Chapter 2
Methods for measuring intelligibility

In investigations set up to measure the level of intelligibility between closely related languages, several different methods have been used to collect data. This chapter describes several of them and discusses their advantages and disadvantages. Section 2.1 discusses questions that should be considered when choosing a method for quantifying the level of intelligibility and carrying out an intelligibility investigation. At the end of Section 2.1, an overview of the considerations is presented. Section 2.2 provides an overview of methods for measuring the level of intelligibility. Each method is exemplified with one or more examples. In Section 2.3, the results of different methods of measuring intelligibility are compared. These comparative analyses provide insights into the significance of selecting a particular method.

2.1 Methodological considerations

Like any other scientific research project, an intelligibility project begins with formulating a research question (and hypotheses) to provide the project with a clear focus (see, e.g., Sunderland 2018). The research question should guide the further process. However, when setting up an intelligibility project, numerous methodological and practical considerations need to be made. These considerations may influence the whole research process, so clearly and accurately defining the research question can become an iterative process. For instance, a researcher may intend to measure the general intelligibility of two languages but discover that it is only feasible to test a limited group of listeners that is not entirely representative of all speakers of a language. This could mean that the research questions would have to be (slightly) reformulated.

Methodological considerations applicable to intelligibility testing share many similarities with those pertinent to experimental investigations in sociolinguistics, psycholinguistics, and language acquisition. The choice of a testing method depends on several factors. There may be practical limitations and hurdles to be faced during the development and administration of an experiment. A limited amount of time or funding may be available, so the time needed to develop the experiment or to process and analyze the data needs to be taken into account. Other factors that may guide the choice of testing method are more fundamental because they depend on the purpose of the investigation (see Chapter 8).

https://doi.org/10.1515/9783111134697-002

It is imperative to plan an investigation well in advance and in detail to get a realistic picture of the amount and nature of the work that needs to be done at every step. Unless the object of investigation is their own native language variety, researchers may be dependent on support from native speakers when preparing stimulus material, for translating words or texts or making recordings. These native speakers should be chosen with care to ensure that they represent the target language optimally. Other phases in the investigation may also need planning. For instance, locating a sufficient number of willing participants for the experiment might require considerable effort. Ethical approval should be given before the investigation starts. This aims to protect both the researcher and the participants in the research by ensuring they have their dignity, rights, safety, and welfare respected. In this chapter, the processing and analysis of the data are not discussed in detail. Still, this part of the investigation should also be considered in advance to make sure that the data will allow for a reliable analysis.

2.1.1 Test material

The choice of test material for an intelligibility experiment depends on the aim of the investigation. It can vary along various dimensions: mode (written or spoken), style (spontaneous or read-aloud, formal or informal, monologues or dialogues), level of analysis (isolated words, sentences, or texts), recordings (audio, video), and speaker characteristics and demography (social background, numbers, voice quality, gender, age, etc.). If the intelligibility of more languages is to be compared, these factors should be kept as consistent as possible across languages when collecting material for tests, unless they are one of the variables under investigation. We will discuss this in further detail below.

2.1.1.1 Mode

The mode is the medium of communication, and it is fundamentally divided into speech and writing. It depends on the purpose of the investigation, whether tests will be carried out in written or spoken mode. Most of this book deals with the intelligibility of spoken language, and this chapter focuses on test techniques that are developed for spoken language. However, it is important to be aware of the differences between the two modes and how they may influence each other. While the spoken mode is coded in sounds, the written mode is coded in graphic symbols. Unlike with the speech mode, there is the opportunity to revise and correct in the written form. Writing is long-lasting, while speech is ephemeral. Listeners can ask speakers to repeat, which may be simulated in a test situation by

presenting spoken stimuli twice (see Section 2.1.3.2), but other than that, they cannot check back that a word or sentence has been correctly identified. In addition, listeners usually cannot control the speed at which listening takes place, whereas readers can modify reading speed according to their needs.

Each mode may range from spontaneous (a casual conversation or a scribbled written note) to planned (a prepared talk or a formal essay). Note also that the introduction of various means of communication via social media has blurred the traditional distinctions between oral and written discourse in some contexts (Sindoni 2013). Iwasaki and Oliver (2003: 63) point out that online chat is similar to a face-to-face interaction because all participants continually take turns to exchange messages and are not given much time to review their written messages.

It is often easier to understand a closely related language in the written form than in the spoken form, at least if the writing system of the target languages is the same as that of the participant. Little research has been conducted to compare the ease of understanding written and spoken language. To compare the ease of comprehension of the two modes, the same tests and textual materials should be used. Gooskens and van Heuven (2017) included written as well as spoken versions of three tests in an investigation, using the same tests and the same textual materials in both modes in 70 combinations of closely related Germanic, Romance, and Slavic languages (see Chapter 3). This allowed them to directly compare mutual intelligibility between the two modes. The results showed that the written modality was easier to understand than the spoken modality in 65 of the language combinations. The written form of a closely related language may be easier to associate with its counterpart in the native language. Especially in the case of well-established languages with a long history of writing, the written language often represents an earlier stage of the language, where the two languages had diverged less from their original common form than in the spoken form (Gooskens and Doetjes 2009; Schüppert 2011, see also Section 4.4). For example, it may not be too difficult for Dutch listeners to recognize the written form of Danish *brød* [b̥ʁœɣˀ] as a cognate of Dutch *brood* [bʁoːt] ‘bread’, but when listening to the Danish pronunciation of the word it may no longer be possible for them to recognize the word. Among elderly participants, the advantage of printed forms may be attributed to a reduction in processing capacity for spoken language (Vanhove 2014). As mentioned above, among both younger and older adults, it may be an advantage that written text can be reread several times. In a natural situation, spoken text is usually presented only once and is always short-lived. This effect may be stronger for elderly participants.

However, the opposite can also be the case. Some closely related languages use the same alphabet but are spelled with such divergent orthographic conventions that it is difficult to read texts that are intelligible in the spoken form. For

instance, it is difficult for Dutch readers to understand written Frisian because the spelling conventions are quite different from those used in Dutch and other languages. Still, Dutch listeners often understand spoken Frisian fairly well (van Bezooijen and van den Berg 1999; van Bezooijen and van den Berg 2000; Gooskens 2007). Dissimilar orthographic and phonological systems may prevent learners from discovering cognates. Helms-Park and Dronjic (2016: 87) illustrated this with Polish-English cognates such as *egzystencja* [ɛgzɨ'st̪ɛnct͡sja] 'existence' and *menedżer* [mɛ'nɛd͡ʒɛr] 'manager' that are presumably easy to understand in the spoken form for a listener who knows English but may be challenging to recognize in the written form.

Sometimes, there is an interaction between the written and the spoken mode. When reading a text in a non-native language, readers often self-pronounce the written words. Vanhove and Berthele (2015) describe how the assumed phonetic similarity to words in their native language can help readers achieve cognate guessing success in the written mode. To illustrate, a Dutch native speaker may have this experience when reading German. The German spelling may be rather deviant from Dutch (e.g., Dutch *brood* vs. German *Brot*, 'bread'). Still, if Dutch readers read aloud the German word to themselves, they may more easily discover the correspondence between the Dutch and the German forms because the Dutch *d* at the end of words is pronounced as a /t/ and because long vowels are not always reflected in the orthographic form (the plural of Dutch *brood* is *broden* while both are pronounced with a long vowel. In a think-aloud task, Möller and Zeevaert (2015) found that native German speakers used intuitions about the pronunciation of cognates in Germanic languages (Dutch, Frisian, Danish, Norwegian, Swedish, Icelandic, Luxembourgish, Low German) when asked to recognize them in the written mode.

At the same time, knowledge about the written form may help the listener to understand spoken words in the target languages. Schüppert et al. (2022) showed that speakers of Danish can use their orthographic knowledge of Danish to decode spoken Swedish because spoken Swedish is close to written Danish, while the pronunciation of words in these two languages is sometimes quite different (see examples in Section 4.4).

When developing an experiment, the researcher must realize that the written and spoken modes may influence each other. Written instructions, questions, and answer alternatives may influence the results of a test set up to measure spoken intelligibility because the participants do not need to rely on spoken information only. For example, the list of words to be filled in the gaps during a spoken cloze test (see Section 2.2.3.6) is usually presented as a written list. This means that the written mode may have some influence on the results of the spoken intelligibility measurements. In other tests, such as the picture-pointing task (Section 2.2.1.2), the sentence verification task (Section 2.2.2.4), picture-to-story matching (Section 2.2.3.5),

and eye tracking experiments (Section 2.2.5), the written mode is less likely to influence the spoken mode.

Many experiments are set up in such a way that participants must be able to read and write to read the instructions, fill in questionnaires, and give written responses to stimuli. However, in cases where the participants cannot read and write (young children and illiterate adults), it is necessary to think of alternative ways of collecting material by recording spoken responses or developing experiments where no reading and writing is involved.

2.1.1.2 Style

If the aim of the investigation is to compare general intelligibility of languages, the material from the different target languages must be matched as closely as possible. One way to control the material is to use translations of the same words, sentences or texts in all the target languages.

It may be necessary to use read speech to control the material maximally, for instance to ensure that differences in word choice or grammatical constructions do not affect intelligibility. Different speakers are unlikely to produce the exact same sentences and texts in spontaneous speech, and as such, it would be necessary to collect a large amount of speech material to gather the same words or concepts from several speakers. When using translations, a text in one of the target languages is often translated into the other target languages. However, there is a risk that the translators may adhere too strictly to the source text when selecting words and expressions for translations, resulting in translations that do not accurately reflect the target languages. To prevent one of the target languages from being given a special status, one solution is to employ translations from a source language that is not one of the target languages or, alternatively, to use source texts from each of the target languages and have them translated into each of the other target languages. By doing so, common words and structures are more likely to be equally represented in all target languages.

It may be preferable to use spontaneous speech since, compared to read speech, this simulates a more natural, real-life situation, making it easier to generalize the results to real-world settings (ecological validity). Pronunciations in spontaneous speech will likely be closer to pronunciations in everyday communication, where speakers often reduce and assimilate sounds. Read speech will usually be closer to orthography and standard variants because speakers tend to pronounce words as they are written. However, since it is challenging to control spontaneous speech, a good compromise may be to use recordings of semi-spontaneous speech that demands speech production in a controlled setting with a predetermined subject of conversation, such as a picture description task (see example in Figure 2.1).

Figure 2.1: A drawing ("the street") used to collect semi-spontaneous speech. Speakers are asked to describe the picture. Source: van Bezooijen and van den Berg (1999: 4).

To elicit content-controlled discourse, Swerts and Collier (1992) had participants describe a network of geometrical figures differing in shape and color and connected by horizontal and vertical lines (see Figure 2.2).[4]

Bulatović, Schüppert, and Gooskens (2019) had speakers watch videos carefully and then describe the events from the beginning to the end. During the recording process, the speakers were presented with ten screenshots of the specific video to help them remember as many details as possible, and to ensure that they did not divert from the storyline too much while narrating the plot. This procedure resulted in recordings in different language varieties that were so similar in content that it was possible to formulate the same questions about the texts. Recordings of dialogues can be elicited with semi-structured games, such as the map tasks, where pairs of speakers are provided with maps, and one speaker gives directions to the other (see Section 2.2.4.2). It should be noted that even if the contents of the recordings of semi-spontaneous speech are similar, there may still be differences in how the information is structured. This can affect intelligibility, not

4 This network was originally developed by Levelt (1989) to illustrate his so-called lineralization model. It shows how speakers order information when they have to express a multi-dimensional structure.

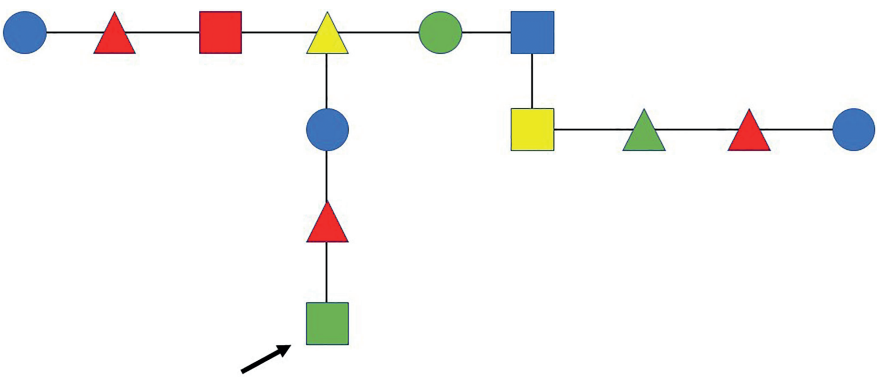

Figure 2.2: A spatial grid-like network consisting of geometrical figures differing in shape and color. Speakers are instructed to describe the network in such a way that a listener can reconstruct it based on the description. Adapted from Swerts and Collier (1992: 465).

because of word recognition or grammatical understanding but because of the variable processing costs of information.

2.1.1.3 Level of analysis

Methods for investigating intelligibility can be characterized by the level of analysis (word, sentence, or text). If the researcher is interested in an overall indication of the level of intelligibility of a whole language (higher-order intelligibility), methods at the level of entire texts are adequate. Investigations at the text level are often more realistic or natural (ecologically valid) than investigations at lower levels, such as sentences or words (lower order intelligibility), since we usually encounter more than only isolated words or sentences when we listen to a language.

The fragments for an intelligibility test are often selected to represent the language as a whole. If the sample is large enough, such as a longer text, it is often assumed to be a representative sample of the target language. However, the text must be selected with care. There are ways to control the complexity of a text, for instance, by counting or quantifying sentence length, number of syllables, word length, word frequency, etc. See Jensen (2009) for a discussion of indicators for text complexity. The topic of a text (daily life, science, society, technique, politics, etc.) should be selected in such a way that listeners will not be able to answer questions about the text based on their knowledge of the world but will need to understand the text itself. It is also important to be aware that one unintelligible word may result in lower intelligibility of the whole text if it is central to understanding the text. To illustrate, one of the texts used to test the mutual intelligibil-

ity of Scandinavian languages by Delsing and Lundin Åkesson (2005) was about frogs. The word for 'frog' is very different in Danish (*frø*), Norwegian (*frosk*) and Swedish (*groda*). Since this word was crucial for understanding the whole story, listeners who did not understand the word for 'frog' in one of the neighboring languages could not answer the open questions about the text.

However, if the investigation aims to pinpoint specific linguistic factors that play a role in explaining the level of intelligibility, methods that involve smaller entities, such as words rather than whole texts, are more suitable. To illustrate, the researcher may be interested in experimentally testing the separate contribution of vowels, consonants, prosody, cognates, or word order to the intelligibility of a language. By keeping all but one factor constant and systematically varying the characteristics of this one target factor, any difference in intelligibility must be the effect of the variations in the target factor (experimental setting). Suppose we want to test the hypothesis that Swedes have difficulties understanding Danish due to the weakening of consonants in Danish. In that case, we can substitute the weakened consonants in recordings of Danish with their Swedish unweakened equivalents. If Swedes understand the manipulated version better than the original version, the consonant weakening must be the cause of the reduced intelligibility of the Danish version.

An advantage of testing isolated words is that the influence of the context or the situational redundancy on the intelligibility of a word can be excluded. This makes it possible to draw conclusions about the role of individual word characteristics in intelligibility. It is more challenging to trace back poor intelligibility to specific sources at sentence or higher levels. If the words are presented in a sentence or whole text, the context or the situational redundancy may make up for the poor intelligibility of single words. The word predictability depends on the content of the context (see Section 2.2.2.2). For instance, the predictability of the word *car* is low in the sentence *I bought a car* but higher in *I brought the car to the garage*. However, this is complicated by the fact that the pronunciation of predictable words is often less clear than that of unpredictable words. Lieberman (1963) found that unpredictable words in sentences tend to have a longer duration, higher amplitude, and more precise articulation than highly predictable words. This may make predictable words recorded in a sentence more difficult to understand than unpredictable words if they are played in isolation to the listener.

Generally, a word test is ecologically less valid than a test involving whole sentences or texts. However, results from word-level tests tend to correlate highly with results from text-level tests (e.g., Gooskens and van Heuven 2017). This seems logical since a listener must understand individual words to understand a text and get help from the context (Field 2019: 290). Results of word intelligibility tests have also been shown to correlate highly with the general intelligibility of

the target language as perceived by listeners in a judgment task (Gooskens and van Heuven 2017). Hence, tests measuring word intelligibility provide a good indication of the overall intelligibility of a language.

Smaller fragments (words, sentences) can be selected randomly under the assumption that a random selection will represent the language as a whole. However, several word characteristics may influence their intelligibility (see Section 5.1.1), and the selection may therefore require some level of control. For instance, one can make sure that the material is phonetically balanced, that is, in accordance with the statistical distribution of phonemes in the language (see Martin, Champlin, and Perez 2000; Field 2019). The concepts need to be known to the listener in order not to test world knowledge. Words can be selected based on a frequency list. The most frequent words in such a list generally cover a substantial portion of ordinary language use. It may be best to avoid homophones since they may confuse the listener. Other word characteristics, such as neighborhood density, false friends, and word length, which are known to influence intelligibility (see Section 5.1.1), should also be controlled for.

2.1.1.4 Recordings

Most of the methods discussed in this chapter are based on listening experiments and, therefore, involve recordings of the target languages. When making recordings, the recording conditions should be as consistent as possible. It is advisable to use the same recording equipment for all recordings and ensure that there is no background noise or bad acoustics. The best recording circumstances are found in a soundproofed recording studio. It may be necessary to normalize the sound level of the recordings afterward to make sure that the level is the same for all recordings. If read speech is collected, it is a good idea to ask the speaker to read the material more than once in case of unwanted effects, such as reading mistakes, background noise, etc. Readers can be asked to repeat if they read too fast, unnaturally, or with a non-optimal intonation. Speakers tend to pronounce word lists with so-called list intonation, using a rising tone on each word except the last one, which is typically pronounced with a falling intonation. It may help to ask the speaker to pause between words or to imagine that someone asks them for each word “How do you pronounce this word?”. The words can also be pronounced in a fixed carrier sentence, such as “I now say . . . . again”. However, it may not be easy to improve the way a speaker reads. Recording several readers and selecting the best one afterward may be a better solution.

When setting up intelligibility experiments, it may be worthwhile using video recordings rather than audio-only recordings as stimulus material. Such stimuli may be more natural and realistic since visual information like gestures and

mouth and face movements may add information that makes it easier to understand the speakers (see Section 5.4.2). Innovative approaches with virtual reality (360° videos) offer the potential for an even higher degree of realism. However, opting for visual stimuli presents its own challenges, as visual information can influence the outcomes of intelligibility assessments. For instance, social information about a speaker can influence how well listeners understand the speaker. Kang and Rubin (2009), further discussed in Section 4.3, showed that visual primes in the form of photos of either a Caucasian or an East-Asian person could result in different levels of intelligibility even though the sentences were read aloud by the same American speaker.

2.1.1.5 Speakers

When selecting speakers for intelligibility experiments, it should be considered that some speakers are more intelligible than others because of differences in voice quality (e.g., hoarseness, nasality, breathy voice, creak), precision of articulation, and reading ability.

The gender of the speaker may play a role in intelligibility. In both British (Markham and Hazan 2004) and American data (Karl and Pisoni 1994; Bradlow, Torretta, and Pisoni 1996; Bent, Buchwald, and Alford 2007), there is a small but significant effect of gender. Women are slightly more intelligible than men, with a difference of 2% in the British word data and 3% in the American sentence data. On the other hand, Smorenburg and Chen (2020) mention that previous research on the role of speaker gender in verbal processing suggests that female voices require more, and thus slower, processing than male voices. However, they found no such gender effects in their own research, and neither did Tielen (1992).

It may also be desirable to control for other speaker characteristics, such as level of education, age, regional background, and socio-economic status, when selecting speakers since sociolinguistic research has shown that language use varies across such dimensions too (Krug 2013).

When comparing the intelligibility of multiple languages, it is crucial to consider the impact that a speaker's voice characteristics and background may have on their intelligibility. If the design of the experiment allows it, several representative speakers could be used per language variety. In that way, effects of variability between speakers will average out. Otherwise, a single speaker per language can be selected based on intelligibility scores as judged by native speakers of the same language. Drawing from a larger group of speakers, the speaker whose intelligibility scores are in the middle of the score range can be chosen to represent the speaker group. In a listening test where the intelligibility of only two languages is

compared, one option is to use a balanced bilingual speaker to record the stimuli (matched-guise design). The same speaker fluent in two languages is used to record both languages, and therefore, the influence of voice characteristics on the intelligibility is matched or kept constant. To ensure that the speaker sounds native-like in both languages, a voice line-up can be arranged similar to the procedure used in forensic phonetics (Broeders, Cambier-Langeveld, and Vermeulen 2002). To select a bilingual speaker for a matched-guise experiment (see Section 4.3.1), Hilton et al. (2022) conducted two voice line-ups, a Danish and a Swedish one. They presented native listeners with a number of recordings of native speakers, including one recording by the bilingual speaker. They instructed them to pick out the one speaker that did not sound native. The results showed that the bilingual speaker was not picked out more often than the distractors in any of the two languages and could, therefore, be regarded as a representative speaker of both native Danish and Swedish.

2.1.2 Data collection

2.1.2.1 Equipment

In the past, the basic equipment needed for an intelligibility investigation was pen and paper and, in the case of a listening experiment, recordings of the target languages involved in the investigation. It was time-consuming to collect data because it was necessary to print forms to be filled in and to bring the tape recorders to the listeners. After the experiment, the researcher had to spend time on data-entering. The introduction of computer-based methods and software for conducting experiments has made life easier for researchers because data can be collected digitally. This makes it easier to collect and process the data. However, researchers should carefully consider whether they are willing to spend time acquiring the necessary skills to use the software (see Section 2.2.5) or whether somebody who already has these skills can support them. Even though software tends to become more and more user-friendly, some software packages still need programming skills. Furthermore, the software may not be suitable for the experiment that the researcher has in mind. Therefore, it is important to check the limitations of the software well before deciding to use it.

The software used for conducting the experiments may have to be installed on individual computers, which makes it necessary to have the listeners do the experiment on a computer that has the software installed. However, the internet has made it possible to carry out web-based experiments. The advantage of a web-based experiment is that the listeners can simply be provided with a link to the investigation. They do not need to meet the researcher or download any software, and the responses will be returned to the researcher in a format that is easily analyzed. The

drawback is that researchers have less control of the experimental situation if they are not present during the sessions. They cannot be sure that listeners do not ask for help during the experiment or are distracted by noisy surroundings or bad equipment. Furthermore, it is not possible to immediately answer questions that the listeners may have, so the instructions must be very clear and provide all necessary information. At the same time, the instructions should not be too long since this may cause the listeners to lose interest even before starting the experiment.

A noisy or distracting test situation can influence the results, so the test location should be chosen carefully. It is a good idea to ask the listeners to listen to stimuli via good-quality headphones rather than over speakers.

2.1.2.2 Listeners

The task performance of listeners is always somewhat variable depending on many different listener characteristics, such as their level of education, intelligence, personality traits, age, gender, socio-economic status, geography, language background, and experience with the target language and other languages. Listeners may also be influenced by other factors, such as their motivation to carry out the test task and their attitude toward the target language. These factors are discussed in more detail in Chapter 4.

Depending on the situation and the nature of the test, listeners can be tested as a group or individually. Groups may be easier to convene than multiple individuals because only one appointment has to be made, such as with a class teacher or the organizer of an event. However, sometimes it is necessary to test listeners individually, which makes it more time-consuming to collect a sufficient number of tests for a reliable statistical analysis.

To ensure that differences in intelligibility between target languages cannot be attributed to unwanted factors, it is important to select a fairly large and well-defined group of listeners that should be comparable across the target groups. To illustrate, the researcher can choose to test a balanced number of male and female students between 18 and 25 years who are born and raised in specific locations and who have no prior exposure to the target language. However, even such precisely defined groups may not be completely comparable. For example, each country has its own school system, which makes it difficult to compare the educational levels in different countries.

To control for all the listener characteristics mentioned above, an intelligibility test is often accompanied by a background questionnaire that the listeners have to fill in. Here, questions are asked to provide demographic information (e.g., age, gender, places the listeners have lived, language background of the listeners and their parents, schooling, etc.). More detailed questions can, of course,

be added if necessary. The questionnaire can also include questions about experience with the target language (see Section 4.1.1) and attitude toward the target language (see Section 4.3.1). If the questions are asked using an online form, multiple-choice menus can be used to make it easier for the listener to fill in and for easier administration and processing for the researcher. The answers to the questionnaires can be used to exclude certain listeners from further analysis because they do not meet all listener criteria or to investigate whether listeners with different characteristics perform differently in the intelligibility test. For example, the researcher may be interested in comparing the performances of listeners with different genders or age groups. Kaushanskaya, Blumenfeld, and Marian (2020) developed the Language Experience and Proficiency Questionnaire (LEAP-Q), a tool for collecting self-reported data about proficiency and experience from bilingual and multilingual participants ranging from 14 to 80 years old. It is available in over 20 languages and can be administered in a digital, paper-and-pencil, and oral interview format. In Figure 2.3, a simple example of a questionnaire is provided.

1. What is your gender?
   - ☐ Male
   - ☐ Female
   - ☐ Other/prefer not to tell
2. How old are you?_____ years
3. What is your highest completed level of education?
   - ☐ None
   - ☐ Elementary school
   - ☐ High school
   - ☐ Bachelor's degree
   - ☐ Master's degree
   - ☐ PhD
4. What is your profession?_____
5. What is your place of birth (city and country)? _____
6. Where did you grow up (city and country)? _____
7. What is your native language or dialect? _____
8. Do you speak other languages at home besides your native language?
   - ☐ No
   - ☐ Yes, namely: _____
9. Which languages have you learned (e.g., in school)? _____

Figure 2.3: Example of background questionnaire.

A frequent stumbling block in investigations involving listeners is the availability and willingness to do the test among listeners that fulfill the criteria formulated by the researcher. Many researchers will recognize the situation where they have

thought of an interesting setup of an experiment, but when they start to look for listeners, they discover that it is difficult to recruit enough listeners for their design. The number of listeners fulfilling the criteria may be small, or the potential listeners may not be motivated to participate. Researchers have to think of ways to get into contact with their target group and ways to make it attractive to participate. A useful point of departure is often to use one's own network and the networks of friends, family, and colleagues. If the experiment is web-based, it is possible to use social media to spread the link to the experiment and maybe benefit from a snowball effect by asking people to spread the link. It may help to offer something in return, such as a small gift or money, or to write an engaging text that makes it clear to the listeners why participating is worthwhile, for instance to help advance science or to test their language knowledge. Recently, it has become popular to present the test as a kind of game that is fun and entertaining and often contains some competitive element (gamification, see Leemann, Derungs, and Elspaß 2019).

When using web-based experiments, the researcher is in less control of who may get the link to the experiment. It may be necessary to remove listeners from the data set because the listeners did not fulfil the criteria for participation or did not take the experiment seriously. It may be challenging to decide whether the listeners made an effort to do the test as well as possible. However, this is sometimes clear, particularly if their score is very deviant from the rest of the group. Brysbaert et al. (2014) ensured that student participants took their task seriously and would not generate numbers at random by telling them in advance that they would only be paid if their results correlated positively with those of the other listeners.

Finally, it is a good idea to check whether listeners have any hearing problems in the case of a listening experiment or bad eyesight in the case of experiments that involve reading or looking at pictures. Also, note that illiteracy can be a problem. In developing countries, there are usually large segments of the population that are illiterate, but also in Western countries a large percentage of the adults are illiterate or have difficulties reading. As an illustration, around 15% of Dutch adults currently are functionally illiterate and may not be able to carry out a test involving reading and writing (Stichting Lezen en Schrijven 2021).

2.1.3 Designing the study

2.1.3.1 Qualitative or quantitative

When setting up an investigation, a major consideration is whether it is necessary to express the results in numbers and use a quantitative approach. This may be the case if the aim is to investigate how well one language is understood compared to another or to carry out statistical tests to show the relationship between

intelligibility and various linguistic and extra-linguistic factors (see Chapters 4 and 5). An alternative to such a quantitative approach is a more qualitative approach where the main interest may be investigating people's strategies when trying to make sense of a closely related language (see Section 4.6).

Both approaches come with their own challenges. Test situations are often somewhat artificial, and the quantitative results from intelligibility tests may not reflect actual understanding optimally. In real life, mutual understanding depends on interactive cooperation, and in a natural context, the number of possible interpretations of an utterance is often reduced. Listeners are often good at achieving pragmatic communicative goals, even if they only understand a little of what is said. Some of the experimental methods for testing intelligibility resemble real-life situations more than others. Still, they may reveal less information to the researcher about the factors that may determine the level of intelligibility.

On the other hand, it can be challenging to determine if listeners genuinely comprehend the target language as they are often proficient at concealing their misunderstandings and adapting their speech to their conversation partner. This may be a potential disadvantage of the qualitative approach unless the investigation aims to study how negotiation of meaning and accommodation works towards successful communication. Moreover, qualitative approaches typically require significant effort and labor as they involve a meticulous examination of conversations, making them time-consuming.

In the rest of this chapter, the main focus is on experimental methods for testing intelligibility that will result in quantitative results. However, even in quantitative investigations, it may be informative to ask the listeners more qualitative questions about their experience with the languages and the particular characteristics of languages that make them difficult to understand in addition to the quantitative measurements. This information may be helpful when interpreting the quantitative results.

2.1.3.2 Ceiling/floor effects

When designing an experiment, it is crucial to consider the constraints of the assigned task. If the task is too simple or challenging for a particular audience, it can impede or prohibit the interpretation of the findings. Also, the listeners may get bored if the test is too easy or frustrated if it is too difficult and, therefore, decide not to finish the whole test. Thus, it is vital to avoid such circumstances.

Some investigations may involve target languages that are very similar to the listener languages. In such cases, the task may be too easy, and (almost) all listeners may get a score close to the maximum. This will result in a ceiling effect, i.e. a situation where a measurement cannot take on a value higher than some limit or "ceil-

ing". It may then be necessary to use more sensitive test procedures. In addition to choosing a more difficult task or text, there are several other ways to avoid ceiling effects. The intelligibility of a spoken text can be reduced by artificially speeding up the spoken text (Janse, Nooteboom, and Quené 2003; Syrdal et al. 2012), adding noise (Gooskens, van Heuven, van Bezooijen, and Pacilly 2010; Lecumberri, Cooke, and Cutler 2010), applying filtering (Wang et al. 2011), or signal compression as used in GSM telephony (Nooteboom and Doodeman 1984). The task can also be made more difficult by putting the listeners under time pressure, either by requesting them to complete the task as quickly as possible or allowing them only a limited amount of time to respond. Another way to minimize a ceiling effect is to measure reaction time. The underlying assumption is that the quicker the listeners respond, the higher the level of intelligibility. Reaction times give a precise and sensitive measurement of processing costs, and even if the listeners answer all questions correctly, there may still be a difference in the time it took the listeners to understand the various stimuli correctly. However, if a task is really easy, listeners will react with a minimum reaction time. In such a situation, reaction time measurements will not solve the ceiling effect problem.

On the other hand, if the task is too difficult, the percentage of correct answers may be so low that it is also difficult to interpret the results (floor effect). To check that the task itself is not too difficult and to test the level of difficulty in advance, a reference condition can be built into the experiment, with a control group of native speakers listening to their own language. If they cannot perform the task in their own language, the task is obviously too difficult. If the memory limitations of listeners are not considered, the task may become excessively challenging. Hence, it is advisable to avoid complex tasks or lengthy sentences that could overburden the memory capacity of the listeners. However, it should be kept in mind that even in optimal situations, native listeners tend to make mistakes. To make the task easier for the listeners, it may help to have a recording played twice (double play). It may be argued that this does not reflect a real-world situation, where a listener is usually not able to hear spoken input more than once. On the other hand, the possibility of listening twice may reduce anxiety among listeners. In addition, it simulates situations where listeners ask speakers to repeat what they said. Field (2019: 306) discusses the pros and cons and the effects of double-play in detail.

2.1.3.3 Priming effects

When the same stimulus is presented more than once, this may result in a priming effect, an effect that occurs when listeners' exposure to a certain stimulus influences their response to a subsequent stimulus, which has something in common (form or meaning) with the previous stimulus. To illustrate, if English listeners first

hear the Danish word *træ* 'tree' and later hear the Swedish equivalent *träd,* there may be a greater chance that they will translate the Swedish word correctly than if they had not heard the Danish word first. Therefore, in an experimental setting (see Section 2.1.1.3), the same (or similar) stimuli should not be presented to the same listener more than once. This contradicts the fact that it is desirable to use identical stimulus material when establishing the listener's understanding of more than one language (see Section 2.1.1).

A solution is to use a Latin square design whereby each listener is exposed to an equal proportion of the stimuli in each target language and never hears the same stimuli (for example a word, a sentence or a text) in more than one version (Godfroid 2020). An example of a Latin square design is presented in Table 2.1. In this simple example, the test material consists of four different stimuli (1 to 4) in four languages (A, B, C, and D). This results in four different test versions (I to IV). Only one of the four versions is presented to each group of listeners. As an illustration, listeners who are tested with version II hear stimulus 2 in language A, stimulus 3 in language B, stimulus 4 in language C, and stimulus 1 in language D. It takes some effort to administer a Latin square design because different listener groups will be presented different stimuli. Many listeners are needed to fill all the cells in such a design with enough listeners for a sufficiently powerful statistical analysis. Four groups of listeners are required for the example in Table 2.1, one for each version.

From Table 2.1 it becomes clear that the different stimuli 1, 2, 3 and 4 are presented at different places in the experiment. Stimulus 2 follows after stimulus 1 in three of the versions. The place and order of the languages is always the same, starting with language A and ending with language D. This may have some influence on the results because of learning effects and fatigue among the listeners. A way to counterbalance this effect is to have eight different versions where versions 5 to 8 are presented in the mirrored order of version 1 to 4. The experiment can also be programmed so that the listeners are presented with a random selection of stimuli in a random order, but without the same words or texts presented twice.

Table 2.1: Latin square design with languages A–D, stimuli 1–4 and test versions I–IV.

Language	Test version			
	I	II	III	IV
A	stimulus 1	stimulus 2	stimulus 3	stimulus 4
B	stimulus 2	stimulus 3	stimulus 4	stimulus 1
C	stimulus 3	stimulus 4	stimulus 1	stimulus 2
D	stimulus 4	stimulus 1	stimulus 2	stimulus 3

2.1.4 Summary

To summarize the methodological considerations discussed in this section, a list of questions researchers may ask themselves before setting up an investigation is presented in Figure 2.4. It is advisable to keep track of all the decisions made in a log book since it is easy to forget why specific choices were made during the development of an experiment. It is also advisable to draft an initial version of the methods section for the paper intended for publication of the results. While writing this section, errors in method and design will often become apparent, and a peer reader may uncover blind spots. A final piece of advice is to carry out a pilot experiment before starting with the actual experiment.

- What is your **research question**? Formulate research question and hypotheses (expected answers to the question) as precisely as possible. (2.1)
- What is the **time frame** and is it realistic to carry out your project within this time frame? Make a timeline. (2.1)
- What are the **costs** for carrying out your project and do you have the necessary funding? Make a budget. (2.1)
- Do you need help and **assistance** from others for setting up the experiment, carrying out the experiment and analysing the data? Are these people available to help you when you need them? (2.1)
- Do you have access to **language expertise** to help you prepare your test material, e.g., making translations? (2.1)
- Do you have the necessary **equipment** (e.g., headsets and computers) and **software**? Do you have the necessary **skills** to use this software? (2.1.2.1, 2.2.5)
- What **method** will you use to test intelligibility? (2.2)
- Do you have **ethical approval?** (2.1)
- What **test material** will you use? (2.1.1)
  - Will you test with pen-and-paper experiments or via the computer? And if you develop a computer-based experiment, will it be stand-alone or web-based? (2.1.2.1)
  - Spoken or written? (2.1.1.1)
  - Spontaneous, semi-spontaneous or read? (2.1.1.2)
  - Words, sentences or texts? (2.1.1.3)
  - Random selection or based on certain characteristics? (2.1.1.3)
  - Topic? (2.1.1.3)
  - Audio, video, or both? (2.1.1.4)
  - Style and complexity? (2.1.1.2)
  - Monologues or dialogue? (2.1.1.2)
  - Is the method appropriate for your target group? (2.1.1.1, 2.1.2.2)
  - Who will be the speakers (gender, age, language background, voice characteristics?) (2.1.1.5)
  - What recording equipment will you use? (2.1.2.1)
  - Where will the recordings take place? (2.1.2.1)

Figure 2.4: List of questions that researchers may ask themselves before setting up an experiment.

- Did you take into account **methodological considerations**? (2.1)
  - Is there a ceiling or floor effect? (2.1.3.2)
  - Is there a priming effect? (2.1.3.3)
  - What does your design look like? (2.1.3.3)
- Who will be the **listeners** (2.1.2.2)?
  - Age, gender, social and geographic background, language background, etc. (2.1.2.2)?
  - How many (2.1.2.2)?
  - How will you approach them and motivate them to participate? (2.1.2.2)
- What **testing procedure** will be followed?
  - Will the listeners be tested individually or in groups? (2.1.2.2)
  - Is there a suitable testing location (quiet, reachable, no distractions)? (2.1.2.1)
  - Will listeners be paid? (2.1.2.2)
  - Are there clear instructions (2.1.1.1, 2.1.2.1)?
- How will you analyse the data statistically? How should you organise the data to do the **statistical testing** necessary to answer your research question (2.1.2.2, 2.1.3.1, 2.1.3.3)?

Figure 2.4 (continued)

2.2 Overview of methods

Listening to a closely related language is similar to other instances of nonoptimal speech input, such as listening to talking computers, foreign accents, individuals with a speech disorder, or native speech in noisy surroundings. Native listeners are generally successful in getting the speaker's intentions in such situations where the input speech is nonoptimal. It can be assumed that similar mechanisms are generally involved in decoding all these kinds of speech. As a result, methods for investigating mutual intelligibility between closely related languages can be adopted from other disciplines, such as speech technology, second-language acquisition, and speech pathology.

The rest of this chapter provides an overview of various common and less common methods for investigating intelligibility, discusses their advantages, disadvantages, and difficulties, and illustrates them with examples. Note that each method has different variants and that the overview of methods is not exhaustive. Each subsection deals with a particular level of analysis (word, sentence, text, and discourse), but some methods can be used for testing at various levels. The focus is on quantitative methods and functional tests, which test how well a listener *actually* understands the target language. However, opinion testing and a few qualitative approaches are also discussed. In addition, the chapter mainly concentrates on techniques developed for testing intelligibility of spoken language. Most techniques can easily be adapted to the testing of written language.

When choosing a method for testing intelligibility, it is essential to realize that each method may produce a different score. It is, therefore, possible to determine differences in intelligibility between different languages with one specific test method, but the absolute level of the scores will differ if the same languages are tested with different (more difficult or easier) methods or tasks (see Section 2.3).

2.2.1 Word level

First, methods that are mostly used to investigate intelligibility at the word level are discussed. As discussed in Section 2.1.1.3, word-level testing is generally less similar to naturalistic situations than testing at higher levels but will often more easily allow the researcher to pinpoint linguistic factors that can explain the intelligibility results. If listeners are tested with whole texts, it is difficult to trace the linguistic characteristics that make the text difficult to understand. Listeners can use the context to reach an overall understanding of the text even though there are individual words that they do not understand. If words (or sentences) are used, it is easier to trace difficulties back to particular characteristics of the test material.

2.2.1.1 Word translation task

The most widely used word intelligibility test is perhaps the word translation task. Listeners hear isolated words in the target language and write down or pronounce translations in their own language that capture the same meaning. The percentage of correctly translated words is a measure of overall word intelligibility.

If the test is administered digitally, the number of correct translations can be counted automatically through a pattern match with expected answers. However, it is generally necessary to check the incorrect translations manually. This may not be a straightforward task since words may have several meanings when they are presented out of context. For example, the Swedish word *brist* 'lack' can be translated into Danish *brist* or *mangel*, both meaning 'lack'. Both translations should be counted as correct since listeners have no way to know which of the two synonyms is the correct translation. In the case of homonyms, all possible translations should also be accepted as correct. An example is the Swedish homonym *här*, which can be translated correctly into Danish *hær* 'army' or *her* 'here'. Listeners often make spelling errors. These should be objectively defined, for example as instances where the listeners only misspelled one letter without this resulting in another existing word. In this definition, the mistake in Danish *ærende* (correct *ærinde* 'errand') is considered a spelling mistake. Therefore, *ærende* is

counted as correct (only one wrong letter without resulting in another existing word), whereas *aske* (correct *æske* 'box') is not counted as correct because the mistake results in an existing word meaning 'ash'.

It is often not completely clear whether a word is correctly translated or not and it is left to the researcher to decide. Semantic overlap may pose a challenge for the researcher to objectively determine if translations should be classified as accurate or inaccurate. For instance, a Danish listener may translate Swedish *piga* 'maid' into the Danish cognate *pige* 'girl,' which has only a partly overlapping meaning (a young female person). There may also be a morphological overlap, such as if a singular noun is translated into the plural. In some cases, a solution may be to give partial points (for example, half a point) to translations that show overlap with the correct translations.

Because of difficulties with objectively correcting translations, an alternative solution could be to have the listeners select the correct translation from a closed list of alternatives (multiple-choice). Creating such a test can be challenging as the complexity of the test is heavily influenced by the range of alternatives presented. For instance, it will be more difficult to choose the correct translation if there are many words to choose from or if the list contains words that are very similar to the correct answer. These are so-called neighbors or false friends (i.e., word forms that closely resemble the stimulus word but have a different meaning, see Section 5.1.1). Furthermore, if the investigation aims to compare the intelligibility of more languages it is often difficult to select equivalent distractors (non-correct answers) for all languages involved. As a result the results may not be comparable.

Example 2.1: Word translation task Danish/Swedish

Kürschner, Gooskens, and van Bezooijen (2008) tested the intelligibility of 384 frequent Swedish words among Danish listeners via the internet. The translations were automatically categorized as right or wrong by the computer through a match with expected answers. The answers that were categorized as wrong were subsequently checked manually by a Danish mother-tongue speaker. Responses that deviated from the expected responses because of a mere spelling error were counted as correct identifications. The word intelligibility results were correlated with eleven linguistic variables that had been shown to contribute to intelligibility in earlier studies (see Chapter 5). The strongest correlation was found between word intelligibility and phonetic similarity.

The errors that listeners make when they translate words may often be mere spelling or typing errors. However, other kinds of errors may be a rich source of knowledge for the researcher who is interested in understanding how closely related languages are processed. The mistakes that listeners make tell something about how they try to match words in the target language with the corresponding cognates in their native language (see e.g., Gooskens and van Bezooijen 2013; Härmävaara and Gooskens 2019). To gain an even better understanding of this process, the listeners can be

asked to reflect on their decisions by thinking aloud while translating (see also Section 2.2.3.4).

Example 2.2: Word translation task European languages

To test word intelligibility in 70 combinations of closely related languages involving 16 languages in Europe, a list of the 100 most frequently used nouns in the British National Corpus (BNC Consortium 2007) was compiled (Gooskens and van Heuven 2017, see Chapter 3). The target words were translated into the other target languages and recorded by four native standard speakers per language. Each speaker contributed a different quarter (i.e., 25 words) of the stimulus words. To ensure that the test did not exceed a reasonable duration, each listener was presented with a randomized subset of 50 words from the larger set of 100 words. Listeners were instructed to translate each stimulus word into their native language using the computer keyboard. The results were correlated with those of a (sentence level) cloze test (r = .73), see Section 3.2, and with judged and perceived intelligibility (r = .72 and .78), see Section 2.2.3.1.

2.2.1.2 Picture pointing task

Sometimes, it is not possible to have listeners write down their answers or choose from a list of words because they cannot read and write, as is the case with young children, oral cultures, and illiterate adults. In such cases, a special version of the multiple-choice test, the picture-pointing task, may be a good alternative (Gooskens and Schneider 2016).

Listeners hear recorded words in the target languages, and for each test word, they are shown a card with four (or more) pictures, one of which depicts the test word. They have to point to the picture that corresponds to the word that they have heard. The responses can easily be noted down on the spot, even by researchers who do not speak the target languages. This makes the test suitable for use in communities without a strong tradition of literacy and where the researcher is interested in the mutual intelligibility of lesser-known languages.

Alternatively, a touchscreen on a computer or tablet can be used to register the responses. This saves time because no data entry needs to be done afterwards and because it allows for automatic data analysis. It also makes it possible to measure reaction times. The time taken to make a decision can be viewed as a reflection of processing complexity. The level of intelligibility can then be expressed as the percentage of correctly identified words as well as the time it takes for the listener to make the decision. Gooskens and Schneider (2016) provided a checklist of steps to take when developing the material for a picture-pointing task. Figure 2.5 shows one of the picture-pointing cards.[5]

5 The full set of cards is available via http://www.let.rug.nl/gooskens/picture_cards_gooskens_and_schneider.pdf

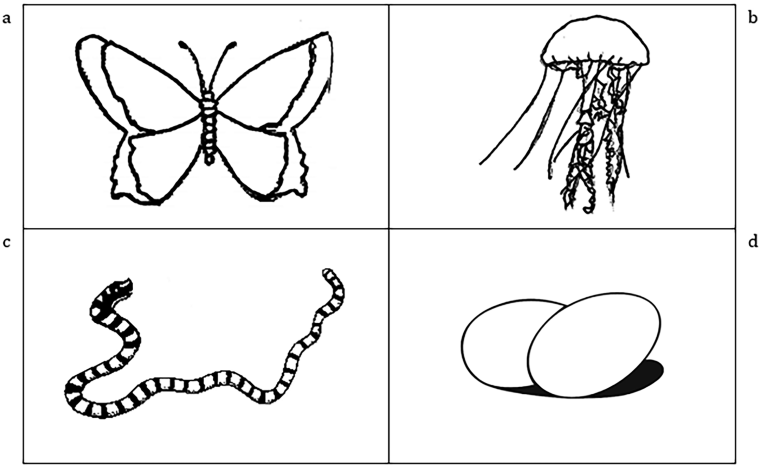

Figure 2.5: A card used for a picture-pointing task with nouns as target word and distractors ('egg', 'butterfly', 'jellyfish', and 'snake'). Source: Gooskens and Schneider (2016: 286).

The words that can be chosen for the test are limited because they need to be 'picturable'. This is difficult for abstract concepts, and nouns are generally easier to draw than other word classes. However, verbs have also been successfully pictured, such as in the Peabody Picture Vocabulary Test (Dunn and Dunn 2007). For her research on the mutual intelligibility between five languages spoken in the Shefa province of Vanuatu, Severin (in progress) created pictures using artificial intelligence. This made it possible to use a uniform style without irrelevant details; see example in Figure 2.6. It may also be an option to picture verbs with short video recordings showing an action.

When choosing pictures to represent the target words, it is necessary to test in advance that they are recognizable to the target group. In some oral societies, there may be a certain lack of picture literacy. It may be necessary to set up a pre-test to check labeling consistency with participants from the target group. They can be asked to label the depicted objects as spontaneously as possible and pictures that are not consistently labeled should be excluded from the experiment. In particular, some target groups might find it challenging to interpret line drawings. It may be better to use photos. On the other hand, photos make it more difficult to abstract from irrelevant details. Brodeur, Guérard, and Bouras (2014) provided a databank of standardized images to use for scientific purposes.[6]

6 Other resources can be found at the internet, see e.g., http://pics.stir.ac.uk/.

Figure 2.6: A card used for a picture pointing task with verbs as target words and distractors ('draw', 'inject', 'call', and 'paint'). Source: Severin (personal communication).

It is equally important to select the distractors with care. The distractors should be just as recognizable as the target pictures. Pictures showing words with high neighborhood density, ambiguous meanings, homonyms, or similar meanings as target words must be avoided. It might be necessary to check this with a native speaker. Both the target and distractor pictures should be presented in the same uniform style and size, preferably drawn by the same artist, so as not to draw more attention to some pictures than others.

Picture pointing tasks are enjoyable and easy to explain, and so far, tests have shown that listeners feel confident when taking part because it does not feel like a real test but rather like a game and because they do not have to produce any spoken or written answer. This makes the test suitable for use with oral communities and for testing children and illiterate adults. The results of the picture-pointing task reflect general intelligibility well (Gooskens and Schneider 2016). However, developing the test is rather time-consuming because of the necessity of preselecting test words, target pictures, and distractor pictures.

Example 2.3: Picture pointing task

Schüppert and Gooskens (2011, 2012) tested the mutual intelligibility of Danish and Swedish preschoolers and adults using a picture-pointing task. The stimulus material consisted of nouns chosen based on high frequency, cognateness, and "picturability". The pictures were pre-tested for labeling consistency among 4-year-old Swedish and Danish children. The scores of the children exhibited a near-normal distribution, with mean scores of 63% for the Danish and 65% for the Swedish listeners. However, the scores of the adults suffered from a ceiling effect (around 90% correct). The ceiling effect stresses the importance of pre-testing the material and recording reaction times in addition to accuracy.

Pinto and Zuckerman (2019) note that the explicit presence of alternatives in the picture-pointing task is ecologically invalid, since, under normal circumstances, listeners do not make a conscious choice between several alternatives when hearing a word in another language. They present an alternative, the Coloring Book. This is a web application where listeners are asked to color items on a coloring page by simply touching them on a touch-screen tablet. An example is provided in Figure 2.7. The subject colors the item in the coloring plate that they thinks corresponds to the word in the prompt. Each coloring plate contains several items (for vocabulary assessment up to 10–12 different items) in a natural context (a classroom, a farm, a birthday party, etc.). The application can also be used to test sentence comprehension.

2.2.1.3 Lexical decision task

Another method that can be used to circumvent the challenges associated with correcting translations and the ceiling effects is a lexical decision task where the listeners are played a word-like sequence of sounds in the target language. They are required to determine, as fast as possible, without making any errors, if the sequence is a word or a non-word in that language. The assumption is that the faster the listeners respond with the correct decision, the easier it is to process the word. Both the existing words and the non-words should be carefully selected. The non-words should be formed by sound sequences that adhere to the rules and restrictions concerning how syllables can be created in the language. In addition, they should comprise an identical number of morphemes as the existing words. Non-words can be generated by replacing certain sounds in existing words or by adding existing morphemes to phonotactically possible non-words. For instance, the long /ee/ in Dutch *komeet* 'comet' can be replaced with a long /oo/ to create the non-existing *komoot,* or the existing morpheme *-ig* can be added to the non-word

Figure 2.7: Example illustrating the Coloring Book method. The listeners hear a prompt containing coloring instructions, in this case *The balloons are red*. Source: Pinto and Zuckerman (2020).

fant to create *fantig.* The non-occurrence should be checked in dictionaries or with a search engine.

Example 2.4: Lexical decision task

Impe, Geeraerts, and Speelman (2008) presented 200 existing and as many non-existing words recorded in ten different varieties of Dutch to listeners from the Netherlands and the Dutch speaking part of Belgium. They were asked to decide as quickly as possible whether the items were existing Dutch words or not. The existing words were subdivided according to nationality-based typicality (typically Netherlandic or typically Belgian), frequency, and word class. In addition to the lexical decision task, the listeners were also asked to decide which of two possible alternatives best reflected the meaning of the stimulus words, one option being a synonym or semantically strongly related word and the other being a semantically unrelated word. In their analysis of the reaction times, they only took into account the stimuli that the listeners categorized correctly in the lexical decision task and identified correctly in the multiple-choice task. The results showed that Belgian listeners had significantly fewer problems understanding Netherlandic Dutch than vice versa. Furthermore, they found a clearly positive effect of the degree of standardness on the ease of word recognition.

2.2.1.4 Semantic classification task

Another method that can involve reaction time measurements is a semantic classification task where listeners hear words and, for each word, have to decide, as fast as

possible while avoiding errors, to which of several pre-given categories the words belong. For instance, in an investigation on the mutual intelligibility of 15 Chinese dialects, Tang and van Heuven (2008) had listeners classify the dialect words that they heard as one of ten semantic categories by ticking one of ten boxes: body parts, plants (sweet fruits and nuts), plants (vegetables), animals (four-legged), animals (other), textiles/fabrics/articles of clothing, orientation in time/space, natural phenomena, perishables (food/drinks other than fruit and vegetables), and verbs of action/things people do. If, for example, the listeners heard the word for 'apple' they should categorize it as a member of the category 'fruit'.

The underlying assumption is that correct categorization can only be achieved if listeners correctly recognize the target word. The percentage of correctly categorized words is a measure of word intelligibility. Reaction time can serve as an indication of processing difficulty. The choice between categories can be binary (e.g., tangible or intangible, animate or inanimate) or multivalued. However, if the list of choices becomes too long, it may result in noise in reaction times because listeners spend time searching for the correct category in the list. On the other hand, the role of guessing is, of course, smaller the more categories the listeners have to choose from. The test affords fast and economical testing of the recognition of a large number of isolated words. Like in the picture-pointing task (see Section 2.2.1.2), responses can be collected and scored without any interpretation by the researcher. This can also be done automatically if a computer-based experiment is used.

Example 2.5: Semantic classification task

To test the mutual intelligibility of Maltese, Libyan Arabic, and Tunisian Arabic, Čéplö et al. (2016) used a classification task with 11 categories: animals, body parts, clothing/jewelry, colors/shapes/properties/, eating/drinking, emotions, family/other people, in the house, orientation in space, time, and world around us. The listeners select the semantic category by tapping one of 11 icons representing that category as both text and as a simple black-and-white image on a touchscreen (see Figure 2.8). The results showed that Tunisian Arabic has the highest level of mutual intelligibility with the other two varieties. Additionally, an asymmetric mutual intelligibility was found, with speakers of Tunisian and Libyan Arabic better able to understand Maltese (40% correct answers), than the other way around (about 30 % for either variety of Arabic).

Figure 2.8: Example of a classification task with 11 categories. From Čéplö et al. (2016).

2.2.2 Sentence level

The sentence level is the level between words and whole texts. The advantage of testing at the sentence level compared to the word level is that it is more natural and includes the syntactic and morphological levels. The advantage of testing at sentence level rather than at text level is that it has fewer memory limitations. The short-term memory has limited capacity, and a listener can only remember short sentences.

2.2.2.1 Full-sentence translation

Full-sentence translation was first used to assess the interlingual intelligibility of native American languages (e.g., Voegelin and Harris 1951; Hickerson, Turner, and Hickerton 1952; Pierce 1952; Biggs 1957). Listeners hear a (recorded) spoken sentence and write down (or say aloud) the translation into their native language.

If the sentences are longer than a few words, it may be necessary to repeat the sentence to reduce memory load (see Section 2.1.3.2). This can be compared to a natural situation where the listener asks a speaker to repeat a phrase or sentence. The intelligibility score is based on the percentage of words in an utterance translated correctly by the listeners (Derwing and Munro 1997).

The scoring of the responses poses problems similar to those of the word translation tasks discussed above (Section 2.2.1.1). It is labor intensive and often challenging for the researcher to decide whether a sentence is correctly translated. However, unlike the word translation task, it is difficult to have the translations corrected automatically. Syntactic, idiomatic, and morphological differences between target language and listener language make it more complicated to determine how large a proportion of the sentence is correctly translated. Content words (nouns, verbs, adjectives, and adverbs) are likely to be more fundamental for the understanding of an utterance than function words (e.g., auxiliary verbs, prepositions, articles, conjunctions, and pronouns) that add little meaning beyond defining the relationship between words (van Bezooijen and Gooskens 2007). As such, it may be reasonable to assign different weights to different word categories, but it is not clear how these differences should be weighted. Like in the case of word translations, the translations can be supplemented with a think-aloud task to gain more information about the translation process (see Section 2.2.3.4).

Example 2.6: Full sentence translation task

Gooskens et al. (2010, see Section 5.4.1) tested the mutual intelligibility between Swedish and Danish with spontaneously produced sentences. Intelligibility was expressed as the percentage of content words correctly translated, disregarding any errors in the recognition of function words. The asymmetry traditionally claimed between Swedish and Danish was indeed found, even when differences in familiarity with the nonnative language were controlled for. In a separate task, the same listeners listened to the recordings in their own native language, presented in descending levels of noise. The results show that Danish is as intelligible to Danish listeners as Swedish is to Swedish listeners, thereby rejecting the hypothesis that Danish is an intrinsically more difficult language than Swedish.

2.2.2.2 Partial-sentence translation

In various types of translation tasks the listeners are asked to translate only part of a sentence. An example is the Speech In Noise (SPIN) test (Kalikow, Stevens, and Elliott 1977) that was developed for testing English speech intelligibility in noise but has also been used to measure the intelligibility between closely related languages (e.g., Tang and van Heuven 2008, 2009, 2015). Results by Wang (2007) showed that the SPIN test is highly sensitive to differences in intelligibility of foreign-accented English as a result of different language backgrounds of both speakers and listeners.

The task requires listeners to write down the translation of the final word they hear in a sentence presented in the target language. The predictability of the final word may vary, and may or may not be predictable from the earlier part of the sentence. An example of a sentence with highly predictability is *He wore his broken arm in a sling* (target underlined), and an example of a sentence where the target word is less predictable from the context is *We could have discussed the sling*. Wang (2007) showed that the highly predictable sentences in the SPIN test were more sensitive to differences between speaker and listener groups with varying levels of proficiency in English.

An advantage of this type of test is that it is easier for the researcher to analyze the data than full sentence translations.

Example 2.7: SPIN test

To investigate the mutual intelligibility of 15 Chinese dialects, Tang and van Heuven (2009) selected 70 sentences based on the high-predictability section in the SPIN test sentence lists and translated them into the 15 dialects. The 70 sentences were selected based on their applicability to the Chinese linguistic/cultural situation. In addition, only sentences that maintained the structure of the SPIN sentences were selected, such that each Mandarin sentence ended with a final noun as it does in English. The SPIN sentences were first translated into Standard Mandarin by the first author and next from Standard Mandarin into the 15 dialects by the designated speakers of each dialect. The results correlated significantly with opinion scores (Section 2.2.3.1) and objective lexical and phonetic distance scores (Sections 5.1 and 5.2).

A related task involves presenting words in a context with part of the message replaced with blanks for selected words only. It is easy for the researcher to process the data, but the disadvantage of this method is that it is uncertain what role the written context plays in the translation results (see Section 2.1.1.1).

Example 2.8: Partial translation task

Van Bezooijen and van den Berg (1999) played semi-spontaneous samples of various Dutch varieties to groups of listeners from the Netherlands and Belgium. The texts were printed in Standard Dutch but dotted lines of uniform length replaced the nouns. The participants were instructed to listen to the recordings and write the missing words on the lines. The results showed that differences in intelligibility between the varieties could be attributed to linguistic differences between the varieties and standard Dutch.

2.2.2.3 Translation of Semantically Unpredictable Sentences (SUS)

If researchers want to exclude the confounding influence of semantic or syntactic contextual cues on the intelligibility of words in a sentence, they may consider using Semantically Unpredictable Sentences (SUS) to determine word-level intelligibility. Such nonsense sentences are sequences of words with normal word order and prosody. They do not provide any clues for the listeners to predict the

identity of content words based on sentence semantics or situational context. As an illustration, the syntactic structures are correct in semantically anomalous sentences such as *He drank the wall*, *The state sang by the long week*, or *Why does the range watch the fine rest?* Intelligibility can be expressed as the percentage of correctly translated words in the sentence or be limited to the percentage of correctly translated content words since they are more central to understanding a sentence than function words (see Section 2.2.2.1). However, the simplest and fastest way to score results is to only count the sentences that are entirely correctly translated as correct. This easy-to-obtain scoring approach is strongly related to measures of word intelligibility (Benoît 1990).

The method is commonly used in speech and language pathology settings to investigate the clarity of disordered speech and in testing text-to-speech synthesis intelligibility, but it has also been used to test mutual intelligibility between closely related languages. It is useful for testing intelligibility when the researcher is interested in testing the role of the lexicon and pronunciation but wants to exclude the influence of semantic contextual cues. Listeners receive cues about the syntactic category and prosody only, but they cannot use semantic contextual cues to make additional predictions about word identity. A disadvantage of the method is that the task is unnatural because of the semantically anomalous sentences.

Benoît, Grice, and Hazan (1996) developed a so-called SUS generator to generate Semantically Unpredictable Sentences for several European languages. These sentences consist of common syntactic structures and words randomly selected from lexicons containing commonly used "mini-syllabic" words (i.e., the smallest words within a given category). The syntactic structures are simple, and the sentence length does not exceed seven or eight words to avoid saturation of the listeners' short-term memory.[7]

Example 2.9: Translation task with SUS sentences

Gooskens et al. (2010) used SUS sentences to assess the (asymmetrical) mutual intelligibility of Danish and Swedish as well as the intrinsic intelligibility of the two languages. They used the SUS generator developed by Benoît, Grice, and Hazan (1996) to generate the Swedish SUS sentences. Since no Danish SUS generator was available, they developed one themselves. To counterbalance possible language-specific influences, such as differences in word frequency, half of the 12 SUS sentences originated from the Swedish SUS generator, and the other half from the Danish SUS generator. The Swedish sentences were then translated into Danish, and the Danish sentences into Swedish. The results showed the same overall pattern as translations of spontaneous sentences (see Example 2.6), confirming the Danish-Swedish asymmetric intelligibility.

7 Newer SUS generators are available online (e.g., https://github.com/ecooper7/SUSgen; Lilley, Stent, and Zeljkovic 2012).

2.2.2.4 Sentence verification task

The sentence verification task resembles the lexical decision task (see Section 2.2.1.3) and can be used to avoid a ceiling effect. Listeners hear a sentence that contains a logical proposition that is either true (e.g., *Cows eat grass*) or false (e.g., *Horses are known to climb up trees*). They have to decide as quickly as possible whether the statements presented are true or false (or plausible/implausible, see Example 2.10). This task requires a listener to process the overall information of an utterance and combine background knowledge and knowledge of word occurrences with their ability to perceive individual phonemes. The assumption is that more intelligible speech will allow listeners to understand the message correctly and that the decision will be faster for more intelligible language varieties than for varieties that are more difficult to understand (Munro and Derwing 1995b; Bohn and Askjær-Jørgensen 2017).

The sentences should not be too long and should require little cognitive processing. Their truth value must be easily determined by the listeners on the basis of everyday knowledge. This means that potentially ambiguous or misleading sentences must be avoided. To check that the sentences used for the experiment are all deemed unmistakably true or false to the listeners, a pilot experiment should be conducted with a group of participants with the same background as the prospective listeners in the real experiment. The participants in the pilot read the sentences in their native language. The real experiment should include only sentences that (almost) the whole group consistently finds either true or false. If higher error percentages are found in the real experiment this must be caused by the sentences being difficult to understand rather than by lack of background knowledge.

Since the response is binary, it is recommended to use a considerable number of test items to minimize the impact of guessing. In addition to error counts, reaction times can be measured. In the latter case, the decision times must be lined up with the earliest point in the acoustic stimulus where sufficient information is available to correctly decide on the truth value of the sentence. Responses given before this moment must necessarily be based on guessing. In the sentence *Cars cannot eat books* the listeners can decide on the truth value after they have heard the word *eat*, and therefore the reaction time should be measured from this point.

The appendix in Munro and Derwing (1995a) provides more examples of true/false test sentences.

Example 2.10: Sentence verification task

Hilton, Gooskens, and Schüppert (2013) used an alternative application of the sentence verification test, where Danish listeners presented with Norwegian test sentences were asked to judge the plausibility of the proposition. The aim was to test the influence of morphosyntax on intelligibility. This approach was chosen over a more traditional sentence verification test (where the content would be deemed as either "true" or "untrue") as it was considered impossible to design enough sentence types for the experiment that are universally true. An example of an implausible Norwegian sentence is Du lukter med hendene dine *'You smell with your hands'. Examples of plausible sentences are:*

1a	*En*	*bonde*	*arbeider*	*på*	*sin*	*traktor*
	A	farmer	works	on	his	tractor
	'A farmer works on his tractor'					
1b	*En*	*bonde*	*arbeider*	*på*	*traktoren*	*sin*
	A	farmer	works	on	tractor+THE	his
	'A farmer works on his tractor'					

While 1a is grammatical in Danish (and archaic in Norwegian), 1b is ungrammatical in Danish (and the unmarked variant in Norwegian). Longer reaction times were found for Danish participants for sentences like 1b, which have an extra marking of definiteness. Therefore it was concluded that this extra marking reduces the intelligibility for Danes.

2.2.2.5 Carry out instructions

A simple approach to access sentence intelligibility involves giving listeners a set of instructions to follow. The success rate in executing the instructions, the time it takes the listener to start carrying out the instructions, and the time it takes to complete the task successfully are measures of intelligibility. Typically, the instructions entail moving or arranging objects within a simulated environment displayed on a computer screen.

This method offers the benefit of assessing intelligibility within a realistic communicative context. However, the requirement for listeners to carry out the specified actions can restrict the range of syntactic structures and vocabulary that can be incorporated.

Example 2.11: Carry out instructions task

To test the relative contributions of pronunciation and morpho-syntax to the intelligibility of foreign-accented speech, van Heuven and de Vries (1981) collected short Dutch utterances in which Turkish speakers of Dutch as a second language spontaneously described simple acts that were performed by the experimenter, such as pouring beer into a glass. Next, for each act, three additional sentences were recorded so that, in total, the following four versions could be presented to native Dutch listeners in a Latin square design (see Section 2.1.3.3):

(i) *Original utterance by a Turkish speaker of Dutch, with non-standard pronunciation and morpho-syntax;*
(ii) *Same utterance repeated literally by a native Dutch speaker, with standard pronunciation, but imitating the non-standard morpho-syntax used by the Turkish speaker of Dutch;*
(iii) *Same utterance by a Turkish speaker, with non-standard pronunciation, but standard morpho-syntax;*
(iv) *Same utterance by a native Dutch speaker, with both standard pronunciation and standard morpho-syntax.*

The listeners were instructed to listen to the utterances and perform the act described in each, as promptly as they could, using the same array of objects that had been provided to the original speakers. Both the number of errors (failure to understand the description) and the listeners' reaction time in case of correct understanding were established. It was concluded that phonetic factors play a more important role than non-phonetic factors in the intelligibility of foreign-accented speech.

2.2.3 Text level

If the main interest of the researcher is to get a general impression of the overall intelligibility of a language, it is more natural and often simpler to test at the text level rather than at a lower level of analysis.

2.2.3.1 Opinion testing

To obtain a rapid assessment of the general level of intelligibility of a language, a simple and effective approach is to request participants to rate their comprehension on scale(s). This method was used by ethnographers in early anthropological fieldwork on intelligibility and referred to as 'ask the informant', as opposed to 'test the informant', where intelligibility scores are based on a functional test (see Voegelin and Harris 1951; Casad 1974; Simons 1979).

The most straightforward opinion testing (judged intelligibility) involves no speech fragments. Participants are asked to indicate the percentage of the words of a language they think that they can understand (for example, by moving a slider on the computer screen) or to judge on a Likert-like scale how well they think that they understand a speaker of the target language (for example using a scale from 1 "not at all" to 10 "everything"). Alternatively, they can be requested to assess the extent to which they believe that monolingual speakers of their own language would understand the speaker. This may provide a better estimate of inherent intelligibility. An advantage of using judged intelligibility is that no speech material has to be selected and recorded. Without recordings it is possible to abstract from individual speakers who may influence the results because of specific voice characteristics and speaking styles (see Section 2.1.1.5) as well as dif-

ferences in recording conditions (see Section 2.1.1.4). On the other hand, it remains unclear whether respondents can consistently assess intelligibility in the absence of speech samples. They may never or rarely have heard the language, may not remember how well they understood the speaker or it may be unclear to them how a typical speaker of the target language sounds. Consequently, respondents may base their opinions on some extra-linguistic factor, such as their positive or negative attitudes toward the country and its speakers (see Section 4.3), political borders, desirable answers, or the geographic distance to the place where the language is spoken.

Often, a better alternative is to present recordings of the target language to listeners before they judge how well they can understand the language **(**perceived intelligibility). However, even with this approach, it is uncertain whether listeners are actually able to make the judgments on an objective linguistic basis without the influence of extra-linguistic factors.

Developing appropriate functional tests and administering them can be time-consuming, since special care must be taken to prevent priming effects, ceiling effects, memory overload, and other unwanted effects (see Section 2.1). Opinion tests are generally easier to set up and carry out than functional tests. In particular, the same material can be presented repeatedly without the risk of a priming effect (see Section 2.1.3.3). However, the results of opinion testing should be interpreted with some care. They provide us with information about individuals' subjective perception of the intelligibility of languages, but people's self-reported language behavior may not be in line with their actual language behavior. In research on the intelligibility of foreign-accented speech, correlations between opinion measures and functional measures are not always significant (see Lindemann and Subtirelu 2013: 579 for an overview). Lindemann and Subtirelu (2013) mention two potential reasons for low correlations found in previous research. First, they may be caused by a ceiling effect in the results of the functional test (see Section 2.1.3.2). Second, they may be caused by inconsistent listeners' biases where some listeners upgrade and/or downgrade their subjective ideas.

Other studies show that listeners are often good at estimating how well they can understand a closely related language. Significant correlations have been found between opinion scores and results from functional tasks. Gooskens and van Heuven (2017) found correlations between r = .72 and .97, and Tang and van Heuven (2009) between r = .77 and .81. Significant correlations were also found in studies carried out to evaluate the quality of speech synthesis, see for example van Bezooijen and van Heuven (1997), and references therein. Therefore, opinion testing may provide a useful shortcut to functional intelligibility. However, many researchers doubt the validity of opinion testing and prefer to test actual speech comprehension with functional testing methods. Tang and van Heuven (2009)

note that since the correlations are not perfect, mutual intelligibility should be tested functionally whenever the resources are available.

Example 2.12: Perceived intelligibility
Tang and van Heuven (2009) played recordings of the same text, the fable The North Wind and the Sun, *in 15 Chinese dialects to 24 listeners from each place where the dialects were spoken. For each dialect, the listeners were asked to indicate how well they believed monolingual speakers of their own dialect would understand the speaker on a scale from 0 ("They will not understand a word of the speaker") to 10 ("They will understand the other speaker perfectly"). The results showed high correlations between perceived intelligibility and intelligibility tested functionally by a sentence and a word intelligibility test (*r = *.77 and .81) and with objective measures (lexical similarity and phonological correspondence,* r = *.83 and .71).*

2.2.3.2 Recorded Text Testing (RTT)

A common method to test overall comprehension of a text is to have listeners listen to a text and have them answer questions about the content of the text. This method was first used in the fifties of the previous century to establish the mutual intelligibility of American Indian languages by Voegelin and Harris (1951), Hickerson, Turner, and Hickerton (1952), and Pierce (1952). They referred to the test as the Recorded Text Testing (RTT) method. The intelligibility of a language is quantified by calculating the average percentage of correct responses provided by the listeners. The questions can be posed after the presentation of the text or interspersed in appropriate places throughout the text. However, it may be more realistic to provide the questions to the listeners in advance, enabling them to concentrate on the relevant aspects of the content since people mostly listen to a text with a specific purpose in a real-life situation. Casad (1974) and Nahhas (2007) provide a detailed overview of the steps that should be taken to carry out a test with RTT.

The questions about the texts should cover the content of the whole text as well as possible, and a correct answer should not depend on understanding a single specific word. To ensure that the evaluation solely measures comprehension of the text (and not that of the questions), it is recommended to present the questions (and multiple-choice alternatives, if applicable) in the listener's native language.

The test should not measure the listener's memory, general knowledge, or intelligence. Hence, it is crucial to pre-test the questions on a separate group of participants not provided with the text to ensure that correct responses are not based on general knowledge or logic.

Defining correct answers to the questions can be challenging, and researchers may need to distinguish between varying degrees of correctness, such as "completely correct", "partly correct", and "incorrect". A more objective solution

is to use multiple-choice questions, where listeners have to choose between a limited number of possible answers. This approach allows for easy manual or computerized correction of answers. However, one drawback of multiple-choice questions is the difficulty in finding plausible distractors that the listener does not easily eliminate. Additionally, multiple-choice questions are rather unnatural since people are usually not given a choice between several possible answers in a natural situation. When comparing intelligibility among different groups of listeners, it may also be important to realize that some listeners are familiar with multiple-choice testing. They may do very well by virtue of understanding how these tests work rather than through actual understanding while others are less familiar with this kind of testing method and therefore do not have this advantage. Kluge (2006) discusses the disadvantages of the RTT method in more detail. She points out that questions may be culturally inappropriate in many societies and that listeners can often infer the correct answers from the questions.

A critical discussion of the reliability of past research with the RTT method is presented by Yoder (2017). The recordings used for the RTT tests have often been different narrative texts that are unique for each speaker and are therefore also followed by different questions. By analyzing 25 papers using the RTT method, Yoder demonstrates that the selection of a text for an RTT has significant implications on the results and could lead to incorrect conclusions.

Example 2.13: Recorded text testing

Delsing and Lundin Åkesson (2005) asked listeners to answer five open questions about short passages of continuous texts. Unlike many other investigations, they used read-aloud texts, which made it possible to use the same texts and questions for all languages in a Latin square design (see Section 2.1.3.3). This is important if results are to be compared (see Section 2.1.1). The texts were short news items about a kangaroo that has escaped from captivity and hops around the streets of Copenhagen and about counting frogs to indicate how well the natural environment is doing. They were read aloud by professional news reporters. Five open questions were asked about the content, such as "For how long was the kangaroo on the run?". A correct answer ("2 months") was given 2 points and a partly correct answer (e.g., "1 to 2 months") was given 1 point.

2.2.3.3 RTT retelling

Kluge (2006) discusses an alternative approach to the standard RTT question format, the RTT retelling method. This method was developed by Ring (1981) and has since been subsequently refined by various researchers. Listeners are required to listen to a narrative divided into natural segments of one or two sentences each and then retell the recorded text in their native language while keeping as much detail of the original as possible. In this way, the listeners do not have to answer specific comprehension questions. For each segment of the text, the number of

correctly retold core elements are counted, and the segment scores are added up to obtain the overall score for a given RTT text.

Kluge (2006) provides the following example to illustrate the RTT retelling method. In this example, the listeners hear the following fragment in the target language:

> *As I pulled in the net, a crocodile was sleeping in the water under a Sago tree. I didn't see it. I pulled in the net with the crocodile behind me.*

The researcher defines three core elements (indicated by the three colors) in the fragment, each counting for 1 point:

- (he/the man) pulled/investigate net (1 point)
- doesn't know/see (crocodile) (1 point)
- crocodile in the water/under Sago tree/between Sago roots (1 point)

In the following examples of retellings by three different listeners, all core elements are correctly retold, and all listeners therefore get the maximum number of points (3) for this fragment:

> Listener 1: He was pulling like this and didn't see the big crocodile between the Sago roots.
>
> Listener 2: While the man was investigating the net, he didn't know that there was a crocodile there under a Sago tree.
>
> Listener 3: While he was investigating, he didn't see that a crocodile was under the Sago (trees).

The RTT retelling method has several advantages over the standard RTT question method. Firstly, it tests comprehension of the entire text rather than selected sections only. Secondly, retelling a story may be more culturally appropriate and less intimidating than answering comprehension questions, especially in traditional societies. Additionally, this method eliminates the need to design and translate comprehension questions into different speech varieties. However, one major disadvantage is that it is time-consuming to develop the test and count the number of correctly retold segments.

As in the case of the original RTT method, it remains a concern that different narratives are mostly recorded in different languages (see Section 2.2.3.2).

Example 2.14: RTT retelling task

To classify Lalo, a Central Ngwi (Loloish) language cluster spoken in western Yunnan, China, Yang (2012) tested the mutual intelligibility of 18 Lalo villages using a RTT retelling task. To prepare the RTTs, native speakers related short narratives, typically one to three minutes long. Next, for each recording, a panel of eight native speakers listened to the whole text and then listened again section by section, with a pause

after each section. Each section was at most 10 seconds. During the pause, the listeners retold the section's content in Chinese, which most Lalo speakers master. Elements that all native listeners retold formed the baseline for scoring responses for listeners from other varieties.

For the actual experiment, eight to ten participants with a minimum of previous exposure to the target language were selected in each village. Each participant followed the same testing procedure as the panel of native listeners. A listener's intelligibility score was the number of core elements mentioned divided by the total number of core elements identified by the panel of native listeners. Not all RTTs were tested at each village due to time constraints and the fatiguing nature of the RTT process for the participants.

2.2.3.4 Text translation

The text translation method is very similar to the word translation task (see Section 2.2.1.1) and the sentence translation task (see Section 2.2.2.1); it has the same advantages and drawbacks. The use of whole texts may provide the listener with more context than a sentence translation task. In the case of spoken language, the text is typically presented one sentence at a time while the listener responds by either writing out or pronouncing the translation as literally as possible. Compared to content questions, an advantage of this method is that the researcher does not have to formulate questions about the text.

Translating sentences and texts may be challenging for certain listeners since it is not a natural task for them. The skill of translation entails much more than intelligibility only, and it may heavily rely on the listener's memory. Therefore, the text must be presented in short segments with pauses between them, during which the listener can write down the translation. To facilitate the writing task and to decrease possible memory problems, Scharpff and van Heuven (1988) printed the function words in the target text on the answer sheets and left blanks for content words to be filled in (one blank of uniform length per content word).

To gain an even better understanding of the process of understanding a text in a related language, the listeners can be asked to reflect on their decisions by thinking aloud while translating (e.g., Kaivapalu 2015; Mieszkowska and Otwinowska 2015; Jagrova et al. 2019) but a more natural task may be what Mercer (2000: 98) refers to as exploratory talk, i.e. talk in which partners engage with each other's ideas about how to translate the text. This kind of reflection may provide even more information about the processes of trying to understand a closely related language.

Example 2.15: Text translation

Gooskens, Heeringa, and Beijering (2008) tested the intelligibility of the fable The North Wind and the Sun *in 18 Nordic varieties (Danish, Norwegian, Swedish, and Faroese) among Danish listeners from Copenhagen using a translation task. The listeners were presented with the six sentences of the fable one at a time. Each sentence was in a different variety following a Latin square design (see Section 2.1.3.3). While*

listening to the six sentences, the listeners had to translate each word they heard into standard Danish. Each sentence was presented twice. The sentence was presented as a whole before being repeated in smaller chunks of four to eight words (depending on the position of prosodic breaks) to make sure that saturation of the listeners' short-term memory would not influence the results and that the listeners had enough time to write down their translations. The percentage of correctly translated words constituted the intelligibility score of a given language variety. The results correlated significantly with lexical (r = −.64) and phonetic (r = −.86) distances.

Example 2.16: Text translation with exploratory talk

Börestam and Hansen (2021) used exploratory talk to investigate the strategies of Danes when trying to make sense of a Swedish spoken text and vice versa. Danish and Swedish participants were asked to listen to a text read aloud in the neighboring language. They were instructed to either translate the text (or parts of it) immediately into their native language or to write down (in any form) what they supposed they had heard. Afterwards, the participants worked together in pairs with the same native language and a recording was made. Based on their notes, they made a translation that was satisfactory to both of them. They were told to motivate their choices to each other. The results give an impression of comprehension of a neighboring language as a dynamic and creative process of toggling between the macro and the micro perspective, i.e., from only looking at portions of words up to considering the full context.

2.2.3.5 Picture-to-story matching

The picture-to-story-matching test involves playing a complete text to the listener, who is then presented with several options of pictures on screen or paper and asked to select the picture that best matches the contents of the story. This is, in fact, a multiple-choice task without any written text involved. No questions are posed in writing, nor do the listeners have to write down answers. Like the picture-pointing task (see Section 2.2.1.2), it is therefore suitable for testing children and illiterate adults. If the test is administered on a computer, the answers can be registered automatically using a touchscreen or having the listeners click on one of the pictures with the mouse.

However, the test is a rather rough measurement that may only provide an impression of whether the listeners got the gist of the story. Therefore, it is important to pre-test the task with a separate group of listeners before administering the actual test. To make the test more sensitive, reaction times (see Section 2.1.3.2) can be measured in addition to the percentages of correctly chosen pictures. It may also be possible to improve the picture task by using a larger number and more complex pictures and by varying more message elements (see Example 2.17). This would require the listener to extract more key elements from the text before being able to choose the correct picture.

Example 2.17: Picture-to-story matching

Gooskens and van Heuven (2017) constructed four pictures to embody the correct or incorrect representation of two key elements in a text passage. An example is provided in Figure 2.9. For instance, if the passage was about driving a car in winter, one picture showed a car driving in summer (with a sunny landscape and trees and flowers in full bloom, picture a), another picture contained a plane in a summer setting (picture b), a third picture showed a car driving in a wintery landscape (picture c), and a last picture showed a plane flying over a wintery landscape (picture d). When both content features were correctly identified (picture c), the listeners got full points; when both aspects were wrongly identified (picture b), no points were given, and when one feature was correct (pictures a and d), the listeners were given half points. The correlations between the picture task and two functional tests at the word and text level showed correlations of = .73 and .71. The task seemed too easy for some participant groups, creating a ceiling effect.

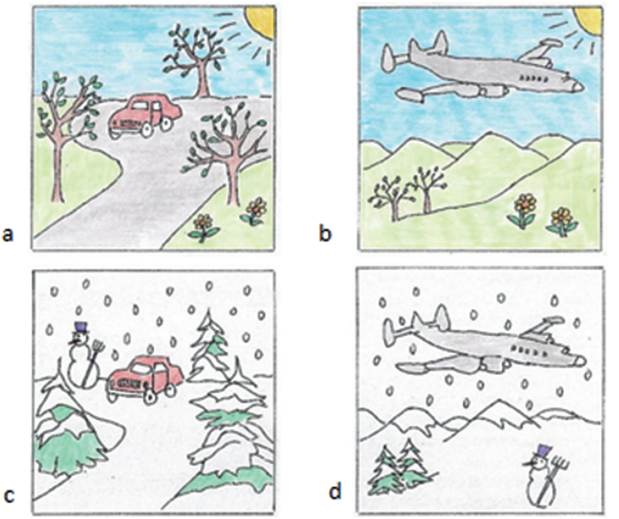

Figure 2.9: Picture used for a picture-to-story matching task. From Gooskens and van Heuven (2017: 28).

2.2.3.6 Cloze test

The cloze test was developed by William Taylor in the United States in 1953 and has since then been used extensively for measuring comprehension in the classroom (Abraham and Chapelle 1992; Keshavarz and Salimi 2007) but also for intelligibility research purposes (e.g., Smith and Rafiqzad 1979; van Bezooijen and Gooskens 2005a, 2005b, 2006). In a cloze test, selected words are removed from the text at regular intervals, for example, every tenth content word, and replaced by gaps, i.e., lines or empty spaces of uniform length (in the case of written language) or by beeps of uniform duration (in the case of spoken language). The percentage of cor-

rectly filled-in words is the measure of intelligibility. The cloze test requires understanding context and vocabulary in the text to identify the correct words or type of words that belong in the gaps. It is, therefore, an easy and useful way to test overall text intelligibility.

It is generally left to the listeners to think of suitable words. The scoring of this version of the test is difficult because the researcher has to decide whether the words offered by the listener are correct. Alternatively, the removed words are placed above the text, and the listeners are asked to put the words back in the correct position. Since the aim is to test the intelligibility of whole texts, a translation of these words should be provided. If some of the response alternatives were unfamiliar to the listeners, they would not be able to place them in the correct gaps, even if they did understand the texts in their entirety. This version of the cloze test can be scored automatically and is, therefore, an efficient and objective way of testing text comprehension, especially when a large number of listeners are involved. Usually, the test is employed to establish the intelligibility of written texts, but it can be adapted to test spoken language. In that case, the listeners will hear only one sentence at a time to avoid the influence of memory limitations. Chapter 3 provides a description of the spoken cloze test used to test the mutual intelligibility of spoken languages in Europe in the MICReLa project.

Example 2.18: Written cloze test

Van Bezooijen and Gooskens (2005a, 2005b) measured the intelligibility of written Afrikaans and Frisian texts among native Dutch participants using a variant of the cloze test. As a basis for assessing the intelligibility of written text, they used two Dutch newspaper articles (of 329 and 317 words) with an average level of difficulty. In both texts, five nouns, five adverbs, five adjectives, and five verbs were randomly selected. These were placed in alphabetic order above the text and replaced by blanks in the text. Next, the two texts were translated into Frisian and Afrikaans. The same words were removed and placed above the texts. Below, one of the Afrikaans cloze tests is shown (from van Bezooijen and Gooskens (2005b: 24). The participants were given ten minutes to put the 20 words back in the right place in the texts. The percentage of words placed back correctly was taken as a measure of the intelligibility of the texts. It appeared that it is easier for Dutch speakers to understand Afrikaans than Frisian.

aantreklike	al	blyk	doen	eksplosief
fantasties	internet	lank	misbruik	moeilike
moet	nooit	onbetroubare	paar	soms
stel	teleurgesteld	verhale	werk	woorde

Nie lank en blond nie maar 'n klein, kaalkop lamsak

Jaarliks soek tienduisende alleenlopers via 'n verhoudingsbemiddelingsagentskap of op die ———————————————————— na 'n metgesel. ———————————————————— met sukses, maar ook dikwels daarsonder. ———————————————————— mans blyk in werklikheid oud en lelik, en ———————————————————— agentskappe verdwyn skoonveld sodra die inskryfgeld betaal is.

Dit het ———————————————— geklink in die kletskamer op die net. Die ou met wie Annemarie op die internet in kontak gekom het, het aan al haar vereistes voldoen: hy was ————————————————, blond en atleties. Het hy geskryf. “Maar tydens die eerste afspraak in die kroeg het daar ’n lamsak op my gesit en wag,” sê die 27-jarige vrou. “Hy was klein, kaalkop en het ook nog tien jaar ouer gelyk as die ouderdom wat hy op my webtuiste opgegee het.” Annemarie het drie ———————————————— gewissel met die man en het kwaad en teleurgesteld vertrek. Die sprokie was verby.

Vir sulke kliënte is Joke Pronk aangestel. Sy ———————————————— vir die Algemene Vereniging Verhoudingsagentskappe (AVV) en bemiddel in geskille tussen kliënt en verhoudingsagentskap. ’n ———————————————— saak. “Die mense is ————————————————,” aldus Pronk, ’n ———————————————— honderd euro terugbetaling bied dikwels min troos, hulle wil veral stoom afblaas.” Pronk ———————————————— vas watter agentskappe nie hul afsprake nagekom het nie en deel “geel kaarte” uit.

Die aantal verhoudingsagentskappe het die afgelope jare – met name op die internet ———————————————— gegroei. Daar is ———————————————— vier – à vyfhonderd (op die net). Met die groei neem die ———————————————— ook toe. Baie internetagentskappe blyk te verdwyn sodra die inskryfgeld betaal is. “Skielik bestaan die webwerwe nie meer nie. Of hulle ———————————————— ’n kantoor in Kazachstan te hê.” Pronk ken die ————————————————, maar dit kom volgens haar nie voor by die vyftien agentskappe met ’n AVV-keurmerk, waarvoor sy werk, nie. Sy raai kliënte aan om ———————————————— ’n eie e-posadres te gee nie, maar alleen e-posse te stuur via die webwerf. “As mense afsprake wil maak, adviseer ons nadruklik om dit te ———————————————— op ’n openbare plek. Dit is regtig nie net psigopate wat op die internet na ’n vrou soek soos sommiges beweer nie, maar jy ———————————————— wel versigtig wees.

2.2.4 Discourse level

The tests discussed so far test cross-language intelligibility in one direction, for instance, the intelligibility of target language A among speakers of language B. However, sometimes researchers’ main interest is to establish how well speakers of different closely related languages can communicate each speaking their own native languages (receptive multilingualism). When two speakers communicate using receptive multilingualism, both the receptive and the productive proficiencies of the individual participants in the conversation determine the success of communication. On the one hand, communication can be compromised by a lack of knowledge on the part of the listener. Communication breakdowns may occur when the speaker uses an idiomatic expression or word unknown to the listener. On the other hand, problems can be found on the part of the speaker. Some speakers may be better at adapting their speech to the listeners than others (see Section 8.5).

So far, most research on receptive multilingualism has focused on testing the receptive proficiency of the listeners by presenting recorded speech stimuli of the target language to the listeners. The advantage of such a procedure is the experimental control that allows the researcher to vary variables systematically. However, since no communication occurs, the authenticity and similarity to real-life settings are low (ecological validity, see Sections 2.1.1.2 and 2.1.1.3). For the listener, there are no opportunities for negotiation of meaning or the use of communicative strategies, such as signaling a lack of understanding or correcting miscommunication (see Lindemann and Subtirelu 2013: 584). The speaker cannot check whether the listener has understood the message and reformulate it if this is not the case. It is currently largely unknown to what extent the productive and receptive proficiencies of the interactants each contribute to the common goal of communication. The success of receptive multilingualism depends on the participating interactants and is not simply the mean receptive proficiency of the two individuals.

Therefore, to model communication, it is not sufficient to test how well listeners understand the speakers' language. There is a need to create experimental situations where interactants communicate, for example, by solving a task together, each using their own language. The results can be analyzed quantitatively by counting the number of tasks carried out correctly in a certain amount of time. A more qualitative analysis can be carried out by looking at the strategies the participants use to communicate successfully. Below, a few of such communicative tasks are discussed.

2.2.4.1 Spot-the-differences task

The spot-the-differences task involves two interactants who each have a copy of a picture that displays a large number of objects arranged randomly. The two pictures differ in various aspects, including shapes, sizes, colors, and the presence or absence of particular objects. The participants cannot see each other's pictures, and their task is to work together to identify as many of these differences as possible as quickly as possible. Dependent variables are the number of differences correctly identified and the time taken to complete the task. The duration of the task is an indicator of the degree of communicative efficiency, with shorter durations denoting greater efficiency and longer durations denoting lower efficiency.[8]

8 For more information and examples of pictures that can be used for a spot-the-differences task see Baker and Hazan (2011).

Example 2.19: Spot-the-differences task

Van Mulken and Hendriks (2015) used spot-the-differences tasks to investigate the effectiveness of language use in computer-mediated communication between native speakers of German and Dutch. The communication took place using English as a lingua franca (ELF) or receptive multilingualism. For each task, they manipulated a picture containing a varied set of objects into two versions. To elicit a conversation, they chose objects with names likely to be unfamiliar to the participants, such as various toiletry items: tweezers, cotton buds, or panty liners (see Figure 2.10). For each task, the two participants were given two different versions of a picture with several objects. Each participant collaborated with a partner in an online chat to find ten differences between the pictures within a time limit of 15 minutes. The participants were instructed to complete the task in their own native language or ELF. The researchers also analyzed the communication qualitatively to gain insight into participants' communicative strategies to gain mutual understanding (see Section 4.6). The findings showed that the participants were more effective in realizing communicative goals in receptive multilingualism interactions than in ELF interactions and that participants made use of different communication strategies in the two kinds of interactions.

Figure 2.10: Pictures used for the spot-the-differences task by van Mulken and Hendriks (2015: 411).

2.2.4.2 Map task

A similar task is the map task, where one interactant is the instruction giver and the other the instruction follower (cf. Brown et al. 1984; Anderson et al. 1991; Grønnum 2009). Both interactants have a map with roads and landmarks. The giver's task is to instruct the follower on how to trace a route between two landmarks that are known to the giver but not to the follower. Introducing subtle differences between the two maps makes the task more challenging. After completing one task, the participants switch roles, and new maps are provided. The dependent variables are success rate and the time taken to complete the task. Example of maps developed for a map task is provided in the HCRC Map Task Corpus.[9]

9 https://groups.inf.ed.ac.uk/maptask/

Example 2.20: Map task

To investigate how well speakers of the two unrelated languages, Russian and Estonian, with receptive knowledge of each other's languages could communicate, Bahtina, ten Thije, and Wijnen (2013) used stylized maps that differed in several features (e.g., non-identical arrows to indicate movement in specific directions or gaps instead of connections, see Figure 2.11). These modifications forced participants to take the longest route to get from Point A to Point B. The participants were seated behind personal computers in separate rooms and connected via an online video link. One of the participants was assigned the role of instruction follower, and the other was assigned the role of instruction giver. The followers were instructed to explain their location (Point A), and the givers were to instruct the followers to get to their location (Point B). They were asked to use their mother tongues as much as possible. The results suggested that efficient communication is reached through the explicit negotiations aimed at establishing mutual understanding and that the level of proficiency in the non-native language did not strongly impact communicative success.

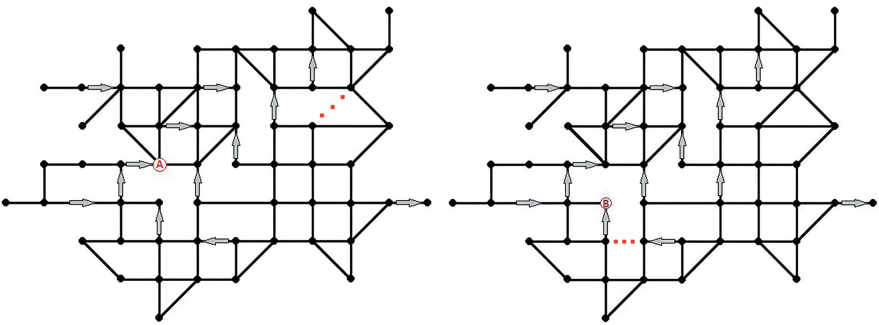

Figure 2.11: Stylized maps for a map task. From Bahtina, ten Thije, and Wijnen (2013: 166).

2.2.4.3 Observations

If we want to get more detailed qualitative insight into the strategies used by participants in a conversation to reach mutual understanding (see Section 4.6), it may be useful to observe real communicative situations. This is, first and foremost, a qualitative approach, but it is also possible to use the results as a basis for a quantitative analysis, for instance, by counting turn-takings, misunderstandings, reparations and length and frequency of pauses (Zeevaert 2004).

Examples 2.21: Observations

Börestam Uhlmann (1994) recorded around thirty inter-Scandinavian arranged conversations between Danes, Norwegians, and Swedes aged 18 to 25 who were unaccustomed to the neighboring languages. The material included conversations between speakers with different native languages as well as monolingual conversations. She was interested in which kind of strategies were used by the participants to enhance mutual comprehension, such as rephrasing, explaining, switching to English, repairing misunderstandings, and interrupting to either clarify something or make certain that the message had been correctly understood. While her data analysis was primarily qualitative, she also quantified the outcomes by counting the

number of repairs and misunderstandings that occurred. The results showed that the communicative flow was more often interrupted for repetition, clarification, confirmation, and paraphrase in, bilingual conversations than in conversations between speakers of the same language.

In another observational investigation (Börestam 2015), the researcher investigated inter-Nordic communication in Iceland. In Iceland, Danish is taught as a foreign language. Since it is closely related to Swedish and Norwegian, Icelandic people are likely to understand these languages based on their knowledge of Danish (mediated intelligibility, see Section 1.1). Pretending to be tourists, fieldworkers who were native Danish, Norwegian, or Swedish speakers approached Icelandic adolescents, asking for directions to the Nordic House in Reykjavík in their mother tongues. The study, conducted across three different periods (1983, 1999/2004, and 2006), revealed a decline in the percentage of young individuals who comprehended the inquiry, dropping from approximately two-thirds in 1983 to just under 40% in 2006. Simultaneously, there was an increase in respondents opting to respond in English, with a significant majority (80%) doing so in 2006 compared to only a third approximately 25 years earlier, in 1983.

2.2.5 Recent methodologies

In most of the intelligibility tests discussed in this chapter, listeners are allowed time to consider their response, and only the result of the response, rather than the time taken to respond, is considered. Such tests have been employed more frequently than techniques that aim to gain insights into the listener's cognitive processes during comprehension, primarily because they do not require any specialized or costly equipment.

In the previous sections, reaction time measurements were mentioned a few times. Such measurements are claimed to offer an indirect measure of the degree of difficulty experienced in processing input. The assumption underlying reaction time measurements is that faster listener reactions indicate better intelligibility of the input being processed. Reaction time can be measured using software applications that register when a listener performs a certain action, such as a vocal response, pressing a button on the computer keyboard, or touching the computer screen. Tasks measuring reaction times, such as lexical decision tasks (see Section 2.2.1.3) or sentence verification tasks (see Section 2.2.2.4), can be employed to assess overall intelligibility. The measurements are temporally accurate to within a few milliseconds. The measuring point should be considered carefully. Generally, it makes most sense to measure from the disambiguation point, in particular, the point in a word where it can be distinguished from all other words or the point in a sentence where it can be decided whether the proposition is true or false (see Section 2.2.2.4). Stimuli are usually responded to faster with the dominant hand, the right hand for right-handed people and the left hand for left-handed people (Rastatter and Gallaher 1982; Shen and Franz 2005). It should be noted that reaction time measurements are very delicate. Many trials fail to record what they are intended

to because of accidental keystrokes, lapses in the listener's attention, and distractions in the laboratory environment. Listeners should, therefore, be allowed ample practice on the task and be given breaks every once in a while. Another procedure for avoiding unwanted influence on reaction time scores is to normalize data by eliminating outliers, for instance, scores exceeding 2.5 standard deviations above a listener's mean reaction time.

Recently, web-based behavioral research is becoming more and more popular. This makes it possible for the researcher to collect large amounts of research data quickly, and programs for preparing and carrying out such research have become increasingly user-friendly. A unique problem with web-based testing is its reliance on listeners' hardware and software. When an experiment is carried out in the lab, the researcher can make sure that the same computer, stimulus software, and hardware are used for the whole response collection. In remote testing (over the internet), it is not possible to control all these conditions. Listeners may use their own computer (desktop, laptop, tablet, or even phone) with their own operating system and access experiments through a variety of web browsers. This may influence the measurements of reaction time. However, modern internet platforms now allow recording the response time with acceptable accuracy and precision (Anwyl-Irvine et al. 2021).

Several other new techniques for measuring cognitive processing during listening have become available in recent years (e.g., van Engen and McLaughlin 2018). Such measurements can serve as a valuable addition to functional intelligibility tests and subjective intelligibility estimates of listeners, as they allow for the detection of variations in listening effort that may not be apparent in the performance level of individuals in traditional functional tests. Eye-tracking measurements can detect the presence, attention, and focus of the users by determining what they are looking at and for how long their gaze is in a particular spot while listening to speech. It thereby provides information about processing difficulties and could be used to determine the moment of word recognition (fast/easy word recognition is the hallmark of intelligibility) by measuring eye fixation on pictures on the screen that correspond to the stimulus. The technique has been used to investigate processing among hearing-impaired patients (Wendt, Kollmeier, and Brand 2015). Since the difficulties encountered when listening to a closely related language can be compared to the challenge for hearing impaired when listening to their native language, it can be assumed that the technique could also be used for intelligibility measurements of closely related languages. As it does not require verbal responses, it is useful for testing various groups of listeners, including children and illiterate adults.

Example 2.22: Eye tracking
Kudera (2022) showed that eye-tracking can be used to measure how well speakers of four closely related Slavic languages (Bulgarian, Czech, Polish, and Russian) understand sentences in the other three non-native languages. He recorded eye fixations from participants who listened to sentences. The participants were instructed to listen to the sentences and look at a visual scene. Then, a pseudo-task involving answering a question in their native language regarding their understanding of the foreign sentence was given. Subjects were informed that pictures provided clues for sentence comprehension and were advised to pay attention to the objects on the screen. The moment of first fixation on a target picture of the direct object was the estimate of the difficulty of sentence processing. The measurements correlated significantly with surprisal scores, a measure of (un)expectedness of the sounds of words in one language given equivalents in another language (see Jagrova et al. 2019).

Pupillometry is a research method used in psychology studies to determine a participant's cognitive effort when listening to a stimulus by measuring the diameter of the pupil (the little black hole in the center of the eye). A dilated pupil and a longer interval between the stimulus onset and the time of maximal pupil dilation have been shown to correlate with lower intelligibility (Zekveld, Kramer, and Festen 2010). Even in situations where non-native listeners perform just as well as native listeners, they may still exert greater listening effort, which may result in increased fatigue and a reduced ability to perform multiple tasks simultaneously. The pupillometry technique can, therefore, be used to demonstrate subtle differences in cognitive effort in cross-linguistic intelligibility testing.

Example 2.23: Pupillometry
Borghini and Hazan (2018) compared the pupil response of 23 native English speakers and 27 Italian speakers of English as a second language. The researchers tested speech intelligibility of English sentences presented in quiet and in background noise at two performance levels. The results indicated that pupillometry is sensitive to differences both across listener types and across the levels of speech intelligibility. Importantly, the study revealed that pupillometry could uncover differences in listening effort even when those did not emerge in the performance level of individuals.

More immediate access to information processing can also be obtained from neurological techniques, such as event-related potentials (ERPs). This technique measures changes in the brain's electrical activity related to cognitive events. They are measured through electroencephalography (EEG) as changes in electrical charges in the scalp. The measurements can tell the researcher exactly when and approximately where in the brain decisions are being made by the listener. The measurements generally support those of behavioral studies. For instance, recognition of cognates is generally shown through a less pronounced peak in the negative wave in the 400 ms window (the N400 component, a negative-going change in the

electrical charges related to semantic processing) than non-cognates. This can be explained by non-cognates requiring deeper semantic processing than cognates (Helms-Park and Dronjic 2016).

Example 2.24: ERP measurements

Goslin, Duffy, and Floccia (2012) used event-related potentials (ERPs) to investigate whether listeners process words spoken with a regional accent or native accent in the same way as word pronounced with a foreign accent. They recorded ERPs during the presentation of foreign and regional accented speech. They compared time epochs in the two most researched ERP components likely to be relevant to accent-related speech, the PMN (Phonological Mapping Negativity, 200–350 ms) and the N400 (350–600 ms) components. In the first epoch, significant differences in average amplitudes indicated that the PMN was greater for regional than native accents. Conversely, foreign accents led to a significant reduction in the PMN when compared to native or regional accented speech. In the later epoch, foreign accents elicited a significantly reduced N400 amplitude when compared to both other accent conditions. However, there was no longer any significant difference in amplitude between regional and native accents. These findings suggest that different strategies were adopted to process regional and foreign accents at a pre-lexical/phonological level.

Various software packages can be used to run the experiments described in this section. E-prime is probably the most commonly used software, but there are also several freely available options, e.g, PsychoPy. There are also options for running experiments over the internet, such as PsyToolkit, jsPsysch, oTree, and PsychoPy. Special equipment is generally still needed for some new techniques, such as eye tracking, pupillometry, and ERP measurements. This makes it difficult to carry out an experiment outside the lab, even though portable equipment has recently become available.

2.3 Cross-validation of methods

The overview of various methods for testing the intelligibility of closely related languages discussed in this chapter has made clear that choosing a method for an intelligibility project is not a straightforward task. The choice of the method to be used in an investigation depends on various practical considerations, such as budget constraints, time limitations, and the listeners' backgrounds. However, even with sufficient time, resources, and listeners willing and able to undergo complicated and lengthy tests, the choice of method should first and foremost be guided by the research question.

But apart from these considerations, does it matter which method is used? To shed more light on this point, we need to determine whether the same people who perform well in one test also attain high scores in another test while controlling for all other factors. Some researchers compared the results of different

methods of measuring intelligibility. Such comparisons are informative, as they offer insight into the importance of choosing a specific method. As noted in the introduction to Section 2.2, it is not possible to compare the absolute level of the scores from tests that use different methods. Cross-validation of methods can, therefore, only be meaningfully done on the basis of correlation coefficients, not by comparing absolute score levels.

Maurud (1976) conducted a study to assess the mutual intelligibility of Scandinavian languages through word and content tests. The study involved soldiers from Denmark, Norway, and Sweden, who were required to read and listen to a text and translate several key words into their native language. The same text was used for both the word and content tests. Afterward, the participants also had to answer questions about the contents of the text. Maurud found correlations between the test results ranging from r = .6 to .8 for various groups of participants. This shows that different tests measure different aspects of intelligibility.

Doetjes (2007) investigated the effect of six different test types on the measurement of the intelligibility of Swedish for Danes. The same text was tested in six test conditions: true/false questions, multiple-choice questions, open questions, word translation, long summary, and short summary. Each test condition was presented to a different but comparable group of high school pupils in two small province towns in western Denmark. The percentage of correct responses decreased from 93.0% for the true/false questions to 66.2% for the short summaries. This shows that test results are strongly affected by the method chosen. There is no clear answer to the question of how well Danes understand Swedish.

Tang and van Heuven (2009) examined the mutual intelligibility of 15 Chinese dialects using functional intelligibility tests at both the word and sentence levels (see Example 2.7 and Section 6.2.2). They compared the outcomes of the tests with each other and with opinion scores. The correlation between word intelligibility and sentence intelligibility was high, with a value of r = .9. They found correlations between the opinion scores and the functional scores ranging from r = .7 to .8. The authors' conclusion was that functional sentence intelligibility tests should be the preferred method for assessing mutual intelligibility.

Gooskens and van Heuven (2017) administered six functional intelligibility tests involving 70 combinations of closely related pairs of European target languages and listener languages, see Chapter 3. The tests were a word recognition test, a cloze test at the sentence level, and a picture-to-story matching task at the paragraph level. Also, tests were presented in spoken as well as written mode. The scores on these functional tests were compared with each other and with measures obtained through opinion testing, i.e., judged and perceived intelligibility. The various tests revealed similar patterns of intelligibility (correlations between r = .71 and r = .79 for tests in the same mode). The correlations with the

picture-to-story matching scores (which suffered from ceiling effects) were generally lower than with the cloze tests and word translation scores. There was a strong correlation between both measures of intelligibility based on opinion testing (judged and perceived), with a value of $r = .99$. These measures also correlated highly with the functional test scores, particularly with the cloze test, with correlations ranging from $r = .94$ to $r = .99$.

Čéplö et al. (2016) tested the mutual intelligibility of spoken Maltese, Tunisian Arabic, and Benghazi Libyan Arabic using three listening tests, a word test where the listeners were asked to perform a semantic classification task with 11 semantic categories (see Figure 2.8), a sentence translation test, and a multiple-choice comprehension test at the text level. The correlations between the individual test results of the three tests were very low and sometimes even negative. This suggests that listeners' performance in either test is not a good predictor of their performance in the other and that the three tests constitute cognitively different tasks.

While comparisons of various tests generally show relatively high correlations, a large amount of unexplained variance is still left. Even though there is considerable overlap, different tests measure different aspects of intelligibility and some tests are less sensitive to interfering factors than others. Therefore, the best approach may be to have listeners carry out more than one task to confirm results or illuminate distinct aspects of the same phenomenon.

It can be concluded that it is not possible to say how much of a language a listener understands. To illustrate, a result of 70% correct answers in a test does not mean that the listener understands 70% of the language. The percentages of correct answers depend on the difficulty of the test material and the test used. Therefore, the same test method must be used to compare the intelligibility of different languages.

In addition, different levels of intelligibility are needed for different communication purposes. In some situations, only a very basic level of understanding is needed, while in others, it is essential to reach a high level of understanding of complex speech. Therefore, at present, it is not possible to give a definite answer to the question of how well speakers of language A understand target language B. Consequently, caution should be exercised when comparing the results of different investigations. So far, it has not been possible to develop one universal test that could be used to test and compare the intelligibility of languages pairs worldwide. Such a test would be instrumental but also difficult to realize because the results of intelligibility tests depend on many factors, as has become apparent from the overview in this chapter and as will be discussed in the rest of the book.