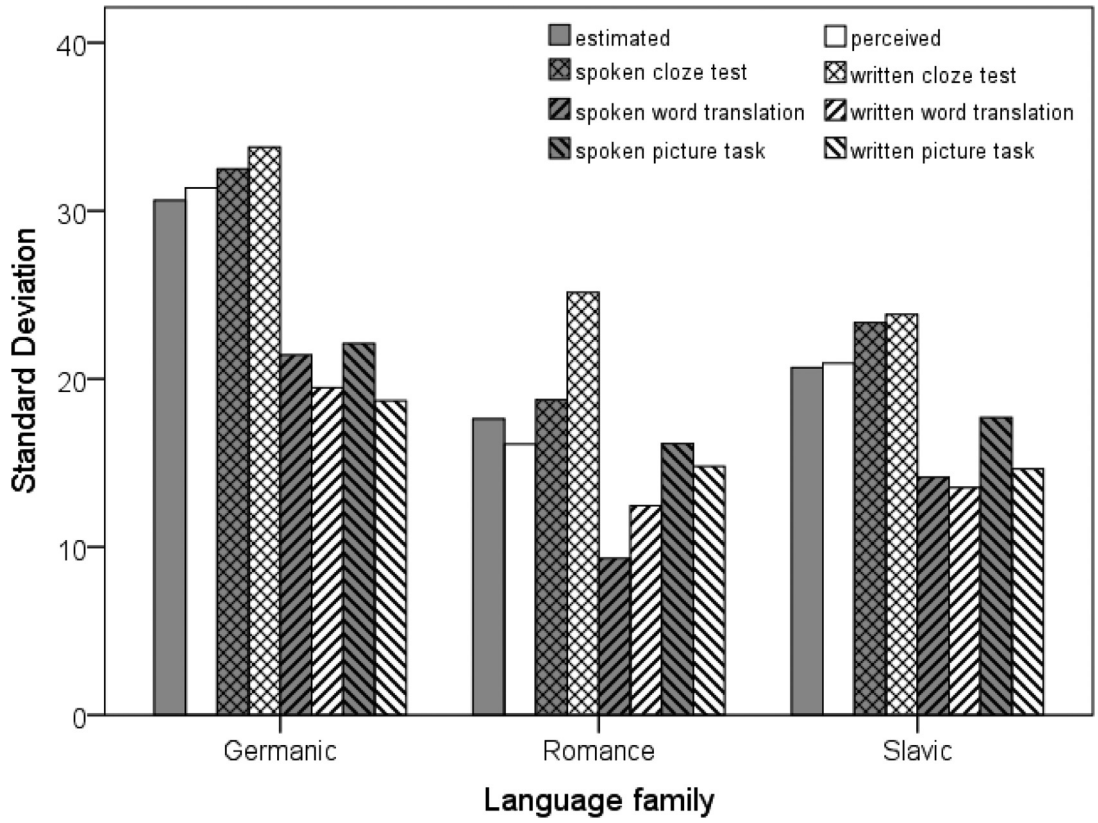

Fig. 5. Standard deviation of cross-language intelligibility scores (%) on eight tests broken down by Test type and Modality for language pairs in three families (see Fig. 4 for details).

derstand the written form of the test language better than the spoken form. This tendency is most obvious for the cloze test, where the written modality is easier in 65 of the 70 language combinations tested, $t(69) = 10{,}7$ ($p < .001$). The same tendency is seen in the word translation task: the written modality is easier in 59 of the 70 combinations (with nine of the eleven exceptions in the Slavic language pairs), $t(69) = 8.3$ ($p < .001$). In the case of the picture task, however, the two modalities are roughly equal (speech is easier in 37 combinations), $t(69) = -.1$ ($p = .930$, ins.).

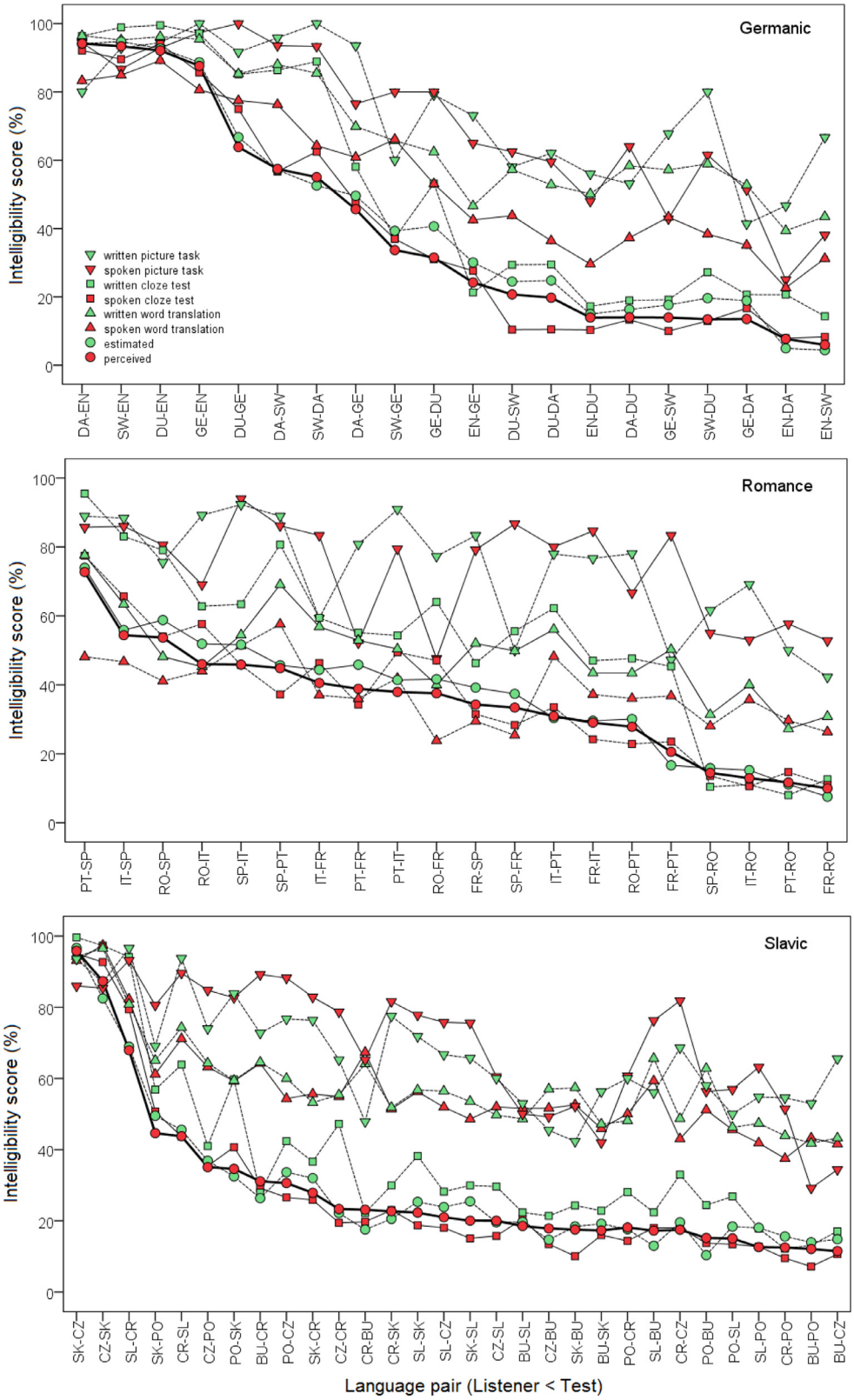

Fig. 6. Mean test scores (%) on six functional intelligibility tests and two judged intelligibility measures (%) per language group. Germanic (DA = Danish, DU = Dutch, EN = English, GE = German, SW = Swedish). Romance (FR = French, IT = Italian, PT = Portuguese, RO = Romanian, SP = Spanish). Slavic (BU=Bulgarian, CR = Croatian, CZ = Czech, PO = Polish, SK = Slovak, SL = Slovene). For each language combination, participant language is given first and test language second, e.g., EN-DA = English native responses to Danish stimuli.

Table 1
Pearson correlation coefficients between mean scores of six intelligibility tests and two judged intelligibility measures (upper triangle). In the lower triangle correlations are listed at the participant level (see text). All correlations are significant at the .01 level.

	Judged intelligibility		Cloze test		Word test		Picture task	
	Estimated	Perceived	Spoken	Written	Spoken	Written	Spoken	Written
Estim. intel.		.99	.97	.94	.72	.82	.72	.74
Perc. intel	.84		.99	.95	.78	.85	.79	.76
Cloze spoken	.75	.86		.94	.73	.81	.71	.73
Cloze written	.63	.76			.66	.79	.73	.77
Words spoken	.50	.57				.90	.63	.55
Words written	.62	.65					.66	.68
Picture spoken	.26	.40						.65
Picture written	.24	.37						

3.3. Correlation structure

We will now analyze the correlations between the eight intelligibility measurements in more detail. Each participant completed just one functional intelligibility test, and therefore we can only correlate the mean result (rather than individual results) of the functional tests for each of the 70 language combinations. The Pearson correlation coefficients for all $(8 \times 7)/2 = 28$ pairs of tests are listed in the upper triangle of Table 1. All participants completed the questionnaire and judged the intelligibility of the test language before and after completing the functional task. Estimated and perceived intelligibility scores can therefore be correlated with each other, and with the functional test scores, at the participant level. These correlation coefficients are presented in the lower triangle of Table 1. Correlation coefficients are lower when computed at the participant level than at the group level, since the individual participant variability is eliminated as a source variance in the mean scores per language combination.

Considering the *functional tests* first, the correlations between the word translation tasks and the cloze tests are highest ($r = .73$ for the spoken versions and $r = .79$ for the written versions). Somewhat lower correlations are observed between the word tests and the picture tests ($r = .63$ for the spoken and $r = .68$ for the written versions). When comparing the written and spoken versions of tests using the same method (cloze test, word translation task or picture task), we find the highest correlations between the cloze tests ($r = .94$), and the word tests ($r = .90$), whereas the modalities are less strongly correlated for the picture task ($r = .65$).

As already noted when discussing Fig. 6, the two opinion scores (estimated vs. perceived intelligibility) are very strongly correlated: $r = .99$ at the language level, and $r = .84$ at the participant level. Correlations with the spoken cloze test are almost as high ($r = .97$ and .75 for estimated intelligibility; $r = .99$ and .86 for perceived intelligibility) and a little lower for the written cloze test ($r = .94$ and .63 for estimated intelligibility; $r = .95$ and .76 for perceived intelligibility).

Correlations between judged intelligibility scores and the word tests are also high when measured at the language level ($.72 \leq r \leq .85$) but lower when measured at the participant level ($.50 \leq r \leq .65$). When looking at the results of the word translation task separately per language group, we find that the correlations are high for the Germanic group ($r = .95$ for both the spoken and the written task). For the Slavic group the correlations are somewhat lower ($r = .88$ for both written and spoken tests). This was due to speakers of Polish, Croatian and Slovenian understanding more words than they themselves expected. The low correlations found in Table 1 for the word translation tasks are mainly due to the Romance group ($r = .60$ for the spoken and $r = .79$ for the written test).

The correlations with measures involving the picture task are lowest ($.72 \leq r \leq .76$ at the language level, and $.26 \leq r \leq .40$ at the participant level). This can be at least partly explained as a ceiling effect for some language combinations. However, inspection of scattergrams (not included here) also proves that the picture task sometimes contradicts judged intelligibility. For example, all participant groups performed more poorly on the spoken French picture task than they expected. The Slavic participants often performed very well on the picture task even though they were rather pessimistic about their understanding of other Slavic languages.

We expected the correlations with the test scores to be higher for perceived intelligibility than for estimated intelligibility because some of the participants might not have a realistic a-priori idea of how well they would understand the test language, for example if they had little experience with the test language. This hypothesis finds support in the data, since all correlation coefficients are better for perceived intelligibility ($M = 60.2$) than for estimated intelligibility ($M = 50.0$), $t(5) = 5.8$ ($p = .001$, one-tailed).

4. Conclusions and discussion

In this paper we presented the results of six functional intelligibility tests (spoken and written versions of a word test, a cloze test and a picture task) and two measurements of judged intelligibility (estimated and perceived intelligibility) of 16 closely related standard languages in Europe (in total 70 combinations of participant groups and test languages). We found that the functional tests displayed similar patterns of intelligibility, where especially the cloze test and the word translation test proved strongly correlated. An exception is the Romance group, where scores on the word test were lower than the cloze test scores for some groups. We have no explanation for the low word test scores obtained by some Romance participant groups. They run counter, for instance, to larger number of cognates in the Romance words lists (85% cognates) than in the Germanic (77%) and Slavic (64%) word lists (see also Heeringa et al., 2013). Other linguistic factors that we have not yet been able to identify may be in effect here.

The correlations between the picture task and the word test and cloze test were rather low. Part of the explanation for this finding is that the picture task was too easy for some participant groups, which creates a ceiling effect. However, visual inspection of scattergrams (not presented) showing the relationship between judged intelligibility and the results of the picture task also reveals a good deal of seemingly random scatter. This leads us to assume that the picture task in its present form is not an accurate way of testing global intelligibility. It may be possible to improve the picture task by choosing a more difficult text and by introducing more and more complex pictures varying more than just two binary message elements so that the participant has to extract more key elements from the text before the correct picture can be chosen.

Our two measures of judged intelligibility (estimated and perceived) correlated strongly ($r = .99$ at language level and .84 at participant level). This may be explained by hysteresis, i.e., the tendency on the part of the participants to stick to their first judgment. Alternatively, our subjects may have known the test languages so well in advance that they did not have to change their judgment after the functional testing session. Both judgment scores were also strongly correlated with the results of the functional tests. The cloze tests reflected judged intelligibility particularly well: not only were the correlations exceptionally strong ($.94 \leq r \leq .99$) but the participants were even able to accurately estimate the percentage of gaps they had correctly filled in. Our results show that naïve language users are quite able to judge the intelligibility of a test language (at least one from within their own language family). Conversely, the results also show that the scores on functional tests, especially those obtained on the cloze tests, reflect the perceptions of naïve language users very well.

In the present experiment, participants were asked to estimate how well they would understand a non-native language variety without hearing (or seeing) a sample. They then completed a functional intelligibility test, after which they were (implicitly) asked to estimate how well they had performed on the test. In previous experiments the judgment task and the functional test were not given to the same groups of participants, and/or the stimulus materials of the two tests were not the same. Tang and Van Heuven (2009) reported correlations between perceived and functionally tested speech intelligibility ($r = .77$ for word intelligibility and $r = .82$ for sentence intelligibility) that are similar to the correlation between perceived intelligibility and the results of the spoken word test in the present investigation ($r = .78$). In the present study we did not include a sentence intelligibility test, but the correlation between perceived intelligibility and the results of the spoken cloze test (text level) in our investigation proved exceptionally strong ($r = 99$). Tang and Van Heuven (2007, 2009) asked their participants how well a monolingual speaker of their own dialect would understand a speaker of the dialect in the recording, if they had never heard that dialect before. In our investigation the participants were asked to judge their own ability to understand the test language. This means that they did not judge inherited intelligibility, i.e., the ability to understand the test language without previous exposure or schooling, but the actual situation which may have been influenced by their personal schooling or exposure to the language.

We included written as well as spoken versions of all three tests in our investigation using the same tests and the same textual materials, so that we could compare cross-language intelligibility in the two modes directly. This has not been possible in earlier investigations because spoken and written intelligibility were generally assessed by means of different tests.[3] We found as a significant overall effect that the written tests were easier than the corresponding spoken versions although the advantage of print is not seen in the picture test (possibly due to a ceiling effect, see above). So, even for young participants such as ours it appears to be easier to process the written form of language than the spoken from. The advantage of printed forms is not likely to be due to a decline in processing capacity for spoken language (as has been found for elderly language users, see Vanhove, 2014). Rather, the written form of a closely related language may be easier to match with the corresponding form in the native language, because written language reflects an earlier form of the language where the two languages had diverged less from the original common form than in the spoken form (Schüppert, 2011; Doetjes and Gooskens, 2012).

Which test is preferable for measuring cross-language intelligibility? Unfortunately, it is not possible to provide a clear and simple answer to this question. As mentioned in Section 1, the choice depends on a number of factors, such as purpose of the measurements and the time, effort and means available. Nevertheless, the excellent match between the scores we obtained on the cloze test and perceived intelligibility (when both are expressed as percentages) shows that this test reflects overall intelligibility best among the three functional tests used in our investigation. The cloze test was also consistently the test which discriminated best between the language pairs, both within and across families. On psychometric grounds, therefore, we consider the cloze test the best option.

An important practical criterion for choosing a test is the ease with which it can be developed and administered. If a large number of languages need to be tested, extensive time and effort would be required to collect stimulus materials and to prepare the test. Preparing cloze tests is relatively simple, if texts can be selected from a database in which they are pre-categorized by level of difficulty, such as the B1 level in the Common European Framework of Reference for languages. Deleting one word per sentence is less work than having an artist draw pictures (at least four, possibly eight) that illustrate essential message elements in an unambiguous manner. The word translation test is not only time consuming to construct and administer, it also has the drawback that the results cannot be evaluated automatically. From a practical point of view, then, we consider the cloze test the superior alternative among the three functional tests we studied.

Our results have shown that young adult Europeans are able to judge the intelligibility of languages from within their own language family quite accurately, even without first hearing or reading any test materials. It is not entirely clear, however, to what extent the participants' opinions were grounded in past experience with the non-native languages. In the case of English, the participants predicted high intelligibility, because it was part of their secondary school curriculum. For many other languages, however, we have reason to believe that the opinion scores were at least partly based on ideas of the intelligibility of related languages that exist within the participant's language community, ideas which are often referred to as folk linguistics (e.g., Niedzielski and Preston, 2000). Predictions of cross-language intelligibility may be based in part on geographic distance. Naïve languages users believe that languages diverge as they move farther away from each other on a map. Moderate correlations have been reported between geographic distance (between the capitals of the countries concerned) and linguistic distance (measures of structural difference between language varieties), i.e., $.36 \leq r \leq .70$ for percentage of cognate vocabulary and $.24 \leq r \leq .81$ for differences in visual word forms, depending on the language family (Heeringa et al., 2013). Cross-language intelligibility, in turn, is inversely correlated with linguistic distance as was shown by Tang and Van Heuven (2015) for Chinese languages, Van Heuven et al. (2015) for Romance languages, Swarte (2016) and Gooskens (2007) for Germanic languages, and Golubović (2016) for Slavic languages. It is beyond the scope of the present paper to differentiate these various explanations in detail; we assume that our participants had reasonably accurate knowledge of the geographic locations of the related languages within their group and included this knowledge in their expectations of cross-linguistic intelligibility.

Although the judged intelligibility scores proved quite accurate predictors of functional intelligibility, we do not advocate the indiscriminate use of opinion testing. Judgments may not suffice, for example in cases where the participants may have incorrect ideas about the intelligibility of a language, for example because of negative or positive attitudes towards the language or a desire to char-

[3] An exception is Vanhove's (2014) study of the recognition of isolated Swedish words by Swiss-German dialect speakers with no knowledge of Swedish, using the same word lists in printed and in spoken form in a cognate-guessing task. This study, however, targeted differences in cognitive flexibility as a function of age and sensory input channel, rather than mutual intelligibility.

acterize their own language as being very different or similar to the test language. We do not know, at this time, whether such attitudes and misconceptions have stronger effects on opinion scores than on functional test scores. Until we do, we recommend functional testing since we have determined it is a fast and effective method of constructing, administering and processing such tests.

Written tests are easier to develop and administer than spoken tests. There is no need to find speakers with equally intelligible voices, nor are any measures needed to ensure equal recording conditions. In our study, the correlation between the modalities was especially strong for the cloze test we developed, with $r = .94$. This leads us to our last recommendation, which is to use written test materials as long as there are no compelling a priori reasons to prefer the oral modality. The obvious exception to this recommendation is when the participant group cannot read the test language, as may happen with pre-school children, illiterate adults, or unknown writing systems. In such circumstances the cloze test, whether written or spoken, cannot be used. Only one alternative remains, which would be a combination of aural presentation of linguistic stimuli and some form of picture matching, either between spoken words and iconic representations of their meanings, or the text-to-picture matching task we used ourselves. The correlation between perceived intelligibility and the spoken picture test ($r = .79$) in our data was substantial. This shows that the picture test, especially if improved as suggested above, may provide a viable alternative in situations where the participants are functionally illiterate.

Acknowledgments

This research was supported by the Netherlands Organisation for Scientific Research (NWO) grant number 360-70-430. We thank Jelena Golubovic, Femke Swarte and Stefanie Voigt and numerous student assistants for collecting the material for this investigation. We thank Aleksandar Mancic for programming the web application.

Appendix. The word list used for the word test. Each participant translated a random selection of 50 out of the 100 words

time	place	development	hour
year	point	room	rate
people	house	water	law
man	country	form	door
day	week	car	court
thing	member	level	office
child	end	policy	war
government	word	council	reason
part	example	line	minister
life	family	need	subject
case	fact	effect	person
woman	percent	use	period
work	month	idea	society
system	side	study	process
group	night	girl	mother
number	eye	name	voice
world	head	result	police
area	information	body	kind
course	question	friend	price
company	power	right	position
problem	money	authority	age
service	change	view	figure
arm	interest	report	education
party	order	face	programme
school	book	market	minute

References

Abraham, R.G., Chapelle, C.A., 1992. The meaning of cloze test scores: an item difficulty perspective. Mod. Lang. J. 76, 468–479.

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

van Bezooijen, R., van Heuven, V.J., 1997. Assessment of synthesis systems. In: Gibbon, D., Moore, R., Winski, R. (Eds.), Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin/New York, pp. 481–563.

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

Bø, I., 1978. Youth and Neighbouring Country. An Investigation of the Influence of School and TV on Inter-Scandinavian Comprehension. Ungdom og naboland. En undersøkelse av skolens og fjernsynets betydning for nabospråksforståelsen. Stavanger: Rogalandsforskning.

Börestam Uhlmann, U., 1991. Language Meetings and Meeting Languages in the Nordic Countries. Språkmöten och mötesspråk i Norden Nordisk språksekretariats rapporter 16. Oslo: Nordisk språksekretariat.

Council of Europe, 2001. Common European Framework of Reference for Languages: Learning, Teaching, Assessment Strasbourg.

Cutler, A., 2012. Native Listening: Language Experience and the Recognition of Spoken Words. MIT Press, Cambridge, MA.

Delsing, L.-O., Lundin Åkesson, K., 2005. Does the Language Keep Together the Nordic Countries? A Research Report of Mutual Comprehension between Young Speakers of Danish, Swedish and Norwegian. Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska. Nordiska ministerrådet, Copenhagen.

Doetjes, G., 2007. Understanding differences in inter-Scandinavian language understanding. In: Ten Thije, J., Zeevaert, L. (Eds.), Receptive Multilingualism. Linguistic Analyses, Language Policies and Didactic Concepts. John Benjamins, Amsterdam, pp. 217–230.

Doetjes, G., Gooskens, C., 2012. Skriftsprogets rolle i den dansk-svenske talesprogsforståelse. Språk och stil 19, 105–123.

Golubović, J., 2016. Mutual iIntelligibility in the Slavic Language Area. (Groningen Dissertations in Linguistics 152). Rijksuniversiteit Groningen, Groningen.

Gooskens, C., 2007. The contribution of linguistic factors to the intelligibility of closely related languages. J. Multiling. Multicul. Dev. 28 (6), 445–467.

Gooskens, C., 2013. Methods for measuring intelligibility of closely related language varieties. In: Bayley, R., Cameron, R., Lucas, C. (Eds.), Handbook of Sociolinguistics. Oxford University Press, pp. 195–213.

Gooskens, C., van Heuven, V.J., van.Bezooijen, R., Pacilly, J.J.A, 2010. Is spoken Danish less intelligible than Swedish? Speech Commun. 52, 1022–1037.

Haugen, E., 1966. Semicommunication: the language gap in Scandinavia. Sociol. Inq. 36, 280–297.

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

van Heuven, V.J., 2008. Making sense of strange sounds: (Mutual) intelligibility of related language varieties. A review. Int. J. Humanit. Arts Comput. 2, 39–62.

van Heuven, V.J., Gooskens, C., van Bezooijen, R., 2015. Introducing MICRELA: predicting mutual intelligibility between closely related languages in Europe. In: Navracsics, J., Bátyi, S. (Eds.), First and Second Language: Interdisciplinary Approaches. Tinta Könyvkiadó, Budapest, pp. 127–145. (Studies in Psycholinguistics 6).

Hickerton, H., Turner, G.D., Hickerton, N.P., 1952. Testing procedures for estimating transfer of information among Iroquois dialects and languages. Int. J. Am. Linguist. 18, 1–8.

Hufeisen, B., Marx, N., 2014. EuroComGerm – The Seven Sieves: Learn to Read Germanic Languages. EuroComGerm – Die sieben Siebe: Germanische Sprachen lesen lernen. Aachen: Shaker.

Keshavarz, M.H., Salimi, H., 2007. Collocational competence and cloze test performance: a study of Iranian EFL learners. Int. J. Appl. Linguist. 17 (1), 81–92.

Klein, H.G., Stegmann, T.D., 2000. EuroComRom – The Seven Sieves: Immediately Being Able to Read Romance Languages. EuroComRom – Die sieben Siebe: Romanische Sprachen sofort lesen können. Aachen: Shaker.

Marslen-Wilson, W.D., Welsh, A., 1978. Processing interactions and lexical access during word-recognition in continuous speech. Cognit. Psychol. 10, 29–63.

Maurud, Ø., 1976a. Neighbouring Language Comprehension of Spoken and Written Language in Denmark, Norway and Sweden. Nabospråksforståelse i Skandinavia. En undersøkelse om gjensidig forståelse av tale- og skriftspråk i Danmark, Norge og Sverige. Stockholm: Nordiska rådet.

Maurud, Ø., 1976b. Reciprocal comprehension of neighbour languages in Scandinavia. An investigation of how well people in Denmark, Norway and Sweden understand each other's languages. Scand. J. Educ. Res. 20, 49–71.

Niedzielski, N.A., Preston, D.R., 2000. Folk Linguistics. Trends in Linguistics. Studies and Monographs, 122. Mouton de Gruyter, Berlin/New York.

Pierce, J.E., 1952. Dialect distance testing in Algonquian. Int. J. Am. Linguist. 18, 203–210.

Scharpff, P.J., Heuven, V.J.van, 1988. Effects of pause insertion on the intelligibility of low quality speech. In: Ainsworth, W.A., Holmes, J.N. (Eds.), Proceedings of the 7th FASE/Speech-88 Symposium. The Institute of Acoustics, Edinburgh, pp. 261–268.

Schüppert, A., 2011. Origin of Asymmetry. Mutual intelligibility of spoken Danish and Swedish. (Groningen Dissertations in Linguistics), 94. Groningen: Center for Language and Cognition.

Swarte, F., 2016. Predicting the (Mutual) Intelligibility of Germanic Languages from Linguistic and Extra-Linguistic Factors. (Groningen Dissertations in Linguistics), 150. Groningen: Rijksuniversiteit Groningen, Groningen.
Tang, C., van Heuven, V.J., 2007. Mutual intelligibility and similarity of Chinese dialects. Predicting judgments from objective measures. In: Los, B., Koppen, M.van (Eds.). In: Linguistics in the Netherlands, 2007. John Benjamins, Amsterdam, pp. 223–234.
Tang, C., van Heuven, V.J., 2009. Mutual intelligibility of Chinese dialects experimentally tested. Lingua 119, 709–732.
Tang, C., van Heuven, V.J., 2015. Predicting mutual intelligibility of Chinese dialects from multiple objective linguistic distance measures. Linguistics 53, 285–311.
Taylor, W.L., 1953. Cloze procedure: a new tool for measuring readability. Journal. Quart. 30, 415–433.
Vanhove, J., 2014. Receptive Multilingualism Across the Lifespan. Cognitive and Linguistic Factors in Cognate Guessing Ph.D. thesis. University of Fribourg.
Voegelin, C.F., Harris, Z.S., 1951. Methods for determining intelligibility among dialects of natural languages. In: Proceedings of the American Philosophical Society, 95, pp. 322–329.
Wolff, H., 1959. Intelligibility and inter-ethnic attitudes. Anthropol. Linguist. 1, 34–41.