Oxford Handbooks Online

Experimental Methods for Measuring Intelligibility of Closely Related Language Varieties

Charlotte Gooskens

The Oxford Handbook of Sociolinguistics
Edited by Robert Bayley, Richard Cameron, and Ceil Lucas

Print Publication Date: Jan 2013
Online Publication Date: Jan 2013

Subject: Linguistics, Sociolinguistics
DOI: 10.1093/oxfordhb/9780199744084.013.0010

[–] Abstract and Keywords

To test intelligibility, a large number of tests have been developed. By means of such tests, the degree of intelligibility can be expressed in a single number, often the percentage of input that was correctly recognized by the subject. This chapter presents an overview of methods for measuring the intelligibility of closely related languages, and discusses their advantages and disadvantages. It focuses on spoken-language comprehension, but many tests can also be applied to the comprehension of written language. Methods for investigating mutual intelligibility can be taken from other disciplines, for example in the area of speech technology, second language acquisition, and speech pathology.

Keywords: intelligibility, spoken-language comprehension, speech technology, second language acquisition, speech pathology

The exact number of known living languages varies from 5000 to 10,000, depending on one's definition of "language." An even larger number of dialects are spoken worldwide. Many of these languages and dialects (from now on taken together as "languages" or "varieties") are so similar that they are mutually intelligible to varying degrees, even without prior contact or formal instruction. Speakers of such different yet related languages sometimes communicate each speaking their own language. Haugen (1966) introduced the term *semicommunication* for this kind of communication. Other terms are receptive multilingualism, semibilingualism, nonconvergent discourse, asymmetric/bilingual discourse, and inherent intelligibility. Examples of observed semicommunication can be found in Zeevaert (2004).

For various reasons it may be interesting to establish the degree to which a speaker of one variety understands the speech of another closely related variety, for instance, to resolve issues that concern language planning and policies, second-language learning, and language contact. Unbiased data about distances between varieties and detailed knowledge about intelligibility can also be critical for sociolinguistic studies. Varieties that have strong social stigmas attached to them could unrightfully be deemed hard to understand (Giles & Niedzielski 1998; Wolff 1959). The relationship between attitudes and intelligibility is not a straightforward one, but advances in the field of linguistic distances and intelligibility testing provide sociolinguists with objective data to resolve conflicts that arise concerning varieties on a standard–nonstandard continuum. Knowledge about mutual intelligibility is also needed for standardization and development of new orthographies in communities where no standardized orthography exists.

To test intelligibility, a large number of tests have been developed. By means of such tests, the degree of intelligibility can be expressed in a single number, often the percentage of input that was correctly recognized by the subject. The aim of the present chapter is to give an overview of methods for measuring intelligibility of closely related languages and to discuss the advantages and disadvantages of the various methods. We focus on spoken language comprehension, but many tests can also be applied to the comprehension of written language.

Human spoken language is extremely robust, and native subjects are generally successful in getting the speaker's intentions even if the input speech is defective, for example, in cases of language or speech pathology, foreign accents, and computer speech, and even in noisy conditions. Listening to a closely related language is similar to other situations where the speech input is nonoptimal and we assume that no special mechanism is involved in decoding this kind of speech. This means that methods for investigating mutual intelligibility can be taken from other disciplines, for example, in the area of speech technology, second-language acquisition, and speech pathology.

The experimental and methodological considerations relevant for intelligibility testing have much in common with those relevant for studies in various other areas of sociolinguistics. For example, efforts must mostly be taken to control the context of speech production or speech perception as much as possible while keeping the recording or listening condition as natural as possible. A number of techniques have been developed within the area of experimental methods for the study of language variation. Often a variety of methods are used to study a phenomenon because each method has its shortcomings. Sociolinguists and dialectologists have devoted much attention to giving technical descriptions of language varieties phonetically and exploring general questions about language attitude and stereotypes. However, in recent years there has been an increasing focus on perceptual sociolinguistics (e.g., Clopper 2004; Long & Preston 2002; Preston 1999; Thomas 2002). Researchers have shown a growing interest in uncovering what linguistic and nonlinguistic features listeners react to. Methods include experimental designs using systematically manipulated speech (e.g., Fridland & Bartlett 2004) or listeners' expectations (Niedzielski 1999), as well as statistical correlations between acoustical measurements and reactions to perception experiments (e.g.,Clopper&Pisoni 2004). For an overview of experimental methods for the study of language variation, see Nagy (2006).

General Methodological Considerations

This section gives an overview of methodological considerations that should be made when designing an intelligibility investigation. Factors are discussed that may influence the results and that should either be avoided or taken into consideration when interpreting the results. Topics dealt with are the test material, selection of speakers and subjects, and the characteristics of the task to be carried out by the subjects.

Test Material

To carry out an intelligibility test, recordings of the languages to be used as listening material in the test are needed. The choice of test material depends on the aim of the investigation and can vary along a number of dimensions: style (spontaneous or read, formal or informal), number of speakers involved (monologues, dialogues), linguistic entity (isolated words, sentences, texts), complexity (difficult, easy) and subject matter (daily life, science, society, technique, politics, etc.). If the intelligibility of more languages is compared, great care should be taken to keep these factors constant when collecting material for tests.

It is important that the texts represent the languages to the same extent. A way to control the material is to use translations of the same text in all the test languages. This makes it necessary to use read speech, even though it may be preferable to use spontaneous speech since this simulates a natural situation to a larger degree. A good compromise may be to use recordings of semispontaneous speech, where the material is controlled to some extent, such as in map tasks (cf. Anderson et al. 1991; Brown et al. 1984; Grønnum 2009) or picture description tasks, where speakers have to carry out some task that demands speech production in a controlled setting. However, the use of (semi)spontaneous speech makes it impossible to use the same texts and questions for the different test languages, so the results are less comparable.

When using translations, often a text in one of the test languages is translated into the other test languages. However, there is a risk that the translators may stick too closely to the original text when choosing words and expressions for the translations. To make sure that one of the test languages does not get a special status, a solution is to use translations from a language that is not one of the test languages or, alternatively, to use source texts from each of the test languages. In this way, frequent words and constructions are more likely to be represented to the same degree in all test languages. Frequent words are more easily recognized than infrequent words (Luce & Pisoni 1998).

Other word characteristics are also known to influence intelligibility and should therefore be controlled for. Words with a high neighborhood density are often more difficult to recognize than words with few competitors. A word's neighborhood density can be defined as the number of words that deviate from the target word by just one sound (Luce & Pisoni 1998). A word like *elephant* has no neighbors and is not easily mistaken for another similar word, while the word *cat* has a total of 30 neighbors, for example, *bat, kit,* and *cap* and can therefore easily be confused with another word.

Word length should also be considered. Studies have shown that longer words are better recognized than shorter words (Scharpff & Van Heuven 1988; Wiener & Miller 1946). This is attributed to the relationship between word length and the number of neighbors. Longer words have fewer neighbors than shorter words (Vitevitch & Rodriguez 2005). Furthermore, redundancy increases with word length, and this is assumed to enhance intelligibility as well.

The speech fragments selected for the intelligibility test are generally supposed to represent the language as a whole. If the sample is large enough, for example, a complete text, one may assume that it represents a random sample of the test language. However, it is important to be aware that one single unintelligible word or sound may disturb the picture so that in fact the intelligibility of the whole speech sample becomes lower. Smaller fragments (word lists, restricted sets of sentences) call for some control. For example, one can make sure that the material is phonetically and lexically balanced, that is, in accordance with the statistical distribution of the words and phonemes in the language.

When the same stimulus is presented more than once there may be a learning effect (priming). Therefore, the same stimuli should not be presented more than once to the same subject. This contradicts the fact that it is desirable to use the same stimulus material when comparing the intelligibility of more languages. The solution is to use a Latin square design where each subject hears a proportion of the stimuli in each of the test languages, and yet hears stimuli in each of the languages in equal proportions, and never hears the same stimuli twice (see table 10.1). A disadvantage is that often many groups of subjects are needed, four groups in the example in table 10.1.

Speakers

Speech understanding is affected by speaker characteristics. Some speakers are more intelligible than others, for example, because of differences in voice quality, precision of articulation, and reading ability (Hazan & Markham 2004). The sex of the speaker seems to play a role as well, female voices in general being more intelligible than male voices (Bradlow et al. 1996). Speaker

Table 10.1. Example of a Latin square design with the languages A–D, stimuli 1–4 and test versions I–IV

Languages	**Test version**			
	I	**II**	**III**	**IV**
A	1	2	3	4
B	2	3	4	1
C	3	4	1	2
D	4	1	2	3

characteristics may also vary across educational level, age, and social class. If the aim of an investigation is to compare the intelligibility of several languages, one should select speakers with similar voice characteristics and background. If the design of the experiment allows it, more than one speaker could be used per language variety, so that effects of variability between speakers will average out. If the intelligibility of only two languages is compared in a listening test one could opt for a bilingual speaker to make the stimuli. To be sure that the speaker sounds native in both languages, a voice lineup can be arranged (Schüppert, Hilton, & Gooskens accepted;

Broeders, Cambier-Langeveld, & Vermeulen 2002).

Subjects

The task performance of human subjects is always somewhat noisy, or variable. Humans may be influenced by unwanted factors, such as motivation to carry out the test task and previous experience with the test language. Also, a certain relationship between attitudes and intelligibility has been found in previous research. The fact that Danes understand Swedish better than Swedes understand Danish, for example, is often explained by less positive attitudes among Swedes toward the Danish language, culture, and people than vice versa (Delsing & Lundin Åkesson 2005; Gooskens 2006). Therefore, if researchers wish to test the intelligibility of a language without the influence from attitudes, they may exclude subjects with strong positive or negative attitudes or aim to match the subject groups so that they have (approximately) the same attitudes toward the test language.

The subjects should be representative of the group of people to be tested as far as educational level, intelligence, age, gender, social class, geography, language background, and experience with the test language are concerned. In order to control for these factors it is important to select a well-defined group of subjects, for example an equal number of male and female high school pupils aged between 17 and 18 years, who are born and raised in a specific place and who have no prior experience with the test language.

To control for all the above-mentioned subject characteristics, an intelligibility test is often accompanied by a questionnaire that the subject has to fill in. Here, questions are asked about personal background (age, gender, places the subject has lived, language background of the subject and the parents of the subject, and schooling, etc.); attitude toward the test language (e.g.,"How beautiful does language X sound on a scale from 1 (beautiful) to 5 (ugly)?"); and experience with the test language (e.g.,'How often do you hear/speak/read/write language X on a scale from 1 (never) to 5 (every day)?'). The answers to the questionnaires can be used to exclude certain subjects from further analysis because they do not meet all subject criteria.

Task

When designing a listening task to establish the level of intelligibility in a group of people, it is important to take into account the limitations of the task offered. The task can be either too easy or too difficult, and both situations should be avoided since they make it difficult or impossible to interpret the results.

If the task is too easy and the subjects answer all questions correctly, this will result in a ceiling effect whereby a measurement cannot take on a value higher than some limit or 'ceiling.' This will make it hard to interpret the results. There are several ways to avoid ceiling effects. The intelligibility of the speech sample can be made more difficult by manipulating the signal by means of filtering or signal-compression techniques or by adding noise. Another way to make the task more difficult is to put the subjects under time pressure, either by asking them to perform the task as quickly as possible or by giving them only a limited amount of time to answer. In addition, reaction time can be measured. This gives a more precise measurement, and even if all subjects answer all questions correctly, there may still be a difference in the time it took the subjects to correctly comprehend the various stimuli.

It is important to build a reference condition into the experiment, with native speakers listening to their own language as control group to check that the task is not too difficult. It should be kept in mind, however, that even under the most favorable circumstances, native subjects will mostly make mistakes. If the task is too difficult, the percentage of correct answers may be so low that it is difficult to interpret the results (floor effect). Furthermore, the subjects may get frustrated and decide not to finish the test. The task is for instance too difficult if it does not take the memory limitations of subjects into account. Therefore, too-complex tasks or too-long sentences should be avoided. Also the limitations of the specific subject group should be taken into account (e.g., illiteracy, hearing loss, or visual handicaps). For some groups of subjects it may be a hindrance to have to use a computer to perform the task. In reaction-time experiments it should be taken into consideration that right-handed persons generally respond faster to verbal stimuli with their right hand than with their left hand, and vice versa for left-handed persons (Rastatter & Gallaher 1982).

Methods for Measuring Intelligibility

In this section I first present methods for measuring overall intelligibility of complete spoken varieties and give examples of investigations in which these methods were used. A major division can be made between investigations in which subjects are asked how well they *think* they understand the other language (opinion testing) and investigations testing how well subjects *actually* understand the other language (functional testing). At the end of the section methods are presented for determining the role of single linguistic phenomena for intelligibility.

Intelligibility can be measured at several levels of the linguistic hierarchy, from sounds to larger entities like words, sentences, and whole texts. When testing overall intelligibility, preference may be given to the text level since this is closer to reality, where subjects are mostly confronted with whole messages. However, word level is very central, since it is the key to speech understanding. As long as the subject correctly recognizes words, he will be able to piece the speaker's message together. By testing isolated words, it becomes possible to pinpoint the role of specific sounds for intelligibility. Therefore, some tests are restricted to the word level.

Opinion Testing

An easy and efficient way to get a quick impression of the intelligibility of a language is to ask subjects to rate along scale(s) how well they think they understand the language at hand. It may provide a shortcut to functional intelligibility tests, and in addition, it provides information about people's subjective ideas about the intelligibility of languages. The results should be interpreted with some care, however, as a person's reported language behavior may not be in line with his or her actual language behavior.

Without speech samples. The simplest kind of opinion testing involves no speech fragments. An example of such an investigation is Haugen (1966). In the first large investigation on the mutual intelligibility between the three closely related Scandinavian languages, Danish, Norwegian, and Swedish, he sent questionnaires to 300 persons in each of the three countries. Three questions explored the informants' opinion concerning the level of mutual comprehension:

1. When you met an X for the first time, how well could you understand him? (not at all—with great difficulty—had to listen intently—all but a few words—understood everything)
2. Do you now understand X speech without difficulty (no—yes—fairly well)
3. When you speak with X, how well do they understand you? (same alternatives as under 1.)

An advantage of this paper-and-pencil method is that no speech material has to be selected. Furthermore, it is possible to abstract from individual speakers who may influence the results because of specific voice characteristics and speaking styles. On the other hand, it is uncertain whether respondents are actually able to judge intelligibility without speech samples. They may never or rarely have heard the language or may not remember how well they understood the speaker. The consequence may be that the respondents base their opinions on some extralinguistic factor, such as their positive or negative attitudes toward the country and its speakers, political borders, desirable answers, or the geographical distance to the place where the language is spoken.

With speech samples. An example of an investigation using speech samples to test intelligibility is Tang and Van Heuven (2007). Recordings of the same text, the fable "The North Wind and the Sun," in 15 Chinese dialects was presented to 24 subjects from each of the places where the dialects were spoken. For each dialect subjects were asked to indicate how well they believed monolingual subjects of their own dialect would understand the speaker on a scale from 0 ("They will not understand a word of the speaker") to 10 ("They will understand the other speaker perfectly"). With this approach it also is uncertain whether subjects are actually able to make the judgments on an objective linguistic basis without being influenced by nonlinguistic factors.

Functional Testing

Doubting the validity of intelligibility judgments, most researchers prefer to test actual speech comprehension. The disadvantage of this approach is that it is generally difficult to abstract away from individual speakers and test situations. In addition, an effort must be made to avoid priming effects, ceiling effects, too-heavy memory load, and

other unwanted effects. These considerations often make it rather time consuming both to develop suitable tests and to carry out the tests themselves.

Content questions. In order to simulate a test situation that is as close to reality as possible, a number of investigators have tested intelligibility by means of questions about the content of a text. The intelligibility of a language variety is expressed as the mean percentage of correct answers given by the participants chosen for the task.

The questions about the texts must be formulated with great care. They should cover the content of the whole text as well as possible, and not measure the memory, general knowledge, or intelligence of the subject. Correct answers to the questions must also be well-defined. This is not always an easy task and may force the researcher to distinguish different degrees of correctness, for example,"completely correct,""partly correct," and "incorrect." A more objective solution is to use multiple-choice questions, where the subject has to choose between a limited number of possible answers. An additional advantage of this method is that the answers can be corrected rather easily, either manually or automatically by computer. A disadvantage of multiple-choice questions is that it may be difficult to find distracters that are not too easily excluded by the subjects. Furthermore, the use of multiple-choice questions is rather unnatural, since people are mostly not given several possible replies in a natural situation where intelligibility is required.

Translations. Another way of testing the intelligibility of a text is to have the subjects translate it. Intelligibility is then expressed as the percentage of correctly translated words. An advantage of this method compared to content questions is that the researcher does not have to formulate questions about the text, which, as previously noted, is sometimes a difficult task. All words in the text count to the same degree even if the subject does not completely understand the text, and general knowledge about the subject only plays a limited role.

For the researcher, it may be difficult to decide whether translations should be counted as correct or incorrect and the choice may be rather subjective. For example, a Danish person may translate Swedish *piga* 'maid' into the Danish cognate *pige* 'girl,' which has only a partly overlapping meaning. Furthermore, some subjects may have difficulty translating since for them it not a natural task to perform. The ability to translate appears to involve far more than mere intelligibility, and it may draw heavily on the subject's memory. Therefore, the text must be presented in short chunks with pauses in between during which the subject can write down the translation.

An example of a translation task is Gooskens, Beijering, and Heeringa (2008) who tested the intelligibility of "The North Wind and the Sun" in 18 Nordic language varieties among subjects from Copenhagen. The six sentences of the fable were presented sentence by sentence to the respondents with each sentence in a different variety. To avoid learning effects, the same respondents should not hear the same sentence twice, and since all sentences from the 18 varieties had to be presented, a total of 18 groups of respondents were tested. In addition to the intelligibility scores, distances between Standard Danish and each of the Nordic language varieties were measured at the lexical level and at different phonetic levels. In order to determine how well these linguistic levels can predict intelligibility, the intelligibility scores were correlated with the linguistic distances, and a number of regression analyses were carried out. The results show that for this particular set of closely related language varieties, phonetic distance is a better predictor of intelligibility ($r = -.86$) than lexical distance ($r = -.64$).

Another possibility for avoiding memory problems is to have the respondents translate a collection of isolated words that are representative for the test language, for example, a random selection or words selected from a frequency list. The correction of the translations may be even more difficult than in the case of whole texts since words may have more meanings when they are presented out of context. Furthermore, respondents often make spelling errors that might make it unclear to the researcher whether the respondent has actually understood the test word.

Kürschner, Gooskens, and Van Bezooijen (2008) tested the intelligibility of 384 frequent Swedish words among Danish subjects via the Internet. The translations were automatically categorized as right or wrong by the computer through a pattern match with expected answers. The answers that were categorized as wrong were subsequently checked manually by a Danish mother-tongue speaker. Responses that deviated from the expected responses because of a mere spelling error were counted as correct identifications. Spelling errors were objectively defined as instances where only one letter had been spelled wrong without resulting in another existing word. So, for example, the mistake in *ærende* (correct *ærinde* 'errand') is considered a spelling mistake and

therefore counted as correct (only one wrong letter without resulting in another existing word), whereas *aske* (correct *æske* 'box') was not counted as correct because the mistake results in an existing word meaning 'ash'. Some Swedish words have more than one possible translation. For example, the Swedish word *brist* 'lack' can be translated into Danish *brist* or *mangel*, both meaning 'lack'. Both translations were counted as correct. In the case of homonyms, both possible translations were accepted as correct. For example, Swedish *här* can be translated correctly into Danish *hær* 'army' or her 'here'. The aim of the investigation was to determine to which degree various linguistic factors contribute to the intelligibility of Swedish words among Danes. The word intelligibility results were correlated with 11 linguistic factors. The highest correlation was found in the negative correlation between word intelligibility and phonetic distances. Also, word length, different number of syllables than in Danish, foreign sounds not present in Danish, neighborhood density, word frequency, orthography, and the absence of the prosodic phenomenon of 'stød' in Swedish had a significant influence on the level of intelligibility.

The words can be presented in a context where part of the message may be printed out with blanks for selected words only. For example, Van Bezooijen and Van den Berg (1999) played semi-spontaneous samples of various Dutch varieties to different groups of subjects from The Netherlands and Belgium. The texts were written down in Standard Dutch but the nouns were replaced by dotted lines of the same length. The subjects were asked to write the missing words on the lines while listening to the recordings. There were considerable differences in intelligibility among the tested varieties, and intelligibility depended to some extent on the geographic background of the listeners. An advantage of this test type is that it is easy to make sure that the correct translation is given. However, this approach makes it uncertain which role the (written) context plays in the interpretation of the words.

To make it easier to correct the responses, multiple-choice tests are often used in which respondents are asked to select the best possible translation from a list of choices. It is difficult to construct such a test, since the choice of distracters determines how difficult the test is, and it is often not possible to select the same distracters if more than one language is involved in the test. To solve this problem, Tang and Van Heuven (2009) determined word intelligibility by having subjects perform a semantic categorization task whereinwords had to be classified as one of ten different pre-given semantic categories, such as 'body part,' 'plant,' 'animal,' and so on. For instance, if the subject hears the word for 'apple,'she or he should categorize it as a member of the category 'fruit.' Here, the assumption is that correct categorization can only be achieved if the subject correctly recognizes the target words. Since there are as many as 10 semantic categories, the role of guessing is negligible. It is a disadvantage of this method that only words from predefined categories can be tested.

Van Heuven and Van Bezooijen (1995) provide an overview of methods for quality evaluation of synthesized speech. Here, it is mostly tested how well subjects understand synthesized speech in their own native language. I will discuss two of the translation tasks that have also been used for testing the intelligibility of natural languages. The advantage of these methods in comparison with the translation task mentioned earlier is that the test words are presented in a controlled spoken context. The results are easy to score by hand or automatically. The tests are easy to adapt to new test languages, but the number of words that can be tested in one test session is more limited than in the case of isolated words.

A set of semantically unpredictable sentences (SUS) was compiled by Benoit, Grice, and Hazan (1996). These sentences consisted of five different, common syntactic structures, and words were randomly selected from lexicons with frequent "mini-syllabic" words (smallest words available in a given category). The SUS sentences can be automatically generated using five basic syntactic structures and a number of lexicons containing the most frequently occurring short words in each language. The syntactic structures are simple and the sentence length does not exceed seven words (eight for English because of the auxiliary in questions) in order to avoid saturation of the subjects' short-term memory. The sentences have normal word order and prosody but do not permit the subject to predict the identity of content words from sentence semantics or situational context. For example, in a semantically anomalous sentence such as *He drank the wall,* the syntactic structure is correct. Subjects receive cues about syntactic category only; other than that they will not be able to make any further predictions about word identity by means of semantic or syntactic contextual cues. Since words are tested in different positions in the sentence, word segmentation is an essential feature assessed by this test. Intelligibility can be expressed as the percentage correctly translated (content) words, but the simplest and fastest way to score results is to only take into account the sentences that are entirely correctly translated. This easy-to-obtain score is strongly related to word score. Gooskens, Van Heuven, Van Bezooijen & Pacilly (2010) presented Danish and Swedish SUS sentences to Danish and Swedish subjects in order to test the mutual intelligibility as well as the intrinsic

intelligibility of the two languages.

The SPIN (Speech Perception in Noise) test is a list of sentences that test word intelligibility (Kalikow, Stevens, & Elliott 1977). The subject translates only the last word in a number of short spoken sentences. Since the position of the target word is pre-given, word segmentation problems are minimal. There are two types of materials in the SPIN test. One type presents target words that are highly predictable from the earlier context, as in *He wore his broken arm in a sling* (target underlined). The other type presents words that are not predictable from the context, as in *We could have discussed the dust*. Wang (2007) showed that the high predictability part of the SPIN test was more sensitive to differences between speaker and subject groups with different degrees of listening comprehension in English than the low predictability part. The test is easy to adapt to new test languages, but the number of words that can be tested in one test session is more limited than in the case of isolated words.

Recorded text testing. A special problem arises when a researcher wants to test the mutual intelligibility of languages that he does not master himself. For such a situation the recorded text testing (RTT) method has been developed. This method was first used in the fifties to establish the mutual intelligibility of American Indian languages (Hickerton, Turner & Hickerton 1952; Pierce 1952; Voegelin & Harris 1951). Casad (1974) and Nahhas (2006) give detailed overviews of the steps that should be taken to carry out a test with RTT. The standard method uses a short text recorded from a speaker of the speech variety to be tested. The subject hears the text, with questions in his own mother tongue about the text interspersed following the portion that contains the answer to the question. The subjects are required to answer these questions.

An alternative approach to the standard RTT question format is the RTT retelling method, which requires subjects to listen to a narrative that has been broken down into natural segments of one or two sentences each and to retell the recorded text, segment by segment, in their L1 (see Kluge 2007). In this way, the subjects do not have to answer specific comprehension questions. For each segment the number of correctly retold core elements are counted and the segment scores are added up to obtain the overall score for a given RTT text.

The main advantage of the RTT retelling method, when compared to the standard RTT question method, is that comprehension of an entire text is tested, rather than selected sections only. A second major advantage is that in many more traditional societies, retelling a story is more appropriate and less threatening than answering questions. An additional advantage is that this method does not require the design of comprehension questions and the translation of these questions into the speech varieties of the communities under investigation. The most important disadvantage is that it is very time consuming both to develop the test and to count the number of correctly retold segments.

Reaction times. In cases where the test language is very similar to the language of the subjects, an off-line intelligibility task where responses are to be given after subjects have heard test passages, may be so easy that most answers are correct, resulting in a ceiling effect. There is a need, therefore, to use more sensitive testing procedures. Reaction time is a possible response measure that might improve the sensitivity of an intelligibility test. The assumption is that the faster the subjects react, the better the intelligibility. To ensure the credibility of the experiment, the lexical decision task needs to be followed by a second meaning-identifying task.

Reaction time can be measured by means of software applications that measure temporally accurately to within a few milliseconds. The application registers when a subject performs a certain action, for example, a vocal response, pressing a button on the computer keyboard, or touching the computer screen, which makes it suitable for various groups of test subjects, including children. Response times cannot be measured precisely via the Internet, and therefore this method is not suitable for web-based experiments.

Various tasks can be used to measure overall intelligibility using reaction times. In a sentence-by-sentence listening task subjects listen to sentences and push a button whenever they are ready for hearing the next sentence (e.g., Ralston et al. 1991). Comprehension is checked afterward. In a sentence verification test (e.g., James et al. 1994), subjects decide whether or not short sentences are true statements (e.g., *Mud is dirty* and *Rockets move slowly*). Impe (2010) used a lexical decision task where the subjects had to decide as quickly as possible—by means of pushing either a yes button or a no button—whether the stimuli (200 existing and as many non-existing words in 10 Dutch language varieties) were meaningless or meaningful Dutch words.

Observations. It can be argued that by its very nature intelligibility is a quality that does not easily lend itself to

quantitative measurement. It is probably possible to achieve certain pragmatic communicative goals even with a low degree of understanding. Comprehension depends on interactive cooperation, something that does not emerge in artificial test situations. Comprehension may be better in its natural context than in an artificial one because a specific setting reduces the number of possible interpretations. Börestam Uhlmann (1994) taped some thirty inter-Scandinavian arranged conversations between Danes, Norwegians, and Swedes aged 18 to 25 who were unaccustomed to the others' languages. She was first of all interested in which kind of strategies the participants used to improve mutual intelligibility, such as rephrasing, elaborate explanation, use of English, repairing and interruptions, to either clarify something or make certain that the message had been correctly understood. Her analysis of the result is mainly qualitative, but she also showed that it is possible to quantify the results by, for example, counting the number of reparations and misunderstandings. Zeevaert (2004) observed real Nordic meetings and quantitatively analyzed turn taking and length and frequency of pauses.

A disadvantage of this approach is that speakers and subjects are well able to conceal misunderstandings and to adapt their language to the conversational partner, so that it may be difficult to express exactly how well the speakers understand each other. It also demands a large effort from the researcher because he has to make a detailed analysis of the conversation.

Performance task. A way of simulating a natural communicative situation is a performance task. For example, Van Heuven and De Vries (1981) tested the intelligibility of various versions of foreign-accented speech by means of a performance task. Dutch subjects listened to recordings of Dutch accented utterances produced by a Turkish speaker who was asked to describe a number of simple actions (e.g., someone puts a spoon in a glass). The subjects were asked to perform the actions described by the speaker as quickly as possible. The mean reaction time of the correctly performed actions was the measure of intelligibility. The aim of the investigation was to investigate the role of phonic and nonphonic factors in the intelligibility of foreign-accented speech using an experimental approach. The results showed that phonic factors are more important than nonphonic factors.

The advantage of this method is that it measures intelligibility in a communicative situation. However, the fact that the subjects have to perform the described actions limits the variation in syntactic constructions and words that can be included.

Testing with the Aim of Determining the Role of Linguistic Factors

So far, methods for measuring intelligibility have been discussed that can be used to measure overall intelligibility, that is, languages as a whole. However, sometimes the aim of intelligibility testing is to assess the contribution of single linguistic phenomena to intelligibility. For example, very little is known as yet on the specific contributions of single sounds to overall intelligibility.

One approach when aiming at identifying specific factors that influence intelligibility is to do an error analysis on the test results. For example, Kürschner et al. (2008) carried out correlations and logistic regression analyses with the results of an experiment on the intelligibility of 384 Swedish words among Danes as the dependent variable, and 11 linguistic factors that have been found to contribute to L1 intelligibility in earlier studies as independent variables. In this way, they could make conclusions about the relative importance of these for intelligibility. Phonetic distance turned out to be the most important predictor of intelligibility, followed by word length.

Another way of investigating the role of specific linguistic factors is the experimental method. By keeping the effects of all factors but one constant, and systematically varying the characteristics of the latter, any difference in intelligibility must be caused by the variations in the target module. If, for example, we wish to test the hypothesis that Danish is poorly understood by Swedes due to the presence of *stød* (a voice characteristic creating phonological contrasts not present in Swedish), we can remove the *stød* from recordings of Danish. If Swedes understand the manipulated version better than the original version, *stød* must be causally related to the intelligibility of Danish.

If diagnostic testing is used to investigate the role of specific sounds, the most purposeful approach is to test the intelligibility of isolated words, since at sentence level or higher levels poor intelligibility is difficult to trace back to specific sources. If the words are presented in a sentence, the context or the situational redundancy is likely to make up for poor intelligibility.

Various diagnostic tests can be used to pinpoint linguistic factors that influence intelligibility. These factors may be found at all linguistic levels (segmental, prosodic, morphological, and syntactic). Many of the functional tests that have been discussed in the previous section may also be used for diagnostic purposes in adapted forms, and they are connected to the same advantages and disadvantages.

The results of investigations of the relative importance of various linguistic factors for intelligibility may be used to develop a model of intelligibility. As we have seen, intelligibility testing involving human subjects is often labor intensive and involves many considerations. It also yields noisy data. It would therefore be helpful to have an objective way of predicting intelligibility that would not involve actual testing. Since languages differ in many dimensions, such as sound inventories, prosody, vocabularies, morphology, and syntax, such a measure would involve linguistic distance measurements at different linguistic levels. However, we still lack information about how to weigh these dimension in order to develop a measure that can predict intelligibility. If, for example, word order differences hardly compromise the communication between speakers of two languages, while small differences between the sound systems make the mutual intelligibility difficult, then differences in phonology must be weighted much more than syntax in the computation of linguistic distance. So far, no complete model of intelligibility exists, but Gooskens et al. (2008) have shown that at an aggregate level phonetic distances measured by means of the Levenshtein algorithm (Heeringa 2004) in combination with lexical distances expressed as the percentage of noncognates (historically nonrelated words) can predict intelligibility to a large extent (.81 percent explained variance). Morphosyntax may also play a role in the intelligibility, though to a smaller degree than phonology (Hilton et al. accepted). A refined model may improve the predictive power, but it should be realized that nonlinguistic factors such as attitudes and previous experience may also play an important role.

Comparing Methods

In the preceding section a number of methods for measuring intelligibility were presented. Unfortunately, it is not possible to give an answer to the question of which method is best. The choice of the method to be used in an investigation depends on a large number of practical factors, such as time and funds available and the background of the subjects. Even with sufficient time and money and subjects who are able and patient enough to undergo complicated and lengthy tests, the choice of method still depends on the precise aim of the investigation.

But apart from these considerations, does it still matter which method is used? In order to shed some more light on this point, we need to know whether the same persons who achieve high scores in one test also achieve high scores in another test when all other factors are kept constant. A few researchers compared the results of different methods of measuring intelligibility. These comparisons are valuable because they give an impression of the importance of choosing a specific method. Doetjes (2007) investigated the effect of six different test types on the measurement of the intelligibility of Swedish among Danes. The same text was tested in six different test conditions: true/false questions, multiple-choice questions, open questions, word translation, summary, and short summary. The percentages of correct answers decreased from 93.0 percent for the true/false questions to 66.2 percent for the short summaries. This shows that at this point in time it is not possible to give an absolute answer to the question of how well subjects understand a language, and caution should be taken when comparing results from different investigations. When comparing various previous investigations on Swedish-Danish mutual intelligibility, for example, we see very different results, probably due to the use of different texts and tasks and the different backgrounds of the subjects. However, it is notable that Danes, for example, always have higher scores on the Swedish intelligibility tests than vice versa. This indicates that it may not be possible to express how well a language is understood in an absolute sense, but that it may be possible to compare the relative intelligibility of various languages as long as the test conditions are kept as constant as possible.

Maurud (1976) tested mutual intelligibility between the Scandinavian languages by means of word tests and content tests on the same texts. He found correlations between the test results between $r = .6$ and .8 for various groups of subjects. Tang and Van Heuven (2009) tested the mutual intelligibility of 15 Chinese dialects by means of functional intelligibility tests at word and sentence level and compared these with each other and with opinion scores and objective distance scores at the lexical and the phonological level. They found correlation between the opinion scores and the functional scores between $r = .7$ and.8. The same results were found for correlations between functional and opinion tests, on the one hand, and objective measurements, on the other hand. The authors conclude that mutual intelligibility should preferably be tested by means of functional sentence intelligibility

tests. The correlation between word intelligibility and sentence intelligibility was very high ($r = .9$) but sentence intelligibility reflected traditional Chinese taxonomy better than word intelligibility does. So, comparisons of various tests show rather high correlations; still, a large amount of unexplained variance is left. Even though there is a large overlap, different tests measure different aspects of intelligibility.

References

Anderson A. H., Bader M., Bard E. G., Boyle E., Doherty G., Garrod S., Isard S., Kowtko J., McAllister J., Miller J., Sotillo C., Thompson H. S., & Weinhart R. 1991. The HCRC Map Task Corpus. Language and Speech 34 : 351–66.

Benoit C., Grice M., & Hazan V. 1996. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. Speech Communication 18 : 381–92.

Börestam Uhlmann U. 1994. Skandinaver samtalar: språkliga och interaktionella strategier i samtal mellan danskar, norrman och svenskar. Uppsala : Uppsala University.

Bradlow A. R., Torretta G. M., & Pisoni D. B. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. Speech Communication 20 : 255–72.

Broeders A. P. A., Cambier-Langeveld T., & Vermeulen J. 2002. Case report: Arranging a voice lineup in a foreign language. The International Journal of Speech, Language and the Law 9 (1): 1350–1771.

Brown G., Anderson A., Shillcock R., & Yule G. 1984. Teaching talk: Strategies for production and assessment. Cambridge : Cambridge University Press.

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

Clopper C. G. 2004. Linguistic experience and the perceptual classification of dialect variation. Ph.D. dissertation, Indiana University.

Clopper C. G., & Pisoni D. B. 2004. Some acoustic cues for the perceptual categorization of American English regional dialects. Journal of Phonetics 32 : 111–40.

Delsing L-O., & Lundin Åkesson K. 2005. Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska. Copenhagen : Nordiska ministerrådet.

Doetjes G. 2007. Understanding differences in inter-Scandinavian language understanding. In J. Ten Thije & L. Zeevaert (eds.). Receptive multilingualism. Linguistic analyses, language policies and didactic concepts , 217–30. Amsterdam : John Benjamins.

Fridland V., & Bartlett K. 2004. Do you hear what I hear? Experimental measurement of the perceptual salience of acoustically manipulated vowel variants by Southern speakers in Memphis, TN. Language Variation and Change , 16 : 1–16.

Giles H., & Niedzielski N. 1998. 'Italian is beautiful, German is ugly.' In L. Bauer & P. Trudgill (eds.), Language myths , 85–93. London : Penguin.

Gooskens C. 2006. Linguistic and extra-linguistic predictors of Inter-Scandinavian intelligibility. In J. Van de Weijer & B. Los (eds.), Linguistics in the Netherlands 23 , 101–13. Amsterdam : John Benjamins.

Gooskens C., Beijering K., & Heeringa W. 2008. Phonetic and lexical predictors of intelligibility. International Journal of Humanities and Arts Computing 2 (1–2): 63–81.

Gooskens C., Van Heuven V. J., Van Bezooijen R., & Pacilly J. 2010. Is spoken Danish less intelligible than Swedish? Speech Communication 52 : 1022–037.

Gr ønnum, N. 2009. A Danish phonetically annotated spontaneous speech corpus (DanPASS). Speech Communication 51 : 594–603.

Haugen E. 1966. Semicommunication: The language gap in Scandinavia. Sociological Inquiry 36 : 280–97.

Hazan V., & Markham D. 2004. Acoustic-phonetic correlates of talker intelligibility in adults and children. Journal of the Acoustical Society of America 116 : 3108–118.

Heeringa W. 2004. Measuring dialect pronunciation differences using Levenshtein distance. Ph.D. dissertation, University of Groningen.

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

Hilton N. H., Gooskens C., Schüppert A., Van Bezooijen R., & Van Heuven V. Accepted. The influence of non-native morphosyntax and phonology on the intelligibility of a closely related language. In J.-M. Tirkkonen & E. Anttikoski (eds.). Proceedings of the 24th Scandinavian Conference of Linguistics. Joensuu : Publications of the University of Eastern Finland. Reports and Studies in Education, Humanities, and Theology.

Impe L. 2010. Mutual intelligibility of national and regional varieties of Dutch in the Low Countries. Ph.D. dissertation, Catholic University of Leuven.

James C.J., Cheesman M. F., Cornelisse L., & Miller L. T. 1994. Response times to sentence verification tasks (SVTS) as a measure of effort in speech perception. Fifth Australian International Conference on Speech Science & Technology II : 600–05.

Kalikow D.N., Stevens K. N., & Elliott L. L. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. Journal of the Acoustical Society of America 61 : 1337–351.

Kluge A. 2007. *RTT retelling method: An alternative approach to intelligibility testing*. SIL Electronic Working Papers, 2007–006.

Kürschner S., Gooskens C., & Van Bezooijen R. 2008. Linguistic determinants of the intelligibility of Swedish words among Danes. International Journal of Humanities and Arts Computing 2 (1–2): 83–100.

Long D., & Preston D. R. (eds.). 2002. Handbook of perceptual dialectology, vol 2. Amsterdam : John Benjamins.

Luce P. A., & Pisoni D. B. 1998. Recognizing spoken words: The Neighborhood Activation Model. Ear and Hearing 19 : 1–36.

Maurud Ø. 1976. Nabospråksforståelse i Skandinavia: en undersøkelse om gjensidig forståelse av tale- og skriftspråk i Danmark, Norge og Sverige. Stockholm : Skandinaviskråd.

Nagy N. 2006. Experimental methods for study of variation. In K. Brown (ed.), Encyclopedia of language and linguistics , 2nd ed., vol. 4, 390–94. Oxford : Elsevier.

Nahhas R. W. 2006. The steps of recorded text testing: A practical guide. Chiang Mai : Payap University.

Niedzielski N. 1999. The effect of social information on the perception of sociolinguistic variables. Journal of Language and Social Psychology 18 (1): 62–85.

Pierce J. E. 1952. Dialect distance testing in Algonquian. International Journal of American Linguistics 18 : 208–18.

Preston D. R. (ed.). 1999. Handbook of perceptual dialectology , vol 1. Amsterdam : John Benjamins.

Ralston J. V., Pisoni D. B., Lively S. E., Greene B. G., & Mullennix J. W. 1991. Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. Human Factors 33 (4): 471–91.

Rastatter M. P., & Gallaher A. J. 1982. Reaction-times of normal subjects to monaurally presented verbal and tonal stimuli. Neuropsychologia 20 : 465–73.

Scharpff P. J., & Van Heuven V. J. 1988. Effects of pause insertion on the intelligibility of low quality speech. In W. A. Ainsworth & J. N. Holmes (eds), Proceedings of the 7th FASE/Speech-88 Symposium (Edinburgh): 261–68.

Schüppert A., Hilton N. H., & Gooskens C.Accepted. Swedish is beautiful, Danish is ugly: Investigating the link between intelligibility and language attitudes. Linguistics.

Tang C., & Van Heuven V. J. 2007. Mutual intelligibility and similarity of Chinese dialects: Predicting judgments from objective measures. In B. Los & M. Van Koppen (eds.), Linguistics in the Netherlands 24: 223–34. Amsterdam : John Benjamins.

Tang C., & Van Heuven V. J. 2009. Mutual intelligibility of Chinese dialects experimentally tested. Lingua 119 : 709–32.

Thomas E. R. 2002. Sociophonetic applications of speech perception experiments. American Speech 77: 115–47.

Van Bezooijen R., & Van den Berg R. 1999. Taalvarieteiten in Nederland en Vlaanderen: hoe staat het met hun verstaanbaarheid? Taal en Tongval 51 (s): 15–33.

Van Heuven V. J., & De Vries J. W. 1981. Begrijpelijkheid van buitenlanders: de rol van fonische versus niet-fonische factoren, Forum der Letteren 22 : 309–20.

Van Heuven V. J., & Van Bezooijen R. 1995. Quality evaluation of synthesized speech. In W. B. Klein & K. K. Paliwal (eds.), Speech coding and synthesis , 707–38. Amsterdam : Elsevier Science.

Vitevitch M. S., & Rodriguez E. 2005. Neighborhood density effects in spoken word recognition in Spanish. Journal of Multilingual Communication Disorders 3 : 64–73.

Voegelin C. F., & Harris Z. S. 1951. Methods for determining intelligibility among dialects of natural languages. Proceedings of the American Philosophical Society 45 : 322–29.

Wang H. 2007. English as a lingua franca: Mutual Intelligibility of Chinese, Dutch and American speakers of English. LOT Dissertation Series, 147. Utrecht : LOT.

Wiener F. M., & Miller G. A. 1946. Some characteristics of human speech. *Transmission and reception of sounds under combat conditions. Summary Technical Report of Division 17, National Defense Research Committee* (Washington, DC), 58–68.

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 : Kovač.

Charlotte Gooskens

Charlotte Gooskens is Associate Professor of Scandinavian Linguistics at the University of Groningen, the Netherlands. Her research deals with perceptual and communicative effects of language variation, for example, language attitudes, speaker identity, and mutual intelligibility of closely related languages. She uses experimental research methods and exact measurement techniques. She is leader of the project Mutual Intelligibility of Closely Related Languages in Europe: Linguistic and Non-Linguistic Determinants, financed by The Netherlands Organization for Scientific Research at the Center for Language and Cognition, Groningen.