# Mutual Intelligibility

Gooskens, Charlotte; van Heuven, Vincent

*Published in:*
Similar Languages, Varieties, and Dialects

*DOI:*
10.1017/9781108565080.006

# 4 Mutual Intelligibility

*Charlotte Gooskens and Vincent J. van Heuven*

## 4.1 Introduction

### *4.1.1 Intelligibility in Speech Communication*

When two persons communicate through spoken language, thoughts that enter into the mind of the speaker first have to be expressed in terms of the vocabulary and grammatical structures of the speaker's language. The mental linguistic structures are then used to make the speaker's vocal organs move so that they produce audible sound. The sounds travel through the air (or some other medium – e.g., a telecommunication channel) and then impinge on the ear of the listener. The result is that the listener hears a stream of sounds. If the listener is familiar with the language, he or she will recognize (the same) linguistic units (e.g., words) in the same order in which they left the speaker's mouth. This part of the communication process is what we call *speech recognition*. If a sufficient number of units have been correctly recognized in the correct sequential order, the listener will be able to reconstruct the original thoughts and intentions of the speaker. This last part of the process is what we call *speech understanding* or *comprehension*. The sequence of events sketched here is known as the *speech chain* (Denes and Pinson 1963), and it has been the blueprint of Levelt's (1989) model of speech production and Cutler's (2012) model of native listening.

The intelligibility of a speaker, or of a speech utterance, is the degree to which a listener is able to recognize the linguistic units in the stream of sounds and to establish the order in which they were spoken. If the listener does not know the language the speaker uses, the speaker's intelligibility is (close to) zero – even if the utterance(s) would be perfectly intelligible to native listeners of the language. The comprehensibility of a speaker (or a spoken text) is the degree to which a listener is able to understand the speaker's meaning and intentions. Intelligibility, then, is the correlate of speech recognition and the comprehensibility of speech understanding. In this view speech understanding is a higher-order process than speech recognition. Different methods are required to assess a speaker's intelligibility than to assess that individual's comprehensibility. For instance, a strict intelligibility test would be ask a

listener to take down by way of dictation, a series of nonsense utterances produced by the speaker. This is what the semantically unpredictable sentences (SUS) test (Benoît et al. 1996) does with utterances like *The state sang through the whole week*. A speech comprehension test would, for example, ask a listener to determine whether a spoken sentence embodies a truth or a falsehood. If a listener would think that *Most human babies are heavier than a full-grown elephant* is true, he or she obviously has not understood the sentence. Section 4.2 reviews a range of experimental methods that have been used to establish the intelligibility of speakers and spoken utterances. It is important here to point out that intelligibility, in our view, is the joint product of the combination of a particular speaker and a particular listener. Speakers may differ from one another not only in their command of the language but they may also differ in the quality of their speech production due to personal habits, such as weak versus loud voice, fast versus slow tempo, and sloppy versus clear articulation. By the same token, listeners may differ in their familiarity with the language being spoken, hearing acuity or even motivation to understand what the speaker is trying to convey.

We spell this out in some detail because other disciplines and other researchers have used the terms differently from the way we do. For instance, applied linguists (Munro and Derwing 1995; Munro et al. 2006) use the term *intelligibility* as the degree to which a speaker can be understood using functional tests and the term *comprehensibility* for the listener's opinion as to how well a speaker (or utterance) can be understood. Our position is that intelligibility and comprehensibility address two different stages in the speech chain and that each can be measured both by functional tests (see earlier) and by opinion tests.

Some researchers also make a further distinction between *comprehensibility* and *interpretability*, where the former concept refers to the ease with which the listener may extract the propositional content of the sentence(s) produced by the speaker and the latter to the ease with which the speaker's intentions can be understood by the listener even if there is a cultural gap between the two interactants (Kachru and Smith 2008, chapter 4). In our own research we tend to ignore this subdivision because the languages and cultures we study tend to be closely related and do not regularly require large cross-cultural gaps to be bridged.

### *4.1.2 Mutual Intelligibility*

Until the 1950s it was assumed that the mutual intelligibility between two languages would be *symmetrical* (also called *reciprocal*). It was also assumed that the structural difference (or *distance*) between two languages A and B would be symmetrical. This, at first sight, is a reasonable proposition, since

generally the distance between any two locations on the map is symmetrical: the distance from London to New York is the same as the distance from New York to London. In Section 4.4, however, we will see that linguistic distance is not necessarily symmetrical.

Given symmetrical differences between two related languages A and B, language A should be as intelligible to native listeners of language B as the other way around. Casad (1974: 73) points out that reciprocity was part of the definition of mutual intelligibility in the work by the American structuralists in the early 1950s. Any deviation from perfect reciprocity would then be the result of either measurement error or of differences in extra-linguistic factors such as previous exposure to the other language. Pierce (1952), in fact, used *mutual intelligibility* for linguistically determined (and necessarily reciprocal) intelligibility versus *neighbor intelligibility* for nonreciprocal intelligibility, where the asymmetry could be due only to extra-linguistic differences (mainly contact). Small deviations from perfect symmetry in intelligibility were averaged out by computing the mean of the intelligibility scores for the two directions, AB and BA, as an index of mutual intelligibility. Such an index, however, will fail to predict the success of communicating if there is a large discrepancy between the AB and BA scores. An intelligibility index of 70 percent, for instance, would suggest reasonably successful communication between speakers A and B. However, if listener A understands speaker B at 90 percent, but listener B gets only 50 percent of speaker A's utterances (yielding an average of 70 percent), communication may well break down, and one-way communication would be the best possible result. Adopting Weinreich's (1957) term *cross-language communication* we prefer to use the term *cross-language intelligibility* when talking about the separate directions AB and BA (see also Ladefoged 1968). We will use *mutual intelligibility* as the mean of the two directions, with the caveat that the measure will fail when the cross-language intelligibility strongly deviates from reciprocity.

### *4.1.3 Inherent versus Acquired Intelligibility*

Listener A may be able to understand a speaker of a related language B for two (sets of) reasons, which are unrelated in principle. When the related languages (or language varieties) A and B are spoken in neighboring countries, or more generally on opposite sides of a shared geographic boundary, there will usually be contact between the speakers of the two languages, caused by, e.g., trade, tourism, personal relationships, and mass media. The speakers may also have become familiar with the other language as the result of formal teaching in a school setting. These reasons can be subsumed under the heading of extralinguistic or social factors. These will be discussed in detail in Section 4.3. Several terms have been proposed to capture the notion of intelligibility

as determined by extralinguistic factors, such as *acquired intelligibility*, *social intelligibility*, and *contact-based intelligibility* (Simons 1979). These terms can be used interchangeably.

When we are dealing with written language (rather than spoken language) a special extralinguistic factor influencing cross-language intelligibility may be the use of a shared ideographic writing system, as in China. Intelligibility between Sinitic languages across the Mandarin–Southern divide may be close to zero (Tang and Van Heuven 2007, 2009, 2015) for the spoken modality. However, since the same characters tend to be used to represent the same concepts across all Sinitic languages, no matter how the words are pronounced, printed text will still be understood. The opposite has also been found. Some closely related languages are spelled with such divergent orthographic conventions (but still using the Roman alphabet) that printed texts cannot be understood even though the spoken forms are mutually intelligible. For instance, Dutch readers generally fail to understand written Frisian (with spelling conventions that depart grossly from those of Dutch – and from any other language), but they will understand spoken Frisian rather well (Van Bezooijen and Gooskens 2005, 2006; Van Bezooijen and Van den Berg 2000).

Even if the interactants have never been exposed to the other language, we may still find considerable cross-language intelligibility, depending on how much the two languages are alike. This part of the cross-language intelligibility would then be based entirely on the degree of linguistic similarity between the two languages. Linguistic similarity is multidimensional and subsumes differences in any of the linguistic subdomains, such as lexicon (shared cognates with shared meanings), phonology (same or similar sound systems, transparent correspondences between the sound systems), morphology (same or similar word structure), and syntax (same or similar word order). The dimensions of linguistic similarity will be dealt with in Section 4.4. The terms *inherent intelligibility* and *similarity-based intelligibility* (Simons 1979) were introduced to represent the theoretical degree of understanding between two language varieties whose speakers have never had any contact.

It has been suggested that communication might be possible between two languages A and B that are not mutually intelligible if the interactants A and B both know a non-native language (variety) C that may bridge the gap between A and B. Language C should then be genealogically "in between" language A and B in terms of shared vocabulary and phonological distance. Shared familiarity with an in-between language helps extend the range of language varieties within a language family (or dialect group) within which speakers may use receptive multilingualism as a means of communication. This is a form of communication in which each interactant speaks his or her own language but is able to understand the language of the other enough to sustain a meaningful exchange of information. The mediating language C (also

referred to as a *bridge language*) is not necessarily genealogically in between the related languages A and B. It may also be the case that interactants are familiar with the mediating language C because it was taught in school. For instance, in eastern Europe Russian used to be a compulsory subject taught at secondary schools. Although Russian is not structurally and lexically halfway between Estonian (a Finno-Ugric language) and Ukrainian (Slavic), Estonian listeners will understand a Ukrainian speaker, simply because Ukrainian and Russian are mutually intelligible, and Estonians have learned Russian at school (Branets et al. 2020). In this example, of course, the Ukrainian listener will not understand the Estonian speaker, so this would end as a case of one-way intelligibility.

### *4.1.4 Symmetry versus Asymmetry in Intelligibility*

As explained in the foregoing, it was assumed in the 1950s that cross-language intelligibility should be reciprocal and that any deviations from perfect reciprocity (or symmetry) would be due to nonlinguistic factors such as exposure. Since then, however, many pairs of languages have been identified in which the cross-language intelligibility was far from symmetrical, and could not be attributed to differences in exposure or some other social variable. For instance, Jensen (1989) showed that Latin American Spanish speakers were better understood by Brazilian-Portuguese listeners (58 percent correct), than Brazilians were understood by the Spanish listeners (50 percent). The discrepancy was even stronger in recent experiments done with European Spanish and Portuguese speakers and materials (Gooskens and Van Heuven 2017). Similarly, Danish speakers are rather poorly understood by other Scandinavian listeners (Norwegians, Swedes), whereas Danes have less difficulty understanding spoken Norwegian or Swedish (see Gooskens et al. 2010; Schüppert 2011, and references therein).[1] Although partial explanations of these asymmetries based on sociological factors (differences in attitude, contact) cannot be ruled out entirely, it is generally agreed that another cause of the asymmetries is in the different linguistic structures of the languages involved. Typically, phonological lenition processes (such as consonant deletion and vowel apocope) have corrupted Danish and Portuguese words beyond recognition by listeners of the neighboring Scandinavian and Romance languages. Danish and Portuguese listeners, however, easily recognize the non-lenited forms in

[1] Kluge (2007: 11) lists the following examples of asymmetrical intelligibility: Gurage speech varieties of central Ethiopia (Gutt 1980; Ahland 2003); the Kaansé, Kpatogoso, Dogosé, and Khisa varieties of southwestern Burkina Faso (Showalter 1994); the Mazatec and Trique speech varieties of southwestern Mexico (Casad 1974: 76 ff.), as well as a number of speech varieties of southern Nigeria (Wolff 1959). The original publications should be consulted to determine the direction of the asymmetries for each language group.

the neighbor languages, especially if the orthography abstracts away from the lenition processes (Schüppert 2011; Schüppert et al. 2016, for the Scandinavian languages; Voigt and Schüppert 2013, for the Iberian languages).[2] If this is indeed the case, then symmetric distance measures quantifying the difference in linguistic structure between Spanish and Portuguese or between Danish and Swedish will be inadequate as explanations of the asymmetries found. In Section 4.4 we discuss a number of attempts to establish structural asymmetries between related languages, which may account for nonreciprocal intelligibility.

### *4.1.5 Why Study Mutual Intelligibility?*

Languages change over time. Changes are driven by internal linguistic forces as well as by social pressures. Sounds that are difficult to articulate or are difficult to hear as different from some other sound may be replaced by, or merged with, easier sounds. The way we speak and pronounce sounds may also be used as an identity marker by which we show that we belong to one particular group of speakers, and not to any other. Every time a group of speakers adopt a new way of speaking (called an *innovation*) and become different from some other group in what up to that moment was one homogenous linguistic community, a new language variety has come about. Over the centuries the language of a single linguistic community may have diversified into a great many varieties, differing from one another in the type and number of innovations that were implemented over the years. It is customary in (historical) linguistics to represent this process of diversification by means of family trees (or cladistic trees). A parent language (or ancestor language) splits up into younger varieties every time an innovation takes place. It is generally held that two languages are related to one another if they can be shown to have descended from the same parent language. Language varieties that differ from one another in only a few (recent) innovations are referred to as dialects of a language (see, e.g., Chapter 5 in this volume). If the number of innovations is large and such innovations were introduced many generations ago, we usually consider the varieties different languages.

[2] These explanation of the asymmetries were proposed in the literature several decades earlier. For instance, Chambers and Trudgill (1980: 4) write:

> Mutual intelligibility may also not be equal in both directions. It is often said, for instance, that Danes understand Norwegians better than Norwegians understand Danes. (If this is true it may be because, as Scandinavians sometimes say, 'Norwegian is pronounced like Danish is spelt', while Danish pronunciation bears a rather more complex relationship to its own orthography. It may be due, alternatively or additionally, to more specifically linguistic factors).

Casad (1974: 73) writes "since Portuguese has undergone a consonant deletion rule that Spanish has not, the surface phonological forms of Spanish correspond more closely to underlying proto-forms than the surface forms of Portuguese do. One might therefore predict that Portuguese speakers can understand Spanish better than Spanish speakers can understand Portuguese."

There is, however, no consistent way to quantify the number, type, and recency of innovations that differentiate two language varieties such that a clear-cut boundary can be drawn between what are different dialects of one language and what are different languages. Inherent mutual intelligibility was introduced by the American structuralists as a practical solution to this problem. If there is mutual intelligibility between the members of two different language varieties, these are considered dialects of one language; if there is no mutual intelligibility, the varieties belong to different languages. Mutual intelligibility as a criterion to distinguish dialects from languages generally works well, but it is known to fail when varieties are arranged along a dialect continuum. In a dialect continuum the geographically adjacent varieties differ only in a few innovations and are mutually intelligible. However, varieties at one extreme of the continuum will not be mutually intelligible with varieties at the other end, possibly hundreds of kilometers away. In the case of a dialect continuum the criterion of mutual intelligibility should then be applied as if it were a transitive relationship: if A understands B, and B understands C, then A also understands C.

It should be noted that neither the innovation-based nor the intelligibility-based language versus dialect dichotomy necessarily corresponds with the status currently attributed to many language varieties. Danish, Norwegian, and Swedish are mutually intelligible (e.g., in general Danes, Norwegians, and Swedes communicate with their own L1s rather than using a lingua franca such as English) and yet are considered different languages. Conversely, there are many pairs of geographically adjacent Chinese varieties that are not mutually intelligible (e.g., Tang and Van Heuven 2009) and yet are called dialects of Chinese. This means that the choice of whether two language varieties are dialects of the same language is basically a practical, political matter and not a question with any scientific import or theoretical status.

Mostly, the study of cross-language intelligibility has no theoretical import but is exclusively motivated by practical questions. It has been suggested, for instance, that spoken language varieties that are mutually intelligible – i.e., are dialects of one language – may well be served by a single orthography, which should then abstract away from superficial differences in pronunciation and realization of lexical tones. A single orthography, of course, would save time, effort, and development of teaching resources.

Testing of cross-language intelligibility is a prerequisite to building a theory that predicts how well speakers of two different but related languages will be able to communicate with each other through receptive multilingualism. Generally, establishing the degree of cross-language intelligibility is a time-consuming process involving large numbers of experimental participants. One of the goals of our work is to build a model that predicts the degree of cross-language intelligibility between two languages from a detailed comparison of the lexicon, phonology, morphology, and syntax of the languages concerned

and, if the written modality is included, also the orthographies. Once such a predictive model is available, it may help policy makers decide whether, for instance, television programs in a neighboring language should be dubbed or subtitled, whether receptive multilingualism would be a viable option for cross-linguistic communication or whether the use of (English as) a lingua franca should be promoted. If cross-language intelligibility is predicted to be insufficient, the model should be able to pinpoint the source of the difficulties that block intelligibility. Dedicated teaching programs can then be devised to help overcome the difficulties and permit receptive multilingualism (e.g., Golubovic 2016, chapter 5). In various parts of Europe, educational programs have been developed to teach receptive multilingualism, mostly in the written modality (e.g., the GalaNet and GalaPro,[3] EuroCom,[4] Linee,[5] and Dylan[6] projects). However, only little research has been conducted to investigate the effects of these programs.

## 4.2 How to Measure Intelligibility

A large number of methods have been devised to determine the degree of intelligibility of a speaker or of a speech utterance. Speech intelligibility testing has seen a wide range of applications, such as quality assessment of talking computers, determining the severity of a patient's speech or hearing defects, foreign language proficiency testing, and mutual intelligibility testing. A typology of test techniques can be given, using a limited number of parameters. Here we will concentrate on test techniques developed for spoken language. It is not difficult to see how the techniques should be adapted for the testing of written language.

### *4.2.1 Typology of Intelligibility Tests*

A first division of techniques is that between *opinion testing* (also called *judgment testing*) and *functional testing*. In an opinion test, listeners are asked how well they think they would understand a speaker or a spoken text. This is what the American structuralists called "ask the informant" (Voegelin and Harris 1951). Opinion testing can be done without using a physical stimulus. Assuming that the informants have had ample experience with another language, they can be asked to indicate on some scale (e.g., between 0 and 100) how well they would understand speech in the target language, where 0 would mean nothing and 100 would stand for perfect intelligibility. However, since it is

[3] http://e-gala.eu/. [4] EuroComprehension, www.eurocomprehension.eu.
[5] https://cordis.europa.eu/docs/results/28/28388/124376831-6_en.pdf.
[6] Dylan, www.dylan-project.org.

unclear what conception the participants have of the typical speaker of the target language, opinion testing is usually based on a selection of speech materials produced by one or several representative speakers of the target language. Opinion tests are relatively simple to carry out and take little time. Moreover, the same materials can be presented to the listeners repeatedly, either spoken by the same talker or by different talkers, without affecting the judgments. In functional testing, listeners have to actually show they have recognized linguistic units in a particular order and/or understood the meaning of what they just heard. This is what the American structuralists called "test the informant." Obviously, functional testing cannot be done without a physical stimulus being presented. A drawback of functional tests is that listeners can be presented with the same word or sentence only once. Once a listener has recognized a word, the same word – even when spoken by a different speaker or in a degraded condition – will be recognized much faster and more effectively. The problem can be circumvented by blocking multiple versions of the same stimulus over different groups of listeners, but this is time-consuming and presupposes the availability of large numbers of listeners. We usually find strong correlations between test scores obtained from judgment tasks and functional tasks. Moreover, the opinion scores are realistic in the sense that respondents do not overestimate or underestimate how well they would do in a functional test using the same materials (Gooskens and Swarte 2017).

The second parameter concerns the *linguistic level* that is tested. A test may address low-level intelligibility, requiring the listener to judge or show how well he or she recognized the words and the order in which they occurred. Alternatively, higher-order speech understanding may be targeted by asking the listener to judge or show how much of the content of what was said he or she understood. Since there may be substantial interaction between lower-order recognition and higher-order understanding processes, it is often necessary to construct the stimulus material in such a way that higher-order processes cannot be employed. Access to the mental lexicon can be blocked by presenting non-words only, as in 'Jabberwocky' (Carroll 1871). Speech understanding based partly on contextual cues can be blocked by using semantically unpredictable sentences (Benoît et al. 1996). The relative importance of semantic context can be assessed by systematically comparing the listener's word recognition scores obtained from constraining and nonconstraining sentences (Kalikow et al. 1977; Wang 2007, chapter 9).

Addressing different linguistic levels overlaps to some extent with another parameter in the typology of intelligibility tests, viz. *black box* versus *glass box* test testing.[7] If the researcher is interested in only the overall interlingual

[7] Also called "white box" testing, "clear box" testing, or "open box" testing. The concept was developed in the software testing industry (see e.g., Ehmer Khan 2011).

intelligibility between interactants, the process by which the communication takes place is considered a black box, the inner workings of which need not be known. However, if the researcher wants to pinpoint the causes of imperfect interlingual intelligibility, some form of diagnostic testing is required. Diagnostic intelligibility testing presupposes a modular view of the communication process, and specific tests that address each of the modules separately. The ultimate black box test would consider only the success rate of some interactive task performed by two participants who use receptive multilingualism – i.e., the interlingual communication type in which each interactant speaks only his or her native language and tries to understand the nonnative language as much as possible, relying on the lexical and structural similarity between the languages. The successful task completion rate in a (simulated) information service game would be an example of such a black box test. One interactant would be the information giver, the other the information requester. Together they must complete a communicative task – e.g., finding out how to travel from A to B, ordering a meal from a menu of choices, booking a seat on a plane, or obtaining the telephone number of a particular person. The first step on the way to glass box testing would be to test the success of the information exchange separately for the A-to-B and B-to-A directions. On a more fine-grained level, the separate contributions of vowels, consonants, prosody, non-cognates, morphological structure, and word order can be experimentally controlled and tested.

The third parameter, which applies only to functional intelligibility tests, is whether the test is online or offline. *Online* test techniques aim to tap into the listener's mind while the recognition and comprehension processes are being carried out. The results of online tests inform the researcher about the speed, sequential ordering, and interaction of modules involved in the processing of the spoken input. Reaction time measurements are claimed to provide an indirect indication of relative difficulty experienced during the processing of the input. More immediate access to the information processing can be obtained from eye tracking techniques (matching pictures on screen with words or sentences) or from neurological techniques such as evoked response potential (ERP) and or functional magnetic resonance imaging (fMRI), which tell the researcher exactly when and where in the brain decisions are being made by the listener, although we are not familiar with such neurolinguistic approaches in the area of mutual intelligibility testing.

Most of the intelligibility tests, however, are *offline*. Here the listener is allowed time to consider a response, and only the result of the response, rather than the time course, is considered. These results are typically the percentage of correctly recognized or translated linguistic units. Offline tests are used much more often than online techniques because they require no special and/or expensive equipment. It is not unusual to run online techniques in an offline mode. For instance, lexical decision tasks (decide whether a string of sounds

or letters exists as a word in the lexicon) or category monitoring tasks (decide whether a word is a member of some semantic category, e.g., denotes a concrete object) are online if the decision time can be measured (with millisecond precision). Such reaction time measurements can be done only by computers that are disconnected from a network. This precludes administering reaction time tests long distance over the internet. In such cases the correctness of the decision is the only measure of intelligibility.

### *4.2.2 Considerations*

In our view, measuring mutual intelligibility between two languages breaks down into separately assessing the cross-language intelligibility of language A for receivers of language B (AB) and of language B for receivers of A (BA). Moreover, measuring cross-language intelligibility is not principally different from measuring the intelligibility of a sender (speaker, writer) to a receiver (listener, reader) who both communicate in the same language. In our own work, we are mainly interested in establishing inherent (similarity-based) cross-language intelligibility. We are not interested in interactive strategies, and therefore assess the speaker's intelligibility in strictly one-way tests in a laboratory setting. How intelligible a sender is, can be determined only by studying the responses of the receiver to the signals (stimuli) produced by the sender.

Even healthy adult speakers of a language differ substantially in the quality of their speech production, depending on their habitual rate of delivery, fluency and pausing strategy, clarity of articulation, liveliness of melody, loudness, and overall voice quality (as determined by the efficiency of the vocal fold vibration). Intelligibility tests (Markham and Hazan 2004: 733) have shown that the scores obtained for a random selection of 33 speakers of British English (18 men) producing simple CVC words (and legal non-words) in a fixed carrier ranged between 82 and 97 percent correct as determined from the responses of 45 adult native listeners. Sentence intelligibility (100 Harvard Sentences; IEEE 1969) scores ranged between 81 and 93 percent correct for 20 American speakers (10 men) (Karl and Pisoni 1994; Bradlow et al. 1996; Bent et al. 2007).[8] If we want to compare the cross-language intelligibility in both directions, the speakers should be matched for intelligibility within their own speech community. This can be ensured either by using a large random sample of speakers (which is unpractical) or by selecting a small number of optimally representative speakers, or even a single one, from the

[8] In both the British and the American data, there is a small but significant effect of gender: women are slightly more intelligible than men, with a difference of 2 percentage points in the British word data and of 3 points in the American sentence data. No gender effects could be found in similar Dutch data (Tielen 1992).

larger group. The representative speaker(s) should be in the middle of the range of intelligibility scores found for the larger peer group. A clever way of circumventing the speaker variable is using one perfectly bilingual speaker. This would be a speaker who has learned both languages under comparison from childhood onward, and who cannot be identified as being different from monolingual native speakers of either language (using a voice line-up procedure as used in forensic phonetics, see, e.g., Broeders et al. 2002).

Listeners, like speakers, differ in how well they recognize and understand speech. Healthy adult native listeners in the British research cited obtained scores ranging between 88 and 96 percent correct across all talkers; no indication of between-listener variation is given in the American publications. When running cross-language intelligibility tests, researchers would therefore do well to recruit a fairly large number of listeners and at some later stage exclude those listeners who find themselves at the extremes of the score distribution.

It is often assumed that communication between native speakers and listeners of the same language is flawless. However, since even normal speakers differ in intelligibility, and listeners vary in their listening skills, we recommend measuring between-native intelligibility of the test materials used in the experiment as a baseline condition.

### *4.2.3 Survey of Intelligibility Testing Methods*

We will now present a nonexhaustive survey of techniques that have been employed in the field of cross-language intelligibility testing, concentrating on functional tests only. For more complete surveys of intelligibility tests, including opinion tests, we refer to chapters in handbooks such as Lawson and Peterson (2011) and McArdle and Hnath-Chisolm (2015) for speech audiometry, Gooskens (2013) for measuring interlingual intelligibility, Van Bezooijen and Van Heuven (1997) on the assessment of intelligibility of text-to-speech systems, and Kang et al. (2018) for intelligibility testing in the foreign language curriculum.

The first set of tests can be, and have been, used to test the intelligibility of single words presented out of context. The tests necessarily involve low-level, signal-driven word recognition and can be used to determine how difficult it is to recognize a (cognate) word in spite of a (strongly) deviant sound shape.

- *Word translation*. Listener hears isolated words in the non-native language, and writes down, types, or pronounces a word in the native language that captures the same meaning. Alternatively, the listener selects the correct translation from a closed list of alternatives.
- *Word-to-picture matching*. As word translation, but used when the listener is not required to respond using language. Listener hears a word and identifies

its meaning with one of (usually) four pictures presented on screen. This is an online technique using either a touch screen or eye tracking.

- *Lexical decision*. Listener hears a word-like sequence of sounds in the non-native language and has to decide as fast as possible, without making any errors, if the sequence exists as a word in the language or not. This is primarily an online word recognition technique. Decision time can be measured as an indication of processing difficulty. The assumption is that the sequence can be identified as a word only if the listener recognizes it. The paradigm can be complicated by presenting a prime word that is or is not (semantically) related to the test word. When related, the response time will be shorter. Impe (2010) used this technique to find very small differences in cross-lingual intelligibility between regional varieties of Standard Dutch as spoken in the Netherlands and Belgium.
- *Word category monitoring*. Listener hears a word and has to decide, as fast as possible while avoiding errors, which of a range of pregiven categories the word belongs to. The choice can be binary (e.g., tangible or intangible; animate or inanimate) or multivalued. Tang and Van Heuven (2009) used ten semantic categories such as body parts, family members, animals, and plants. Assigning category membership presupposes word recognition. Online decision time can be used as an additional indication of processing difficulty.

The next group of test techniques measures the intelligibility of words at the sentence level. Sentence context makes the target words predictable to a greater or lesser extent (if the context part is understood).

- *Full sentence translation*. Listener hears a (recorded) spoken sentence, possibly repeated at regular intervals to reduce memory load, and produces a translation in the native language, typically by writing or typing. This is an off-line task. The scoring of the response may be a problem. The technique was first used by the American structuralists (test the informant) to assess the interlingual intelligibility of Native American languages (e.g., Voegelin and Harris 1951; Hickerton et al. 1952; Pierce 1952; Biggs 1957).
- *Partial sentence translation*. Listener hears a sentence in a non-native language. The task is to write down the translation of the last word heard. The target word may or may not be highly predictable from the earlier part of the spoken sentence. Comparing the difference between the two conditions provides an indication how much of the semantic context was used by the listener to recognize the target. This is a interlingual adaptation of the Speech in Noise test (Kalikow et al. 1977), and was used by Tang and Van Heuven (2009) to measure the interlingual intelligibility of fifteen Chinese languages, and by Wang (2007) and Wang and Van Heuven (2013) for non-native Englishes.

- *Cloze test with written gaps*. Listener hears a spoken (sequence of) sentence(s) and sees a printed translation of the speech utterance in the native language. One or more words in the translation are replaced by blanks. The task is to write down (or choose from a list of alternatives) the blanked-out word(s). The task is easier (faster completion times, fewer errors) as the blanked-out words are contextually more constrained. Cloze testing was used by, e.g., Smith and Rafiqzad (1979) to assess the cross-lingual intelligibility of Asian World Englishes. It was also used to test the intelligibility of Frisian for Dutch listeners (Van Bezooijen and Van den Berg 2000).
- *Cloze test with spoken gaps*. Listener hears a spoken sentence in the non-native language in which one word is replaced by a beep. The task is to select one word from a printed list of alternatives (in the listener's L1) such that it optimally expresses the meaning of the missing word. This technique was used to assess the cross-lingual intelligibility in seventy pairs of European languages by Gooskens and Van Heuven (2017), and Gooskens et al. (2018).
- *Translation of semantically unpredictable sentences*. SUS sentences are quasi-random but syntactically grammatical sequences of short (monosyllabic, high-frequency) words. Five basic syntactic frames are used to generate sentences up to eight words long, as, e.g., *The state sang by the long week* or *Why does the range watch the fine rest?* There are SUS generators for most European languages (Benoît et al. 1996). The score is the percentage of correctly translated (content) words. Word recognition potentially benefits from top-down information on lexical category and sentence prosody. Responses are not constrained by semantic dependencies. SUS sentences were used by Gooskens et al. (2010) to assess the (asymmetrical) cross-lingual intelligibility of Danish and Swedish.
- *Sentence verification*. Listener hears a sentence that contains a logical proposition that is either true or false, e.g., *Horses are known to climb up trees*. The listener's task is to decide, as quickly as possible and without making any errors, whether the proposition is true or false. The technique can be used either as an offline task or as an online measure of sentence processing. In the latter case, the decision times have to be lined up with the earliest moment in the acoustic stimulus where sufficient information is available to correctly decide on the truth value of the sentence. Responses given before the temporal alignment point should be discarded as guessing. Since the response is binary (true/false), the number of test items should be large so as to reduce the influence of guessing. In an alternative application of the test, the listener is asked not to judge the truth of the proposition but its plausibility (e.g., Hilton et al. 2013 on the interlingual intelligibility of Scandinavian languages).
- *Carry out spoken instructions*. A straightforward method of measuring sentence understanding of a non-native language is having the listener carry

out instructions. Dependent variables are (1) the success rate with which the instructions are carried out, (2) the time it takes the respondent to start carrying out the description, and (3) the time it takes to successfully carry out the instruction. The instructions usually ask the listener to move or arrange objects in a virtual world on a computer screen. Van Heuven and De Vries (1981) used this technique to measure the intelligibility of Dutch spoken by Turkish immigrants (see also Van Heuven 1986).

Speech understanding is generally measured at the text level, using short texts composed of several sentences making up a coherent story or reasoning. Comprehension is then tested either by asking questions or by having the listener retell (i.e., translate or interpret) the text in his or her own language.

- *Text comprehension*. The participant reads or listens to a text of some length and answers questions about the contents. Usually the questions are asked after the presentation of the printed or spoken text, but some researchers maintain that it more realistic to present the questions beforehand so the participant knows what aspect of the contents to focus on. The questions are typically presented in multiple-choice format (three or four alternatives, only one of which is correct), which facilitates the scoring of the responses. Questions should be about the general ideas developed in the text and should not hinge on one specific word. As a precaution the questions should be tried out on a separate group of participants without any text presented to make sure that correct responses cannot be chosen on the basis of world knowledge or on logical grounds. The questions and the alternatives should be presented in the participant's L1, so that only the comprehension of the text (and not that of the questions) is measured. The recorded text testing (RTT) technique (Casad 1974) is an example of this type of test. In the RTT-Q version the questions are in open format. Interlingual comprehension testing based on this method was done in Scandinavia. Delsing and Lundin Åkesson (2005), for example, asked participants to answer just five open questions about short passages of continuous text.
- *Text translation*. This is the same technique, with the same advantages and drawbacks as sentence translation. Typically the text is presented sentence by sentence, while the participant responds by either reading out or typing the translation. Again, the scoring of the translations is a problem. In the RTT-retelling technique (Kluge 2007) the participant is asked to listen to a story in the non-native language and then to retell the story in the L1, keeping in as much detail of the original as possible. The result of the retelling is scored in terms of number of propositions in the original that are reflected in the retelling. This places a burden on the fieldworker, who has to analyze the original text and the retold versions in terms of propositions and then assess how well each of the propositions is maintained in the retelling. Retelling a

spoken story is basically the same task as *consecutive interpreting* (also called *conference interpreting*).

- *Story to picture matching*. A text is shown or played to the participant in its entirety after which the participant has to select one of four pictures shown on screen such that the picture chosen optimally matches the contents of the passage presented. In Gooskens et al. (2018) the four pictures were constructed such that they embodied the correct or wrong representation of two key propositions in the passage. For instance, if the passage was about driving a car in winter, one picture showed a car driving in a wintery landscape, another picture showed a car driving in summer (with a sunny landscape and trees and flowers in full bloom), a third picture would show a plane flying over a wintery landscape and a last picture contained a plane in a summer setting. When both content features were correctly identified the participant got full marks, when both aspects were wrongly identified no mark was given, when one feature was correct, the participant was given half marks. The technique is very fast and scoring is done automatically. In Gooskens et al. (2018), however, the test insufficiently discriminated among languages with high interlingual intelligibility, due to ceiling effects.[9]

We end this survey with a few examples of recent attempts to determine interlingual intelligibility at the discourse level. These tests involve live interaction between two participants, each using his or her own language, who have to solve a problem together.

- *Map task*. One interactant is the instruction giver and the other, the instruction follower. Both interactants have a map with roads and landmarks. The giver's task is to tell the follower how to trace a route between two landmarks that are known to the giver but unknown to the follower. To complicate the task, the two maps may differ in subtle ways. After completion of one task, new maps are provided and giver and follower switch roles. Dependent variables are success rate and time to completion (Anderson et al. 1991).
- *Spot the differences task*. Each of two interactants (who cannot see one another) has a copy of a picture that displays a large number of objects (e.g., toiletry articles) arranged in arbitrary order. There are differences between the two pictures in shapes, sizes, colors, and presence/absence of the objects. The participants' task is to identify as many of these differences as possible within

[9] There are many ways to systematically reduce the intelligibility of a spoken text to avoid ceiling effects, such as artificially speeding up the spoken text (Janse et al. 2003; Syrdal et al. 2012), adding (babble) noise (Gooskens et al. 2010), applying filtering (Wang et al. 2011), or using signal compression as used, for instance, in GSM telephony (Nooteboom and Doodeman 1984) or by varying the number of electrodes in simulated cochlear implants (Friesen et al. 2001).

a certain time frame. Dependent variables are the number of differences correctly identified and the number of spoken words used in the interaction (Van Mulken and Hendriks 2015).

## 4.3 Extra-linguistic and Para-linguistic Factors Influencing Intelligibility

The methods for establishing the level of intelligibility discussed in the previous section were developed for various purposes and capture the extent to which speakers of language A understand language B. Of course, linguistic overlap between the language of the listener and that of the speaker plays an important role in explaining how well listener A will understand speaker B. However, since intelligibility measurements are based on experiments with living persons, the results of the measurements depend on a large number of extra-linguistic and para-linguistic factors. An overview of such factors is provided in Gooskens (2019). Extra-linguistic factors include personality traits that have been identified within psychology to influence language learning, such as the ability to adapt to new situations, knowledge of the world, and access to sociocultural and cognitive resources.

Also age of the listener has been shown to affect the intelligibility of a related language. Vanhove and Berthele (2015) showed that in the written modality, cognate guessing skills, i.e., the ability to recognize words that are related to the historically related word in the L1, improve throughout adulthood while in the spoken modality, cognate guessing skills remain fairly stable between ages twenty and fifty but then start to decline. The speech-specific decline in cognate-guessing ability was tentatively attributed to different reliance on fluid intelligence (reasoning and problem-solving skills) and crystallized resources (in particular L1 vocabulary knowledge). Fluid intelligence tends to increase sharply into young adulthood and then declines, while crystallized resources stay stable or even increase throughout adulthood. Possibly, this interaction with modality is due to the fact that sounds differ between languages but letters do not. Older people are used to recognizing letters even though type fonts and personal hand-writing styles differ widely. But they no longer have the cognitive flexibility to accept atypical exemplars as tokens of their native sound categories. Alternatively, the authors suggest that it may be the time pressure associated with auditory stimulus presentation that caused the difference between the modalities. Spoken items were presented only once and thus required the quick application of cognitive flexibility, whereas speed was a lesser issue in the written mode because the words remained on the screen until the participants had entered their translations.

Attitudes toward the language and country of the speakers may affect the listener's willingness and motivation to understand an L2 speaker. Negative

attitudes or social stigmas attached to languages are often seen as a potential obstacle for successful communication between speakers of different languages. If people do not have the will to try to understand each other, linguistic similarity between languages is of little help. However, experimental support for the relationship between attitude and intelligibility has been rather weak (e.g., Gooskens and Van Bezooijen 2006; Impe 2010; Schüppert et al. 2015) probably due to the fact that it is difficult to elicit (subconscious) attitudes in experimental settings.

An important factor in explaining the level of intelligibility is the nature and amount of previous exposure to the language of the speaker. The more exposure listeners have had to a language, the more likely they are to understand it. Listeners who have learned the language in a formal setting will generally understand the language better than listeners who have not, but also exposure outside the class room (e.g., via television, music, social media, and personal contact) may improve intelligibility, because the listeners will learn some of the vocabulary and become conscious about sound correspondences between the L1 and the L2 (Gooskens and Swarte 2017).

Most listeners have at least some knowledge of other languages or dialects than their own L1. Often, this knowledge can be used to understand a closely related language. Listeners may understand some non-cognate words in the language of the speaker because they are loanwords from a language that they are familiar with. In addition, multilingual listeners tend to have a higher level of metalinguistic awareness and are better able to use cross-linguistic similarity to understand a language. Listeners with experience in listening to other languages are also likely to develop strategies to guess the meaning of cognates in a related language (inferencing strategies; Berthele 2011). Examples of competences for good guessing capacities are the ability to make a flexible and selective comparison of features and patterns, focusing on consonants and neglecting or systematically varying the vowels, and the ability to use contextual information to make decisions. Listeners should know when to stop searching for correspondences between the L1 and the L2 in order not to waste time. They can also make clear when they do not understand the speaker and provide feedback to show whether they have understood or not (back-channeling). The speakers on the other side can also use various strategies to improve intelligibility, such as speaking slowly, reformulating sentences, and avoiding words they know to be difficult in their own language and using words known to be cognates in the two languages.

Orthographic knowledge may play a role in the intelligibility of a closely related language, even when the interaction takes place in the spoken modality. For instance, divergence between orthographic and spoken similarity between the two languages has been suggested as the explanation for the asymmetric mutual intelligibility between Danish and Swedish (Chambers and Trudgill 1980; Schüppert 2011). Danes understand spoken Swedish better

than Swedes understand Danish, as has been borne out by an abundance of studies (see Gooskens et al. 2010) for a summary). How orthography helps can be illustrated by the following example. Literate Danes confronted with the spoken Swedish word *land* /land/ "country" can probably use their orthographic knowledge to match this word to their native correspondent *land* /lan$^{2}$/. On the other hand, this is not the case for Swedes listening to the Danish word because of the absence of the phoneme /d/, which is present in Swedish pronunciation as well as orthography. Gooskens and Doetjes (2009) showed that there are more Swedish words that Danes can understand by means of the orthography in the corresponding Danish cognates than Danish words that Swedes can use their orthography to recognise. This difference can be explained by the fact that spoken Swedish is close to both written Swedish and written Danish, whereas spoken Danish has changed rapidly during the last century and has undergone a number of reduction processes that are not reflected in the orthographic system. This means that Danes can often understand spoken Swedish due to its close similarity to written Danish, while Swedes get less help from written Swedish when understanding spoken Danish. Schüppert (2011) used event-related brain potentials (ERPs) to collect evidence that online activation of L1 orthography enhances word recognition among literate speakers of Danish who are exposed to samples of spoken Swedish. On the basis of these investigations, it can be concluded that Danish listeners indeed seem to make more use of the additional information that the L1 orthography can provide when listening to Swedish than Swedes when listening to Danish.

Paralinguistic factors include speech phenomena such as pitch, volume, speech rate, modulation, and fluency and nonvocal phenomena such as facial expressions, eye movements, and hand gestures are often included in the list of paralinguistic factors (Lyons 1977). Many linguists stress the importance of such factors for successful communication (Crystal 1975), but little research has been carried out to experimentally test the role of paralinguistic factors for the intelligibility of a closely related language.

The short overview of extra linguistic and para-linguistic factors provided in this section makes clear that predicting the level of intelligibility between languages is a complicated matter involving a large number of factors that may influence intelligibility to varying degrees. Simons (1979) notes that such factors may often explain asymmetric intelligibility between language pairs and he suggests that "discrepancies larger than 10% are due to social factors rather than linguistic factors" (quoted in Grimes 1992: 26). Grimes therefore suggests this threshold as a way to recognize such factors. However, he continues by noting that for some language combinations specific areas of phonology may play a role in explaining asymmetry. In Section 4.1.4 we provided examples of languages that are known to show asymmetric intelligibility. We will discuss possible linguistic explanations for asymmetric intelligibility in Section 4.4.4.

## 4.4 Linguistic Determinants of Intelligibility

The multilingualism factor discussed in Section 4.3 places many situations of intelligibility somewhere on the scale between inherent and acquired intelligibility. However, often the main interest of the researcher is to establish inherent intelligibility, i.e., the level of intelligibility that is linked to linguistic factors only, without any influence from previous exposure to the language of the speaker (acquired intelligibility) or another related language (mediated intelligibility). However, in practice inherent intelligibility is almost a theoretical construct since most listeners have had at least some exposure to the language of the speaker or some related language. In addition, some researchers note that functional testing is often very labor intensive and that the wide varieties of tests, test situations, and personal backgrounds of listeners involved in intelligibility research make it hard to compare levels of intelligibility between different language pairs. A way to circumvent these problems may be to measure objective linguistic distances by means of methods that have been developed for dialectometric research. By means of such measures, the degree of linguistic overlap at various linguistic levels can be expressed. Various investigations have shown that linguistic distance measures correlate with measures of inherent intelligibility. In this section, a number of computational approaches to measuring linguistic overlap between closely related languages are presented and discussed, and in particular it is shown how the measurements have been used to model intelligibility.

### *4.4.1 Lexical*

Many researchers have argued that the degree of lexical overlap between two languages is likely to be very fundamental for predicting the level of intelligibility. If two languages share no vocabulary, the languages are in principle not mutually intelligible, and the larger the lexical overlap, the larger the mutual intelligibility will be. A simple way of measuring lexical distance between two languages is to calculate the percentage of non-cognates. Cognates are historically related words in the vocabularies of the two languages. Cognates share form and meaning even though both may have changed so much across time that they are difficult to recognize as cognates. For example, the cognate word pair English *fish* and Danish *fisk* obviously has the same origin, but the word pair Eng. *year* and Da. *år* may be difficult to recognize since the forms have changed more. Note that lexical distance from language A to language B maybe be different from that from B to A. This can be part of the explanation for asymmetric mutual intelligibility. For example, A might have two synonyms for a concept that has only one equivalent in B. An example is *rom* (“room”) in Swedish, and *rum* or *værelse* in Danish. On first confrontation, a Swede will probably understand the Danish cognate word *rum* but not the non-cognate *værelse*. On the other hand, a Dane will easily understand Swedish *rom.*

To measure lexical distance between two languages the percentage of non-cognates needs to be established. This is not always a straightforward task and a number of decisions need to be taken.

First, in the strict definition, cognates are such word pairs that have developed from the same word in a common ancestor language, but for the purpose of predicting intelligibility it makes sense to also count borrowings that have the same origin as cognates, since they are often easily recognizable for a listener. They are generally more similar to the corresponding word in the L1 because they have had less time to change than inherited words that have been part of the lexicon for a much longer time than loanwords. For example, many Low German words were borrowed into Danish during the Middle Ages while more recently French and English were sources of borrowing into the languages. Examples are Da. *køkken* and Ge. *Küche* ("kitchen"), Da. *kusine* and Fr. *cousine* ("cousin"), and Da. *teenager* ("teenager"). In addition, many loanwords have specific segmental and/or prosodic properties that make them resistant to linguistic changes that affected inherited words (Gooskens et al. 2012). For example, French loans are stressed on the final syllable, cf. Sw. *mil'jö* and Da. *mil'ieu* ("environment"), whereas Germanic languages stress the stem-initial syllable. While in Germanic languages vowels in unstressed final syllables are often reduced, final syllables mostly maintain the full vowel in French loans.

Second, when measuring the lexical distance between two languages it is important to consider carefully what data set will be used for the measurements. It should contain enough words for a stable measurement. Furthermore, the selection of words used for the calculations depends on the purpose of the measurements. In traditional research on glottochronology and lexicostatistics the Swadesh list has often been used to calculate the genealogic relationship between languages (Swadesh 1971). However, to model intelligibility it is important to base the measurements on lists of words that represent the modern languages well. In recent years many corpora have been compiled for larger languages. Some researchers base their measurements on the most frequent words in such corpora, assuming that this is a good representation of the language as a whole. The 1000 most frequent words in a large corpus generally cover more than 70 percent of the word tokens in running English text (Nation and Waring 1997). Other researchers use running texts as a basis for measurements. Gooskens and Van Heuven (2020) established the degree of mutual intelligibility of sixteen closely related spoken languages within the Germanic, Romance, and Slavic language families in Europe using the same uniform methodology (cloze tests based on translations of the same four texts of in total 800 words). They measured the lexical distances between all language pairs within the same language family on the basis of the test material and found high correlations with intelligibility scores of listeners with little or no previous exposure to the test language ($r = -.69$ for the Romance languages, $r = -.80$ for the Slavic languages, and $r = -.95$ for the Germanic languages).

To establish whether the results could be generalized, i.e., whether the results would be the same if the intelligibility scores were predicted by means of linguistic distance measurements based on another data set they repeated the analysis with distance measures based on translations of a list of the 100 most frequently used nouns in the British National Corpus.[10] The correlations with intelligibility were just as high as the correlations with the distances based on text data. This shows that inherent intelligibility can be predicted quite well by lexical distances and that a short word list provides sufficient input for computing the distance measures needed.

Third, the researcher needs to consider what kinds of words to include in lexical distance calculations. Some non-cognate words in a text can easily be interpreted from the context or have little negative influence on intelligibility. The meaning of other words may be more difficult to predict or be more important for understanding the text. It is often assumed that content words (nouns, adjectives, numerals, main verbs) are more important for intelligibility than are function words (articles, conjunctions, prepositions, pronouns, auxiliaries, modals, particles, adverbs) because they express the content of the message (Van Bezooijen and Gooskens 2007). The importance of content words becomes clear when looking at the vocabulary in telegrams and newspaper headlines. To express a message as shortly as possible, most function words are left out; yet it is possible to understand the message. And even within the group of content words, some words are more important than others in certain contexts. Salehi and Neysani (2017) found that Turkish listeners had more difficulties guessing the meaning of Iranian-Azerbaijani verbs and nouns than the meaning of adjectives and adverbs. They explain this by the higher semantic load of nouns and verbs. This means that it may be possible to improve lexical distance measurements as predictors of intelligibility by weighing differences in verbs and nouns more heavily than differences in function words, adjectives, and adverbs. On the other hand, the results by Gooskens and Van Heuven (2020) summarized earlier show that measurements based on whole texts including all word classes are equally good predictors as measurements based on frequent nouns.

Fourth, it is not a straightforward task to decide what words are cognates. They can be coded qualitatively by the researcher on the basis of etymological knowledge. The information can be found in etymological dictionaries for the largest languages of the world. Ciobanu and Dinu (2014) describe a method that can do this manual work automatically by means of electronic dictionaries. However, when such etymological information is not available or if the researcher wishes to measure distances on the basis of large numbers of words, the researcher may use a quantitative method where string

[10] See British National Corpus at www.natcorp.ox.ac.uk.

distances are automatically computed (McMahon and McMahon 2005; Holman et al. 2008). Schepens et al. (2013) compared how well qualitatively (cognates in the Swadesh-200 word lists) and quantitatively established percentages of cognates predict speaking proficiency scores among 30,066 immigrants with thirty-five different mother tongues and found that the qualitative expert scores were better predictors ($r = -.77$) than quantitative measures ($r = -.66$). The intercorrelation between the qualitative and quantitative distance measures was $r = .90$. Partial semantic overlap can be taken into account when coding words lists for cognacy. For example, the English word *queen* is historically related to Danish *kvinde* ("woman") but shares only part of the meaning. Also compounds may cause coding problems. For example, in Danish *barnevogn* only the second part of the word is cognate of Dutch *kinderwagen* ("baby pram"). A pragmatic solution to such coding problems is to count such words as half cognates and assign them half a point for cognacy.

Finally, when calculating lexical distances with the aim of modeling intelligibility, the researcher also needs to decide how to deal with so-called false friends – i.e., words that sound similar but are not historically related and mostly have different meanings. An example of a false friend is German *Dach* /da$\chi$/ ("roof") Dutch *dag* /dɑ$\chi$/ ("day") by the German subjects. The German word is more similar to the Dutch word than the German cognate *Tag* /tak/. A false friend cannot be recognized by a listener with no previous knowledge of the language of the speaker. While regular non-cognates will in principle hinder intelligibility, false friends may cause even larger problems because they may actually mislead the listener. In addition, listeners are less likely to use contextual cues to guess the meaning of false friends than in the case of other unknown words because they do not realize that they are non-cognates. For this reason, it may be sensible to give such words an extra (negative) weight when coding for cognacy.

Neighborhood density is another lexical property that may influence intelligibility. Neighbors are defined as word forms that are similar to the stimulus word but differ from it in the presence, absence, or substitution of just one sound (or letter). A large number of neighbors broadens the pool of recognition candidates, causing delay or even failure of successful word recognition (see Luce and Pisoni 1998). Kürschner et al. (2008) found neighborhood density to be a significant predictor of intelligibility of Swedish words for Danish listeners. For instance, the Swedish word *säng* ("bed") was less often correctly translated (as *seng*, which has four Danish neighbors: *syng* ["sing"], *senge* ["beds"], *hæng* ["hang"], and *stæng* ["close"]), than the Swedish word *adress* ("address"), which has no neighbors. A measure of lexical distance might be refined by taking neighborhood density into consideration.

Lexical distance measurements are generally good predictors of experimental measurements of intelligibility, as was shown in early publications by, for

example, Bender and Cooper (1971), who found an *r* of .67 between morpheme cognateship (established on a variant of the Swadesh-100 word list) and interlingual intelligibility for all twenty-five combinations of five Cushitic languages.[11] Results of more recent investigations showing the relationship between lexical distances and intelligibility measures are summarized in Sections 4.4.2 and 4.4.5.

### *4.4.2 Phonetic*

The research discussed in the previous section shows that there is a strong relationship between lexical similarity and intelligibility but that lexical distances are not perfect predictors of intelligibility. As discussed, the lexical distance scores themselves could be improved, but also other linguistic levels might play a role in predicting intelligibility. The fact that a word is a cognate does not mean that the listener will always be able to match it with the counterpart in his or her own language. Two cognate words may have changed beyond recognition (see Tatman, Chapter 2 this volume, on phonetic variation in dialects). Various methods have been developed within dialectometry to measure dialect distances and draw dialect maps. These distance measurements can also be used to predict the intelligibility of cognates.

An early investigation was carried out on Chinese dialects on the basis of phonetic transcriptions of over 2,700 cognate words in seventeen dialects (Cheng 1997). The complexity of the correspondence patterns needed to convert the word strings from one dialect to their counterparts in the other was computed (systemic mutual intelligibility). Arbitrary reward and penalty points were assigned to sound correspondences in onset consonants, post-onset glides, nuclear vowels, coda consonants, and tones. Frequent sound correspondences (above the mean frequency for a particular sound) were assigned positive values, while relatively rare correspondences were negatively weighted. Cheng reasoned that the larger the complexity of the rule system needed to convert cognate strings between dialects, the lower the cross-dialect intelligibility would be, but he never tested the prediction against experimental results. Moreover, Cheng's phonetic distance measure is asymmetrical when fewer and simpler correspondences are needed in one direction than in the other. Tang and Van Heuven (2015) correlated Cheng's phonetic distance measure

[11] In Biggs (1957) lexical distance and interlingual intelligibility (averaged over AB and BA pairs and excluding AA pairs) for six Yuman languages (spoken in Arizona, USA) are even correlated at $r = .990$ (computed by us). This correlation is inflated, however, due to the bimodal distribution of the intelligibility scores. When computed separately for language pairs above 70 percent mutual intelligibility and those below 20 percent (there are no scores between 20 and 70 percent; see Figure 4.1a), we find $r = .882$ and $r = .533$, respectively.

with functional sentence intelligibility scores for 210 Chinese dialect pairs, and confirmed this prediction ($r = .772$).

The complexity of Cheng's computations makes it difficult to apply them to other language situations. However, other dialectometric distance measurements have been used successfully to model intelligibility. The Levenshtein distance measure has become the most widely used algorithm for predicting intelligibility. Phonetic distance between two language varieties is computed for aligned cognate word pairs as the smallest number of string edit operations needed to convert the string of phonetic symbols in language A to the cognate string in B. Possible string operations are deletions, insertions, and substitutions of symbols. Each string edit operation needed incurs a penalty of one point. The total number penalty points is then divided by the length of the alignment (number of alignment slots) to yield a length-normalized Levenshtein distance. The overall phonetic distance from language A to language B is the arithmetic mean of the normalized distances for all cognate word pairs in the research corpus (Nerbonne and Heeringa 2010). The Levenshtein algorithm was used to explain mutual intelligibility between various Germanic language varieties – e.g., Gooskens et al. (2008) for eighteen Scandinavian languages and dialects among standard Danish speakers, Gooskens and Swarte (2017) for twenty Germanic language combinations, Vanhove and Berthele (2017) for intelligibility of Swedish among German participants – and recently the algorithm has been used for intelligibility research in a large number of language combinations in other language areas – see, e.g., Golubovic (2016) for Slavic, Kaivapalu and Martin (2017) for Finnish-Estonian, Tang and Van Heuven (2015) for Chinese dialects, Salehi and Neysani (2017) for Turkish-Azeri, Čéplö et al. (2016) for Arabic dialects, Gooskens and Schneider (2019) for Pacific dialects, and Feleke et al. (2020) for Amharic and two Tigrigna varieties spoken in Ethiopia. All of these investigations found high correlations between intelligibility measurements and Levenshtein distances, typically at $.7 < r < .9$. Many of the investigations combine the Levenshtein measurements with measurements of lexical overlap, as described in Section 4.4.1 in regression analyses. For example, in an investigation by Beijering et al. (2008) and Gooskens et al. (2008) a regression analysis including lexical and Levenshtein distances resulted in a proportion explained variance of $R^2 = .81$ for the intelligibility of seventeen Scandinavian language varieties and standard Danish as assessed among young Danes from Copenhagen.

The simplest version of the Levenshtein algorithm uses binary differences between alignments; more advanced versions use graded weights that express acoustic segment distances. For example, the pair [i, o] is seen as being more different than the pair [i, ɪ]. However, for the purpose of modeling intelligibility, it is not clear how the differences should be weighted. Gooskens et al. (2015) found that minor phonetic details that could hardly be captured by Levenshtein

distances may sometimes have a major impact on the interlingual intelligibility of isolated words. The optimal weighing is likely to differ for different language combinations and depends on predictability and generalizablity of sound correspondences. Improvements of the algorithm should take into account the human decoding processes. For example, Gooskens et al. (2008) tested the intelligibility of eighteen Scandinavian language varieties among Danish listeners and correlated this with Levenshtein distances split up into consonant and vowel distances. Their results showed higher correlations with consonant distances than with vowel distances, suggesting that consonants convey more lexical information than vowels and therefore play a more important role in predicting intelligibility. However, the relative contribution of consonants and vowels to intelligibility may be different across languages since the size of consonant and vowel inventories can vary considerably and so can the number of vowels and consonants used in running speech. Čéplö et al. (2016) tested mutual intelligibility between three Arabic varieties and found vowel differences to affect mutual comprehension more than consonants. They explain this finding with the large interdialectal and allomorphic variation in consonants that listeners seem to be well able to deal with.

Gooskens et al. (2008) also found that insertions are better predictors of intelligibility than deletions. This is confirmed by Kaivapalu and Martin (2017) who found that Finns perceive more similarity between Finnish and Estonian than Estonians do. They explain this by the fact that Finnish word forms often contain material that is not present in the corresponding Estonian form, both within the inflectional formative and within the stem. They therefore conclude that the fact that something is missing compared to the L1 results in a larger perceived similarity than when something is added. We will come back to this point in Section 4.4.4.

Kürschner et al. (2008) correlated the intelligibility of 384 frequent Swedish words among Danes with eleven linguistic factors and carried out logistic regression analyses. Phonetic distances explained most of the variance. However, they also found that individual characteristics of words influence intelligibility. Word length, different numbers of syllables in L1–L2 words pairs, Swedish sounds not used in Danish, neighborhood density (see Section 4.4.1), and word frequency also correlated with intelligibility. Berthele (2011) and Möller (2011) note that listeners rely more on word beginnings than on later parts of words, and similarities of word onsets have been found to be more important than similarities in the rest of the word. Van Heuven (2008) showed that correct recognition of words synthesized from low-quality diphones was severely reduced if stress was shifted to an incorrect position in Dutch words. He therefore assumes that unexpected stress positions play a negative role in understanding speech in a closely related variety. Wang et al. (2011) monotonized Chinese sentences and presented these (and the original sentences as

well) to listeners in versions with high, medium, and low segmental quality. The results showed that lexical tone information is important, especially when the segmental quality is poor. Tone is therefore a potentially important factor in the interlingual intelligibility of tone languages. Yang and Castro (2008) computed tonal distance between dialects of tone languages spoken in the south of China in several different ways and found substantial correlations with functional intelligibility scores around $r = .7$. Tang and Van Heuven (2015), however, correlated similar tonal distance measures with functional and judged intelligibility measures for fifteen Mandarin and non-Mandarin Chinese dialects but found no significant correlations.

### *4.4.3 Morpho-syntactic*

Most investigations on linguistic determinants of intelligibility have focused on lexical and phonetic distances. It is generally assumed that these two linguistic levels are most important for intelligibility and in addition, most dialectometric measures for other levels have only recently been developed. However, there is evidence that morpho-syntax plays a role in predicting intelligibility and should therefore not be ignored. For example, by means of reaction time and correctness evaluation experiments, Hilton et al. (2013) investigated whether certain Norwegian grammatical constructions that are not used in Danish may impede Danes' comprehension of Norwegian sentences. Their results showed that when Danish listeners were presented with sentences with Norwegian word orders and morphology not used in Danish, they needed longer decision times and made more errors in a sentence verification task.

Recently various methods for measuring morphological and syntactic distances have been developed and applied in intelligibility research. Nerbonne and Wiersma (2006) introduced the "trigram measure," a measure of aggregate syntactic distance. Trigrams (different sequences of three lexical category labels) are inventoried and counted. Syntactic distance is then defined as 1 minus the Pearson correlation coefficient between the trigram frequencies. Heeringa et al. (2017) developed two additional measures, the "movement measure," which measures the average number of words that has moved in sentences of one language compared to the corresponding sentences in another language, and the "indel measure," which measures the average number of words being inserted or deleted in sentences of one language compared to the corresponding sentences in another language. Swarte (2016) measured mutual intelligibility between five Germanic languages by means of a spoken cloze test. She correlated the intelligibility scores with the three syntactic distance measures. The trigram measure showed the highest correlation with intelligibility ($r = .26$). Gooskens and Van Heuven (2020) found significant correlations between syntactic trigram distances and inherent intelligibility

($r = .72$ for fourteen Germanic language combinations, $r = .77$ for fifteen Romance language combinations, and $r = .53$ for twenty-nine Slavic language combinations).[12]

Heeringa et al. (2014) measured orthographic Levenshtein distances between five Germanic languages separately for stems and affixes. They found that orthographic stem variation among languages does not correlate with orthographic variation in inflectional affixes. This suggests that a distinction needs to be made between stem and affix distances. Gooskens and Van Heuven (2020) found significant correlations between affix distances and intelligibility for fifteen Romance ($r = .54$) and twenty-nine Slavic ($r = .81$) language combinations. The correlation for fourteen Germanic language combinations was insignificant.

### *4.4.4 Asymmetric Intelligibility*

As discussed in Section 4.1, mutual intelligibility may sometimes be asymmetric. This asymmetry is often caused by social factors. However, as we will demonstrate, the linguistic relationship between two languages may be asymmetrical and can therefore be part of the explanation for asymmetric mutual intelligibility.

The following example of the lexical correspondences between Dutch and German shows that lexical relationships can be asymmetric. A word in language A may have a cognate in language B, but a word in language B need not have a cognate synonym in language A. For instance, the Dutch word *plek* ("place, spot, location") has no cognate in German. The equivalent for *plek* in German would be *Ort*, which is cognate to Dutch *oord*. A German person may be able to understand the Dutch cognate *oord* but not the non-cognate *plek*. On the other hand, a Dutch person will probably understand *Ort*. Gooskens et al. (in preparation) modeled this asymmetry in an investigation of the mutual intelligibility of seventy closely related languages in Europe. The texts used for the functional intelligibility experiments (Gooskens and Swarte 2017) were all translated from the same original English text into each of the test languages. However, when calculating the lexical distances, they translated the words in the texts of each of the test languages to the corresponding cognates in the languages of the listeners if such cognates existed. The lexical distances were expressed as the percentages of non-cognates for each combination of stimulus text and the corresponding translations. This sometimes resulted in different distances from language A to language B and from language B to language A.

[12] The correlations are lower in Swarte (2016) because the participants had often had exposure to the test language. In the study by Gooskens and Van Heuven (2020) only the results of participants with very limited or no exposure were included.

For example, the distance from French to Romanian is 49 percent while it is 58 percent from Romanian to French. This would predict that Romanian is more difficult to understand for French speakers than the other way around and this is also what Gooskens and Van Heuven (2017) found.

Grimes (1992: 26) noted that asymmetric intelligibility between Spanish and Portuguese and between Chinese dialects can be traced to specific areas of phonology, and also for other language pairs it has been suggested that characteristics of the pronunciation may cause one language to be more difficult to understand for speakers of a closely related language than the other way around. For example, Bleses et al. (2008) have shown that the early language development of Danish children is somewhat slower than that of children with other mother tongues, such as English and Swedish. Bleses et al. attribute this result to the poor segmentability of Danish, which is caused by prosodic phenomena such as lack of specific juncture cues, compulsory sentence accents, and local signals to utterance function. At the segmental level, lenition of consonants and other reduction phenomena, in particular schwa assimilation and schwa deletion, would result in poor segmentability. These characteristics of Danish may be part of the explanation for the Swedish–Danish asymmetric intelligibility and ideally phonetic distance measurements should be able to capture such asymmetries.

The Levenshtein algorithm does not capture asymmetric phonetic relations between language varieties; the distance from language A to language B is equal to the distance from language B to language A. However, as discussed, it may be possible to improve the Levenshtein algorithm in such a way that it takes into account the human decoding process by assigning different weights to different operations. If insertions are given higher values than deletions, the distances measured may be asymmetric.

Other algorithms have been developed that are able to express phonetic asymmetries. The complexity scores developed by Cheng (1997) for Chinese dialects (see Section 4.4.2) result in different scores between AB/BA pairs of dialects. Somewhat misleadingly, Cheng calls the computed mean of the phonetic complexity scores "mutual intelligibility." It should be kept in mind that this computed mean is a prediction of mutual intelligibility at best, but that the actual mutual intelligibility can only be obtained from experimental results with live listeners and speakers. Tang and Van Heuven (2007) renamed Cheng's computational phonological distance measure the Phonetic Correspondence Index (PCI). Tang and Van Heuven (2009) established cross-lingual intelligibility scores at the word and sentence level for fifteen Chinese languages. Six of these languages were members of the Mandarin group. Standard Chinese, which is based on Beijing Mandarin, is a compulsory subject in primary education throughout China, so that Mandarin languages are not a good testing ground for exploring the potential of the PCI as predictor of asymmetrical

cross-lingual intelligibility. The cross-lingual intelligibility within any language pair involving Beijing (acquired intelligibility) or any other Mandarin language (using the Standard Mandarin as a bridge language) will be substantially better than can be predicted from linguistic distance measures. However, if we limit the comparison to only the nine non-Mandarin languages in the sample, the PCI asymmetry (the AB score minus the BA score) correlates significantly with the asymmetry found in the functional word and sentence scores for the thirty-six language pairs, where the correlation is predictably better at the word level ($r = .454, p = .003$, one-tailed) than at the sentence level ($r = .331, p = .024$, one-tailed).[13]

Moberg et al. (2007) used conditional entropy for accounting for asymmetric phonetic relations. This algorithm measures the complexity of a mapping, and is sensitive to the frequency and regularity of sound correspondences between two languages. Moberg et al. used conditional entropy measures in an attempt to explain the asymmetrical intelligibility between Danish, Swedish, and Norwegian by measuring the amount of entropy in both directions for each language combination. Conditional entropies do not measure how similar two languages are, but how predictable the correspondence is in a certain language pair. Given a certain sound in language A, how predictable is the corresponding sound in language B? The more predictable this sound is, the lower the entropy. Higher predictability aids intelligibility; therefore, the hypothesis is that a low entropy measure corresponds well with a high intelligibility score. The results suggest that conditional entropies correspond well with the asymmetric results of intelligibility tests that have been carried out between the three languages. Other researchers found similar results. Kyjanek and Haviger (2018) measured entropies between Czech, Slovak, and Polish that correspond with most of the results found in previous intelligibility research. Stenger et al. (2017) calculated entropy scores for written Czech–Polish and Bulgarian–Russian and refined these scores with the information-theoretic concept of surprisal. On the basis of the measurements they predict that no asymmetry can be expected for Russian–Bulgarian and, like Kyjanek and Haviger (2018), they predict that Czech readers may have more difficulties reading Polish than a Polish reader reading Czech.

No entropy-based asymmetry measures have been tested on experimental intelligibility data. To be more certain about the relationship, intelligibility experiments should be carried out that test the hypothesis that the most regular

[13] Tang and Van Heuven (2015) limited the prediction of word and sentence intelligibility from objective linguistic distance measures only after eliminating asymmetries from the data by taking the mean scores of the AB and BA pairs. The analysis of the asymmetry given here is new.

and frequent correspondence rules are more transparent for listeners. If this is indeed the case, then we would also expect the results of listeners who have had exposure to the test language to show higher correlation with high entropy measures since the listeners will have had the chance to discover correspondence rules. Listeners with no previous exposure can recognize words in the test language only on the basis of similarities with their native language and on their intuitions of possible sound correspondences. In this context the distinction between item similarity (similarity of individual forms such as sounds, morphemes, words or phrases) and system similarity (a set of principles for organizing forms paradigmatically and syntagmatically) introduced by Ringbom and Jarvis (2009) is relevant.

Another point to consider when applying the asymmetry measures discussed here is the fact that a large corpus is needed for stable results. According to Moberg et al. (2007) at least 800 words are needed for entropy measures, while measures with the Levenstein algorithm have given stable results with only 100 words.

Also morpho-syntactic asymmetries may play a role in explaining asymmetric intelligibility. For example, Gooskens and Van Bezooijen (2006) found that Dutch speakers tend to understand Afrikaans better than vice versa. They showed that one of the reasons for this is the simplified grammar of Afrikaans resulting in a higher degree of morphological and syntactic transparency for speakers of Dutch than for speakers of Afrikaans. Heeringa et al. (2017) measured mutual intelligibility between five Germanic languages on the basis of four texts of in total 800 words. The distances were measured from the source texts in the five languages to the five languages of the listeners, resulting in a total twenty distance measurements involving two distances per language combination, from language A to language B and from Language B to language A. They used the three different methods described in Section 4.4.3. Because of the nature of the database used for measuring the distances, they were able to detect asymmetric relationships between the languages. For example, all three measures suggest that asymmetric syntactical distances could be part of the explanation why native speakers of Dutch more easily understand German texts than native speakers of German understand Dutch texts (Swarte 2016).

### *4.4.5 Modeling Mutual Intelligibility*

Languages do not differ along just one dimension but may differ at all linguistic levels (lexical, phonetic/orthographic, morpho-syntactic, prosodic), and at each of the linguistic levels, languages may furthermore vary on many different parameters. Ideally, we would like to express the linguistic distance between language varieties using a single number on a one-dimensional scale. However,

there is no a priori way of weighing the different linguistic dimensions. Intelligibility is mostly expressed in one number (for example, the percentages of correct answers in an intelligibility test) and intelligibility measurements are an adequate way of determining the relative importance of the various linguistic dimensions. On the other hand, correlations between intelligibility measurements and linguistic distance measurements reveal which linguistic dimensions are most important for the intelligibility of a closely related language.

The investigations presented show that all linguistic levels can play a role in predicting how well speakers of closely related languages can understand each other's languages. Often there is a strong relationship between the various levels, but this does not always have to be the case. In their investigation of seventy European language combinations Gooskens and Van Heuven (2020) found that especially lexical distances show overlap with other linguistic distances. However, the correlations differ between different linguistic levels and languages. For example, Gooskens and Swarte (2017) found a small lexical distance between Danish and Swedish (5 percent) but a relatively large phonetic distance (46 percent), while for Dutch and German, the lexical distance was larger (20 percent) and the phonetic distance smaller (37 percent). They found a correlation of $r = .95$ between intelligibility scores and lexical distances, while the correlation with phonetic distance was nonsignificant ($r = .28$). Salehi and Neysani (2017) also found lexical distance to be more important than phonetic distance for explaining the intelligibility of Turkish among Iranian-Azerbaijani speakers. They explain this finding by the fact that the phonetic distances between Turkish and Azerbaijani are small and highly rule governed. Gooskens et al. (2008), on the other hand, correlated lexical and phonetic distances with the intelligibility of seventeen Scandinavian language varieties among Danes and found phonetic distance to be a better predictor of intelligibility ($r = .86$) than lexical distance ($r = .64$) for this particular set of closely related language varieties.

Differences at the morphological and syntactic levels are generally assumed to be less important for intelligibility than lexical and phonetic differences. Hilton et al. (2013; details in Sections 4.2.3 and 4.4.3) found that the non-native phonology impeded comprehension to a larger degree than morpho-syntactic differences, confirming the important role of phonetic similarities besides lexical similarities for comprehension. Gooskens and Van Heuven (2020) included lexical, phonetic, orthographic, and syntactic distances in a regression analysis in order to explain mutual intelligibility among closely related Germanic, Slavic, and Romance languages. Lexical distance was the best predictor of intelligibility in the case of Germanic and Slavic. However, for Romance languages, syntactic distance was the only predictor included in the model, and it was also included in the model for Slavic languages.

## 4.5 Relationship between Intelligibility and Language Trees

Generally, the greater the historical depth (also called *glottochronology*: how long ago did language A undergo an innovation that language B was not part of?), the less the two languages resemble one another, and the more difficult it will be for speakers of language A to be understood by listeners of language B and vice versa. Computational linguists have developed algorithms to assemble phylogenetic trees that represent hypotheses about the evolutionary ancestry of languages (see Dunn 2015; Bowern 2018). In this section we review some attempts at testing how well present-day intelligibility patterns of (inherent) cross-language intelligibility reflect genealogic taxonomies established by comparative linguists.

In terms of data processing, cross-lingual intelligibility scores are best presented in a matrix with speaker language in the rows against listener language in the columns. The scores in the cells can be seen as a distance measure between the two languages. Within-language intelligibility scores, which should be near perfect, are in the cells along the main diagonal of the matrix. The full square matrix can be simplified to a triangle by computing the mean of the contra-diagonal cell contents, which abstracts away from any asymmetries (see above). The matrix data can be converted to either hierarchical tree structures or to maps. The maps are generally two-dimensional and can be compared with geographic maps, which tend to congrue with intelligibility-based maps. Three-dimensional maps may be drawn, which are a combination of the geographic maps, with hierarchical tree structures superposed as contours or colors (delineating islands of equal interlingual intelligibility). Matrices, trees, and maps can be constructed with any kind of distance measures, including the various objective linguistic distance measures discussed in Section 4.3. In the present section we will only compare the congruence between interlingual intelligibility and diachronic distance using hierarchical tree structures.

The comparison, however, is not without problems. Linguists often disagree on the exact family relationships among languages. Cladistic trees they suggest may be based on synchronic counts of lexical correspondences rather than on glottochronology, which is indeed difficult to establish. Also, the fact that two languages split away from another a long time ago need not by itself compromise mutual intelligibility. If the innovation involved only a small detail of the phonology, for instance the change of all stem-initial voiceless plosives to affricates, and is not followed by many other changes in later years, intelligibility will hardly be affected. Moreover, there is no agreement among linguists that glottochronology is the only or even the preferred criterion for establishing the genealogy of languages. An alternative approach is based on counting the number of differences among (related) languages and so defining isogloss bundles that set apart groups of language varieties from

one another. But then again, there will be disagreement on which isoglosses should be considered important and which ones are insignificant (Chambers and Trudgill 1980: 112).

In the following we give the family tree structure suggested for the Yuman languages by Kroeber (1943) and for the Iroquoian languages by Julian (2010: 10). The trees we present have been pruned so as to contain only the languages that were tested (in italics) for mutual intelligibility.

| Yuman | Northern Iroquoian |
|---|---|
| North-West Arizona | *Onondaga* |
| *Walapai* | *Seneca* |
| *Havasapai* | *Cayuga* |
| *Yavapai* | Mohawk-Oneida |
| Colorado River | *Mohawk* |
| *Yuma* | *Oneida* |
| *Mohave* | Tuscarora-Nottoway |
| Gila River | *Tuscarora* |
| *Maricopa* | |

Figure 4.1 shows the distance matrices that can be filled with the cross-lingual intelligibility data reported by Biggs (1957) for six Yuman language varieties (with filling in of two missing cell contents by imputation) and for six Iroquoian languages by Hickerton et al. (1952, detailed tests only). Hierarchical trees were drawn below the matrices using average linking between groups.

The intelligibility-based tree conforms closely to Kroeber's (1943) linguistic tree, with the exception of Maricopa, which Kroeber placed in a separate (Gila River) group. Kroeber's family structure, however, was mainly based on

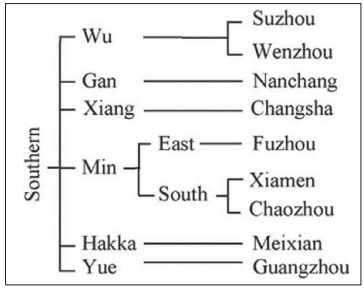
(a)

| Speaker language | Listener language Ma | Yu | Mo | Wa | Ha | Ya |
|---|---|---|---|---|---|---|
| Maricopa | **96** | 96 | 77 | 14 | 11 | 12 |
| Yuma | 98 | **92** | 49 | 11 | 9 | 15 |
| Mohave | 67 | 49 | **84** | 27 | 18 | 20 |
| Walapai | 18 | 11 | 11 | **96** | 98 | 82 |
| Havasapai | 10 | 4 | 16 | 91 | **91** | 78 |
| Yavapai | 13 | 4 | 13 | 96 | 96 | **94** |

(b)

| Speaker language | Listener language Sen | Cay | Ono | One | Moh | Tus |
|---|---|---|---|---|---|---|
| Seneca | **83** | 54 | 27 | 17 | 2 | 0 |
| Cayuga | 82 | **80** | 2 | 18 | 0 | 5 |
| Onondaga | 30 | 32 | **100** | 43 | 10 | 0 |
| Oneida | 30 | 7 | 0 | **46** | 45 | 0 |
| Mohawk | 0 | 20 | 0 | 50 | **82** | 0 |
| Tuscarora | 0 | 0 | 0 | 0 | 0 | **83** |

Figure 4.1 Cross-lingual intelligibility scores for all combinations of six Yuman languages (a) and six Iroquoian languages (b). The better the intelligibility, the darker the shading of the cell. Affinity trees are given below the matrices.




| Speaker language | Listener language | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | We | Gu | Xi | Chz | Fu | Chg | Su | Me | Na |
| Wenzhou | **93** | 5 | 12 | 2 | 3 | 2 | 5 | 7 | 10 |
| Guangzhou | 7 | **92** | 10 | 25 | 20 | 13 | 5 | 55 | 22 |
| Xiamen | 5 | 8 | **97** | 28 | 23 | 13 | 13 | 13 | 18 |
| Chaozhou | 0 | 3 | 52 | **98** | 13 | 3 | 7 | 3 | 12 |
| Fuzhou | 3 | 2 | 17 | 7 | **92** | 5 | 3 | 3 | 8 |
| Changsha | 3 | 8 | 23 | 3 | 17 | **93** | 12 | 17 | 25 |
| Suzhou | 7 | 5 | 18 | 5 | 13 | 13 | **77** | 7 | 13 |
| Meixian | 2 | 12 | 28 | 20 | 17 | 18 | 13 | **70** | 25 |
| Nanchang | 13 | 20 | 25 | 17 | 27 | 32 | 28 | 33 | **50** |

Figure 4.2 Cross-lingual intelligibility scores for all combinations of nine southern Chinese languages, based on functional sentence intelligibility test. Affinity tree (average linking between groups) is drawn below the matrix. The diachronic classification (left panel) is based on Li (1987a; 1987b: A1–2).

counting synchronic sameness in the lexicon and phonology, without much attention for historical depth, which strictly speaking renders this an invalid test. Similarly, the Iroquoian family tree predicts the intelligibility data reasonably well, although Mohawk and Oneida should have shown better cross-lingual intelligibility. Also, Cayuga has clearly better cross-lingual intelligibility with Seneca than with Onondaga, which is not predicted by the family tree.

Tang and Van Heuven (2007, 2009) established cross-lingual intelligibility (using opinion tests and functional tests) for 15 × 15 = 225 pairs of Chinese languages. Leaving out the six Mandarin languages (which pollute the results due to acquired intelligibility; see Section 4.4.4), Figure 4.2 shows the cross-lingual scores for the remaining 9 × 9 = 81 pairs of non-Mandarin languages based on functional sentence intelligibility.

Congruence with the diachronic taxonomy (Figure 4.2, left) is found only for the pair Xiamen–Chaozhou, which both belong to the South Min branch in the family tree. The close-knit pair Meixian–Nanchang, however, cannot be predicted from the genealogical tree.[14]

Gooskens et al. (2018) compared the congruence between cross-lingual intelligibility scores obtained from experiments with genealogies of Germanic (five), Romance (five) and Slavic (six) languages (seventy language pairs in total). In Figure 4.3 we show the genealogies, intelligibility data, and the affinity trees based on the experimental data. The intelligibility scores are based on a

[14] The affinity tree in Tang and Van Heuven (2009: 722) contains one incorrectly drawn branch, by which Wenzhou forms an early cluster with Suzhou. The tree presented here (and also in Tang 2009: 81) is correct.

| Speaker language | Listener language | | | | |
|---|---|---|---|---|---|
| | Da | Sw | Du | Ge | En |
| Da | | 56.0 | 9.9 | 12.5 | 7.9 |
| Sw | 43.8 | | 10.4 | 10.0 | 8.7 |
| Du | 13.3 | 13.0 | | 25.5 | 9.6 |
| Ge | - | 13.1 | - | | 9.5 |
| En | - | - | - | - | |

| Speaker language | Listener language | | | | |
|---|---|---|---|---|---|
| | Sp | It | Pt | Ro | Fr |
| Sp | | 56.0 | 62.0 | 46.6 | 11.0 |
| It | 38.2 | | 44.1 | 47.2 | 22.9 |
| Pt | 35.7 | 23.4 | | 20.7 | 18.0 |
| Ro | 13.7 | 8.7 | 14.7 | | 10.5 |
| Fr | - | 18.6 | - | - | |

| Speaker language | Listener language | | | | | |
|---|---|---|---|---|---|---|
| | Cz | Sk | Pl | Cr | Sl | Bu |
| Cz | | - | 24.0 | 18.1 | 18.1 | 10.8 |
| Sk | 87.5 | | 40.6 | 23.0 | 18.8 | 16.0 |
| Pl | 34.3 | 48.7 | | 9.5 | 12.6 | 7.1 |
| Cr | 19.9 | 24.5 | 14.6 | | 71.8 | 29.2 |
| Sn | 16.7 | 16.0 | 13.4 | 41.3 | | 20.2 |
| Bu | 13.4 | 10.1 | 13.7 | 19.7 | 18.6 | |

Figure 4.3 Cross-lingual intelligibility scores (cloze test) for all combinations of five Germanic, five Romance, and six Slavic languages. Affinity trees are given below the matrices.

cloze test, done by listeners who indicated to have had minimal prior exposure to the test language. For some language pairs, no such listeners could be found. For instance, since English is a compulsory school language in Scandinavia, Germany, and the Netherlands from age ten or earlier, the cells for English as a test language remain empty for listeners with some other Germanic L1. In order to compute the affinity trees, such empty cells were given the value of the contra-diagonal cell, assuming reciprocity for cross-lingual intelligibility.

The family trees for the languages are given in the following list. They are excerpted from Harbert (2007), Hall (1974), and Sussex and Cubberley (2006), for Germanic, Romance and Slavic, respectively.

| Germanic | Romance | Slavic |
|---|---|---|
| West | Italo-Western | West |
| North Sea | West | Northern |
| *English* | Ibero | *Polish* |
| Continental | *Portuguese* | Southern |
| *Dutch* | *Spanish* | *Czech* |
| *German* | Gallo | *Slovak* |
| North | *French* | South |
| East | Italo | Western |
| *Danish* | *Italian* | *Slovene* |
| *Swedish* | Eastern | *Croatian* |
| | North | Eastern |
| | *Romanian* | *Bulgarian* |

For the Germanic and Slavic groups the intelligibility-based trees are isomorphic with the linguistic taxonomy. The Romance tree, however, is a rather poor predictor of the Romance intelligibility results. The Ibero cluster (Spanish–Portuguese) is not reproduced, and French, which should be in a cluster with the other Italo-Western languages, is more remote even than Romanian. We computed correlation coefficients for the distances within language pairs based on the intelligibility data and on the linguistic taxonomies, using the cophenetic

(tree distance) measure as the criterion. Correlations were $r = .75$ for the Germanic group, $r = .41$ for the Romance group, and $r = .86$ for the Slavic group.

We conclude from the data presented here that the correlation between intelligibility-based distances and linguistic tree distance (based on the ordering of innovations in the history of the languages), is substantial but clearly imperfect – as was only to be expected given the disparity between the historical innovation-based approach, the synchronic shared vocabulary approach, and the rather noisy effects of these factors on contemporary intrinsic cross-lingual intelligibility. Nevertheless, if decisions about genealogic taxonomies are hard to make, present-day intelligibility results may well be given the casting vote.

## 4.6 Conclusions, Discussion, and Desiderata for Future Research

In this chapter we provided an overview of research on mutual intelligibility between closely related languages. We defined intelligibility as the degree to which a listener is able to recognize the linguistic units in the stream of sounds and to establish the order in which they are spoken. Mutual intelligibility is the mean of the two directions, i.e., the degree to which listener A understands speaker B and vice versa. It should be noted that the two directions can be asymmetric, i.e., can yield different scores. We gave an overview of methods for measuring intelligibility and considerations that have to be made when choosing a method.

One thing that strikes us is that it is very difficult to find cross-lingual intelligibility studies across widely differing language families that use the same methodology. Even the recorded text retelling technique used worldwide by the SIL researchers does not use standardized materials. The stories that are recorded and have to be retold differ from language group to another, thereby precluding cross-family comparisons. Methods used to study cross-lingual intelligibility among European languages (as in Gooskens et al. 2018) use rather different experimental methods than, e.g., Tang and Van Heuven (2009) for Chinese languages. If the same experimental method and the same materials (translations of some language and culture neutral text) were used, we could begin to say with some authority that the differences in cross-lingual intelligibility between certain European languages are larger than those between some Chinese "dialects." This would be a step on the way toward an internationally respected linguistic criterion to distinguish between languages and dialects.

For practical and theoretical reasons it is interesting to be able to explain and predict the results of intelligibility measurements and to understand why mutual intelligibility can be asymmetric. Extra-linguistic factors such as previous exposure to the language of the listener play an important role in predicting acquired intelligibility. However, often the main interest of the researcher is

inherent intelligibility and the relationship with linguistic factors. We showed how linguistic distances can be measured at different levels. The methods have been developed for dialectometric purposes, but the results of the investigations discussed in this chapter show that we are able to a large extent to predict mutual intelligibility of closely related languages by computational distance measurements.

However, since correlations between intelligibility measurements and linguistic distances are not perfect there are more linguistic and extra-linguistic factors that should be taken into consideration. We gave some suggestions for improvements of the computational algorithms for measuring linguistic distances, but future research should target more detailed knowledge of the mechanisms behind the intelligibility of language varieties. Methods that have been developed by experimental linguists and psycholinguists should be exploited when setting up controlled experiments that will give us more detailed insight into the relative importance of various linguistic and extra-linguistic factors that impact the intelligibility of language varieties.

## References

Ahland, Colleen. 2003. Asymmetry in the interlectal intelligibility of Gurage lects: Can implicational patterns explain and predict this phenomenon? Paper presented as coursework for a historical linguistics course at the University of Texas-Arlington.

Anderson, Anne H., Miles Bader, Ellen G. Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, & Regina Weinert. 1991. The Hcrc map task corpus. *Language and Speech* 34(4). 351–366.

Beijering, Karin, Charlotte Gooskens, & Wilbert Heeringa. 2008. Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm. *Linguistics in the Netherlands* 25(1). 13–24.

Bender, Marvin L. & Robert L. Cooper. 1971. Mutual intelligibility within Sidamo. *Lingua* 27. 32–52.

Benoît, Christian, Martine Grice, & Valérie Hazan. 1996. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication* 18(4). 381–392.

Bent, Tessa, Adam Buchwald, & Wesley Alford. 2007. Inter-talker differences in intelligibility for two types of degraded speech. *Research of Spoken Language Processing, Indiana University, Progress Report* 28. 316–330.

Berthele, Raphael. 2011. On abduction in receptive multilingualism. Evidence from cognate guessing tasks. *Applied Linguistics Review* 2. 191–220.

Biggs, Bruce. 1957. Testing intelligibility among Yuman languages. *International Journal of American Linguistics* 23. 57–67.

Bleses, Dorthe, Werner Vach, Marlene Slott, Sonja Wehberg, Pia Thomsen, Thomas O. Madsen, & Hans Basbøll. 2008. Early vocabulary development in Danish and other languages: A CDI based comparison. *Journal of Child Language* 35. 619–650.

Bowern, Claire. 2018. Computational phylogenetics. *Annual Review of Linguistics* 4. 281–296.

Bradlow, Ann R., Gina M. Torretta, & David B. Pisoni. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication* 20(3). 255–272.

Branets, Anna, Daria Bahtina, & Anna Verschik. 2020. Mediated receptive multilingualism: Estonian-Russian-Ukrainian case study. *Linguistic Approaches to Bilingualism* 10(3). 380—411.

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

Carroll, Lewis. 1871. *Through the Looking-Glass, and What Alice Found There.* London: Macmillan.

Casad, Eugene H. 1974. *Dialect Intelligibility Testing.* Norman, OK: Summer Institute of Linguistics of the University of Oklahoma.

Čéplö, Slavomír, Ján Bátora, Adam Benkato, Jiří Milička, Christophe Pereira, & Petr Zemánek. 2016. Mutual intelligibility of spoken Maltese, Libyan Arabic, and Tunisian Arabic functionally tested: A pilot study. *Folia Linguistica* 50(2). 583–628.

Chambers, Jack K. & Peter Trudgill. 1980. *Dialectology*. Cambridge: Cambridge University Press.

Cheng, Chin-Chuan. 1997. Measuring relationship among dialects: DOC and related resources. *Computational Linguistics and Chinese Language Processing* 2(1). 41–72.

Ciobanu, Alina M. & Liviu P. Dinu. 2014. On the Romance languages mutual intelligibility. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, et al. (eds.). *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC*, 3313–3318. Reykjavik: European Language Resources Association (ELRA).

Crystal, David. 1975. Paralinguistics. In Jonathan Benthall & Ted Polhemus (eds.). *The Body as a Medium of Expression*, 162–174. London: Institute of Contemporary Arts.

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

Delsing, Lars-Olof & Katarina Lundin Åkesson. 2005. *Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska* Copenhagen: Nordic Council of Ministers.

Denes, Peter B. & Elliot N. Pinson. 1963. *The speech chain. The physics and biology of spoken language*. Murray Hill, NJ: Bell Telephone Laboratories.

Dunn, Michael. 2015. Language phylogenies. In Claire Bowern & Bethwyn Evans (eds.). *The Routledge Handbook of Historical Linguistics,* 190–211. London: Routledge.

Ehmer Khan, Mohammed. 2011. Different approaches to white box testing technique for finding errors. *International Journal of Software Engineering and Its Applications* 5(3). 1–13.

Feleke, Tekabe L., Charlotte Gooskens, & Stefan Rabanus. 2020. Mapping the dimensions of linguistic distance: A study on South Ethiosemitic languages. *Lingua* 243, 1–31. DOI: 10.1016/j.lingua.2020.102893.

Friesen, Lendra M., Robert V. Shannon, Deniz Baskent, & Xiaosong Wang. 2001. Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *Journal of the Acoustical Society of America* 110. 1150–1163.

Golubovic, Jelena. 2016. *Mutual Intelligibility in the Slavic Language Area.* Groningen: University of Groningen.

Gooskens, Charlotte. 2013. Experimental methods for measuring intelligibility of closely related language varieties. In Robert Bayley, Richard Cameron, & Ceil Lucas (eds.). *The Oxford Handbook of Sociolinguistics*, 195–213. Oxford: Oxford University Press.

Gooskens, Charlotte. 2019. Receptive multilingualism. In Simona Montanari & Suzanne Quay (eds.). *Multidisciplinary Perspectives on Multilingualism*, 149–174. The Hague: De Gruyter.

Gooskens, Charlotte & Gerard Doetjes. 2009. Skriftsprogets rolle i den dansk-svenske talesprogsforståelse. *Språk och stil* 19. 105–123.

Gooskens, Charlotte, Wilbert Heeringa, & Karin Beijering. 2008. Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing* 2(1–2). 63–81.

Gooskens, Charlotte, Wilbert Heeringa, & Vincent J. van Heuven. In preparation. Comparing Germanic, Romance and Slavic: Relationships among linguistic distances. *Quantitative Linguistics*.

Gooskens, Charlotte, Sebastian Kürschner, & Renée van Bezooijen. 2012. Intelligibility of Swedish for Danes: Loan words compared with inherited words. In Henk van der Liet & Muriel Norde (eds.). *Language for Its Own Sake*, 435–455. Amsterdam: Amsterdam Contributions to Scandinavian Studies.

Gooskens, Charlotte & Cindy Schneider. 2019. Linguistic and non-linguistic factors affecting intelligibility across closely related varieties in Pentecost Island, Vanuatu. *Dialectologia* 23. 61—85.

Gooskens, Charlotte & Femke Swarte. 2017. Linguistic and extra-linguistic predictors of mutual intelligibility between Germanic languages. *Nordic Journal of Linguistics* 40(2). 123–147.

Gooskens, Charlotte & Renée van Bezooijen. 2006. Mutual comprehensibility of written Afrikaans and Dutch: Symmetrical or asymmetrical? *Literary and Linguistic Computing* 23. 543–557.

Gooskens, Charlotte, Renée van Bezooijen, & Vincent J. van Heuven. 2015. Mutual intelligibility of Dutch-German cognates by children: The devil is in the detail. *Linguistics* 53(2). 255–283.

Gooskens, Charlotte & Vincent J. van Heuven. 2017. Measuring cross-linguistic intelligibility in the Germanic, Romance and Slavic language groups. *Speech Communication* 89. 25–36.

Gooskens, Charlotte & Vincent J. van Heuven. 2020. How well can intelligibility of closely related languages in Europe be predicted by linguistic and non-linguistic variables? *Linguistic Approaches to Bilingualism* 10(3). 351–379.

Gooskens, Charlotte, Vincent J. van Heuven, Jelena Golubovic, Anja Schuppert, Femke Swarte, & Stefanie Voigt. 2018. Mutual intelligibility between closely related languages in Europe. *International Journal of Multilingualism* 15(2). 169–193.

Gooskens, Charlotte, Vincent J. van Heuven, Renée van Bezooijen, & Jos J. A. Pacilly. 2010. Is spoken Danish less intelligible than Swedish? *Speech Communication* 52(11–12). 1022–1037.

Grimes, Joseph E. 1992. Correlations between vocabulary similarity and intelligibility. In Eugene H. Casad (ed.). *Windows on Bilingualism*, 17–32. Dallas, TX: Summer Institute of Linguistics and the University of Texas at Arlington.

Gutt, Ernst-August. 1980. Intelligibility and interlingual comprehension among selected Gurage speech varieties. *Journal of Ethiopian Studies* 14. 57–85.

Hall, Robert A. 1974. *External History of the Romance Languages.* New York: Elsevier.

Harbert, Wayne. 2007. *The Germanic Languages.* Cambridge: Cambridge University Press.

Heeringa, Wilbert, Femke Swarte, Anja Schüppert, & Charlotte Gooskens. 2014. Modeling intelligibility of written Germanic languages: Do we need to distinguish between orthographic stem and affix variation? *Journal of Germanic Linguistics* 26(4). 361–394.

Heeringa, Wilbert, Femke Swarte, Anja Schüppert, & Charlotte Gooskens. 2017. Measuring syntactical variation in Germanic texts. *Digital Scholarship in the Humanities* 33(2). 279–296.

Hickerton, Harold, Glen D. Turner, & Nancy P. Hickerton. 1952. Testing procedures for estimating transfer of information among Iroquois dialects and languages. *International Journal of American Linguistics* 18. 1–8.

Hilton, Nanna H., Charlotte Gooskens, & Anja Schüppert. 2013. The influence of non-native morphosyntax on the intelligibility of a closely related language. *Lingua* 137(4). 1–18.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, & Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica* 42(3–4) 331–354..

IEEE. 1969. IEEE recommended practices for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics* 17. 227–246.

Impe, Leen. 2010. *Mutual Intelligibility of National and Regional Varieties of Dutch in the Low Countries.* Leuven: Catholic University of Leuven.

Janse, Esther, Sieb Nooteboom, & Hugo Quene. 2003. Word-level intelligibility of time-compressed speech: Prosodic and segmental factors. *Speech Communication* 41. 287–301.

Jensen, John B. 1989. On the mutual intelligibility of Spanish and Portuguese. *Hispania* 72(4). 848–852.

Julian, Charles. 2010. *A History of the Iroquoian Languages.* Winnipeg: Department of Linguistics University of Manitoba.

Kachru, Yamuna & Larry E. Smith. 2008. *Cultures, Contexts and World Englishes.* New York/London: Routledge.

Kaivapalu, Annekatrin & Maisa Martin. 2017. Perceived similarity between written Estonian and Finnish: Strings of letters or morphological units? *Nordic Journal of Linguistics* 40(2). 149–174.

Kalikow, D. N., K. N. Stevens, & L. L. Elliott. 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. 1337–1351.

Kang, Okim, Ron I. Thomson, & Meghan Moran. 2018. Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning* 68(1). 115–146.

Karl, John R. & David B. Pisoni. 1994. The role of talker-specific information in memory for spoken sentence. *Journal of the Acoustical Society of America* 95. 2873.

Kluge, Angela. 2007. *"Sorry, Could You Repeat That, Please?!" – Where Does One Language End and the Next Begin?* Jakarta: SIL International-Indonesia.

Kroeber, Alfred L. 1943. *Classification of the Yuman Languages.* Berkeley/Los Angeles: University of California Press.

Kürschner, Sebastian, Charlotte Gooskens, & Renée van Bezooijen. 2008. Linguistic determinants of the intelligibility of Swedish words among Danes. *International Journal of Humanities and Arts Computing* 2(1–2). 83–100.

Kyjanek, Lukáš & Jiří Haviger. 2018. The measurement of mutual intelligibility between West-Slavic languages. *Journal of Quantitative Linguistics* 26(3). 205–230.

Ladefoged, Peter N. 1968. *The measurement of Cross-Language Communication.* Washington, DC: Office of Health, Education, and Welfare.

Lawson, Gary & Mary Peterson. 2011. *Speech Audiometry.* San Diego CA: Plural Publishers.

Levelt, Willem J. M. 1989. *Speaking: From Intention to Articulation* (ACL-MIT Press Series in Natural-Language Processing). Cambridge, MA: MIT Press.

Li, Rong. 1987a. Chinese dialects in China. In Stephen A. Wurm, Benjamin T'sou, David Bradley, et al. (eds.). *Language Atlas of China*, map A-2. Hong Kong: Longman.

Li, Rong. 1987b. Languages in China. In Stephen A. Wurm, Benjamin T'sou, David Bradley, et al. (eds.). *Atlas of China*, map A-1. Hong Kong: Longman.

Luce, Paul A. & David B. Pisoni. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19. 1–36.

Lyons, John. 1977. *Semantics,* vol. 2. Cambridge: Cambridge University Press.

Markham, Duncan & Valérie Hazan. 2004. The effect of talker- and listener-related factors on intelligibility for a real-word, open-set perception test. *Journal of Speech, Language, and Hearing Research* 47(4). 725–737.

McArdle, Rachel & Theresa Hnath-Chisolm. 2015. Speech audiometry. In Jack Katz, Marshall Chasin, Kristina English, Linda J. Hood, & Kim L. Tillery (eds.). *The Handbook of Clinical Audiometry*, 61–75. Philadelphia: Wolters Kluwer.

McMahon, April M. S. & Robert McMahon. 2005. *Language Classification by Numbers* (Oxford Linguistics). Oxford: Oxford University Press.

Moberg, Jens, Charlotte Gooskens, John Nerbonne, & Nathan Vaillette. 2007. Conditional entropy measures intelligibility among related languages. In Peter Dirix, Ineke Schuurman, Vincent Vandeghinste, & Frank Van Eynde (eds.). *Computational Linguistics in the Netherlands 2006: Selected Papers from the 17th CLIN Meeting*, 51–66. Utrecht: LOT.

Möller, Robert. 2011. Wann sind Kognaten erkennbar? Ähnlichkeit und synchrone Transparenz von Kognatenbeziehungen in der germanischen Interkomprehension. *Linguistik Online* 46(2). 79–101.

Munro, Murray J. & Tracey M. Derwing. 1995. Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning* 45. 73–97.

Munro, Murray J., Tracey M. Derwing, & Susan L. Morton. 2006. The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition* 28(1). 111–131.

Nation, Paul & Rob Waring. 1997. Vocabulary size, text coverage, and word lists. In Norbert Schmitt & Michael McCarthy (eds.). *Vocabulary: Description, Acquisition and Pedagogy*, 6–19. Cambridge: Cambridge University Press.

Nerbonne, John & Wilbert Heeringa. 2010. Measuring dialect differences. In Jürgen E. Schmidt & Peter Auer (eds.). *Language and Space: Theories and Methods*, 550–567. Berlin: Mouton De Gruyter.

Nerbonne, John & Wybo Wiersma. 2006. A measure of aggregate syntactic distance. In John Nerbonne & Erhard Hinrichs (eds.). *Linguistic Distances Workshop at the Joint Conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July 2006*, 82–90.

Nooteboom, Sieb G. & Gert J. N. Doodeman. 1984. Speech quality and the gating paradigm. In M. P. R. Van den Broecke & A. Cohen (eds.). *Proceedings of the Tenth International Congress of Phonetic Sciences, Utrecht, August,* 481–485. Dordrecht: Foris.

Pierce, Joe E. 1952. Dialect distance testing in Algonquian. *International Journal of American Linguistics* 18. 208–218.

Ringbom, Håkan & Scott Jarvis. 2009. The importance of cross-linguistic similarity in foreign language learning. In Michael H. Long & Catherine J. Doughty (eds.). *Handbook of Language Learning*, 106–118. Oxford: Blackwell.

Salehi, Mohammad & Aydin Neysani. 2017. Receptive intelligibility of Turkish to Iranian-Azerbaijani speakers. *Cogent Education* 4(1). 1–15.

Schepens, Job, Frans Van der Slik, & Roeland van Hout. 2013. The effect of linguistic distance across Indo-European mother tongues on learning Dutch as a second language. In Lars Borin & Anju Saxena (eds.). *Approaches to Measuring Linguistic Differences,* 199–230. Berlin: De Gruyter Mouton.

Schüppert, Anja. 2011. *Origin of Asymmetry: Mutual Intelligibility of Spoken Danish and Swedish*. Groningen: University of Groningen.

Schüppert, Anja, Nanna H. Hilton, & Charlotte Gooskens. 2015. Swedish is beautiful, Danish is ugly? Investigating the link between language attitudes and intelligibility. *Linguistics* 53(2). 275–304.

Schüppert, Anja, Nanna H. Hilton, & Charlotte Gooskens. 2016. Why is Danish so difficult to understand for fellow Scandinavians? *Speech Communication* 79. 47–60.

Showalter, Stuart D. 1994. How attitude, use, and bilingualism data help define language shift and dialect choice in a rural cluster. *Notes on Sociolinguistics* 40. 3–26.

Simons, Gary F. 1979. *Language Variation and Limits to Communication.* Ithaca, NY: Department of Modern Languages and Linguistics, Cornell University.

Smith, L. E. & K. Rafiqzad. 1979. English for cross-cultural communication: The question of intelligibility. *TESOL Quarterly* 13. 371–380.

Stenger, Irina, Klára Jagrova, Andrea Fischer, Tania Avgustinova, Dietrich Klakow, & Roland Marti. 2017. Modeling the impact of orthographic coding on Czech–Polish and Bulgarian–Russian reading intercomprehension. *Nordic Journal of Linguistics* 40(2). 175–199.

Sussex, Roland & Paul Cubberley. 2006. *The Slavic Languages.* Cambridge: Cambridge University Press.

Swadesh, Morris. 1971. *The Origin and Diversification of Language: Edited Post Mortem by Joel Sherzer*. Chicago, IL: Aldine.

Swarte, Femke. 2016. *Predicting the Mutual Intelligibility of Germanic Languages from Linguistic and Extra-Linguistic Factors.* Groningen: University of Groningen.

Syrdal, Ann K., H. Timothy Bunnell, Susan R. Hertz, Taniya Mishra, Murray Spiegel, Corine Bickley, Deborah Rekart, & Matthew J. Makashay. 2012. Text-to-speech intelligibility across speech rates. In *Proceedings of Interspeech–2012*, 623–626.

Tang, Chaoju. 2009. *Mutual Intelligibility of Chinese Dialects: An Experimental Approach.* Utrecht: Netherlands Graduate School of Linguistics (LOT).

Tang, Chaoju & Vincent J. van Heuven. 2007. Predicting mutual intelligibility in Chinese dialects. In William J. Barry & Jürgen Trouvain (eds.). *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken*, 1457–1460. Saarbrücken: Universität des Saarlandes.

Tang, Chaoju & Vincent J. van Heuven. 2009. Mutual intelligibility of Chinese dialects experimentally tested. *Lingua, an International Review of General Linguistics* 119(5). 709–732.

Tang, Chaoju & Vincent J. van Heuven. 2015. Predicting mutual intelligibility of Chinese dialects from multiple objective linguistic distance measures. *Linguistics* 53(2). 285–312.

Tielen, Mirjam. 1992. *Male and Female Speech: An Experimental Study of Sex-Related Voice and Pronunciation Characteristics.* Amsterdam: Universiteit van Amsterdam.

Van Bezooijen, Renée & Charlotte Gooskens. 2005. Intertalig tekstbegrip. De begrijpelijkheid van Friese en Afrikaanse teksten voor Nederlandse lezers. *Nederlandse Taalkunde.* 10(2). 129–152.

Van Bezooijen, Renée & Charlotte Gooskens. 2006. Waarom is geschreven Afrikaans makkelijker voor Nederlandstaligen dan andersom? In Tom Koole, Jacomine Nortier & Bert Tahitu (eds.). *Artikelen van de Vijfde Sociolinguïstische Conferentie in Lunteren*, 68–76. Delft: Eburon.

Van Bezooijen, Renée & Charlotte Gooskens. 2007. Interlingual text comprehension: linguistic and extralinguistic determinants In Jan D. ten Thije & Ludger Zeevaert (eds.). *Receptive Multilingualism and Intercultural Communication: Linguistic Analyses, Language Policies and Didactic Concepts,* 249–264. Amsterdam: John Benjamins.

Van Bezooijen, Renée & Rob J. H. van den Berg. 2000. Hoe verstaanbaar is het Fries voor niet-Friestaligen? In Piter Boersma, Pieter Breuker & Lammert G. Jansma (eds.). *Philologia Frisica Anno 1999*, 9–26. Leeuwarden: Fryske Akademy.

Van Bezooijen, Renée & Vincent J. van Heuven. 1997. Assessment of speech synthesis. In Dafydd Gibbon, Roger Moore, & Richard Winksi (eds.). *Handbook of Standards and Resources for Spoken Language Systems,* 481–653. Berlin/New York: Mouton de Gruyter.

Van Heuven, Vincent J. 1986. Some acoustic characteristics and perceptual consequences of foreign accent in Dutch spoken by Turkish immigrant workers. In Jeanne Van Oosten & Johan P. Snapper (eds.). *Dutch Linguistics at Berkeley, Papers Presented at the Dutch Linguistics Colloquium Held at the University of California,*

*Berkeley on November 9th, 1985*, 67-84. Berkeley, CA: Dutch Studies Program, University California, Berkeley.

Van Heuven, Vincent J. 2008. Making sense of strange sounds: (Mutual) intelligibility of related language varieties. A review. *International Journal of Humanities and Arts Computing* 2(1–2). 39–62.

Van Heuven, Vincent J. & Jan W. de Vries. 1981. Begrijpelijkheid van buitenlanders: de rol van fonische versus niet-fonische factoren. *Forum der letteren* 22. 309–320.

Van Mulken, Margo & Berna Hendriks. 2015. Your language or mine? or English as a lingua franca? Comparing effectiveness in English as a lingua franca and L1–L2 interactions: Implications for corporate language policies. *Journal of Multilingual and Multicultural Development* 36(4). 404–422.

Vanhove, Jan & Raphael Berthele. 2015. The lifespan development of cognate guessing skills in an unknown related language. *International Review of Applied Linguistics in Language Teaching* 53(1). 1–38.

Vanhove, Jan & Raphael Berthele. 2017. Interactions between formal distance and participant-related variables in receptive multilingualism. *International Review of Applied Linguistics in Language Teaching* 55(1). 23–40.

Voegelin, Charles F. & Zellig S. Harris. 1951. Methods for determining intelligibility among dialects of natural languages. *Proceedings of the American Philosophical Society* 45. 322–329.

Voigt, Stefanie & Anja Schüppert. 2013. Articulation rate and syllable reduction in Spanish and Portuguese. In Charlotte Gooskens & Renée van Bezooijen (eds.). *Phonetics in Europe: Perception and Production*, 317–332. Frankfurt a.M: Peter Lang.

Wang, Hongyan. 2007. *English as a Lingua Franca: Mutual Intelligibility of Chinese, Dutch and American Speakers of English.* Utrecht: Netherlands Graduate School of Linguistics (LOT).

Wang, Hongyan & Vincent J. van Heuven. 2013. Mutual intelligibility of American, Chinese and Dutch-accented speakers of English tested by SUS and SPIN sentences. In *Proceedings of Interspeech 2013, 2 July 2013, Lyon, France*, 431–435.

Wang, Hongyan, Ligang Zhu, Xiaotong Li, & Vincent J. van Heuven. 2011. Relative importance of tone and segments for the intelligibility of Mandarin and Cantonese. In Wai S. Lee & Eric Zee (eds.). *Proceedings of the 17th International Congress of Phonetic Sciences*, 2090–2093. Hong Kong: City University of Hong Kong.

Weinreich, Uriel. 1957. Functional aspects of Indian bilingualism. *Word* 13. 203–233.

Wolff, Hans. 1959. Intelligibility and inter-ethnic attitudes. *Anthropological Linguistics* 1. 34–41.

Yang, Cathryn & Andy Castro. 2008. Representing tone in Levenshtein distance. *International Journal of Humanities and Arts Computing* 2(1–2). 205–219.