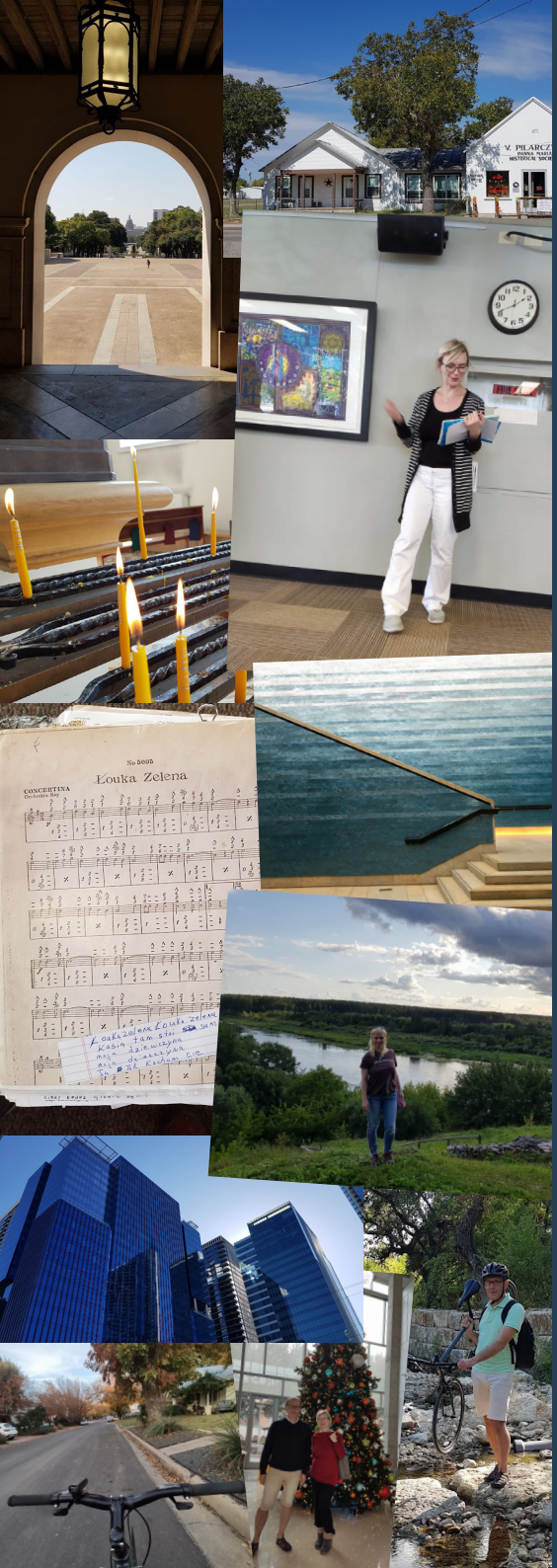

MACIEJ
KARPIŃSKI

KATARZYNA
KLESSA

LINGUIST IN THE FIELD

A PRACTICAL GUIDE TO SPEECH DATA COLLECTION, PROCESSING, AND MANAGEMENT

Linguist in the field: a practical guide to speech data collection, processing, and management

Maciej Karpiński
Katarzyna Klessa

Linguist in the field: a practical guide to speech data collection, processing, and management

wydawnictwo

Poznań 2021

Cover design:
Wydawnictwo Rys

Cover photos:
authors' archives

Review:
dr hab. Anita Lorenc, prof. UW

Copyright by:
Maciej Karpiński
Katarzyna Klessa

Copyright by:
Wydawnictwo Rys

Language editing:
John Catlow

Edition I

Poznań 2021

ISBN 978-83-66666-89-4

DOI 10.48226/978-83-66666-89-4

Publishing house:

wydawnictwo

Wydawnictwo Rys
ul. Kolejowa 41
62-070 Dąbrówka
tel. 600 44 55 80
e-mail: tomasz.paluszynski@wydawnictworys.com
www.wydawnictworys.com

TABLE OF CONTENTS

INTRODUCTION

Research is driven by questions emerging from scientists' own observations, from previous studies, as well as from theoretical considerations which, at some point, may require external verification. We observe in order to be inspired, to formulate hypotheses about salient phenomena, and to discover whether they are true. Researchers observe using the senses, mostly vision and hearing, but sometimes also olfaction, touch, or taste. When the study of human behaviour is considered, real-time observations of any kind, participating or not, may be quite difficult and limiting because of the overwhelming complexity of the processes involved. Even skilled and experienced observers can hardly keep track of the numerous aspects of behaviour that all happen at the same time. Taking notes in real time is another challenge. One must interrupt observation or at least redirect one's attention in order to make reasonable notes. Moreover, if some key aspects of behaviour are not immediately noticed, they can hardly be reconstructed later on. Such phenomena may be overlooked even by the most observant researchers, as it is extremely difficult to expect and capture the unexpected.

For these reasons, sound and image recording technologies have had a profound impact on studies of human behaviour, including communication and language use. While recordings still do not offer a complete picture of the communication process, its settings and contextual grounding, they provide the possibility of the same material being viewed many times, in many different ways (e.g., word by word, frame by frame, slowed down or speeded up, on a small or very large screen), and by many different people (experts, but also naive observers who may be invited to take part in the experiment). They may enable researchers to notice things that remained unnoticed by many others before. Certain limitations of recordings cannot be denied, but still, a vast amount of valuable information is stored there. Now, many instrumental analyses may be carried out using such recordings. Selected parameters of speech can be measured (like speech rate or pitch height). Recordings can be processed and become a source of stimuli in experimental perception-based research. Finally, even if recordings are merely left untouched, in "raw" state, with a minimum description, they may take on value after some time, as they offer a picture of language use at a given moment of history, in a given place and context.

Collecting audio material is a vital part of linguistic fieldwork. It is both demanding and rewarding. One may have the opportunity to meet unique personalities, to record amazing native speakers of small, endangered languages or dialects, but also to witness or become involved in odd or funny situations. This is because recording in the field is not only a technical activity. It often

means close relationships with speakers, with the community. Researchers should learn about them and their history before they even arrive at the location. It is important to remain conscious of basic rules of politeness, of social relations and structures. Interacting with children and elderly people may require a special social disposition, skills and competences, and certainly a degree of sensitivity to cultural differences and local customs. And all this applies not only to research done in distant (from us) parts of the world and (relatively) different cultures – it is of equal importance when we record our neighbours, because for anyone the situation of being recorded may be new, embarrassing or distressing.

When thinking carefully about data collection, processing or management activities, we should not only respect the specifics of a particular language community, but also learn what is actually legal. That may greatly influence the steps that we subsequently take. For example, well before starting the recordings, a good practice is to define the types of information that we wish to collect – whether we need to store more than just recordings, perhaps personal information about our speakers, their family, history, etc. A good practice is to discuss such topics with community members – for example, to clarify how the recordings and the additional data will be stored. What can (and cannot!) be done with the materials? Who will be allowed to access the data? Will data users be allowed to work with the data for any purpose or only for specific kinds of use?

At this point, we arrive at the technological aspects of speech recording, which are often perceived as extremely complex and difficult, especially by researchers in the humanities. While this book has a slightly wider scope, we devote a substantial part of it to this area. We believe that although it is of great help to have a recording engineer or a competent technician as a collaborator, in many situations, researchers specializing in the humanities can easily proceed on their own if they only follow some simple guidelines. On the other hand, even if we have someone "tech-savvy" to advise us, we should be able to understand what can and what cannot be done or achieved. This will help us immensely in the design of our studies. The same applies to further steps of data processing. With some basic knowledge and skills, we are able to build language resources (corpora, databases, etc.) on our own. At some stage, help from computer professionals may be beneficial or even necessary. In that case, again, it is good to understand the basic terms, notions and mechanisms so that we can precisely convey our expectations or requirements.

More and more commonly, research teams involved in fieldwork linguistics are interdisciplinary in character. Therefore, basic mutual understanding is indispensable, as it will surely become useful at some point when dealing with the

fieldwork data. Collaboration within such interdisciplinary groups may involve experts in IT, speech sound acoustics, data processing, statistics, law, ethics, and many other fields. With this in mind, social skills and openness to other people, ideas, and needs appear to be crucial for linguistic fieldwork ventures.

In sum, there are numerous reasons to record and store sounds and images of humans speaking and interacting. (1) We document a communicative event, certain specific behaviour, an act of communication, use of language, facial expression or gesture which may be rare or unique even if it is produced by an ordinary speaker of a widely used language and a member of a huge culture – not only in the obvious case of the last speaker of a language, or a representative of a culture that is fading away. (2) Having recorded a communicative event, one may scrutinize it in depth, view and listen to it many times, and find details which perhaps would not be noticeable in real-time observation. (3) Thanks to recordings, communicative events can be inspected not only on location but also through off-line instrumental analyses involving advanced computation, sometimes too complex to be performed in real time. (4) Recordings can be used as stimuli for various types of listening tests and experimental studies. (5) Potentially (depending on the agreements between researchers and participants) such materials can be used not only for research, but also for education, art, advertising, social campaigns, and other purposes.

This book is conceived as a concise guide to field recording and related field data collection which also includes a portion of methodological considerations and references to certain subfields of linguistics and their particular requirements regarding speech material. In the following chapters, you will find hints and suggestions on how to prepare a plan for your recordings and to design a recording scenario, what equipment you need and how to use it, what you can and what you should not do when it comes to audio editing, how to transcribe and annotate recordings, and how to deal with them and the resulting data and metadata. We are sharing twenty years of our experience in recording speech, in the belief that this book will save you time and stress, remind you about certain things that are perhaps obvious, and allow you to understand some technicalities which may not be so obvious, at least for those with no technological background.

1. PLANNING AND DESIGN

In this chapter, we explain how to prepare for field recordings. By following some simple guidelines and rules, one can avoid making costly mistakes and errors, and wasting time and funds. Nevertheless, although a lot can be foreseen and planned, there is always a certain amount of unpredictability involved, and some flexibility is required to adapt to changing circumstances.

Planning should obviously cover much more than what is going to be done in the field. It is necessary to take a wider perspective, where speech recording is just a component of the entire research process. The starting points are, as always, your **research question**, the **theoretical framework** within which you operate, and the **method** you intend to apply in your research (Labov 1972, 1984; Podesva & Sharma 2014; Kibrik 2017). Big questions and detailed methods should be checked early for technological and organizational feasibility. Can you collect recordings of the required quality and in the required amounts in the given circumstances? Can you find enough speakers that meet your criteria? How much time and money are needed to complete the project? Sometimes great research ideas must wait, or at least be trimmed down or adjusted to actual possibilities in a way that does not significantly compromise the quality of the results. This is often done consciously at the expense of the range or depth of exploration: researchers design narrower studies, ask "smaller" questions, gather smaller groups of speakers, focus on fewer data points.

PLAN AND CHECKLIST

In field linguistics, plans must be flexible, adjustable, and have **optional emergency paths**. This applies to the procedures themselves as well as to certain equipment-related issues. If you are to collect unique material in what may be a once-in-a-lifetime opportunity, it is not unreasonable to take a full spare set of equipment. If people who have had little contact with modern technology are to be recorded, plan additional time to explain how your equipment works, and to allow the speakers to get well acquainted or even play with it. It may be worthwhile to think about alternative spaces for recordings, as you may find the initial one inappropriate. The speaker you want to record may invite you to his/her largest room, while you may actually prefer the smallest one. Sometimes it is also worth considering preparing an alternative scenario, a "plan B", in case your speakers, for example, perceive a taboo in the original scenario or simply find it uninteresting.

It is always recommended to test the entire procedure by making **pilot recordings** with speakers similar to those you will record for your project (age, gender, education, etc. – the more the better). This helps to "debug" and adjust the procedures, to make their description clear and precise, to test the recording setup and conditions, and to diagnose the behaviour of the speakers. Are they distressed, tired, or maybe too relaxed and lacking concentration? In some of our studies, we tested several recording scenarios to establish whether they evoked the communicative behaviour we wanted to analyse; for example, the use of gesticulation, emotional speech, or a particular speaking style (Klessa & Karpiński 2018; Czoska et al. 2015). The insight from the pilot recordings enabled us to adjust, trim, or even redefine the recording procedure. Another important component to take into account is **test recordings** – these immediately precede the actual recording and serve to check that everything works (see Chapter 2). They are not intended to be used to adjust the method and technique, but merely as a **sound check** to adjust the setup to a particular speaker (microphone distance, sensitivity level, etc.).

In sum, **three documents** will be useful. You should have a **plan** which includes all the steps involved in the entire process leading up to the recording session, and the session itself. This is essential when there are many sessions that should be carried out in the same way. It is obviously required when there are multiple people making recordings. But even if you are the only person doing this, it is still necessary to make sure that all the steps are taken in the same sequence each time. Another document that you should certainly prepare is a **checklist** to tick before you leave for recordings (see CHECKLIST window). It will guide you through the tedious and dull process of preparation. It should include equipment, additional artefacts, tools or objects to take, and documents (like instructions or questionnaires for speakers). Finally, you need the **scenario for speech elicitation**, which also includes instructions for speakers. For example, you may want to collect read speech, to interview people, or to arrange some kind of interaction among them.

These documents will be immensely helpful during the project. When already in the field, it might be surprising how fast one gets deeply engaged in interaction with the community members, how absorbing the recording surroundings often are. Consequently, we might very easily forget about certain crucial steps or technical details. The written procedures will surely help us to avoid many smaller and bigger issues, and save time for all participants in the sessions. Afterwards, the documents will be of great assistance in sorting out the information at the stage of creating project documentation, conducting promotional activities, writing research articles and project reports, or preparing a specification of your resource when you decide to share it.

PLAN

Below you will find a scheme for preparing a plan of your own, not a ready-made universal plan. Much depends on the profile of your research, on your preferences, resources, and research questions.

1. WHERE? Do you have a place for the recordings? What do you know about the place? Does it require any adjustments?
2. WHEN? Is this time good for the kind of recordings to be made? This may concern the time of day as well as the daily schedule of the speakers, so that they are not too tired or distracted.
3. WHAT? What is actually to be recorded? Dialogues? Monologues? Read texts? Theatre performances?
4. HOW MUCH? HOW MANY? How much material should be collected? From how many speakers?
5. TOOLS (WITH WHAT?) What kind of equipment is necessary? Is it robust and safely stored and transported?
6. HOW? Recording procedure. Is it precisely described and flexible at the same time?
7. WHO? What kind of SPEAKERS do you need? Do you have any arrangements with the speakers? What do you need for/from them?
8. WHO? Who is going to make recordings? Sometimes it is essential to have certain skills in order to make recordings: to speak a given language, to be able to enter a certain community or place, to know the cultural background and customs of the community we want to explore. And last but not least, basic technical skills are useful if not necessary.

The purpose of the checklist is mostly to reduce memory load and to extend the group of team members who will be able to take care of preparing expeditions. A checklist helps immensely to deal quickly and effortlessly with the process of preparation. If it is well-designed, it should be easy to understand and use for any team member, including those less experienced with speech recording and fieldwork. Sometimes checklists are associated only with equipment, but they may (and should) include all the items to be collected, tested, prepared, and packed, as well as things to be done on location (setting up and testing the equipment, artefacts, recalling the recording scenario, saving backup copies).

In the blue box below, you will find a list of sections that your checklist(s) should include. Everything can be easily adjusted to the particular needs of your study. Prepare it, consult it. And if you plan a series of recording sessions, test and modify it, if necessary, at the very beginning. As always, take into

account that you may not be the only one who uses it. Some members of your team may be less acquainted with particular aspects, such as the equipment.

CHECKLIST(S)

This list may be a starting point for preparing something better adjusted to your actual needs.

1. **Recording equipment.** Check your equipment before you leave (your recorder, microphones, perhaps including stands, headphones; don't forget about batteries, memory cards, cables, power supply).
2. **Documents**. You may need some questionnaires for metadata collection or other documents, such as pre-prepared agreements for the speakers, financial documents (if the speakers are to be paid). The documents you take should include the description of your recording procedure, printed instructions for the speakers, and probably the checklist itself.
3. **Artefacts**. Sometimes your recording scenarios involve using artefacts (e.g., pictures, maybe some toys if you record children). You may also think about some small gifts for your speakers. If the recording is going to be time-consuming, consider whether there is any access to water for the speakers. If not, take some with you.
4. **Acoustics check**. You should examine the acoustics of the location and maybe also somehow improve it. Checking for electric interference may be important as well. (Read more in Chapter 2.)
5. **Speakers.** Are the arrangements with the speakers precise enough? Make sure they will find you or you find them. Are they prepared to be recorded or do they need some introduction? You may need something to immediately reward the speakers or make them more comfortable and relaxed during the recording.
6. **Arrangement.** Think in advance how to arrange the recording. Is there enough room to sit comfortably, to install the microphones, or just to safely put the recorder in a good place?
7. **Equipment setup and test on location.** Once you have connected everything, test if it works, if the power supply is properly and safely connected, if batteries have enough power, and if sound is actually recorded. If your recording setup is complex (e.g., many microphones, headphones used by speakers, multichannel recording software), you should certainly test it before you start, and listen to a sample recording (see more in Chapter 2).
8. **Safety of the recordings.** Are memory cards locked? Are there any safety copies of the recordings? Are the copies kept separately? (Storing them in one bag won't help you much if the bag is lost or stolen.)

Depending on the circumstances, it might be useful either to use an electronic version of the documents or to print out the plan and the checklist so as to use a paper copy on location (or a number of paper copies that can be also used to take notes regarding the conduct of each session).

ELICITATION TECHNIQUES

In this section, we briefly overview basic speech elicitation techniques and explain how to use them in linguistic fieldwork. Here, by "elicitation" we mean the ways in which researchers prompt or induce people to speak or communicate. Elicitation techniques may strongly influence not only the content (words and sentences produced by the speaker), but also the way of speaking as well as extra- and paralinguistic behaviour (facial expression, gesticulations, body movements).

Some elicitation techniques may give much freedom to the speaker and not limit him or her regarding the way of speaking or the topic, while others may be very "restrictive" and involve precise instructions on what the speaker should do, leaving little or no room for spontaneity (Podesva & Zsiga 2013: 176-180). This may include keeping a constant distance from the microphone, staying in a given position during the recording, or being under the influence of some additional factors (e.g., strong or weak light, temperature). In some cases, we just want our informants to speak, whatever they are inclined to say.

When designing the elicitation procedure, one should take into account the planned amount of material (e.g., the number of utterances, their duration or length), and the time needed to acquire it: speakers can get tired or lose concentration and interest in the task. Therefore, unless the aim is to explore how fatigue influences speaking, breaks should be planned. It is important to provide drinking water.

READING

Reading has obvious advantages: Readers will say almost exactly what you want them to say, namely, what is in the written text they are handed and when they read the same text, you obtain easily comparable material. It also has some limitations. People tend to speak in a different way when reading; for example, they tend to be hyper-correct in their pronunciation, and be less fluent with a text they do not know. Moreover, if we ask them to read from a sheet of paper, they may tend to look downwards and lower their

heads, which leads to changes in the recording quality. A sheet of paper in the hands of the speaker is often a source of additional noise. It seems that using a computer display (LCD, for instance) positioned in front of the speaker is the most convenient solution. It also allows you to easily adjust the size and typeface, which can be chosen in the pilot study. Nevertheless, computer displays may also produce acoustic and electrical noise that may be captured by the microphone. The medium of presentation might depend on the recording environment, the group of speakers and the scenario. Sometimes even a cell phone will be sufficient. In any case, it should be taken into account that the method of text presentation may have some impact on the manner of reading (speaking) (e.g., Giles & Coupland 1991).

Reading lists are especially common in phonetic and phonological research. The lists may contain not only words, but also word-like units (pseudo-words) that follow (or not) the phonotactic rules of the language, some word parts or syllables, or entire phrases or paragraphs (sometimes referred to as *paratones* in speech). Depending on the purpose, lists may contain repetitions of the same unit, usually not consecutive, but distributed through the entire reading list. A popular type of list is referred to as a *Swadesh list* (e.g., Swadesh 1955); this includes a set of "universal" vocabulary and, when translated, is often used to track relatedness between languages.

Reading "plain" texts. Speakers are often asked to read portions of prose or poetry, or texts that are specifically adjusted to the requirements of the research. For some purposes, texts are designed from scratch in order to control their structure and content at multiple levels. If the recorded texts are to be used in listening comprehension and memory testing, we may need a few texts with the same structure, built of words of similar frequency.

"Special speech". For certain purposes, one may need recordings of "special" speech realizations. In the case of read speech, for example, the text to be read may be gradually shown in portions to the reader, may be presented in particular colours, sizes or even locations on the display, or may be accompanied by additional stimuli like sounds or images. We may want to record the spoken utterances (read or spontaneous) of people who are tired after physical exercise or have consumed psychoactive substances (this kind of project undoubtedly requires special agreements and often medical supervision).

FIRST/LAST ITEM EFFECT

Most readers tend to read the final item from a list of words or phrases with a different ("closing") melody. One method to overcome this is simply to add an additional item to the end of the reading list – an item which will not be analysed, but will prevent distortion of the pronunciation of our actual last item of interest by making it the penultimate one. Of course, the additional item should not be very different from the others in the list, to avoid surprise when the reader sees it.

The first few items on the list may also be read in a slightly different manner. The speaker may want to test his/her voice, tempo or pitch, making sure that it sounds appropriate. Even though this effect may be very weak, it is not difficult to start the list with a few "dummy items" to give readers space to test their voices.

Another method to deal with such phenomana may be nesting target words in the clausal or sentential frames (e.g., "She recently bought a new + TARGET WORD"). It should be remembered, however, that each frame may impose a certain melody on the target word as well.

Note that these phenomena may occur in the main reading session even if we start with a sound-check or "pre-recording" session to provide a warm-up for our speakers.

REACTION (SPEAKERS REACT TO A STIMULUS OR A SERIES OF STIMULI)

This is especially common in psycholinguistic research. Participants in experiments are presented with precisely adjusted stimuli and are asked to react by saying something. The stimuli may be, for example, questions to be answered, utterances to be completed, or sounds or images to be named. These approaches to speech data collection are not rare in fieldwork either. The category also includes naming people in pictures or video recordings, or naming songs or objects. Naming scenarios can be particularly useful when we record speech in a less well-documented language and wish to elicit names of specific objects, but prefer to avoid translation from some other language. However, many abstract notions may be difficult to depict in graphics, and the same image may be named in different ways by different speakers.

Here we often need special equipment and/or software for presentation of the stimuli. Software like E-Prime (commercial) (Schneider 2012), Open Sesame (Mathôt 2012), or PEBL (Mueller & Piper, 2014) (free) may be of immense help in preparing the presentation of stimuli, as well as in performing necessary reaction measurements.

INTERACTION (interaction with speakers or among speakers is planned in order to encourage them to speak)

Scenario-based interaction. Providing interlocutors with a scenario for their interaction may sound like depriving them of spontaneity and freedom of conversation. This is true to some extent: they are given a topic, a goal, sometimes even suggestions on how they should interact. On the other hand, a well-designed task may evoke lively, engaging, and expressive conversation which will also meet our requirements regarding content and speaker behaviour. The Map Task is probably the best known and most widely used dialogue task (e.g., Koiso et al. 1998; Grabe & Post 2002; Makarova & Petrushin 2003; Karpiński 2007; Hedeland & Schmidt 2012). Two speakers are given maps of the same area, but a trail is represented on one map only. The speaker with this map is asked to guide his/her conversational partner along the trail, so that the latter can copy it and draw the same path on his/her copy. Sometimes the maps are designed to differ slightly, to make the task more complex and challenging. Among other scenarios we have used in our projects are origami (one person tries to reconstruct a paper figure seen only by the other person) (Jarmołowicz-Nowikow & Karpiński 2011) and tower-building (two participants build a tower using imaginary blocks and try to remember the entire construction, which grows solely in their minds). Tasks can be collaborative, where the participants have the same goal, or competitive, when they must interact, but have different goals. Example scenarios can be inspected in the description of the Paralingua corpus, which describes several conditions for recordings (Klessa et al. 2013).

Interview (controlled interaction). A good journalist is always well prepared for an interview, not only in terms of the list of questions to be asked, but also in terms of knowledge about the people to whom she/he intends to talk. The same applies to a linguist doing fieldwork. Some knowledge on the background and history of the speakers, as well as their cultural identity, may be of importance to maintain respect and delicacy while keeping the conversation going (Codó 2008; Talmy 2010; Decker & Nycz 2013). It is good to start with a relaxing conversation as a warm-up and, if there are any more serious topics to discuss, to build a closer, more intimate relationship with speakers before addressing sensitive issues. Sometimes it is good to take into account that some team members may be better suited than others as interviewers for certain types of speakers. This may be just a matter of personalities, shared interests or some other factors. It is not bad for an interview gradually to turn into a spontaneous talk, as long as you can achieve what you plan with the recordings (Mann 2011).

In certain situations, one may want the speakers to produce longer speeches or tell certain stories. This goal can be achieved gradually, by starting with

a regular interview, giving more and more space to the speaker, or by simply telling them at the very beginning what they are expected to do. In any case, we may consciously support the speaker when she/he lacks words, or – if this is our strategy – we may leave them to seek their own ways of expression. With narratives, much depends on the personality of the speaker. Some people are able to speak fluently for hours, while others need some support, or at least feedback from listeners. Another problem in recording narratives may be the behaviour of the speakers when they are no longer constrained or directed. It is even more difficult when we need the same story to be told by a few different people in a relatively similar way. In sum, it turns out that while recording narratives may seem to be easy and fun, in fact it requires the most experience and sometimes also specific personality traits on the part of the person making the recording.

RECORDING PROCEDURE

Here is a scheme you can use to build your own recording procedure (for more details, refer to Chapter 2).

1. **Meet the speakers**, introduce yourself and do everything to make them feel comfortable (unless you actually want to record distressed speakers for some reason).
2. Ask them to sign **agreements** for the recordings and to complete **questionnaires** (see Chapter 5) Note: Sometimes it is better to interview speakers after the recordings, as the interviews may additionally distress them or give some hints on what is going to happen during the recording when this is not intended to be revealed.
3. Give them **instructions** if necessary.
4. **Start the recording**. You may announce that you are starting the recording but, if possible, you may still have some free conversation while the recorder is on.
5. **Monitor the recording.** It is not easy to do this while you are interviewing people, but if there are more people in the team, one of you should have the headphones on to listen to what is being recorded.
6. **Finish the recording**. Again, you may announce that you are going to switch the recorder off.
7. **Release the participants** if they are physically attached to your equipment, for example by headphone or head-worn microphones.
8. Express your gratitude to the speakers and **interview** them (if you have not done this earlier). Let them sign any additional documents.
9. **Debrief them**
10. Let them leave or leave them (when you are the guest) in a **good mood**. Some sweets, fruits, or a small gift may be appropriate.
11. **Secure the recordings :-)**

A different kind of longer elicited speech consists of utterances obtained within a scenario based on retelling a given story. For example, speakers can be shown a (silent) movie, a comic book or a picture series and be asked to retell the story in their own way. Some stories, such as *The Pear Story* (Chafe ed. 1980) have already been retold in very many different languages, and the text thus collected can be compared and analysed for various aspects of grammar, vocabulary, phonetics, or prosody.

SPEAKERS

Speakers most often have to be (pre)selected and, in principle, the process is determined by what is to be achieved in your study, and what your research questions are. You may need only female speakers or only young ones, or maybe only those who come from a certain region, belong to a certain culture or subculture, and are older than a certain age (Labov & Boberg 2008; Buchstaller & Khattab 2013; Buchstaller & Alvanides 2013). Sometimes you can hardly apply any conditions – for example, because you have access only to a few speakers of a given language or people suffering from a certain medical condition that influences language use, and you do your best to record them all. Sometimes you may have a huge group of potential speakers even after you apply all the criteria. In such a case, random preselection may be a solution. But, again, there may be some additional conditions to meet. For example, you may need equal numbers of female and male or young and elderly speakers. In such situations, you may use a random procedure within groups. And, again, this may be problematic when the available groups are of very different sizes. For instance, there may be only 5 male and 50 female speakers of a given language left. If you take all the male speakers and an equal number of female speakers (i.e., five of them), the first sample will be equal to the whole population of male speakers, while the second will just be (hopefully!) a representative sample of the female population. Note that in the domain of speech corpora and many other research areas, representativeness is essential. The sample that you explore should represent, at least in terms of some relevant features, the entire population (e.g., Biber 1993; Sankoff 2008; Raineri & Debras 2019). This may apply to the speakers, but also to the recorded material. Sometimes, however, you can be more flexible and focus just on looking for some interesting phenomena in speech or some peculiar communicative behaviour, and describe it as part of qualitative research, without reaching the stage of hypothesis testing.

Age as a factor. The age of our speakers may be, on one hand, a criterion for choosing them (e.g., we record young children) or something found or given

(e.g., when we record an endangered language and find that all the speakers are elderly). The age of the speakers determines our approach to the recording procedure. Cognitive abilities change during a person's lifespan. Instructions or tasks for adults and children may need to be different because of their different levels of cognitive and language skills. Moreover, the organizational arrangements for children may be completely different. Depending on local law and regulations, as well as culture-specific customs, children may be required to be accompanied by an adult. We may need special agreements from their parents or guardians. Finally, the recorded material itself may prove to be completely different because of the age factor. The acoustic quality of a person's voice changes significantly with age. Elderly people may have weaker voices, less stable phonation and less precise articulation, while children may have problems with amplitude and breath control (e.g., Hawkins & Midgley 2005; Harrington et al. 2007; Walker et al. 1992; Walker & Archibald 2006). Changes in voice may indicate certain serious medical conditions; you may suggest to your speakers that they see a doctor if you suspect something. The age of the speakers may also influence voice quality (creaky voice in elderly people or squeaky voice in some children), speaking style, and genre (small children may not be able to build long narratives).

Many other, sometimes less obvious, culture- or gender-related factors require similar sensitivity or flexibility from field linguists. This applies not only to the recording procedure itself, but also to the way we arrange it, how we contact and address speakers, or how we gather metadata.

SPEAKERS' METADATA

Information about speakers, along with information on the recording location or equipment, is considered as metadata. Information on speakers' gender, age, dialect spoken, or place of birth may be useful in organizing our archive. On the other hand, this kind of information can also be used as 'direct' data, depending on the purpose of our study. For example, we may regard age information as peripheral in studies of certain kinds of phonetic phenomena, while in others, we will use age as a significant factor. The latter situation happens quite often. Therefore, when we have our (only) opportunity to interview our speakers, we should do so in depth (see also Chapter 5).

PREPARING FOR RECORDINGS

We influence speakers from the very first contact, and it is our behaviour that shapes their attitudes towards the recording session and ourselves. The recording procedure, therefore, may include instructions on how to deal with speakers before we press the red button or invite them into the recording booth (if we have a mobile one).

Some potential speakers may withdraw their cooperation when they learn that the recordings will be made public or that they will be listened to by many people during our experiments. Sometimes, even if they agree to proceed, they may be quite distressed and speak in a different voice and manner than normally. In some cases, it helps to explain everything in more detail, e.g. "You will be one of two hundred speakers in our database. It's hours and hours of recordings"; "Your speech will be cut into very short pieces and your voice will be almost impossible to recognize"; "The database will be used by professionals – they won't focus on what you were saying, they will just extract some parameters to make calculations for speech synthesis." Some speakers may appreciate the fact that they are taking part in an "exclusive" project and that their voice will be saved for future generations, while others may feel even more distressed by the "weight of responsibility". Once we know the potential speaker better, it might be easier to explain the situation to them and find convincing arguments. Let us emphasize (although this should be absolutely obvious) that it is not acceptable in any event to trick or mislead the speakers. We are dealing with people and often with sensitive, emotional topics, and building trust between us and the speakers is of great importance. Even though some recording scenarios assume a kind of "misleading information" (see below) in order to obtain a specific outcome, in the end, everything needs to be explained to the speaker, and it is the experimenter's responsibility to make sure that the participant does not feel any inconvenience after the experiment.

While sometimes it is essential to give detailed instructions on what to say and how to speak, in many situations you just want to encourage spontaneous, free speech. And in such a case, it is important to stress that you are interested in how people normally speak and there is no such thing as a better or worse way of speaking: they should remain themselves and should not care about any mistakes.

DECEPTION AND DEBRIEFING

In many research contexts we avoid sharing full information on our intents and purposes with the participants before the experiments, because to do so

might thwart the aim of the study. Participants may start to act according to a strategy devised in response to what they know or imagine about the study, in order to "do better". Many of them, regardless of the type of recording, ask afterwards whether they were good enough. Except in a rather narrow range of studies, our speakers are usually "good enough" (or simply excellent) because we are interested how THEY speak, with all the failures, mistakes, disfluencies, and other peculiarities. We should let them know that they are exactly what we have been looking for.

It is customary to **debrief** participants immediately or shortly after the recording session (Holmes 1976; Brody et al. 2000). If everything was obvious and overt from the beginning, we may just convey brief information on what we are going to do with the recordings (even though speakers should know this from the agreements they have signed). If the recording scenario included any kind of hidden goal or planned deception, we should explain what was concealed and what was the actual idea behind what we were doing. Some participants may express disappointment with themselves: "If only I had known…!" But we need to reassure them that none of the participants knew and there is no reason for misgivings, because we take people as they are. Although this may not actually be quite true at the stage of recruitment, once they meet our conditions, all participants are treated the same way. It is very important to avoid being judgemental at this stage. While in our study we may distinguish, for example, between fast and slow readers, it should be clear that "fast" is not better than "slow" or the other way round. It is just different, and that is what is interesting for us to study.

HUMAN FACTOR

You will deal with humans. Some of them may have a bad day. Not all of them are outgoing extroverts. Even if they are willing and interested to participate, you may need to try and convince some of them to speak into the microphone in a certain way, and this may take time.

LEGAL AND ETHICAL ISSUES

Legal and ethical issues should be taken into account from the very beginning, that is, at the stage of the design of the study (e.g., Lehmberg et al. 2008; Rice 2012; Eckert 2013; Mallinson 2018). Detailed regulations vary from

country to country, and local law and customs may also be of importance. In general, you should pay attention to the following aspects:

1. Are your speakers aware that they will be recorded? Do they accept it? Are they of legal age and can they decide themselves about participating in your recordings?
2. What kind of use and publication or dissemination of the recordings would they accept? Can you make the recordings public?
3. If the speakers are not inclined to allow publication of the recordings, maybe they will agree to the use of some selected portions for anonymous analyses (publications, conference presentations, etc.). If not, try to obtain agreement for the use of authorized transcripts. Offer further anonymization of the recordings and sensitive metadata.
4. Under some circumstances it is better to deal with the documents (detailed questionnaires, metadata forms, etc.) after the recording. However, you need to obtain formal consent for the recordings before you start the first session.
5. If you record elderly people or children who may not be fully conscious of what the consequences may be, the problem becomes quite complex. To work with children, you may need (signed) agreements from their parents, teachers or guardians, and also from the children themselves. Sometimes you may need a formal agreement from institutions that act in a given country to prevent child abuse. Even when you deal with small children who may not be fully conscious of the meaning of the agreement, make sure they feel comfortable when being recorded, and if not, just end the session.
6. You will probably be in possession not only of the recordings but also of some other personal (sensitive) data. You will need agreements from the speakers regarding these data. Even if you promise to code and anonymize the data, ways of doing this may vary and may provide different levels of safety or anonymity. Be ready to explain this to the speakers in simple words.
7. You may want to archive the recordings and metadata for a prolonged period of time, maybe for future generations. Be sure that you are allowed to do so and to make it clear how the data should be treated (select or formulate an adequate licence, make additional comments or instructions for future users). In principle, it is essential to decide who is and who will be the owner of the data, and what are the rights of the owner. In some countries and organizations there are legal restrictions on the time for which data may be stored, which may be independent of what the speakers themselves declare.
8. The actual text of a legal agreement/consent form may be very complex. Be prepared to explain or summarize it to the participant. Make sure that at least the main questions are formulated in a very straightforward way,

e.g. “Do you agree that your voice will be recorded during the entire interview?”, “Do you agree that we keep the recordings for future studies?”, “Do you agree to make the recordings public?”

A number of useful examples and guidelines for creating the consent forms are shared by researchers and institutions. For example, Hannesschläger et al. (2020) propose an online creator of consent forms: DARIAH ELDAH Consent Form Wizard available at: https://consent.dariah.eu/ The wizard takes into account aspects of sensitive data protection formulated in the General Data Protection Regulation (*GDPR*, cf. e.g., Nautsch et al. 2019). The DELAD research group (Database Enterprise for Language And speech Disorders; Lee et al. 2021) publishes example GDPR-compliant consent forms from different institutions at: www.delad.net. The forms include examples for clinical data that are especially sensitive as they may include health information and other personal details of the speakers.

The recording procedure should meet all the legal conditions which apply to working with people and collecting personal data. In principle, it should be understood by each participant in your recordings that she/he can quit at any moment. On the other hand, you may always try to convince her/him to stay. You should avoid any pressure on the speakers (unless it is a part of the scenario, and you have their agreement to behave like that). Touching them, getting too close, speaking to them too loudly or harshly, even if legal, may be destructive to your relationship with the speakers and spoil the recordings.

It may be the case that the presence of parents may somehow sooth the child during the session, and put you in a safer situation as you are not the only person responsible for the child at that time. Although much depends on the age of the child and the recording situation, our experience shows that the presence of parents or guardians may not in fact always be favourable. Children often act to meet parents’ expectations, and turn to look at them to seek acceptance for their actions. If the presence of a child’s parents is for some reason necessary or recommended, a solution might be that they stay nearby, in the same or a neighbouring room, and read newspapers or books, not paying (or at least pretending not to pay) too much attention to the recording session.

RIGOUR AND FLEXIBILITY

Planning itself is a tedious process that requires methodological, technical and organizational skills and experience. Detailed planning should involve a walk through all the stages of the experimental procedures, a simulation of potential challenges or issues. It is absolutely essential that the session docu-

ments (plan, checklists, procedures) can be consulted at any moment by any of the team members; that is, they should be formally prepared and (preferably) easily accessible in both electronic and printed versions.

Last but not least: It would be ideal to adhere rigorously to the procedures, while simultaneously remaining spontaneous, relaxed, open and flexible in communication with the speakers. Keeping a reasonable balance between the two goals is a valuable skill that may be achieved with experience gained over numerous projects, and based on respectful relationships between experimenters, speakers, and language communities.

2. RECORDING

Linguists often look for a set of concise instructions or hints on the technicalities of field recording. Sometimes they do this shortly before leaving for fieldwork or immediately after they have made some recordings. However, it may happen at that point that some necessary equipment is missing, the procedure is not clear enough, or even worse: something has already gone wrong with the sound during the recording session. In other words, frequently, questions about technicalities are asked too late. In this chapter we will briefly explain some of the technical and organizational aspects of speech recording.

Nowadays, sound recording equipment that meets most linguistic and phonetic requirements is relatively cheap, readily available, and easy to operate. There is a wide choice of technical solutions on offer, but one should be aware that they may be better or worse suited to the purpose of a given study (e.g., Campbel 2002; Vogel & Morgan 2009; Podesva & Zsiga 2013; Barsties & De Bodt 2015). Therefore, we believe that some technical knowledge on what is possible – and how – can be of great benefit to researchers, even if they have experienced technicians in their teams. Such knowledge may be essential in the very early stages of planning a research project, throughout the recording sessions, and also when the recordings are already stored in an archive.

Below you will find a brief overview of topics and issues to which you should probably give a thought at the stage of research project design. They include the characteristics of speech material, the settings (environment) and its potential adjustments, the design and realization of the recording procedure (which involves setting up your equipment), as well as some hints regarding metadata (which are discussed in more detail in Chapter 6).

> The technical and organizational aspects of speech recording should be part of your research project design **from the very beginning**, as they may influence or even determine your approach to planning.

WHAT IS TO BE RECORDED

To prepare for field recordings, it is essential to take a closer look at what is to be recorded and what potential challenges are associated with the particular kind of recordings. It may be necessary to answer the following questions:

– **How many speakers do you need** to (or can you) record?

– **How long** will the recording(s) be? How much time will each session take?
– Will you record a **single speaker** or **multiple speakers** at a time? Monologues? Dialogues? Polylogues?
– What **degree of spontaneity** do you expect from the speakers?
– **How** will the people (probably) speak? Do you expect people to whisper, scream, speak very loudly or quietly? This may be important for technical reasons.
– How is the session going to be **organized**? For example, are there any breaks planned? Do any changes need to be made to the recording setup during the session?

All these (and more) factors may be important for the design of the recording procedure and audio equipment setup. For example, with multiple speakers (when recording discussions) you may want to use more microphones and even try to somehow isolate the speakers from each other acoustically. For spontaneous speech, you will probably be more cautious with the sensitivity level, as speakers may tend to speak louder or quieter at different points, which can result in distorted recordings. Moreover, the aforementioned factors may also influence the way you process and archive your recordings. You may want to cut structured sessions into a number of separate audio files, for example, according to the topics you introduced as the animator of the discussion or the questions you asked the speakers. If you expect people to both whisper and shout during the same session, you may consider using condenser microphones, which can deal with such sound dynamics.

SETTING (RECORDING ENVIRONMENT)

There are good reasons to record people in their natural, everyday environment, and sometimes there are no other options. On most occasions, however, this comes with certain acoustic and organizational challenges. Even if the circumstances are adverse – for example, we find that the place is noisy, has poor acoustics, and is inconvenient for operating the recording equipment – it is nearly always possible to make some small adjustments that significantly improve the comfort and quality of the recordings.

If possible, test the room's acoustics. Place the microphone exactly where it will be located during the recordings, run the recorder, set the sensitivity to a high level, and monitor the signal. Listen for background noises (such as a fridge, a clock, or trams and cars outside) and try to eliminate them if possible. Close the window, move the clock to another room, and so on.

– **Good space for recordings**: furniture, shelves filled with books on the wall, heavy curtains, carpets, soft floor, relatively high ceiling.

– **Poor space for recordings**: empty walls, little furniture, hard floor with no carpet.

Then listen to the quality of the sound of speech: Is it natural or coloured? Can you hear echoes or, in general, sound reflections? Sound reflections (reverberation) may be partially or almost completely eliminated by hanging heavy, thick fabrics, blankets, duvets, or just winter clothes around the speaker and the microphone. Smaller rooms have lower potential for reverberation, but if it occurs, it may still cause issues if we are unlucky and, for example, the phase of the signal reaching the microphone from the speaker's mouth is opposite to the one reflected from the walls.

It is much easier to stop high frequency noises. You can experience it when listening to music through a folded blanket. Low frequencies are still perceivable, while the high ones tend to be significantly reduced. To isolate a room from low frequencies coming from outside, one needs very thick acoustic screens – something rather impracticable when doing fieldwork.

To sum up, you may want to deal with at least two factors when trying to improve the recording location acoustically: (a) noise, both internal and external; (b) the acoustic characteristics of the room. Certain actions may help to simultaneously solve problems in both of these categories.

AN "AUDITIVE LOOK" AT THE ROOM

Connect headphones (and microphones if needed) to your recorder, position the microphone for recording, and try to listen to the room using headphones. It may help to increase the sensitivity beyond the regular level. In this way, you will easily hear sounds that may be irritating for listeners to your recordings, but normally escape your attention. The sounds of an old refrigerator, a faulty air conditioner, an oven, a fan, a squeaky floor, windows or doors are frequent components of home recordings.

Recording equipment is sensitive not only to acoustic noise collected via microphones. The electric fields and radiation that are present almost everywhere may also cause problems, as they are picked up by cables and other items of equipment. When **recording in old houses with old electrical installation**, it is highly recommended to test the quality of electric current. If a relatively low buzzing noise is heard in the headphones attached to your recorder, try to disconnect it from the power supply and switch to batteries. If this does not help, make sure that the

recorder and microphone cables are not too close to any power supply cables, and even to walls where electric cables may be hidden. Noise from a home electrical network may actually be caused by old or defective home appliances, such as a fridge, a washing machine, or an air conditioner. If possible, switch them off. If no other solution is possible, you may use power supply filters, which help to eliminate or at least significantly supress such noises. However, these are often very expensive and heavy. The cheapest solution may be to use battery-operated equipment and high-quality symmetrical shielded cables. Mobile phones, when placed too close to the recording equipment, may also be a source of interference. Therefore, in any case, keep them far away from the recorder, cables, and, obviously, from the microphones. A conventional telephone with a wireless handset may also produce electric noise that can be captured by your recording equipment.

When recording in the **open air**, you can hardly influence the acoustics of the environment (insects, birds, wind, flying planes, passing cars, etc.). For such a recording environment, make sure to have a microphone windscreen, also called a "dead cat" or "wind muff", designed to minimize or eliminate such noise. Another useful piece of equipment will be a small acoustic screen that partially surrounds the microphone, reducing the sounds coming from any other side than the speaker's. It may also be reasonable to use a **unidirectional or even shotgun microphone**, which is significantly less sensitive to sounds coming from any other direction than the microphone's axis. All of these pieces of equipment are described below in more detail.

EQUIPMENT AND HOW TO USE IT

Questions relating to equipment are important, but this importance is sometimes overestimated. Often much more can be achieved by preparing the room appropriately for recordings and by setting up the microphones correctly than by buying more expensive equipment. This is especially true in the case of field recordings. When buying equipment, do not focus solely on its recording parameters – take a look at the build quality and robustness, which may be of great importance in the field. Below is a rudimentary description of the most important components of the field recording setup.

PORTABLE DIGITAL AUDIO RECORDER

Nowadays you can buy a reliable hand-held digital audio recorder with good-quality internal microphones for a very decent price. Most semi-profes-

sional models offer more than enough for speech recording, in terms of both recording quality and options. When buying a digital audio recorder, you may want to take the following parameters into consideration:

- **uncompressed** recording capability – may mean higher recording quality (some recorders use only compressed mp3 files; see below for more details)
- **high-quality built-in microphones** (preferably condenser microphones)
- **inputs for external microphones** supplying phantom power (which can feed condenser microphones)
- optional **manual sensitivity adjustment** (sliders or knobs); using automatic recording level adjustment can result in phonetically useless recordings
- **powered by easily replaceable standard batteries** (you may want to avoid keeping speakers waiting for a few hours while you recharge your device)
- **accessories included in the package** and those available on the market (windscreens, additional microphones, bags or cases, etc.); more expensive models tend to have a wider range of accessories available for later purchase
- **humidity/moisture protection, dust protection**
- **optional remote control –** especially important when you rely on built-in microphones and you want to avoid reaching for the unit every time you need to stop or re-start recording
- number of **simultaneously recorded tracks**; some mobile recorders offer four or even more tracks that can be fed by internal or external microphones – a useful feature if you intend to record a few speakers via separate microphones
- **simple and comfortable operation** – large buttons and knobs, most important functionalities easily accessible (not via complex menus)
- last but not least: **overall build quality**; some recorders are clearly built to last: you can easily feel that they are made of high-quality materials, and this is often reflected in the price.

Among dozens of functionalities offered by digital recorders, **markers** are especially useful in field recording. By pressing a button, one may add markers during the process of recording in order to highlight certain events (e.g., the speaker got very emotional or mentioned an important fact) or to mark stages of the session (e.g., here he was talking about his home, here about his job). This can save much time later, at the stage of editing and description of the material.

DIGITAL AUDIO RECORDER TECH TALK

Voice activated recording – some voice recorders feature a system that automatically starts recording when the sound level exceeds a certain value. This is rarely useful in our area of interest! Normally, be sure to switch it off before the session.

Preamplifier – the electric signal coming from the microphone is weak and must be amplified before further processing and digitalization. "Preamps" determine or at least have a great influence on the overall sound quality. Some of them are well-known in the music industry for their unique, "beautiful" sound. But what linguists expect from the recordings is just that they should be neutral and transparent.

Limiter – sometimes useful, sometimes undesirable. When the sound amplitude reaches high levels, the limiter reduces it so that it never exceeds the maximum safe level and you never have any overload. But the original dynamics may be reduced and distorted.

Compressor – changes the dynamics of the signal, applying less amplification when the amplitude is higher and less when it is lower. As a result, the recording is more consistent in terms of amplitude: the range from the lowest to the highest levels is reduced. Obviously, it distorts the original dynamics of the signal and, in principle, should be avoided when recording speech. (See the note on *compression* below!)

Automatic recording level control – may work as a limiter or a compressor and increase recording level (sensitivity) when the incoming sound is too quiet. This feature, although tempting in some circumstances, deforms the amplitudinal characteristics of your recordings, making them useless for a range of phonetic analysis.

High pass filter – some recorders have filters that eliminate or damp lower frequencies. This may be useful for outdoor recording and in some other conditions, but remember that it will also eliminate a part of the speech signal.

Mark-up – adding markers (time stamps) to your recording. This may be very useful. You just press a certain button whenever you hear something of interest in what is spoken, and this will let you find those moments immediately in the recording. It works just like "live" or "real time" tagging.

Sensitivity – how sensitive the microphone input is; set it manually by testing the voice to be recorded and the position of the microphone.

Sampling frequency – how many measurements of signal amplitude are made in one second. The more the better, but 44.1 kHz is usually more than enough for linguistic applications (some other typical values are 22.05 kHz, 88.2 kHz and 96 kHz; some speech technology tools, e.g., automatic segmentation software, use 16 kHz).

Sampling resolution – how precisely the measurement of amplitude is made and coded: how many bits/bytes are allotted to each sample, e.g., 8-bit (rather low quality), 16-bit (CD standard), 24-bit or 32-bit (increasingly often used). With more bits per sample you can cover higher dynamics of the signal (capture quiet and loud portions of sound without much problem).

Bitrate – refers to the number of bytes of information allotted to the recording of each second of the signal (e.g., 128-kbit or 192-kbit per second). Some compression algorithms are intelligent enough to detect that certain portions of the signal are monotonous while others are more complex, requiring more bits for coding, and reduce or increase the bitrate respectively (adaptive bitrate). The bitrate can also be considered in the case of regular, uncompressed PCM recordings.

Note that the term *compression* is used in two very different meanings. The first refers to reduction of the dynamic range of the signal, while the second refers to reduction of the data used to represent a portion of the signal (see also RECORDER TECH TALK above, and Chapter 3 on audio signal processing). An audio signal is transformed by analogue-to-digital (A/D) converters into a digital form and stored as audio files. There are many audio file formats in use, including WAV, AIFF, and MP3. The technologies used for A/D conversion may differ in some respects. Most often, they are based on very frequent amplitude measurements. The frequency and precision of these measurements contribute to the quality of the recording. The initial result of conversion may be transformed on-the-fly or afterwards by applying compression, that is, reduction of the amount of the data representing a unit time of recording. For example, the widely used MP3 format (short for MPEG-1/MPEG-2 Audio Layer 3) is a compressed format in which the user may decide how much space should be saved, for example, by selecting the bitrate. (Some equipment may offer only fixed parameters.) In principle, using compressed formats is not recommended in phonetics, as you can lose some important components of the signal. Accordingly, among the most popular, WAV and AIFF are normally recommended. More on file formats can be found in Chapter 3 (for technical data, you can also check this link: audio file formats).

MICROPHONES

Two basic types of microphones can be distinguished from a practical point of view: external and internal (built-in). A significant step towards

higher-quality recordings is often the use of external microphones. Internal microphones in portable recorders tend to have certain limitations when it comes to capturing lower frequencies; they may also offer lower sensitivity than full-size, external microphones. If you use built-in microphones, operating the recorder becomes less convenient. You may need to remove it from the stand or at least reach it to replace the card or change the sensitivity level. If you operate it during recording, all the noises you produce by turning knobs or pressing buttons may be recorded as well. External microphones – hand-held or on stands – offer much more flexibility. The quality of the microphone (often simply referred to as the "mike") is in most cases more important than the quality of the recorder itself. A good external microphone may raise the recording quality to a new level.

If you record people in a room, a condenser microphone (Rumsey & McCormick 2006: 45-46; Boré & Peus 1999: 32-39) may be a good choice. It is delicate, and requires a so-called phantom power supply, but it is also very sensitive and has a very wide frequency response, which means that it can often deal better with lower frequencies than other types of microphones. Since condenser microphones are very sensitive to pressure changes, a door closing or a window being opened will be surprisingly audible in your recording. As they are also sensitive to shakes and hits, they require **shock mounts** to isolate them from what can be transmitted by the microphone stand from the surface on which it is placed (for example, the steps of a person passing by). Large-membrane condenser microphones require solid, heavy stands, and are not designed to be held in the hands. However, you can also buy a small-membrane condenser mike that looks just like a dynamic microphone, while still working on the condenser principle. Due to the smaller membranes, they often perform slightly worse with low frequencies, but since the membrane is lighter, the overall precision of recording may be higher. However, this may be of importance only if you decide to work on acoustic details. A practical issue important for fieldworkers is that using phantom power means higher consumption of electric current. If your equipment is running on batteries, they may run down earlier than with other types of microphones.

If you want to make real field recordings (in the open air, or in a noisy environment), a condenser microphone may be too sensitive. In such cases you may prefer to choose a dynamic microphone (Rumsey & McCormick 2006: 41-42). Typically used as vocal stage microphones, these are usually rugged and sturdy, often solid like tanks. They do not need an external power supply, but are less sensitive and have a narrower band of efficiently transduced frequencies (typically 150 Hz to 15 kHz, compared with the range 20 Hz to 20 kHz typical of large-membrane condenser microphones). However, as

pointed out above, they may in fact be better for field recordings, as they are less sensitive to wind, are not as delicate as condenser mikes, and do not run down batteries as they do not require a phantom power supply.

For some applications, lavalier microphones, attached to the speaker's clothing or head, are highly recommended. Many of them are electret microphones, which work on a similar principle to condenser mikes, but do not need phantom power. Lavalier microphones are often tiny and may not be perfect for capturing the lower end of the spectrum. However, as they stay a constant distance from the speaker's mouth and relatively far from other speakers, they may be especially good for simultaneous recording of multiple speakers.

Condenser or dynamic microphones are also available as **head-worn** models. While these are very practical in that they remain at a fixed distance from the speaker's mouth and do not touch the clothing, they may be somewhat difficult to use. Few speakers find it easy and natural to be spontaneous and relaxed with appliances attached to their heads and fixed just by their mouths (these mikes are not placed directly in front of the mouth, but to one side).

CONDENSER OR DYNAMIC?

When you work in the (real) field, in changeable conditions, in the open air, in low or high temperatures, etc., a regular dynamic microphone should be your choice. It is lighter and more robust. For recording in homes, libraries, and other relatively quiet indoor locations, a condenser microphone may be a better choice.

Where to place the microphone? In a studio, you just place the microphone at a recommended distance from the mouth of the speaker, usually almost straight in front of him or her (although there are some exceptions). A distance of about 20–30 cm is perfect in most situations. In some cases, you may want to reduce this distance (for example, when the speaker is very quiet, or you are recording whispered speech) (see, e.g., Corbett 2021: 216-219; Rayburn 2012: 366). In the field, when recording free speech, spontaneous conversations, and emotional monologues, one can hardly tell the speakers to remain a fixed distance from the microphone during the entire session. This is one of the reasons to use **more microphones**. They will help to capture what the speaker says even if she/he turns left or right. You may also consider using lavalier microphones, attached in most cases to the speakers' clothing. However, it may happen that there is nothing you can attach the microphone to or clothing produces noises with each movement of the speaker.

Another reason to use multiple microphones is when you need to have each speaker recorded on a separate track, acoustically isolated from the others. Perfect isolation is impossible if speakers are not placed in separate acoustically isolated rooms. But even partial isolation, achieved just by distance between the speakers (using individual microphones), may help in many situations, for example, when there is overlapping speech and it is difficult to comprehend. Use lavalier or head-worn microphones and try to reduce their sensitivity, seat the speakers as far as possible from each other, and place some acoustic obstacles between them (for example, if they are sitting at a table, you may place piles of books or magazines on it). Even flowers may help, as long as they do not reduce mutual visibility (unless this is what you want to achieve). If regular tripod-mounted microphones are used, small acoustic screens may be beneficial (see below). You may also switch the microphone to a more unidirectional setting, if such an option is available. Another reason to use more microphones is to implement a backup strategy – to ensure that even if you have clipping or other issues on one of the mikes, the others will capture a clean signal. One microphone may be placed closer to the speaker, and another further away; then even if the closer one is overloaded, the further one can still handle the signal. When speech is quiet, the more distant microphone may capture it only to a limited extent, while the closer one captures it successfully. However, this solution is rarely used nowadays, when microphones with extremely wide dynamic ranges are available.

For more technical information on choosing and using microphones, please refer to professional literature on sound recording (e.g., Pawera 2010; Rayburn 2012; Corbett, 2021) and materials provided by the manufacturers.

ACCESSORIES

Most portable recorders can be easily attached to a **stand** (a tripod, a table stand, etc.), and such a stand may be the first accessory you want to buy. It can help to improve the quality of your recording, because (a) the hand-held recorder will be better isolated from the table (which may transmit noises easily), and (b) it can be directed and positioned closer to the speaker's mouth. You can also use it for an **external microphone** if you decide to buy and use one. Note that high-quality microphones (especially condenser microphones) tend to be heavy and require not only strong but also heavy stands to maintain balance, which may be a disadvantage in fieldwork conditions, when portability is often preferred. Headphones should not be considered as an option – they are a must (Poldy 2001; Huber & Runstein 2018: 500-501). Before you start

MICROPHONE TECH TALK

Type – dynamic and condenser (including electret) are two major types that interest us

Sensitivity – how efficient the microphone is at transforming acoustic waves into electrical impulses

Self noise – powered microphones tend to produce noise of their own; usually it is at a very low level, but it may become relevant when you work with whispered speech and quiet speakers. Note that some cult microphones used for music recordings are not especially quiet, as this is not of primary importance in that field: singers are significantly louder than the noise level. Self noise is produced by the device and is largely independent of the input; and as it is constant, one obtains a higher signal-to-noise ratio (which is actually what one desires) (Boré & Peus 1999: 69).

Signal-to-noise ratio (S/N) is the relationship of the voltage delivered by the microphone at 1 Pa sound pressure and 1 kHz frequency to its self-noise voltage (Boré & Peus 1999: 70; Rumsey & McCormick 2006: 61)

Frequency range – the range of frequencies (expressed in Hz) that are captured by the microphone efficiently enough (precise definitions used by manufacturers may vary), for example, from 150 Hz to 15 kHz (typical of dynamic microphones) or from 20 Hz to 20 kHz (to be expected in good condenser microphones). Note that some data provided by manufacturers may be misleading. A microphone may be described as capable of capturing, for instance, the 20 Hz component of the signal. But if it works with low efficiency for the this band, low frequencies will be very quiet on the recording. More detailed information on how a given microphone reacts to various frequencies is referred to as its *frequency characteristic* or *response.* It is often represented as a graph showing how sensitive the microphone is to signals of different frequencies (and sometimes also coming from different directions), from the lowest to the highest (usually from 10 or 20 Hz to 20 or 30 kHz).

Dynamic range – the range of signal amplitudes that can be handled by the microphones, from quiet to loud sounds, given in decibels (Rumsey & McCormick 2006: 536). While speech itself does not have an especially wide dynamic range, when speakers get closer and further from the mike, additional dynamic capability may be very useful: even if the speaker leans close in to the mike, the recording won't be distorted if a microphone with a wide dynamic range is used.

Polar (directional) characteristics – this concerns the directions from which the microphone picks up sound more efficiently. Cardioid is the most popular polar, but there are also omnidirectional microphones that are equally sensitive to sound from all directions, as well as strongly unidirectional ones that focus on sounds coming from one particular direction. Polar characteristics are not binary: even unidirectional microphones collect signal from various directions, and omnidirectional microphones are rarely equally efficient for all directions. It is useful to look for the polar characteristics graph for the microphone you plan to buy, to check that it meets your requirements (Rumsey & McCormick 2006: 36-59).

recording, listen to what your microphones can hear. During the recording, check that there is no clipping or other technical problems. In principle, **closed headphones** may often be better in the field, as they isolate you from the acoustic environment and allow you to focus on what is coming in via the microphone. But what is an advantage in many situations may be difficult to accept in others. Closed headphones are not recommended for prolonged use, for higher temperatures, and for situations where you have to control the overall situation during the recording (for example, when you are alone and you are not only recording but also interviewing people at the same time). Sometimes people use closed headphones but only on one ear, leaving the other uncovered for monitoring of the environment. The important parameters of headphones include their impedance (in principle, low-impedance headphones are more "easily" fed by the amplifier and more often found in mobile equipment), the acoustic pressure they can produce, their efficiency in the transformation of electric current into sound pressure, and their frequency response (the range of frequencies they can produce).

Another useful accessory you may take into consideration is a small, portable **acoustic screen** (sometimes referred to as a reflection filter, microphone isolation filter, etc.). This helps to reduce the amount of sound reaching the microphone from directions other than the selected one where the speaker is seated Most studio-type acoustic screens are not convenient to transport; you can hardly put them in your backpack. However, you may find foldable ones. It is also not that difficult to prepare something yourself ad hoc. In a noisy or very "echoey" environment, this may significantly improve the quality of your recordings. On the other hand, if you want to use a camcorder simultaneously or to record dialogues, screens may become obstacles, preventing eye contact or reducing face visibility.

When it comes to microphone accessories, you should consider at least two more categories of items. A **pop filter (pop screen)** can be useful to slightly disperse the energy of plosions produced by speakers. It also helps to keep the microphone clean by protecting it from drops of saliva. A **dead cat** looks like a piece of fur and is put on the mike in order to limit the noises caused by the wind and other external noises (Corbett 2021: 102-104).

Cables are sometimes neglected as a part of field recording equipment. We recommend buying high-quality, reliable cables from renowned manufacturers. Cheap cables are usually not only worse in terms of electrical properties, but also get damaged easily and cannot generally be repaired. On the other hand, we do not need or want costly audiophile cables for linguistic recordings.

Note that recording equipment is rarely waterproof and can be damaged by humidity, not to mention rain. Sand and dust also have the potential to cause

damage. For some pieces of equipment you can buy waterproof and dustproof cases, but sometimes a transparent plastic bag will do.

MORE ON THE PROCEDURE

As explained in Chapter 1, a well-thought-out and precisely described procedure may dramatically increase the chances for a successful recording session, provided that it is diligently followed. We are always tempted to take shortcuts and skip some of the steps ("obviously there is enough space on the memory card", "I checked the battery level ten minutes ago, it must be okay", "people were asked once to switch their phones off, why bother them again!?"). In addition to what was discussed in Chapter 1, here we make a few more comments that apply to the recording phase itself.

HOW SHOULD I SPEAK?

Speakers are often unclear about the speaking style or "tone" they should use. To avoid unnecessary hesitations during the session, and if the aim is not to ensure total freedom in this respect, the instructions may include a hint or a precise request concerning the way of speaking, for example, "speak as you normally do" or "speak in your own, relaxed way", but also "speak as clearly as you can", or "speak as fast as you can". For some types of recordings, speakers and situations, a hint on how to speak may be indispensable. Hints like those above may also include suggestions relating to emotional aspects of speech ("Speak in a soothing, nice way").

Note that speakers are not always fully conscious what their "regular speaking style" is. They may be unable to feel or hear tension in their own voices. After a few minutes of recording in a relaxed atmosphere, they become relaxed as well, and start to speak in a different way. Accordingly, it is a good idea to have a pleasant talk before the actual recording starts. This is a sort of vocal warm-up, which is nearly always helpful. Sometimes it can be an official part of the scenario (a few articulatory exercises before the recording session).

HOW SHOULD I SIT?

Another question frequently asked by speakers concerns their position and distance from the microphone. Therefore, this information should also

be included in the instructions. We know that the voice of a sitting person differs from the voice of someone lying down or standing. While often subtle, these differences may be relevant in certain kinds of studies. If speakers are to be recorded for a long time in tasks like reading lists of words, they should certainly try to find a comfortable position on the chair, as we do not want them to move at all for a prolonged time. It is quite common for them to bow towards the microphone from time to time, or to lean back.

NOISE IS EASY TO PRODUCE AND DIFFICULT TO REMOVE

If the acoustic quality of the recordings is of highest importance, you should make the speakers aware that touching the microphone, tapping on the table, moving the sheet of paper on the table or in front of the microphone, or even making rapid movements on the chair may make the recording less useful (Figure 1). However, this kind of instruction usually distresses speakers. They may become stiff and be afraid of speaking or moving at all. Depending on the speaker (young, older, experienced or novice), you should find an appropriate way of suggesting or even training the most desirable behaviour in front of the microphone.

If the entire procedure is flexible and the recording is continuously monitored (which is highly recommended), it may be easy to ask the speaker immediately to repeat an utterance that was interrupted by noise produced by the speaker, or for example mark it for later repetition after the main session is completed.

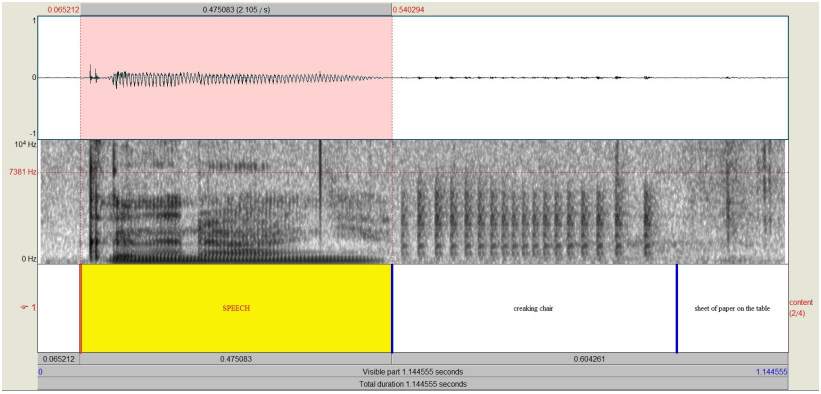

Figure 1. Speech and noises produced by the speaker (studio recording). The oscillogram, representing amplitude changes, is visible in the upper panel, the spectrogram is in the middle one, and transcriptions are the bottom panel.

A COMMENT ON AUDIO PROCESSING

You cannot get from the sound file more than is already in it. It is digital. There is a fixed number of bytes and nothing more. For linguists, the magical "audio cleaning" or "recovery" software is of limited use, usually restricted to very old recordings, as these techniques often involve adding to signals components that are not there or removing components that may prove not to be redundant. By using some of these techniques, you may improve the perceived quality of recordings and make them more intelligible, which may be especially useful when the data need to be transcribed. Do not hesitate to use them if they might help, but always keep the original recording for further analyses. In publications, always mention if you worked with pre-processed material. Sound restoration practices are common in the music industry, and many people like to listen to improved (remastered, remixed, etc.) versions of The Beatles or The Rolling Stones. In the realm of research, this kind of approach is, with few exceptions, highly questionable. (More details can be found in Chapter 3)

3. PROCESSING

Once recordings of speech have been made or found in archives, a few operations normally follow or are considered as next steps:

- **format conversion** and/or change of parameters, for example, to unify the data format and parameters in a way that will be compatible with the phonetic software being used (in the case of archives, digitalization may also be necessary, but we skip this step here as our focus is on the acquisition of new material);
- **cutting recordings** into shorter, manageable pieces, suitable for further applications;
- **processing** of the recordings to adjust them to the requirements of the planned applications (e.g., preparing them for auditive or instrumental analysis, or use as stimuli in experimental studies).

To avoid distracting the speakers, especially in field recordings, the digital recorder is often switched on at the beginning of the meeting and, if possible, the entire session is recorded as a single file. This is not always convenient for further operations or even for archiving, because the resulting file may be very large and difficult to search through, and may include material irrelevant for our purpose. Therefore, as a rule, the resulting audio files require cutting or trimming. We may also find that, in order to maintain compatibility with existing archives to which we intend to upload our material, or just to make it readable for the phonetic software being used, it is necessary to convert it to a different format (e.g., from MP3 to WAV) or change some of its parameters (such as sampling frequency or bitrate). Unfortunately, the quality of field recordings tends to be lower than we would like it to be. When the quality of the material at hand is not satisfactory, it is often tempting to "polish" or enhance the recordings, filter out unwanted noise, or reconstruct what is not there but presumably should be.

In this chapter, selected techniques of digital audio signal processing – those which seem to be most useful for linguistic material – will be briefly overviewed. Before we start, it is extremely important to realize that some of these operations are destructive and many of them change the signal in ways that make it useless for certain types of analyses. Both instrumental and auditory analyses of "processed" speech recordings may give unreliable or distorted results. Therefore, original recordings should always be archived, and any changes to the signal should be documented and reported to potential users.

Although we will refer to certain software products, we will not discuss them here in detail, as such programs evolve, change, emerge and disap-

pear. However, in Appendix 1, we list example software tools that may be useful for empirical phonetic studies and related applications. They include not only those designed specifically for phoneticians, but also some dedicated to instrumental acoustics, music production, as well as databases and data management.

DIGITAL AUDIO SIGNAL

Digital audio file formats and conversion of audio files

Some issues related to audio file format have been mentioned in Chapter 2, which dealt with speech recording. Here, the nature of the digital representation of signals is briefly described.

Analogue-to-digital conversion of signals is based on sampling. In the case of sound, the idea is very simple. Measurements of amplitude are performed at a high frequency, and their values are stored in digital form. The more measurements are made per time unit, and the more precise they are, the higher will be the quality of the recording – or, more exactly – the mapping of the analogue input to the digital representation will be more accurate (Rumsey & McCormick 2006: 200-221). However, even excellent analogue-to-digital conversion will not help much if the microphones used for the recordings were of low quality or acoustic conditions were adverse.

In order to capture enough details of the signal, the **sampling frequency** (the number of measurements made per second) should be at least twice as high as the highest component frequencies of the signal that we want to capture, possibly plus a 10% margin. Now it is easy to understand why 44.1 kHz has been a standard for audio CD and many other types of recordings. The upper frequency limit of human hearing is generally agreed to be around 20 kHz (for very good, young ears). By doubling this value and adding 10%, we obtain 44 kHz. Of course, higher sampling frequencies may be used. For example, 48 kHz is often used in camcorders, as it is better to have a whole number of samples per movie frame, and 48,000 is divisible by 24, 25, 30, 50 and 60, which are the most typical frame-per-second rates. Then we have 88.2 kHz, 96 kHz, and even 192 kHz (e.g., Huber & Runstein 2017: 207). Although sometimes these high frequencies are justified in top-end sound production, they are more than is needed for speech recording for linguistic purposes. In fact, for certain speech technology applications, sample rates of 16 kHz or even less are still used.

The second important parameter is **sampling precision** (also referred to as **sample resolution** or **bit depth**). Each amplitude measurement is taken and stored in a limited amount of memory, which determines its precision. For example, one may decide that each sample will be stored using 8 bits (one byte). This offers 256 values for the coding of 256 different levels of amplitude. It would seem that 256 levels is a large number, when we consider, for example, the number of output levels that would be sufficient for step-by-step (as opposed to continuous) control of a radio or CD player. Here, however, the situation differs – we are considering amplitude changes in the signal, not the global level that is set when using hi-fi equipment. The ear is quite precise, and in certain situations, for instance, when the sound level is gradually increasing or decreasing, we will be able to hear that it is not truly continuous – there are tiny "steps". If one uses 16-bit precision, one can potentially represent over 65 thousand amplitude levels. This sounds impressive, and for many years it was a standard for high-quality audio. Now, when we do not need to save on storage space, higher values are often used, and 24 bits per sample is a more and more common standard. While 16-bit resolution is perfect for most linguistic applications, 24 bits may provide some advantages. Since you can encode more amplitude levels, you have a wider dynamic range at your disposal (Huber & Runstein 2017: 208). If you record people speaking spontaneously, from whispering to shouting, higher sample resolution may be useful. But, again, using a 24-bit file format means nothing by itself: you need a microphone, a preamp, and an analogue-to-digital converter that can cope with a wide dynamic range.

In Figure 2, the idea of analogue-to-digital conversion is explained in the most common form. You can see the "steps" that are actually a feature of the signal re-created from a digital representation. And, in accordance with our intuition, the smaller the steps are, the better is the signal representation.

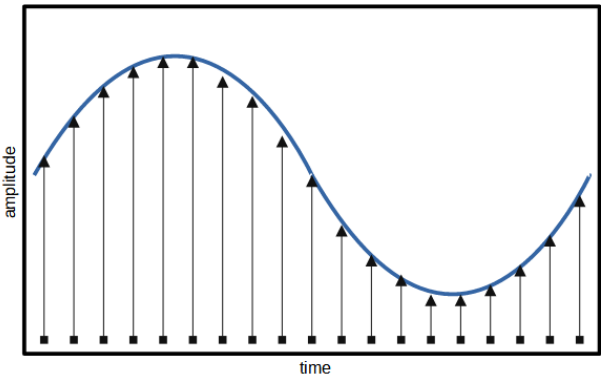

Figure 2. Analogue signal (for clarity, shown as a simple sinusoid wave) and its periodic measurements in the process of analogue-to-digital conversion. Note that the arrows do not always precisely touch the line representing the original signal; this shows how a lower sample resolution leads to less precise representation of the original signal.

Independently of how the signal is converted into a digital representation, it can be stored in a variety of forms, often standardized as audio file formats. There are a limited number of popular audio file formats (WAV, MP3, AIFF, AU, etc.) and a certain number of parameters that can be adjusted for each of these formats (Rumsey & McCormick 2006: 264-275; Huber & Runstein 2018). These parameters often determine, or at least influence, the precision of digital representation of the sound (indirectly, the quality of our recordings).

All audio file formats can be roughly divided into two groups: compressed and uncompressed. File compression emerged from the need to save storage space when computer memory was extremely expensive. When advanced web services became popular and the question of multimedia transfer became important, file compression issues again started to take on significance. However, in linguistic research applications today there is no need to use compressed file formats. Why? Many compression techniques (like MP3) are "lossy", which means that some information is lost when the file is compressed and cannot be restored when it is played back. This information may concern some aspects of the signal that are difficult to perceive, but may be relevant in close listening or instrumental analyses. There are also techniques of lossless compression (like FLAC; cf. Coalson 2002–2009) which guarantee full restoration of the initial information when you play back the file. This is just like zipping a text

document: once you unzip it, it is full and complete. However, lossless compression is not very popular, as high compression rates are more difficult to achieve and it may be computationally more demanding.

For uncompressed formats, we can normally decide on the number of channels (mono/stereo/two channels/more channels), sampling frequency (how many times per second the amplitude is measured), and bit resolution (how precisely the results of these measurements are stored). For compressed files, we may sometimes select the **bitrate**, which means how many bits of memory are to be devoted to one second of recording (for example, 192 kbps, i.e., 192 thousand bits per second; Rumsay & McCormick 2006: 235, 239-242). Note, however, that some pieces of equipment or software will not give you much freedom in this respect. For example, simple hand-held audio recorders may have the bitrate locked to a certain value (such as 128 kbps).

As mentioned above, it is sometimes necessary to convert audio files from one format to another. It should be remembered, however, that conversion never helps to increase the quality of the recording. If you convert mp3 files with a low bitrate (e.g. 64 kbps) to wav file with high sampling parameters (e.g. 24 bit/96 kHz), the quality will not increase. Conversion, even to a "better" format, may actually degrade the quality to a certain extent, due to the way the conversion algorithm works. This is why, if possible, you should record directly in the format that you want to use later – both to save time and to avoid quality issues. Also, if conversion is necessary, it is always good to preserve the original files.

Professional audio processing software is often capable of reading and writing a wide range of audio file formats. In some programs, batch processing mode or scripts to convert multiple files at a time are available. There are also some programs designed solely for the purpose of audio format conversion. Free versions may have limits on the size of converted files or on the input/output formats.

FILE FORMAT "UPGRADE"

Conversion to "better" (uncompressed) formats or to better parameters (e.g., higher sampling rate) within the same format will not change the quality of the recording, even if it may sometimes give an impression of better dynamics or wider frequency range. The very process of conversion may actually even degrade the quality of recordings. In principle, such "up-scaling" conversion is justified only if we need to adjust the format of our files to the standards of an existing collection, tools or archives, or to meet other technical specifications or data sharing standards.

CUTTING AND SPLITTING SPEECH RECORDINGS

When doing fieldwork, we often collect relatively long recordings (e.g., narratives, conversations, or interviews), while for many purposes we need only fragments of these whole sessions, in the form of shorter, manageable files. This applies not only when preparing stimuli for experiments, but also when adjusting material for processing or storage, to make it fit into corpora or databases. Therefore, the amount of time and effort spent on the cutting of signals and related activities was always substantial. Increased processing power and the vast memory of contemporary computers have somewhat changed the situation. Increasingly often, long audio files are stored, processed and made available to users "as they are", accompanied by synchronous annotations that can be used to easily extract interesting portions of the signals or reject those that are irrelevant for current purposes. Notably, by offering segmentation coded in annotation tiers, we in fact offer a virtual, non-destructive cut. Accordingly, most of the guidelines described below hold even if we do not touch the audio file and work only with annotation.

In principle, most of the boundaries that we mark during the process of speech segmentation are arbitrary, because we deal with continuous articulation, with continuous transitions between successive sounds. It takes time for the articulators, even in their most rapid movements, to change positions. As a result, it is difficult to precisely define the boundaries between individual sounds or other segments in running speech (see also Port 2008; Machač & Skarnitzl 2009). In some cases, the task is relatively easy, but when we deal with the segmentation of vocalic clusters or approximants (see below), it becomes extremely challenging. Another approach would thus be to think of the "boundaries" as regions (areas, intervals) instead of markers (related to fixed points in time). Software solutions to support analyses based on such approaches (taking into account the uncertainty about the exact boundary position) have also been proposed (for example, SPPAS (Bigi & Bertrand 2016); see also Bigi 2015).

As already mentioned, to avoid too much technical activity during the recording session, it is advisable to switch on the recording devices before the substantial part of the talk or tasks begins, and to stop recording some time after the main part is finished. When we follow this advice, we may well expect that these additional pre- and post- interview parts of our recording will include some private, informal, incidental conversations that should in principle be removed and not kept even as a part of the original source file. On the other hand, both the pre- and post-recording parts may contain valuable linguistic material. They may also provide us with additional information about the speakers and the circumstances of the recordings. To use them, we

need an additional agreement, but before signing it, the speakers may also want to listen to what they actually said. A separate issue is that the initial and final parts of recordings are often filled with noises (moving chairs, touching microphones), accidental utterances or greetings. Although cutting seems to be the simplest editing operation, there are still some important guidelines to follow, depending on the requirements for the final form of the material. Some linguistic competence is also required. A naive listener can hardly identify word boundaries in an unknown language. With some phonetic experience, you might be able to work with speech recordings in a language that you do not know, but when even the inventory of phonemes is unknown, it may be a challenge.

General rules

The most universal rule (although exceptions can still be found) for cutting signals is to cut them at **zero-crossing points**, that is, where the amplitude is equal to zero. If you cut the signal at any other point, especially where the amplitude is relatively high, a click will be heard at the beginning (or the end) of the newly extracted signal (Figure 3). This can be repaired by onset/offset adjustment (amplitude shaping), but this takes time and, although microscopic, it is still a manipulation of the signal (and as we said earlier, such manipulation should be avoided unless clearly justified). Some software may be set up so that the cursor automatically stops at the nearest zero-crossing.

SILENCE IN RECORDINGS?

There is no absolute silence in any recordings, unless you use a noise gate or reduce the amplitude to zero using sound editing software. This obvious fact is important for many aspects of sound processing, including automatic cutting of the signal. When using a "silence detector" or any segmentation function based on identifying silence in audio files, we normally have to define the "silence level", which is the noise level that we actually consider to be silence in our recording.

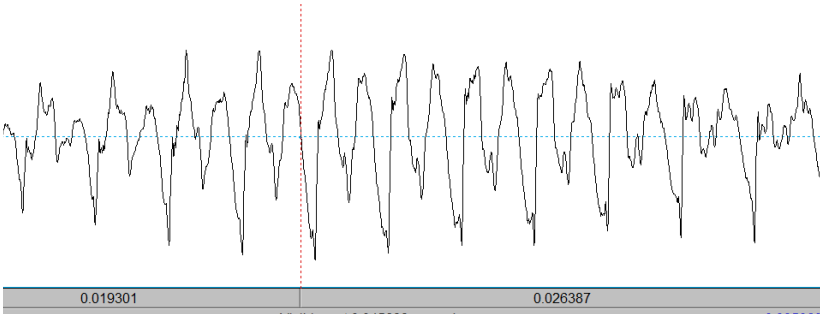

Figure 3. Cutting signals at a zero-crossing point (the red vertical line passes through a zero-crossing point).

Cutting the speech signal where there are no pauses

Segmental boundaries are always somewhat artificial ("phonemes are only ghosts of letters"; Port 2008). Co-articulation results in the "spread" of acoustic features so that a trace of a given sound can be detected around it, beyond its traditionally defined boundaries. Boundaries between units like words or phrases in continuous speech are often not easy to detect (Figure 4). Sometimes this is very striking in the process of segmentation based on close listening: you listen to a portion of a signal up to the boundary you have marked, and you can hear a trace of the first segment that occurs after the boundary. You move the boundary backwards to get eliminate this; but then when you listen to the portion of the signal starting from the marked boundary, you can hear a trace of the last sound before the boundary. You may try to move the boundary so that the effect is minimized, but in many situations, it cannot be eliminated completely.

Speech signal contains acoustic pauses, but they often remain unnoticed by listeners in natural communication conditions. The most often discussed case concerns the pauses preceding plosive consonants. Even though they are acoustically silent, they are traditionally treated as parts of those consonants, as they belong to them from the articulatory perspective: the pause is in fact used to accumulate the air pressure which is necessary for producing a plosion (Figure 5). Another reason for such an "internal pause" may be a hesitation or an articulation problem, but these are very rare situations.

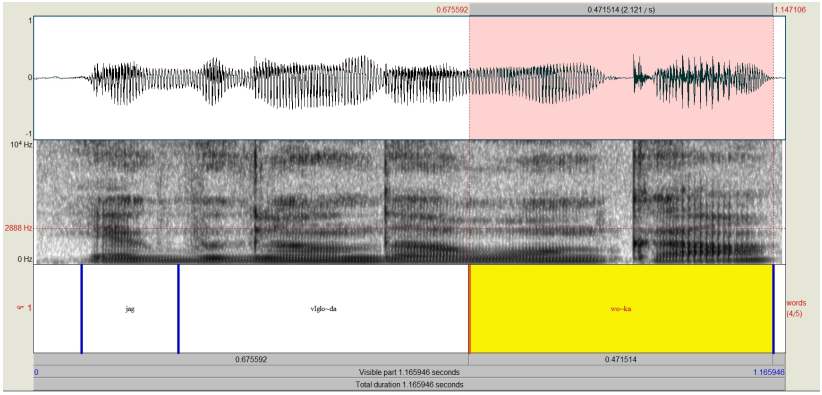

Figure 4. Inter-word boundaries in continuous speech often do not have explicit acoustic correlates.

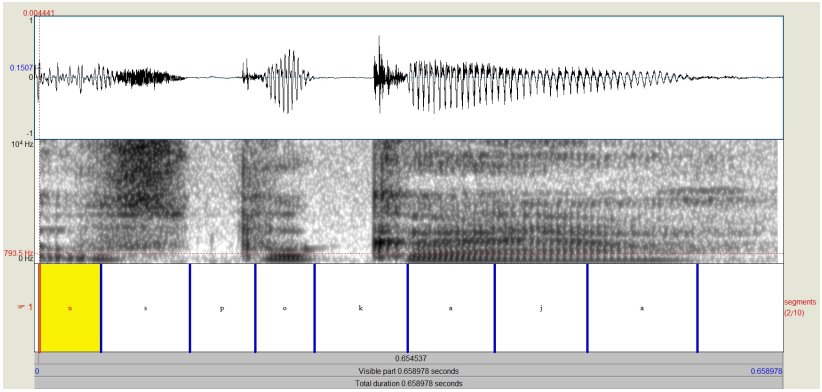

Figure 5. There is a silent pause before /p/ and /k/. In segmentation, such pauses are traditionally regarded as parts of the respective consonants, and no pause labels are included in the time-aligned transcription.

Adding silence

If, due to cutting or some other factors, the speech signal begins immediately at the start of the recording, with no preceding silence, or when it stops only at the end of the sound file, it may be useful to add silent pauses at the beginning

and at the end of the file. For example, some audio file players tend to start and stop playback with a clicking or popping sound, and this may actually mask the beginning or the end of the speech recording. The unwanted sound may also be a side-effect of the sound card being switched on and off (the card may be switched off while unused, for energy saving purposes). Adding silent segments may be of importance when recordings are to be used as stimuli in a perception experiment. If you add 50 or 100 ms of silence at the beginning and end of the file, the device click will occur at a safe time interval from the sound itself. Another technical reason to add a stretch of silence at the beginning of the signal may be the adjustment of pauses in playback during experiments.

Vowels (vocalic clusters) and approximants

Vowels and approximants ("glides") are always a challenge in the process of segmentation, as their boundaries (with other vowels and approximants, as well as with other types of neighbouring segment) tend to be difficult to detect, both by ear and instrumentally. There are certain techniques that may help with placing the boundary, but it will nearly always remain a somewhat arbitrary decision, because often the "boundary" is not a fixed point in time but a transitional region.

You may select the phrase or the word where the boundary occurs, place the cursor at the hypothetical location of the boundary, and listen to the signal up to the cursor and from the cursor onwards. You then shift the boundary left or right until you find an optimal position where you can hear as little as possible of the subsequent segment in the previous one, and as little as possible of the preceding segment in the subsequent one. With fast speech, slowing down the playback may help.

You can also base your decisions on visual scrutiny of the spectrogram, or combine the two methods. In the case of visual scrutiny, it is often useful to look at the intensity, formants or other subtle spectral features, or even at the pitch trace or details of the oscillogram. Sometimes, combining cues from various sources is helpful.

More on speech segmentation from both phonetic and technological perspectives can be found in Chapter 4.

Amplitude normalization

Amplitude normalization is a relatively "neutral" and usually fully reversible operation. It changes the amplitude by the same value in the entire

processed signal. In most cases, it serves to increase the amplitude to a reasonable limit, just below the ceiling defined by the available dynamic range. From the perspective of perception, this will result in a louder signal. We often just want to have a similar amplitude level in all of the signals in a corpus or a collection of stimuli. The parameter to be set is the resulting amplitude, often relative to the available top (ceiling) level (e.g. –1.5 dB, which means 1.5 dB below the ceiling) or as a proportion of the entire available range (e.g. 97%). The most common normalization algorithm looks for the highest peak in the signal and determines by how much it should be increased to achieve the desired top level. Then the amplitude of the entire signal is increased by this factor. This kind of normalization is often referred to as "normalization by peaks". One may also encounter "normalization by the mean" or some other central tendency values. Figure 7 shows oscillograms of a speech signal before and after normalization.

Even though amplitude normalization seems to be both conceptually and technically straightforward, a few questions and issues deserve consideration:

– Amplitude is increased for the entire signal; if there is noise, the noise will be louder as well.
– If there are high peaks (e.g., plosions or knocking sounds), and you use normalization by peaks, the effect will be limited, as the peaks may already be close to the ceiling anyway, and the multiplier value will be very small – accordingly, the overall amplitude will not increase by much.
– If the peaks are not parts of the speech signal (e.g., the speaker was tapping the microphone, or noises occurred in the environment) and they do not overlap with it, they can be easily deleted. Nevertheless, to preserve the temporal structure of your recording, you may decide simply to silence the peaks (that is, reduce the amplitude to zero or to the level of the background noise) instead of cutting them out.
– If the peaks belong to the speech signal (e.g., plosions or high intensity vowels), you can still adjust them (e.g., reduce the amplitude of the peak itself), but this may have serious consequences.
– If you have several recordings of a certain type of speech (e.g., texts read by the same speaker), fully independent normalization by peaks may give unwanted results. For example, some of the recordings may have extreme peaks in the speech signal, and after normalization the peak-containing signals will be quieter on average than those with no peaks, because the peaks (as described above) will result in lower multiplier values for the entire signal.

Note that amplitude normalization is different from loudness normalization. The latter often refers to more complex operations resulting in changes

of perceived loudness. Although amplitude and loudness are connected, their relationship is quite complex. Another notion that may arise in this context is the intensity of the sound wave, which refers to the average amount of energy passing through a unit area per unit of time in a given direction (Roederer 2008; Mihajlovic & Todorovic 2011).

Compression

What we will discuss below is dynamic range compression, and not the compression of the size of an audio file (the difference is explained in Chapter 2).

Compression of an acquired signal is destructive and cannot be easily undone. This is because compression involves a number of parameters and different approaches. For example, sometimes different compression is applied to the low and to the high band. Nevertheless, it is somewhat similar to normalization in the sense that it operates on the amplitude and changes its value. The difference is that while normalization involves equal change to the amplitude throughout the signal, compression entails dynamic adjustment of the amplitude level (Rumsey & McCormick 2006:361-362; Hubner & Runstein 2018:421-428; Réveillac 2017:156-178). In many cases, it involves leaving the areas of high amplitude untouched and "pumping up" the regions of low volume. As can easily be predicted, such a modification to the signal seriously compromises its further use in phonetic research. Compression, however, may increase speech comprehensibility to a certain degree, which might be crucial for field linguists interested in other aspects of the recordings, such as typological or sociolinguistic information contained in an interview.

There is a huge range of compression algorithms and hardware compressors on the market. Many of them are highly appreciated as tools for music recording and editing. However, these are not necessarily best for speech.

The simplest compression is linear – the relationship between corresponding input and output amplitude levels is expressed by a straight line (Figure 6). However, in real applications it is often better to use non-linear compression. In advanced software compressors, one can design customized compression curves, that is, define a function which translates the amplitude of the input into the amplitude value of the output. In the bottom panel of Figure 7, changes to the amplitude resulting from compression are contrasted with the results of normalization (middle panel).

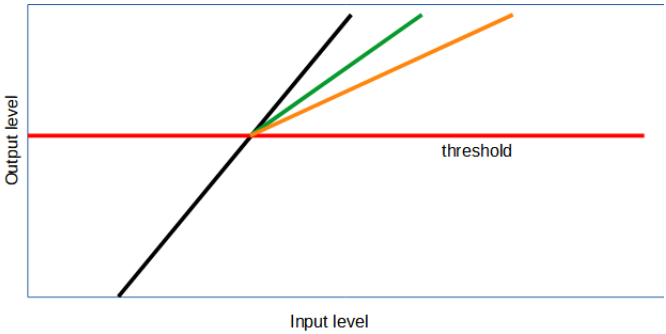

Figure 6. Input level and output level in linear compression.

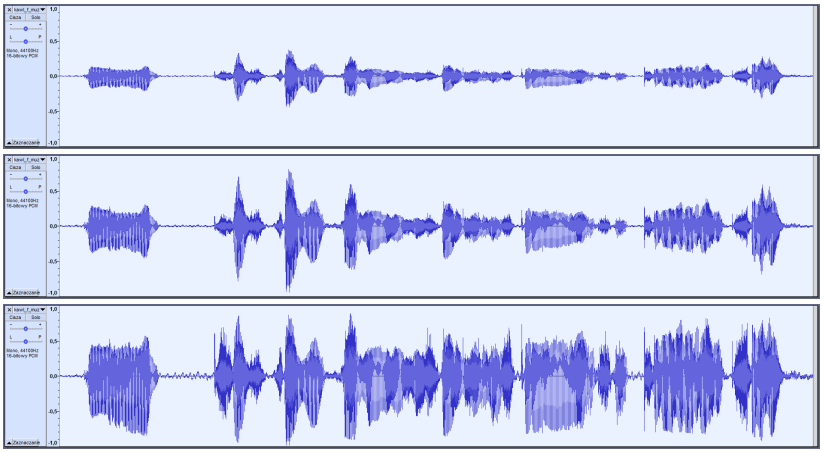

Figure 7. Speech recording: original (upper panel), after normalization by peaks up to –0.5 dB (middle panel), and after compression (bottom panel).

Filters (low-pass, high-pass, band-pass, band-reject, de-noising, etc.)

Filtering refers to a wide range of tools which are primarily designed to eliminate or at least dampen certain frequencies in the signal (Shenoi 2006; Réveillac 2017: 82-91). Filtering is, in general, irreversible – information is permanently lost from the signal. One of the dangers of filtering is that if you want to eliminate unwanted frequencies from your recording (e.g., a buzzing sound in the background or electrical network noise), you will eliminate those

frequencies from the speech signal as well. If the material is to be analysed phonetically, you should certainly refrain from any kind of filtering unless you actually want to filter something as a part of your research plan.

Filtering, even digital, can hardly be perfect, in the sense that one can hardly eliminate all of the unwanted frequencies and leave the rest untouched. Filters' characteristics include not only frequency but also slope, which indicates how steep the filtering is. Most often, the slope or steepness is expressed in dB per octave. The higher the value, the steeper (more "abrupt") is the filter.

A low-pass filter stops the frequencies above the cut-off frequency, while for high-pass filters the frequencies below the cut-off frequency are blocked. Middle-pass (pass band) filters have two cut-off frequencies, and admit frequencies between those values while filtering out those outside the range (see Figure 8).

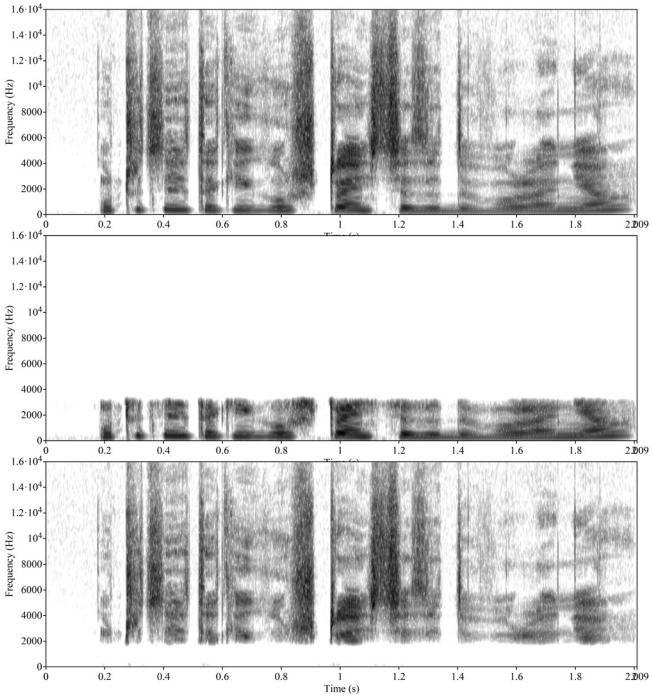

Figure 8. Spectrograms of an unfiltered utterance in Polish (top), and of its versions following treatment with a high-pass filter (middle) and a low-pass filter (bottom).

Noise gate

A noise gate is a particular type of filter that lets the signal through only if it exceeds a certain level (Réveillac 2017: 180-184). As a result, only the sections that match the condition remain in the recording, while any other portions are reduced to silence. A noise gate can be used in real time (during the recording session) or once the recording has been made. The first option has some advantages; for example, when more microphones are used, noise gate parameters can be set separately for each microphone input. On the other hand, there are also good reasons to apply the noise gate to complete, archived recordings. First of all, the original recording remains "raw" and unprocessed, with no filtering and no data loss. Secondly, one can analyse the entire recording to adjust optimal parameters for the noise gate.

Voice-activated (or sound-activated) recording systems, available in many hand-held voice recorders, are a kind of noise gate: they record a signal only when it exceeds a certain intensity level. A classic noise gate does not stop recording – it just reduces the amplitude to zero when it is not high enough.

Often, a noise gate does not react immediately (especially when working in real time). Even if the amplitude is high enough, it may take a few milliseconds for the gate to work. The same applies to the offset of the signal: even if it is below the defined level, it may still stay in the recording due to the inertia of the entire process. In some software and hardware noise gates it is possible to configure these parameters, but there are still limitations, especially regarding initial reaction time (latency).

Typically, noise gate parameters include:

– threshold level – if the amplitude exceeds this level, the gate opens;
– attack (duration of the increase stage);
– hold (duration of the hold stage);
– release time (the time after the gate closes when the amplitude drops below the threshold).

If a noise gate is to be used at all for linguistic (phonetic) purposes, it is usually better to set the shortest possible attack times and relatively short or medium release times.

Advanced de-noising

Sound editing software is more and more often equipped with advanced de-noising tools (e.g., Réveillac 2017: 267; Haque & Bhattacharyya 2018). They can be used freely, as long as the aim is to achieve an improvement in

listening comfort and comprehension. Avoid them, as a rule, when the material is meant to be analysed phonetically, auditorily or instrumentally.

Some programs offer filters that can be "taught" which patterns to eliminate from the signal. If you have the same noisy pattern throughout your entire recording, you may be able to find a moment of "relative silence" where there is no speech but the noise is present. Then you can sample the noise pattern and eliminate that pattern from the entire signal. Even such an "intelligent" function is still destructive to the speech signal itself. Nevertheless, it may improve its comprehensibility, and if the material is to be used only for listening to its content, this approach may be acceptable and desirable.

Reconstruction of the signal

Short portions of the signal may be automatically reconstructed on the basis of the context and general characteristics of the signal (Réveillac 2017: 266-271; Stoian-Irimie & Irimie 2017). This method may work fine for quasi-stable portions of the signal, but is less efficient when rapid signal changes occur, although machine learning approach to this issue still seems to be promising (e.g., Godsill at al. 2002).

A typical and relatively safe reconstruction procedure is the repair of short overloads where the amplitude exceeds the dynamic limits of the software or hardware, and clipping occurs (Figure 9). In the clipping portion of a signal, one may notice that the curve representing amplitude changes resembles a sinusoid with its peaks chopped off at the point where the maximum allowable amplitude is reached. If this phenomenon occurs in larger portions of the signal, the sound becomes very unpleasant. In some sound editors, the problem can be repaired automatically. In some others, it may be possible to manually "re-draw" the problematic portions of the waveform. Manual repair requires patience and precision, and as in any other reconstruction procedure, the sound may sound better afterwards but not necessarily more similar to the original. Its dynamics will probably be distorted, as if it had been compressed to reduce the amplitude range.

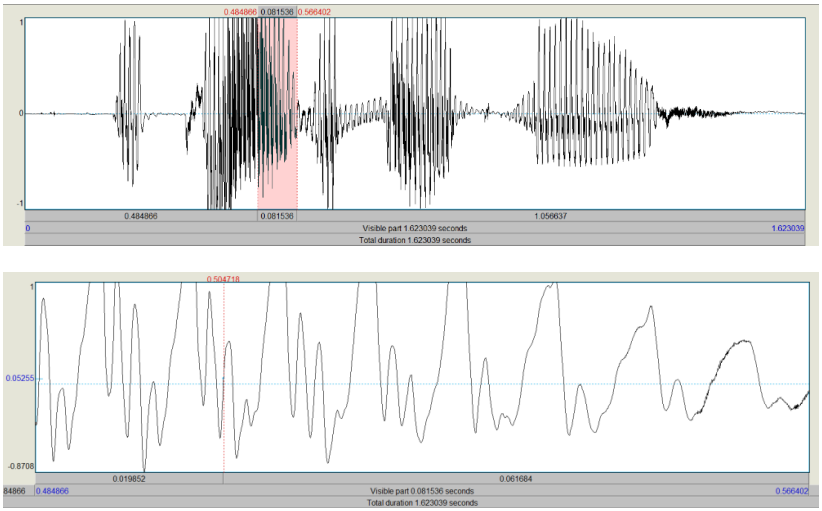

Figure 9. Overload (clipping) in a recording of the utterance “It was a big mess.” Clipping is visible in many areas of the oscillogram (upper panel). The pink section is enlarged (bottom panel) to show the chopped off peaks in the clipping area.

DILLEMAS AND GOOD PRACTICES

We assume that the basic sound editing operations discussed in this chapter will cover the vast majority of requirements for preliminary adjustment of speech material in field linguistics. Needs may be much greater in speech technology, experimental psycholinguistics, or even in some areas of acoustic phonetics.

The fundamental question is always “to process or not to process”. When filtering unwanted noises from audio recordings, we most often also remove some components of the speech signal itself. When portions of the speech signal are reconstructed, we introduce new information that may distort the final picture produced by our study. When distortion or low dynamics of the recording are repaired, we change some characteristics of the signal, which may take us further from the original sound than the technically imperfect recording would. Therefore, all processing should be limited to the necessary minimum, strictly controlled, and reported to potential users of the data.

Another frequently faced dilemma is whether to process the entire material (e.g., a set of sound files from a given recording session) in the same way, using exactly the same sequence of operations, with the same parameters, to ensure its "sonic coherence", or to treat each file individually, to get the best from each of them. It is helpful to know precisely all of the future applications of the material at hand, but that is not always possible. In such a situation, there are good reasons to keep sound processing to a minimum or to preserve backup copies of the original material, uncut and unfiltered.

A third important issue is that different programs may use different processing algorithms, and the results of both processing and analysis may differ significantly even if we use functions that are identically named (e.g., Oğuz et al. 2011). Therefore, it may be important to establish how a given processing function is implemented and whether it differs from what is available in other software tools. It is also important to note that processing functions may be available in several variants in a single piece of sound processing software. Using different variants may give different results. In such a case, it is essential to report clearly which pieces of software, and which of their functions and parameters, have been used for sound processing. In that way, your procedures will be replicable and testable by other researchers.

GOOD PRACTICES

- Always keep a copy of the original recording ("as is", "as recorded") without any cuts or edits. Some editing operations are difficult or impossible to reverse. And even if they are reversible, you may have to remember all of the parameters of the initial operation to return the signal to its original form. Computer memory is cheap – and patient – so it is no problem to save backup copies of your material before each irreversible ("destructive") processing step.
- Limit sound processing to the necessary minimum. You can do incredible things with audio recordings using editing software, but the results will be far removed from the original sound. You may be able to make your recording more comprehensible, easier to listen to, better sounding, but you can hardly reconstruct it.
- Understand what you are doing: use processing that you understand well enough. Some processing functions may be harmless when used knowledgeably, in a certain way, with a certain set of parameters, but may be unacceptable in other situations.

4. TRANSCRIPTION, SEGMENTATION, ANNOTATION

Human communication can be described as a many-layered process (e.g., Laver 1994: 13–23). People simultaneously transmit and receive various kinds of signals using the available communication channels. We use hearing, vision, and other senses for the purpose of communication. Different signals are transmitted and perceived simultaneously or sequentially by the same participants. Furthermore, communicative events always occur in some external environment that influences the communication process. The extent of this influence may vary substantially, especially when we consider fieldwork recording conditions. Audio and video recordings of spoken communication are the source of a plethora of information. Once we have the recordings, we surely wish to use the information efficiently. It is possible to extract some of it directly from the signal – for instance, simply to listen to a recording of a conversation and understand the message, or to measure certain acoustic parameters such as the overall mean fundamental frequency, amplitude or other spectral features. However, in many cases, additional description is required to make analyses possible. The description procedure usually involves three main tasks: transcription, segmentation and annotation. To reflect the many-layered nature of the communication process, multi-layered approaches to description have been proposed and applied for speech corpus development.

In this chapter, we discuss some of the approaches to transcription, segmentation and annotation of speech recordings. The three notions – transcription, segmentation and annotation – are closely related, but still have distinctly different meanings, which will be discussed below and illustrated with examples.

Transcription

TRANSCRIPTIONS OF A PHRASE IN ENGLISH AND IN POLISH

Orthography: the north wind and the sun
IPA: ðə nɔθ wɪnd ən ðə sʌn
SAMPA: D@ nOT wInd @n D@ sVn

Orthography: północny wiatr i słońce
IPA: puwnotsnʲ vjatr i swoɲtse
SAMPA: puwnotsnI vjatr i swon’tse

The simplest definition of **transcription** of speech recordings may be that transcription refers to the written form of what is being said. To create the written form, a certain graphical representation, such as an alphabet, must be used. The alphabet may be simply the standard orthographic alphabet used for a given language, in which case we will speak of **orthographic transcription**. The orthographic transcription is a useful graphical representation of spoken text, especially when our focus is on the meaning of the utterances, for instance when we record an interview and we are interested in the story being told by the speaker, his/her experiences, opinions, etc. Sometimes it might be profitable to use **transliteration**. i.e., swapping the signs of one orthographic script with letters of another orthographic script, e.g., Greek or Cyrylic script with Latin letters or vice versa. However, when we look closer, orthographic transcription (and transliteration alike) is quite ambiguous and restricted in representing peculiarities of the spoken text. Plenty of examples can be found in the world's languages. Take the English letter "o" and consider its different pronunciations in the common words *north*, *who*, *come.* Such graphical representation is ambiguous and does not reflect the actual pronunciation.

Therefore, in experimental phonetics, we will often find it more useful to apply **phonemic or phonetic transcription**, better representing spoken utterances in terms of the features of various components of the speech signal: individual phonemes (or phones), syllables, words, phrases, etc. Phonetic transcription refers to the actually realized speech sounds (phones), while phonemic transcription is more general, as it refers to phonemes, understood as classes of phones (cf. e.g. here: http://languagesindanger.eu/book-of-knowledge/the-sounds-of-language/#ch4 for a brief account of the distinction between phones and phonemes; Karpiński 2014).

In most languages, there is no simple one-to-one correspondence between graphemes and the phonemes they represent. One grapheme can stand for several phonemes, and a given phoneme can be symbolized by different graphemes. A well-established and commonly used alphabet is the **IPA** (the International Phonetic Alphabet) (*International Phonetic Association* 1999). Charts showing the IPA labels, together with examples and illustrations for a number of world languages, are provided online by the International Phonetic Association, for example here: https://www.internationalphoneticassociation.org/content/ipa-chart. Using a phonetic alphabet helps to represent words in such a way that their pronunciation becomes unambiguous for readers who know that alphabet, even if they do not know the language. Let us take the English words from the example above: *north*, *who*, *come.* The respective IPA notations (for standard British English pronunciation) would be /nɔθ/, /hu/, /kʌm/. As can be seen, the letter "o" is converted to three completely different

phoneme labels depending on the context: /ɔ/, /u/, /ʌ/. Note that when cited as part of written texts, phonetic transcription is typically provided in slashes or brackets. Slashes / / are used when providing broad, phonemic transcriptions, while square brackets [ ] indicate phonetic transcription that typically contains more details (additional features of the sound pronunciation or even features resulting from the individual realization by a given speaker). The phonemic transcription is based on a set of characters representing all of the phonemes of a given language; hence, it is language-specific in a sense. Phonetic transcription, on the other hand, is more universal, because it precisely conveys articulatory features, only indirectly referring to the linguistic categories of phonemes. In other words, the phonetic notation will usually be more universal and closer to the acoustic signal, while a phonemic notation will be more abstract and general, but also language-dependent.

The process of converting the written representation from orthographic to phonetic transcription is referred to as **grapheme-to-phoneme conversion**, often abbreviated either as **GTP** or **G2P**. Various strategies are used to perform the conversion process, but usually a transition table of some kind is needed as a starting point to translate the graphemes (letters of the orthographic alphabet or another writing system) to phonemes (units representing spoken sounds). Examples of such tables can be found online for **SAMPA** (Speech Assessment Methods Phonetic Alphabet; Wells 1997).

SAMPA & IPA

Go to the SAMPA website: https://www.phon.ucl.ac.uk/home/sampa/. Inspect the table on the home page showing the mapping between SAMPA, IPA and Unicode.

Compare the tables mapping between orthography and phonetic notation for different languages, for example: Polish (https://www.phon.ucl.ac.uk/home/sampa/polish.htm) vs. Swedish (https://www.phon.ucl.ac.uk/home/sampa/swedish.htm).

SAMPA is especially popular in the contexts of applied linguistics and speech and language technology, where machine-readable encoding of phonetic notation is needed. A practical aspect of using SAMPA that might be important for fieldwork linguists is that the SAMPA notation is based on fonts and characters that are available on any standard keyboard. There is no need to use special keyboard settings or install dedicated font packages. Therefore, a SAMPA transcript can be easily typed with any text editor, using most available devices, including portable ones such as mobile phones or tablets, regardless

of their operating systems. We shall use SAMPA throughout this chapter when providing example transcriptions. The tables (or lists) illustrating the mapping between graphemes and phonemes are often accompanied by language-specific descriptions, including sets of G2P rules relevant for a particular language and its phonotactics (see for example the "Illustrations" of the phonemic systems for a number of languages in: *International Phonetic Association* 1999).

For a given language, one can define a finite number of G2P rules, and these can be a very useful starting point for computer systems enabling automated grapheme-to-phoneme conversion. For Polish, as an example, an important contribution in this respect is the work of M. Steffen-Batogowa, who formulated a concise description of the G2P rules as early as 1975 (Steffen-Batogowa 1975).

GRAPHEME-TO-PHONEME CONVERSION

Inspect the following examples of orthographic notation and corresponding phonemic transcripts in SAMPA. Focus on the grapheme "i".

(1) "witam" (Eng. *welcome*) – /vitam/
(2) "nikt", "sito", "zima" (Eng. *nobody, sieve, winter*) – /n'ikt/, /s'ito/, /z'ima/
(3) "wiatr", "miód" (Eng. *wind, honey*) – /vjatr/, /mjut/
(4) "nie", "sień", "zioła" (Eng. *no, vestibule*, *herbs*) – /n'e/, /s'en'/, /z'owa/

Four sample conversion rules corresponding to the above contexts of usage for the Polish letter (grapheme) "i" are summarized below. The grapheme "i" will be converted to several different phoneme labels depending on the context of its appearance in the orthographic notation. Only in some cases will it denote the vowel /i/ (and the syllable nucleus). (1) When "i" occurs between two consonant letters such as "w" and "t", it will be converted to a vowel label /i/. (2) However, if "i" occurs between two consonants and the preceding consonant belongs to a certain group such as "n", "s", "z", as in "nikt", "sito" or "zima" (Eng. *nobody, sieve, winter*), then two things will happen in the phonemic output as the result of conversion: "i" will still be converted to /i/, but also the preceding consonant label will be converted to a palatalized sound label, respectively: /n'ikt/, /s'ito/, /z'ima/. (3) Another case is that when the letter "i" follows a consonant letter such as "w" or "m" and precedes a vowel letter in the orthography, it will be converted to the approximant /j/ label in the phonemic notation, as in the words "wiatr", "miód" (Eng. *wind, honey*), which will be converted to /vjatr/, /mjut/. (4) However, when the preceding

consonant letter is e.g. “n”, “s”, “z” and the following letter denotes a vowel, as in “nie”, “sień”, “zioła” (Eng. *no, vestibule*, *herbs*), the notation will by default include the palatalized sound label only: /n’e/, /s’en’/, /z’owa/. A special group will be words of foreign origin, for some of which exceptions will need to be implemented; for example, “sinus” is pronounced without phonological palatalization and is therefore converted to /sinus/, contrary to rule (2) above.

There exist a number of software tools supporting automatic G2P conversion, and some of the freely available ones are listed in the table below (see also: Bigi 2015; Reichel & Kisler 2014; Koržinek et al. 2017). Rule-based automatic G2P systems are quite popular and robust; however, when it comes to exceptional pronunciations, it might be difficult to include all of them in the rule set (there might be many exceptions, and new ones may arise due to language change, borrowing of foreign words, etc.). In such cases, statistical systems might be more feasible. These systems are trained: they “learn” about both standard pronunciation and exceptions from transcribed corpora. An example open-source toolkit used for training statistical systems is Phonetisaurus G2P (Novak et al. 2012).

Time alignment and segmentation

For many purposes it might be sufficient to use orthographic or phonemic transcription as the basis for analyses. This may be true for studies of vocabulary usage, some aspects of regional pronunciation variants, or morphology in spoken language. Another example might be recordings of interviews collected primarily to learn about certain social or cultural facts or opinions. However, if the recorded material is to be useful for instrumental phonetic analyses or for speech and language technology applications, a crucial factor that comes into play is time.

Units of language such as words, phrases or individual sounds can all be located on a timeline. To include the time-alignment information in our data, a common practice is to use software tools that display a graphical representation of the recorded signal along the time axis. A basic segmentation task is to detect at which points in time speech actually occurs, that is, to define speech and non-speech segments on the timeline. One of the fundamental techniques is based on pause detection. Many software tools (e.g., Praat (Boersma & Weenink, 1992–2021), Annotation Pro (Klessa et al. 2013), Audacity) help to perform such segmentation tasks automatically or semi-automatically. Figure 10 shows an example result of automatic sound and silence detection in Annotation Pro. The pink stripes with the “SOUND” label show the areas for which

sound was automatically detected by the system. The white rectangles below them represent segments for which speech segments were finally inserted, based on manual adjustment. The blue vertical lines are segment boundary markers. As can be seen, the final result of segmentation differs slightly from the automatically generated one. This is because (1) the recorded audio signal contains not only the target utterances but also certain other sounds that need to be excluded from further analyses, and (2) the automatic detection tool has an adjustable intensity threshold, and with the settings used for the present example the intensity of some parts of the speech signal turned out to be too low to be included in the suggested "SOUND" areas.

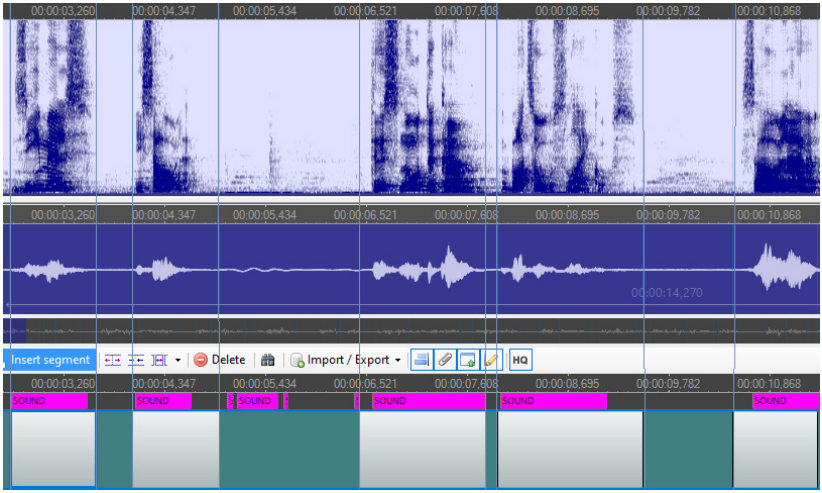

Figure 10. Audio signal segmentation into speech and non-speech areas

The recording used in the above segmentation example comes from a small corpus of Latgalian speech recordings (Klessa et al. 2017; see also: http://inne-jezyki.amu.edu.pl and Klessa & Wicherkiewicz 2015. The recording sessions took place in several different locations, usually quiet office environments. It needs to be taken into account that regardless of the environment, especially when sessions take place in the field, recordings may include certain unwanted noises. This is true even though the recording sessions are usually designed in such a way that the speech signal will be recorded with optimal intensity and unwanted noises will be eliminated as far as possible (Chapter 1). As has already

been mentioned, fieldwork recording conditions are not always 100% predictable, and can thus contain more unforeseen sounds than studio recordings do.

The additional noises present in the recorded material may be of external origin (see also Chapter 1) or may come from the speakers themselves. It is often the case that while speaking, we produce not only speech but other sounds as well: we breathe audibly, cough, laugh, hesitate, mispronounce, move our body, or move objects around us. Certain events are identified only in order for them to be excluded from subsequent steps, although sometimes they may become the focus of analysis. The features of hesitation markers, filled and silent pauses, laughter, breathing patterns and other paralinguistic or non-linguistic phenomena can be a rich source of information that completes the picture for the description of communicative events (see e.g. Winkworth et al. 1995; Karpiński 2013; Bigi & Bertrand 2016).

The time-alignment and segmentation tasks may apply to a variety of events observed in the recorded material, and thus segmentation strategies may differ accordingly. In general, what we are looking for is a change in the spectral pattern. For some uses, an acoustic pause of a certain duration will be the crucial indicator, just as for the detection of silence and speech segments. In a similar manner, acoustic pauses are used as boundary indicators to identify so-called "time groups" (Gibbon 2013), which are continuous stretches of speech delimited by acoustic pauses. Within such time groups, boundaries between individual phones, syllables or words are usually identified. Sometimes the boundary positions are relatively easy to define because of clear discontinuities between the neighboring segments. An example may be a sequence of a fricative and a vowel, such as /z/ and /i/ (see the first two segments in the Phone layer in the multilayer annotation example in Figure 11). But there are also many less obvious ones, due to the continuous nature of speech sound production and acoustics; see for example the transition between /a/ and /w/ below. Such a sequence can in fact be treated as a single diphthong /aw/ composed of a vowel part and an approximant part, depending on the phonological approach. As we have already said in Chapter 3, most often we talk of intervals or continuous transitions rather than fixed points on the timeline. Therefore, indicating a fixed boundary position is a matter of arbitrary agreement based on certain pre-defined criteria. Such an agreement will include specifications for all categories of speech sounds existing in a language. Notably, some of those categories will involve phones that are quite heterogeneous in structure. One such category is the stop consonants, for which the airflow is temporarily blocked during production, resulting in an acoustic pause that is usually regarded as part of the consonant and not a "silence" segment (for illustrations of segmentation, see also Machač & Skarnitzl 2009).

When performed manually, segmentation tasks may be very time-consuming and prone to human error. Fortunately, a number of software tools exist that not only support the time-alignment of different kinds of segments, but also include G2P functionalities, and thus automatically assign transcription labels to particular segments of speech.

EXAMPLE TOOLS SUPPORTING AUTOMATIC G2P AND TIME-ALIGNMENT

Example freely available tools for automated grapheme-to-phoneme conversion can be found at:

- SPPAS: http://www.sppas.org/
- WEB Maus: https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Grapheme2Phoneme
- CLARIN-PL Align https://clarin-pl.eu/index.php/mowa/

See also: Bigi 2015; Reichel & Kisler 2014; Koržinek et al. 2017.

(Multilayer) annotation

In general, speech annotation can be understood as the process of adding information to the recording. Often-desired features of speech annotations are that (1) they should be descriptive and include information important for our goals, for example phonetic transcriptions for phoneticians; (2) they should be time-aligned at as many levels as possible. Annotations are usually saved in separate files; XML-based file formats (see e.g. https://www.w3schools.com/xml/xml_whatis.asp) are popular, but other formats are also used. The visualization of annotated data is one of the fundamental features of software tools used in phonetic research. Since many levels of analysis are usually included in the description, the annotation format uses multiple layers (tiers).

Multilayer annotation typically includes time-aligned (synchronized) information about the events observed in the recorded session at various levels of its structure, for example, phrases, words, syllables, and phones. Figure 11 shows a sample multilayer annotation for the Latgalian utterance "Zīmeļs i saule" (Eng. *The north wind and the sun*) in Annotation Pro. All layers (tiers) provide time-aligned information.

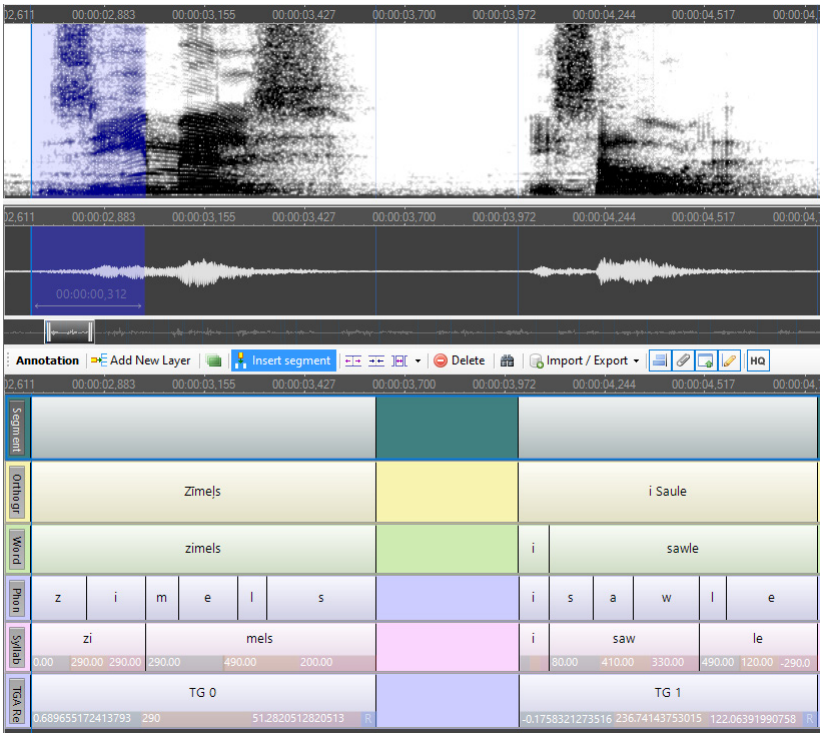

Figure 11. Multilayer annotation of the Latgalian utterance “Zīmeļs i saule” (Eng. *The north wind and the sun*)

Representations of two interpausal stretches of speech are visible in Figure 11. The top layer shows the results of segmentation (speech vs. non-speech; white segments denote speech areas). The second layer contains orthographic transcription at the phrase (time group) level. In the third layer, word-level segments are distinguished and labelled using SAMPA. The next layer provides time-aligned phone-level segmentation together with SAMPA labels. It is followed by a layer with syllable-level segmentation and SAMPA transcription. The last layer includes segments labelled TG 0 (Time Group 0) and TG 1 (Time Group 1) respectively. For the two bottom layers, additional numerical information can be seen at the bottom of each segment. The numbers are the output of the Annotation Pro+TGA module (Klessa & Gibbon 2014) and represent linear regression parameters for syllable duration patterns (as in Gibbon 2013)

and nPVI (Low et al. 2001). They can be useful for analysing speech timing variability within interpausal time groups, because the linear regression function yields local and overall values of segment (here: syllable) duration slope, which expresses an approximation to tempo acceleration and deceleration.

As already mentioned, multilayer annotations of speech can include synchronized labels of different kinds. Probably the most common ones will be the time-aligned speech transcription labels, for example, at the phone or syllable level. Other popular labels will represent more or less directly the events present in the signal (noise labels, pause labels), but they can also include additional information or analysis results, such as the output from Annotation Pro+TGA or specialized linguistic tags (for example, part of speech or morphological glossing).

SPECIFICATIONS FOR ANNOTATION OF PARALINGUISTIC OR NON-LINGUISTIC FEATURES

The level of detail in the annotations of paralinguistic and non-linguistic events in available corpora or archives varies to a great extent. In cases when speech corpora are not annotated specifically to analyse such types of features, they are simply labelled as something different than a regular utterance, for example as "Other", "Non-speech", or they are not labelled at all and are simply excluded from further analysis. When needed, additional categorizations are introduced, to indicate the types of entities more precisely with labels such as "Filled pause", "Silent pause", "Unintelligible stretch of speech", "Mispronunciation", "Speaker noise", "Intermittent noise", "Stationary noise" (see e.g. Fischer et al. 2000). Each of the categories can be treated as a whole, but they might also be subcategorized into particular types of noises; for example, for "Speaker noise" we might distinguish lip smack, breathing in/out, sneeze, cough, etc. Gibbon et al. (1997) list the following items at the level of "non-linguistic and other phenomena": omissions in read text, verbal deletions or corrections, word fragments, unintelligible words, hesitations and filled pauses, non-speech acoustic events, simultaneous speech, speaking turns. In the Polish–German *Borderland* multimodal corpus, in the acoustic domain, the annotations included labels for: incomprehensible stretch of speech, transcription of utterances about which the transcriber was uncertain, filled pauses, cough, laughter, sighs, breaths, groans. In the case of fillers or hesitation markers, the labelling scheme included information about the position of the event and, where possible, the closest approximate "transcription" of the filler (Karpiński & Klessa 2018).

Although most of this chapter is dedicated to the transcription and labelling of speech events, it should be noted that much information can be extracted directly from the acoustic signal, independently of any transcriptions or annotations (examples can be found in the works published as part of the INTERSPEECH Paralinguistic Challenge series, e.g. Schuller et al. 2013; Schuller et al. 2015). For many applications, however, multilayered annotations are an indispensable input – for example, making it possible to investigate the links between perceived and objectively measured, automatically extracted values. In the case of paralinguistic signals, when human interpretation is needed, the annotations are created either manually or semi-automatically.

ANNOTATION MINING

Time-aligned annotations of recordings make it possible to extract structured information about the recorded material, and also to automate the process of extraction using software tools. The process of extraction is often referred to as "annotation mining".

Since it is possible to convert between most annotation file formats, it is also possible to apply the same analytic tools or algorithms to other (even very large) speech corpora. Your original data might be annotated using one of the annotation formats such as TextGrid (Praat), .EAF (ELAN (Wittenburg et al. 2006)), .ANTX (Annotation Pro) or .XRA (SPPAS), and they can all be used as input for automated annotation mining, even if another format is required by the analytic tool. A number of freeware file format converters can be found online; also, Annotation Pro enables conversion of its native format .ANT or .ANTX to and from any of the other above-mentioned formats.

Typically, speech annotation file formats include timestamps attributed to particular annotation labels. The timestamps inform us (directly or indirectly) about the duration of each segment, as well as the exact moments of its beginning and end. Such information allows us to inspect speech timing variability based on segment durations. It is possible, for example, to extract and analyse information about the timing of particular components within a syllable structure, or to study syllable durations depending on the syllable's position within the phrase or other unit of the utterance structure. The timing information can be used in combination with other cues extracted from the speech signal, such as measures of fundamental frequency or amplitude, which may be used as correlates of word or phrase accents (cf. e.g. Jassem 1999; Francuzik et al. 2005).

Moreover, thanks to the interoperability of file formats, one can compare results for data sets created using different tools and different annotation for-

mats. For example, Yu et al. (2014) examined syllable duration variability in three different data sets for typologically different languages: Chinese, English, and Polish. The corpora were annotated using different annotation tools and then analysed using a uniform approach to annotation mining, namely, time group analysis (TGA) for speech acceleration and deceleration patterns (Gibbon 2013). Another example of multi-layered annotation mining is a study of local and global convergence in the temporal domain in Polish task-oriented dialogue (Karpiński et al. 2014) based on two corpora of Polish dialogues: *Paralingua* (Klessa et al. 2013) and *DiaGest2* (Karpiński & Jarmołowicz-Nowikow 2010). One of the corpora was annotated using Annotation Pro, and the other with ELAN. All annotations were then converted to .ANT format and explored using the SRMA (Speaking Rate Moving Average) Annotation Pro plugin. The plugin enabled automatic inspection of the possible alignment in speaking rates of interlocutors for two large data sets, to test the hypothesis that conversational parties tend to mutually adapt their communicative behaviour (here: their speaking rates).

Automated quantitative analyses such as those described in the two examples above would be very difficult or even impossible to perform without using standardized annotations as input.

CROSS-MODAL INTERACTIONS

Since speech and gesture are used simultaneously for the purposes of communication, substantial efforts are made to investigate the nature of the interactions between them, or even to demonstrate the unity of speech and gestures as parallel means of performing pragmatic and semantic functions (McNeill 1985).

Gestures are regarded as an inseparable part of spontaneous discourse, and it has been observed that they do not occur in the absence of speech. Gestural behaviour can play various roles when co-occurring with speech units, for example, to convey meaning, to modify it similarly to discourse or prosodic features of speech, and to regulate the flow of conversation. Some of the cross-modal interactions might be language- or culture-specific. For example, English speakers tend to use more gestures anticipating speech, while Chinese speakers use more gestures synchronized with speech (Ferré 2010).

An example of gesture annotation mining is displayed in Figure 12 that illustrates the idea behind the *Re-occurrence (mimicry)* plugin for Annotation Pro (Karpiński et al. 2018). The plugin enables automated calculation of the number of occurrences of an annotation label (e.g. ‘x1’) found in one annota-

tion layer (e.g., including gesture annotation for Speaker A) that occurs also in another annotation layer (e.g., including gesture annotation for Speaker 2). The number of re-occurring (repeated) labels is calculated within n segments after the end boundary of the original segment. The *Re-occurrence (mimicry)* plugin was used for gesture annotation mining of the *Borderland* multimodal corpus. Among others, we compared the mimicry strategies in Polish and German speakers. The results indicated that the durations of original and repeated strokes for gestures of the same function are similar for Poles, while for Germans, the repeated strokes of the same gesture functions were shorter than the original ones. Polish speakers showed significantly higher mean durations of original and repeated strokes in referential gestures in the one of the dialogue tasks (Karpiński et al. 2018).

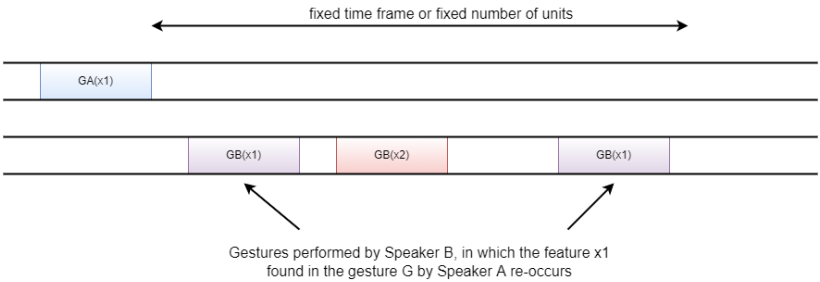

Figure 12. The illustration of *Re-occurrence (mimicry)* plugin for Annotation Pro (Karpiński et al. 2018).

LABELS, CATEGORIES AND DIMENSIONS

Many components of the communication process can be quite accurately described using categorical labels. We can attach time-aligned phonetic or phonemic transcription labels to utterances, categorize speaker noises (e.g., hesitation markers, breathing, sneezing), label basic emotion categories (e.g., joy, sadness) or indicate particular types of gestures, gesture phases or gesture functions (e.g. Ferré 2012; Jarmołowicz-Nowikow 2019). At the same time, we are conscious that some of these categories are fuzzy, created only because of practical needs and only weakly motivated by the nature of the categorized phenomena, and sometimes used mostly because of a long tradition. When a study starts with a fixed set of categorical (discrete) labels, it may limit

its explorative power. Certain research questions may be difficult to answer because of the very nature of the pre-defined categories or vagueness of the distinctions between different features. This is why for some purposes it may be preferred to use continuous (dimension-based) scales for feature description.

Even the question of whether an utterance is or is not well-formed may raise doubts. It may be linguistically coded, or may be more indirect and thus closer to language-external events, as for example with onomatopoeic words. Wharton (2009) discusses several types of continua from "natural" to "properly linguistic" or "linguistically coded" behaviour. Depending on the approach and the particular kind of events, non-verbal communicative events may be located at different points of those continua. For example, interjections that are observed to carry both a coded and natural component might be seen as belonging to different parts of the continuum between "saying" (more coded) and "meaning" (more natural) than other parts of speech (Wharton 2003). A visualization of such a continuum is shown below.

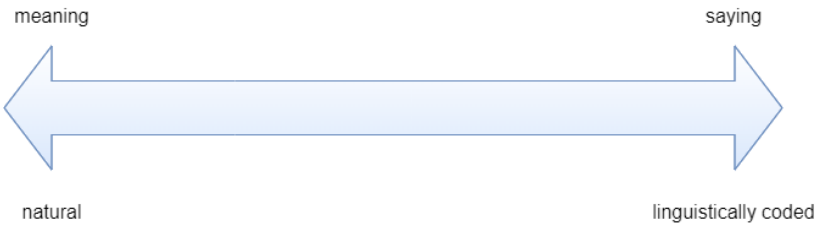

Figure 13. "Meaning" vs. "Saying" continuum; based on: Wharton (2003; 2009).

A topic that engenders a wide variety of different views regarding the use of discrete or continuous rating scales is the description of emotions. The discrete/categorical approaches are often expected to be useful for basic emotions or emotion families (Ekman 1992). Distinguishing between full-blown emotion categories such as anger or joy is usually possible, even when we judge the emotion as non-native speakers, based on foreign language speech. However, full-blown, prototypical emotions are not so frequent in everyday communication. More often, we deal with more subtle states or attitudes for which the dimensional approach might be a justified choice (Laukka 2004). Furthermore, the dimensional approach has been observed to be more suitable

for reflecting the fluctuations of emotion over time (Cowie & Cornelius 2003; Cowie et al. 2000).

Example visualizations of two types of emotion rating scales are shown in Figure 14. The images are examples of the graphic controls developed for the experiments realized with the use Annotation Pro. They can be used either in annotation tasks performed with Annotation Pro or as part of a perception experiment set-up. Labelling is performed by clicking on the picture. As a result, the Cartesian coordinates of the clicked point are saved as the label of the respective time-aligned segment in the annotation layer. This makes it possible, among other things, to track changes in emotion over time.

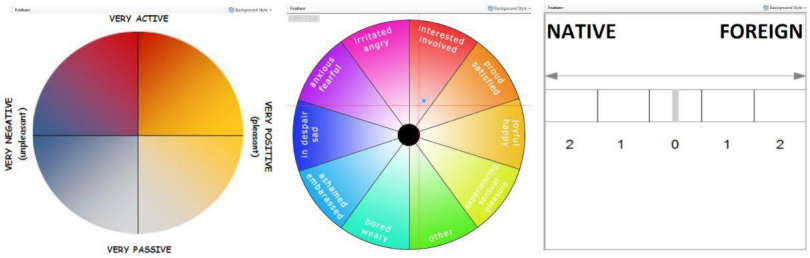

Figure 14. Example visualizations of three different emotion rating scales (cf. Klessa et al. 2013; Klessa et al. 2015).

The left-hand image is an example of a dimensional feature space that can be used for evaluation of the perceived emotion valence/activation. The middle image represents a mixed rating scale including nine categories of emotion families and one additional category, "other". The visualizations were created based on the subject literature (Banse & Scherer 1996; Bänziger et al. 2006; Cowie et al. 2000) and used for perception-based tests of classification of emotional speech. The rating scale in the middle picture is mixed, in the sense that it involves both discrete labels of emotion categories (e.g., irritated-angry, joyful-happy, proud-satisfied) and a dimensional element, namely the distance from the middle of the circle, which can be used to rate the intensity of emotion. The third picture can be treated as a representation of either a mixed or discrete rating scale, depending on the instructions provided to the annotators. Subjects may be informed that the feature space represents a continuous rating scale with additional 5-point scaling (2-1-0-1-2) information, where the middle (zero) means that there is no certainty what the answer should be. The more certain

the participant is about the answer, the further from the middle the answer point should be located, either to the left (when the sound is judged to be the same as the listeners' language, i.e., native) or to the right (when perceived as foreign). Alternatively, the rating scale may be used as a categorical one, with five major points or levels (cf. also Likert 1932).

Example annotation specifications

The tables below present inputs to the specifications of annotation of speech (Table 1) and paralinguistic features of speech (Table 2).

Table 1 contains a part of the segmentation and orthographic transcription guidelines for the MultiCo multimodal corpus (developed within the DARIAH.PL project; see acknowledgements). The description is slightly adjusted to make it clear to readers without access to the entire specification. Successive steps of segmentation and transcription are described in detail and mapped to specific tiers in the multi-tier annotation. The example refers only to the annotation of monologues. In case of more interlocutors, additional layers should be added.

Table 2 lists a selection of features potentially useful for paralinguistic annotation, together with example values and possible types of description using discrete, continuous or mixed rating scales. Depending on the feature type and the nature of the data, paralinguistic information may be included either in the form of time-aligned annotations or as part of the corpus metadata. For some of the features the extraction of acoustic correlates of the perceived phenomena is a standard and common procedure; examples are indicated in the table, for example as "objective measure".

	Activity	Tier
0	Transcripts are available as PDF files in case of some of the recordings. However, they are stylistically adjusted and they do not strictly reflect what was said. If you use them, please check them against actual spoken utterances, normalize and convert to a plain text format.	**n/a**
1	Mark interpausal units (IU), which are parts of utterances separated by a silent pause with a minimum duration of 100 ms.	**Phrase**
2	When the duration of IUs exceeds 5 seconds, divide them into smaller ones on the basis of prosodic criteria.	**Phrase**
3	Transcribe the utterances orthographically, according to what you hear. There is no need to use any punctuation.	**Phrase**
4	Transcribe abbreviations and acronyms as they are pronounced by the speaker, e.g., "HBO" as "age bee oh".	**Phrase**
5	Transcribe numbers as they are pronounced by the speakers, e.g., "in the year two thousand ten".	**Phrase**
6	Transcribe unfinished words as well, replacing the missing parts with a tilde (swung dash), e.g., driv~ (probably unfinished "driven"). If there are doubts regarding the spelling of an unfinished word, stick to the rule of "readability": what you write should sound, when read, as close as possible to what was in the recording.	**Phrase**
7	In case you are uncertain of what has been said by the speaker, transcribe the closest approximation of what you can hear.	**Phrase**
8	Filled pauses or hesitations should be transcribed as close as you can to how you hear them, as if they were regular words, but always add an asterisk as the last character (ehm*). If the sound is very unclear, barely articulated, or cannot be easily categorized as a hesitation (like a prolonged final sound of a word), just ignore it.	**Phrase**
9	Mark the places where the transcription is uncertain on a separate tier with the tag UNCLEAR.	**Issue**
10	Ignore stretches of speech that are totally incomprehensible and impossible to transcribe. Mark them on the Issue tier with the tag UNINTELLIGIBLE.	**Issue**

	Activity	Tier
11	Non-speech sounds are transcribed on a separate tier, using tags referring to the type of sound or distortion (categories and abbreviations based on Fischer et al., 2000): – SPK (speaker noise – sounds like coughing or yawning) – INT (intermittent noise – sudden sounds not coming from the speaker, e.g., knocking, beating) – STA (stationary noise – continuous, relatively stable sounds not coming from the speaker, e.g., continuous buzzing).	**Noise**
12	Other comments that you want to submit regarding a given portion of the recording can be made in the Comment tier. If the problem requires immediate discussion, contact the coordinator and formulate your doubts in the Comment field of the current session in Corpus Mini.	**Comment**
13	Once the orthographic transcription is finished, duplicate the orthographic transcription tier, name it "Phone", and proceed with automatic segmentation into phones, using the ANNPRO (CLARIN-PL Align) module (Koržinek et al. 2017; Klessa & Koržinek 2019).	**Phone**
14	Once the automatic segmentation and phonetic transcription is finished on the level of Phones, duplicate the Phrase tier again, and name it "Syllable". Proceed with automatic transcription and segmentation into syllables, using the ANNPRO module (syllabification algorithm uses a rule-set based on the rules defined by Śledziński (2007)).	**Syllable**
15	Once the orthographic transcription is finished, duplicate the orthographic transcription tier, name it "Word", and proceed with automatic segmentation into words, using the ANNPRO module.	**Word**
16	Review the transcripts, check if the files are complete, check them for obvious, major mistakes, and mark the session as finished in Corpus Mini.	**all the tiers**

Table 1. A definition of possible procedures and specifications of multilayer annotation (based on specifications of MultiCo corpus; DARIAH-PL project, cf. acknowledgements).

Feature		Example values, attributes, parameters
Speaker age	age in years / date of birth / approximate evaluation (in case of missing information)	Metadata, discrete
Speaker gender	male/female	Metadata, discrete
Speaker region of origin	name of the geographic region	Metadata, discrete
Language used	– native/non native – language ISO code(s) – code-switching (alternating between two or more languages)	Metadata and/or time-aligned annotation, discrete
Perceived voice quality	– categories (e.g., harsh, whispery, creaky, modal) – degree of nasalization – stability over a period of time (variability within utterances) – voice quality changes as related to the utterance structure or its components – overall voice quality judgement	Time-aligned annotation, discrete and/or continuous Objective measures: acoustic correlates of perceptually judged voice quality features
Affect/emotion	– labels (happiness, joy, sensual pleasure, surprise, fear, disgust, sadness, irony, …) – judgements of emotion in terms of valence, activation, potency, emotion intensity	Time-aligned annotation or metadata, depending on the approach; discrete labels or dimensions

Feature		Example values, attributes, parameters
Perceived expressivity	– stability over a period of time (variability within utterances) – overall judgement of speaker's expressivity	Time-aligned annotation, discrete and/or continuous
Non-verbal fillers	– can be annotated as "fillers" in general or subcategorized (and transcribed with approximate phonetic labels), e.g., vowel-like, nasal-like, compound (vowel–nasal), quasi-verbal ("hmm", "mhm"), non-verbal interjections	Time-aligned annotation, discrete
Self-repairs	– phrase level repairs – word level repairs	Time-aligned annotation, discrete
Non-speech speaker noises	– laughter, cough, yawn, breath, sigh, lip smack, sneeze, swallow, other	Time-aligned annotation, discrete and/or continuous (for some of the events intensity might be rated on a continuous scale)

Feature		Example values, attributes, parameters
Speech rate	– subjective judgement of speech rate	Time-aligned annotation, discrete and/or continuous (perception-based judgements of tempo might be judged on either a discrete or a continuous scale) Objective measure: number of speech units (e.g. speech sounds or syllables) per unit time (e.g. per second)
Voice pitch	– perceived height of voice	Time-aligned annotation, discrete and/or continuous (perception-based judgements of pitch might be judged on either a discrete or a continuous scale) Objective measure: fundamental frequency mean / variability fundamental frequency mean / variability
Voice intensity	– perceived intensity	Time-aligned annotation, discrete and/or continuous (perception-based judgements of intensity might be judged on either a discrete or a continuous scale) Objective measure: long-term intensity mean / variability

Feature		Example values, attributes, parameters
Idiosyncratic linguistic behaviour	– repeated word or grammar errors – items (words, phrases) repeated unconsciously, functioning as verbal fillers or adding emphasis – speaker-characteristic lexical item(s) – speaker-characteristic syntactic structures – speaker-specific repetitions after the interlocutor	Time-aligned annotation, discrete labels, time-aligned comments

Table 2. Selected paralinguistic features, their example values or attributes, types of description and possible rating scales (see also: Klessa 2013).

5. DATA AND METADATA MANAGEMENT

DATA AND METADATA

The materials collected during fieldwork expeditions may include audio or video files, images, pictures of artefacts related to language (e.g., inscriptions on tombstones), and printed or handwritten documents. Most often, two categories of content can be distinguished in the collection: data and metadata. In general, the term "data" can be defined as „representations of properties of the object area of a science that serve certain purposes for their users" (Lehmann 2004). In practice, it usually refers to the content that we planned to gather and use as the subject of our further analysis; as our main study material. It might also be defined as the "objects to be computed upon" (Borgman 2019). Metadata, on the other hand, is defined as data describing other data, or just "data about data" (see also: Good 2002). From the perspective of quantitative studies, one can interpret data as the variables, and metadata as possible factors in the analysis.

The more detailed meaning of the two notions depends on the design and purpose of the corpus. For example, the basic type of data for a phonetician will usually be acoustic information extracted from a sound file together with its time-aligned transcriptions (acoustic-phonetic data). The accompanying metadata may include various types of information about the speakers (such as their sex, age, region of origin, health condition, education, languages spoken, family information, social and family status), recording conditions (environment, background noises), session moderators, technical details (equipment specification and configuration, software used), etc.

The distinction between data and metadata is not always obvious, because metadata can become data and vice versa, depending on the aims of the study. For example, a corpus of quasi-spontaneous dialogues could be used by phoneticians to study features of conversational utterances with respect to timing or intonation, or perhaps paralinguistic features (see Chapter 4). Additional information about the speakers would be regarded as helpful metadata. However, the same corpus could be analysed from a different perspective, such as cultural anthropology; in this case, the focus would likely shift to such information as the descriptions of the speakers, their family relationships, education or other social information. Consequently, these components of the collection would be treated as data rather than metadata.

In any case, both data and metadata play important roles and may in fact be equally significant for research purposes. Therefore, equal care should be taken to collect, organize and protect them.

DATA SAFETY

Before anything else takes place after you have collected your data and metadata, make them safe. Write-protect the memory card (this is often done by means of a small slider) immediately after removing it from the audio recorder and before placing it in the socket of the computer or card reader. Remember that data loss may by caused by various factors, including hardware or software failures, technical problems, or human error. Preserve your original recordings and metadata, preferably in at least two copies that match the quality of the original and are stored separately (in physically distant places – not in the same drawer of your desk, for example).

Prepare backup copies in such a way that were you to decide to spend the rest of your life in Goa and never to contact any of your colleagues again, linguists would be able to determine what the content of the disk was. A very brief **metadata summary**, a "label", in a simple .TXT file in the main folder, may be enough. You should mention:

– the time and location of the recording sessions;
– the participants (how many speakers are recorded, who they are in terms of gender, nationality, age, or other relevant features);
– the authors (who collected the recordings – you, your collaborators) and contact information, if possible;
– file format definition (what is the format of the recordings and annotations or other files included in the collection);
– the structure of the archive (for example: original signals are stored in the folder "ORIG", processed signals are in the folder "PROC", metadata are in "META").

When you are not making new recordings but are dealing with already existing ones (for example, when you process historical data or annotate old recordings), an important precaution is to preserve the original naming conventions and other details related to the source materials. Keeping this information can be very useful should you or other users wish to go back to the very first version of the data.

APPROACHES TO DATA MANAGEMENT

Small linguistic datasets, designed and curated by individual researchers, can be managed using the file and folder structures typically available under any operating system. To organize them, a systematic naming convention, a logically designed hierarchy of folders, a metadata summary and a corpus

documentation file might be sufficient. Annotation tools usually support working with file collections, using annotation templates, which might also be helpful to systematically arrange the data.

In case of larger collections, however, completing the tasks of processing, annotation, and analysis of speech recordings, as well as metadata curation, often requires the collaboration of specialists with various backgrounds and technical qualifications. To support their collaboration and satisfy the diversified expectations of data users, it is reasonable to consider employing a data management system that will automatically supervise certain processes and prevent data loss.

AN EXAMPLE SOLUTION

Corpus Mini (Karpiński & Klessa 2018) is an example of a speech and video data management system. It is designed to deal with data and the corresponding metadata in a controlled manner (providing supervision over who accesses the data, when it happens, and in what way). It ensures remote access to the corpus for many users, and provides the possibility of managing the annotation and analysis processes even under adverse working conditions (often encountered during fieldwork), such as a weak or unstable Internet connection, or the need to use different operating system versions to connect to the database.

The system was designed using the client–server methodology (with Microsoft SQL Server; Karpiński & Klessa 2018). The client application can be installed on any desired number of personal computers connecting to the same central database, where all linguistic data and metadata are stored. The database is installed on a server where automatic backup copies are created as often as required, depending on the user's needs. Typically, during corpus creation (data upload, annotation and processing) backup copies are created frequently, perhaps once or twice a day. After the corpus creation work has been completed, when the files are not being modified so much, there is no need to save backup copies so often.

The model of integrated linguistic data, metadata and annotation workflow management with Corpus Mini is shown in Figure 15.

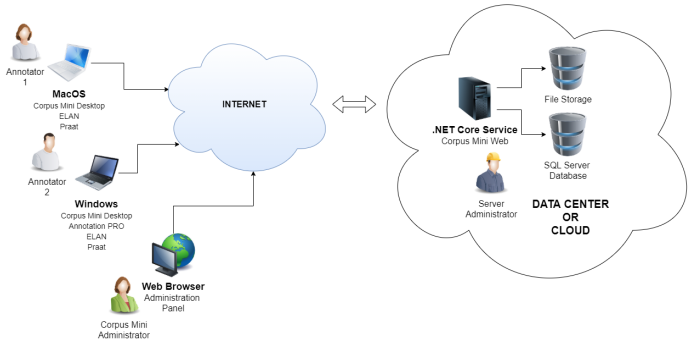

Figure 15. Integrated linguistic data, metadata and annotation workflow management with Corpus Mini (including updates introduced within the Mumostance project; cf. acknowledgements).

In the initial version of the system, the client application was implemented using Visual Studio .NET WinForms C#, which limited its use to Windows OS only. For the upgraded version, developed within the *MumoStance* project (see: Acknowledgements), the client application was ported to a platform-independent technology with JavaFx (https://openjfx.io/), an open-source client application platform for desktop, mobile and embedded systems. In this way, it became possible to access the central database using client computers with any operating system. Additionally, the upgraded version of the system makes it possible to access the database using an administration panel accessible via any web browser (also from mobile devices). The administration panel is designed for administrators and coordinators, and includes options for file and user management. It is also possible to monitor annotation workflow and progress, thanks to a comments panel and session status flags such as 'done', 'accepted', or 'locked' (unavailable for download).

The Corpus Mini management system helps to supervise annotation and analysis workflow in large corpora, containing both audio and video recordings. Usually, materials for a recording session include several files, such as the .WAV or .MP4 files, text files with a metadata summary, or annotation files. In Corpus Mini, all of the files are treated as one data bundle related to that particular recording session. The administrator can grant access to each session to one or more individual users. The system prohibits simultaneous use of the same data by different or unauthorized users, and therefore prevents data loss.

To work with a particular recording session, the user needs to download it to his or her local disk, using the desktop client application. After down-

loading, the annotations and multimedia can be inspected using ELAN, Praat or Annotation Pro. The external tools are launched by pressing the respective buttons in the Corpus Mini desktop client interface.

The most basic metadata (e.g., session or speaker identifier, language of the recording) are displayed both in the client application and in the web administration panel. The full metadata form (currently comprising up to twenty configurable fields) is available in the web panel. Administrators can define the types of metadata stored in the database, and fill in the metadata forms. The metadata field specifications can be defined by the user, and can refer to data features, speaker information, recording environment, equipment used, or any other properties. The metadata sheets can also be created in external tools (e.g., online forms filled in by interested parties) and imported from .CSV file formats to the relational database. Conversely, it is possible to export the metadata fields from the database to a standardized file format, for example, one compatible with the recommendations of the Dublin Core Metadata Initiative (Weibel & Koch 2000).

The functionalities of Corpus Mini were tested in several projects and improved based on the experience gained from those projects. Features of the program that might be useful in fulfilling the functional requirements of speech data management systems are listed below:

- annotation file management; including version control and backup copies;
- annotation and analysis of speech recordings with external annotation tools run through the program (e.g., Praat, Annotation Pro);
- annotation of one or more associated video files with an external annotation tool (ELAN) run through the program;
- user accounts for database users;
- access rights management;
- assignment of selected sessions to particular users;
- remote or local access to the database;
- automatic blocking of data currently in use by another user (only one person can edit data at a time);
- work-time statistics;
- (meta)data searching and filtering;
- each session bundle must include certain files (e.g. a .WAV file), while other files are allowed but optional (e.g., documentation files, photographs, consent forms);
- centralized data storage on a server;
- flexible metadata configuration;
- bulk annotation file import and export;
- bulk metadata sheet import and export.

DATA SHARING AND PUBLICATION

Corpora can be shared using infrastructure designed specifically for the purpose (using the organization's own servers and applications) or via existing repositories. Creating custom infrastructure will require time, effort and money, but in some cases it will be necessary, as it may ensure better control over the data. Frequently, however, an adequate choice will be one of the well-established online repositories that provide long-term hosting of linguistic data without any fees. In the case of some of these repositories, deposits can be made only on condition that some of the data are shared on an open access basis or other kind of access licence.

The access rules in language repositories vary and are often defined individually for each individual deposit. Data depositors usually need to configure the access rights and provide licensing information when uploading data to the repository. Some materials are made publicly available for anyone, while various limitations may apply to others. For example, with some data, the user may be asked to contact the authors or contributors to obtain permissions, and in other cases they are required to explain the purpose for which the data is to be used. Much depends on the type of data, and the access rules are usually more restrictive when sensitive content is involved.

ONLINE REPOSITORIES – EXAMPLES

Below is a list of links to some online language repositories. Each of the repositories enables the submission of one's own materials for the purpose of **archiving, sharing and publication**.

- Documentation Of Endangered Languages (DOBES): https://dobes.mpi.nl/
- The Endangered Languages Project: www.endangeredlanguages.com
- The Archive of the Indigenous Languages of Latin America (AILLA): www.ailla.utexas.org
- The Language Bank of Finland (Kielipankki): www.kielipankki.fi
- The Talkbank Project: www.talkbank.org
- Digital Repository for Data Depositing and Archiving (DSpace): www.clarin-pl.eu/dspace

See also: Drude et al. 2012; Kung & Sherzer 2013; Pol et al. 2018; MacWhinney 2007.

ARE THE SPEAKERS REALLY ANONYMOUS?

Dealing with sensitive data requires appropriate legal and ethical solutions to be applied (Chapter 1). It also poses technological challenges. In some cases, for example, speakers agree to participate in the recordings only if they are anonymized, that is, processed in such a way that it is not directly possible to identify the speakers.

Even if speakers' names, surnames and exact dates of birth are encrypted, any piece of information that we have collected and published might help to identify the speaker (e.g., gender, age, nationality). Finally, the speaker's voice itself – not to mention video images – contains information that can make it possible to identify the speaker. Anonymizing audio and video recordings is challenging from a technical point of view, but also in many cases it causes so much information loss that the anonymized data may become useless for certain purposes. It is thus important to inform the speakers that even if certain steps are taken to obscure identities, full anonymization might not be possible. It is the researcher's responsibility to clarify the situation for participants and obtain their informed consent. A useful overview of practical advice on dealing with linguistic data, including anonymization aspects, can be found in the document *Recommendations on Good Practice in Applied Linguistics* (*BAAL* 2021). Even though *BAAL* is a forum with a British focus, the above document can be applied by a broader audience; many recommendations are universal, and others may be adapted to the specific needs of the researcher.

Sometimes you need to keep speakers' names and addresses for further contact. Even if you are allowed (or obliged) to keep these data, never keep them together with a coding table that translates the names into codes. Such a table may contain entries like "Peter Newman -> PENE_M", "Penelope Newman -> PENE_F"; with more "PENE" speakers you may consider additional indexing, e.g. PENE_01_F, or using the next letter in the string, e.g. PNNE_F for Penelope and PTNE_M for Peter. Any rule can be used, but it should be consistent for the whole corpus and generate unique ID names. If a data management system is used, it will often support control over the session names and will not allow the creation of two sessions with exactly the same names.

The method of coding presented above as an example is not in fact very safe. You may want to create a more elaborate system, using only numbers or alphanumeric symbols in a less obvious way.

Even if your recordings are described in detail and systematically processed, it is quite important to keep the material in its original form whenever possible. WAV files will still be readable for many years, while databases and corpus management systems may evolve, be abandoned by producers, become

very expensive, or be difficult to run under new operating systems. Whatever happens, you should still be able to access the original material.

An important step in conducting a recording session is to obtain formal consents from the speakers (see Chapter 1 on legal issues). From the legal point of view, the written consent of each participant in a conversation is usually sufficient. Sometimes the consent can also be part of the recorded file. In any case, the consent will never be anonymous. Also, even for anonymized data, researchers should be able to identify the material provided by particular speakers, because it may happen that at some time after the recording, they might wish to modify or withdraw their consent. Therefore, another technical requirement is to archive the consent agreements in such a way that they remain available in the long term, but do not compromise sensitive information. Most institutions and companies provide support in case of ethical or legal questions regarding sensitive data management. In European organizations, so-called data protection officers (DPOs) are appointed; their role is to ensure that the personal data of the organization's staff, customers, providers or other data subjects are processed in compliance with the applicable data protection rules, primarily the General Data Protection Regulation or **GDPR** (see: https://gdpr-info.eu/).

DATA RECYCLING. INTEROPERABILITY AND RE-USABILITY ISSUES

Anyone involved in speech data collection and description will understand how expensive and time-consuming it is to compile a fully annotated corpus. This is just one of the reasons to give the corpus more than one life, or as Borgman (2019) puts it, an "after-life". Access to different linguistic resources increases research potential. For example, we might find more answers by applying current techniques of data analysis or processing to older corpora, or by replicating earlier experiments using newer data. Shared linguistic corpora initially created for one purpose (e.g., phonetic experiments or language documentation) may in their next life-cycle become very useful for another (e.g., development of speech and language technology applications, education purposes or dissemination of knowledge).

The efficient re-use of resources is easier to achieve thanks to good data and metadata organization, as well as to the development of data collection techniques and the growing capacities of digital repositories. Standardized metadata formats and structures support resource discovery and **re-usability** (e.g., Bird & Simons 2001; Bird & Simons 2003; Weibel & Koch 2000; Nathan & Austin 2004). They also enhance the **interoperability** of various linguistic

resources. Ide & Pusteyovsky (2010) define interoperability as "a measure of the degree to which diverse systems, organizations, and/or individuals are able to work together to achieve a common goal".

For such collaboration to become possible, it is indispensable to share and exchange information. Interoperability may thus be considered in relation to many different areas, such as metadata structures, specifications of datasets and software, or the standards of archiving and sharing.

METADATA STRUCTURES

Example approaches to structured metadata for linguistic resources can be viewed at:

OLAC, The Open Language Archives Community: http://www.language-archives.org/
DCMI, Dublin Core Metadata Initiative innovation: https://www.dublincore.org/

On one hand, the potential of linguistic data re-usability and interoperability has been appreciated in many initiatives, including DoBes (http://dobes.mpi.nl/) and CLARIN (https://www.clarin.eu/) (see also: Brugman et al. 2002; Váradi et al. 2008). For example, see the Polish Cued Speech Corpus of 20 Hearing Impaired Children at https://phonbank.talkbank.org/access/Clinical/PCSC.html (Trochymiuk 2008; Lorenc 2019-2020). After more than 20 years from its creation, the corpus was curated and made accessible through Talkbank in the USA and stored at the The Language Archive (TLA) in The Netherlands. The data curation process was supported by the team of CLARIN K-Centre for Atypical Communication Expertise (https://ace.ruhosting.nl/) and the DELAD group (Lee et al. 2021).

On the other hand, the need for re-usability, sharing, and interoperability poses new technological challenges. For example, file formats and data structures differ between corpora, which often makes it difficult to apply the same tools to analyse or process different datasets with the same software tools. Various attempts are made to overcome the problems caused by such diversity. Some of them result in proposals for completely new common standards, while others focus on convertibility features and better support for information exchange (Ide & Romary 2007).

One of the crucial conditions for shared resources to be made re-usable is the transparent definition of copyright licences. Once you publish a resource on the Internet and enable downloads (even without any technical restrictions), questions will usually arise about the actual status of the data, for example:

– what can or cannot be done with the data?
– how should we refer to the resource?
– can we modify the data after downloading?
– can we re-publish the data?
– can we publish analysis results based on the data?

Many potential users of the resource will be hesitant to work with data for which answers to the above questions are not clear. A convenient and widely recommended way to provide answers to such questions is to include them in the copyright licence and to publish the licence together with the resource. Many types of licences can be distinguished, ranging from very restrictive, proprietary ones to open access and open-source licences. A detailed discussion of different licence features can be found at the *GNU* (1998) project website: https://www.gnu.org/licenses/license-list.en.html.

CREATIVE COMMONS (CC) COPYRIGHT LICENCES

Creative Commons copyright licences are among the most popular in many domains, including linguistic research and education.

Read more at: https://creativecommons.org/

All Creative Commons licences are designed in such a way that for each licence three "layers" are available, representing three formats of the licence for three different groups of addressees: the legal code for lawyers, a human-readable description for researchers, creators and educators, and finally, a machine-readable format that can be used to automatically interpret the licence with software systems or search engines. The latter is vital for the so-called CC-search engine, available at the https://creativecommons.org/ website. This engine supports Internet searches where content (such as images) is filtered based on the licences assigned to it. Therefore, your online content will be found only if it has been attributed with a CC licence; otherwise, it will be invisible to the search engine.

An example application of Creative Commons licences is the Open Educational Resources (OER). The OER are either (1) in the public domain or (2) licensed in a manner that provides everyone with free and perpetual permission to engage in the following five activities, defined under "Open" in *Open Content and Open Educational Resources* (originally written by David Wiley and published freely under a Creative Commons Attribution 4.0 licence at http://opencontent.org/definition/):

1. Retain – make, own, and control a copy of the resource (e.g., download and keep your own copy)
2. Revise – edit, adapt, and modify your copy of the resource (e.g., translate into another language)
3. Remix – combine your original or revised copy of the resource with other existing material to create something new (e.g., make a mashup)
4. Reuse – use your original, revised, or remixed copy of the resource publicly (e.g., on a website, in a presentation, in a class)
5. Redistribute – share copies of your original, revised, or remixed copy of the resource with others (e.g., post a copy online or give one to a friend).

A number of centres and organizations have been established whose mission is to develop and promote OER. An example of a centre dedicated specifically to building OER based on linguistic resources is COERLL (Center for Open Educational Resources & Language Learning). Ready-to-use resources and details of ongoing projects are shared via the COERLL website at www.coerll.utexas.edu. They include educational materials for many languages, including English, French, Arabic and Chinese, but also under-resourced languages such as Bangla, K'iche', Malayalam and Nahuatl (see also: other OER centres at nflrc.org, and Blyth 2013).

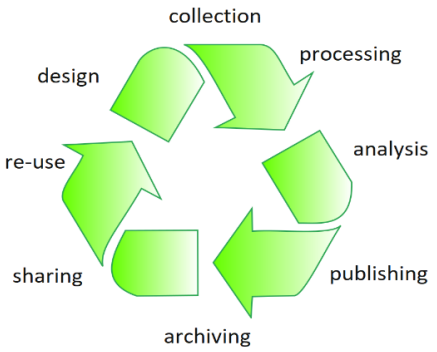

Figure 16. A visualization of the (meta)data recycling process.

REFERENCES

Audacity Team (2021). Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 3.0.0 retrieved March 31st 2021 from https://audacityteam.org/

BAAL (2021). Recommendations on Good Practice in Applied Linguistics. 4th Edition. Retrieved on 30 December 2021 from: www.baal.org.uk

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70(3)*, 614.

Barsties, B., & De Bodt, M. (2015). Assessment of voice quality: current state-of-the-art. *Auris Nasus Larynx, 42*(3), 183-188.Benson, P. (2014). Narrative inquiry in applied linguistics research. *Annual Review of Applied Linguistics, 34,* 154-170.

Bänziger, T., Pirker, H., & Scherer, K. (2006). GEMEP – GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions. *Proceedings of LREC (6),* 15-19.

Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing, 8*(4), 243-257.

Bigi, B. (2015). Uncertainty-tolerant framework for multimodal corpus annotation. Retrieved on 30 November 2021 from: https://hal.archives-ouvertes.fr/hal-01455310

Bigi, B., & Bertrand, R. (2016). Laughter in French spontaneous conversational dialogs. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16),* 2168-2174.

Bigi, B., & Meunier, Ch. (2018). Automatic speech segmentation of spontaneous speech. *In Revista de Estudos da Linguagem*. International Thematic Issue: Speech Segmentation. Editors: Tommaso Raso, Heliana Mello, Plinio Barbosa, vol. 26, no 4, e-ISSN 2237-2083.

Bird, S., & Simons, G. (2001). The OLAC Metadata Set and Controlled Vocabularies. In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources*. Retrieved on 30 November 2021 from: https://aclanthology.org/W01-1506.pdf

Blyth, C. (2013). Open Educational Resources (OER). In C. Chapelle (Ed.). *The Encyclopedia of Applied Linguistics*. Blackwell Publishing. See also: https://www.academia.edu/12888147/Blyth_C_2013_Open_Educational_Resources_OER_In_C_Chapelle_ed_The_Encyclopedia_of_Applied_Linguistics_Blackwell_Publishing?from=cover_page

Boersma, P., & Weenink, D. (1992–2021). Praat: doing phonetics by computer [Computer program]. Version 6.2. Retrieved on 15 November 2021 from: https://www.praat.org

Boré, G., & Peus, S. (1999). *Microphones for Studio and Home-Recording Applications.* Berlin: Druck-Centrum Fürst.
Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review, 1(1).* Retrieved on 15 December 2021 from: https://escholarship.org/content/qt0zp8k7rs/qt0zp8k7rs.pdf
Brody, J. L., Gluck, J. P., & Aragon, A. S. (2000). Participants' understanding of the process of psychological research: Debriefing. *Ethics & Behavior*, *10*(1), 13-25.
Brugman, H., Levinson, S. C., Skiba, R., & Wittenburg, P. (2002). The DOBES archive: Its purpose and implementation. In *The 3rd International Conference on Language Resources and Evaluation (LREC 2002). Workshop on Tools and Resources in Field Linguistics*. European Language Resources Association.
Buchstaller, I., & Alvanides, S. (2013). Employing geographical principles for sampling in state of the art dialectological projects. *Journal of Linguistic Geography*, *1*(2), 96-114.
Buchstaller, I., & Khattab, G. (2013). Population samples. *Research methods in linguistics*, 74-95.
Campbell, N. (2002, May). Recording techniques for capturing natural every-day speech. In *LREC*.
Chafe, W. (Ed.) (1980). *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood, New Jersey: Ablex.
Coalson, J. (2000–2009), *FLAC format Xiph. Org Foundation Std.* Retrieved on 30 November 2021 from: https://xiph.org/flac/format.html and https://xiph.org/flac/documentation_format_overview.html
Codó, E. (2008). Interviews and questionnaires. *The Blackwell guide to research methods in bilingualism and multilingualism*, 158-176.
Corbett, I. (2021). *Mic It!* New York: Routledge.
Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication, 40(1-2)*, 5–32.
Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). FEELTRACE: An instrument for recording perceived emotion in real time. *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle. Retrieved on 20 February 2018 from: http://www.isca-speech.org/archive_open/archive_papers/speech_emotion/spem_019.pdf
Creative Commons Copyright Licenses homepage. Retrieved on 30 October 2021 from: https://creativecommons.org/
Czoska, A., Klessa, K., & Karpiński, M. (2015). Polish infant directed vs. adult directed speech: Selected acoustic-phonetic differences. In *ICPhS Proceedings.*
Decker, P. D., & Nycz, J. (2013). The technology of conducting sociolinguistic interviews. In *Data collection in sociolinguistics,* 134-146. Routledge.

DoBes: Dokumentation bedrohter Sprachen / Documentation of Endangered Languages Project (2000-2013). Available on-line at: http://dobes.mpi.nl/

Drude, S., Trilsbeek, P., & Broeder, D. (2012). Language Documentation and Digital Humanities: The (DoBeS) Language Archive. In Digital Humanities Conference 2012 (pp. 169-173).

Eckert, P. (2013). Ethics in linguistic research. *Research methods in linguistics*, 11-26.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6(3-4)*, 169-200.

Ferré, G. (2010). Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French. Language Resources and Evaluation, Workshop on Multimodal Corpora, May 2010, Malta. W6, 86-91 <hal-00485797>. Retrieved on 10 March 2018 from: https://hal.archives-ouvertes.fr/hal-00485797/document

Ferré, G. (2012). Functions of three open-palm hand gestures. *Journal Multimodal Communication, 1(1),* 5-20.

Fischer, V., Diehl, F., Kiessling, A., & Marasek, K. (2000). *Specification of Databases – Specification of annotation*. SPEECON Deliverale D214.

Francuzik, K., Karpiński, M., Kleśta, J., & Szalkowska, E. (2005). Nuclear melody in Polish semi-spontaneous and read speech. Evidence from the Polish Intonational Database PoInt. *Studia Phonetica Posnanensia (7)*, 97-128.

GDPR (2016). REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Retrieved on 30 November 2021 from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679. See also: https://gdpr-info.eu/

Gibbon, D. (2013). TGA: a web tool for Time Group Analysis. *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop.* Aix-en-Provence, 66-69.

Gibbon, D., Moore, R., & Winski, R. (Eds.) (1997). *Handbook of standards and resources for spoken language systems*. Walter de Gruyter. See also: http://wwwhomes.uni-bielefeld.de/gibbon/Handbooks/gibbon_handbook_1997/index.html

Giles, H., & Coupland, N. (1991). *Accommodating language.* Open University Press.

GNU (1998). The GNU project home page. Retrieved on 30 December 2021 from: http://www.gnu.org

Godsill, S., Rayner, P., & Cappé, O. (2002). Digital audio restoration. In *Applications of digital signal processing to audio and acoustics* (pp. 133-194). Springer, Boston, MA.

Good, J. (2002). A gentle introduction to metadata. Retrieved on 30 November 2021 from: http://www.language-archives.org/documents/gentle-intro.html

Grabe, E., & Post, B. (2002). Intonational variation in the British Isles. In *Proceedings of Speech Prosody 2002, International Conference*.

Hannesschläger, V., Scholger, W., & Kuzman, K. (2020). The DARIAH ELDAH consent form wizard. *DARIAH Annual Event 2020: Scholarly Primitives, 46.* Retrieved on 30 November 2021 from: https://dariah-ae-2020.sciencesconf.org/data/BookofAbstracts.pdf (see also: https://consent.dariah.eu/).

Haque, M., & Bhattacharyya, K. (2018). Speech Background Noise Removal Using Different Linear Filtering Techniques. In *Advanced Computational and Communication Paradigms* (pp. 297-307). Springer, Singapore.

Harrington, J., Palethorpe, S., & Watson, C. I. (2007). Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. In *Proceedings of Interspeech,* 2753-2756.

Hawkins, S., & Midgley, J. (2005). Formant frequencies of RP monophthongs in four age groups of speakers. *Journal of the International Phonetic Association*, *35*(2), 183-199.

Hedeland, H., & Schmidt, T. (2012). Technological and methodological challenges in creating, annotating and sharing a learner corpus of spoken German. *Multilingual Corpora and Multilingual Corpus Analysis*, *14*, 25.

Holmes, D. S. (1976). Debriefing After Psychological Experiments. *American Psychologist*, 859.

Huber, D. M., & Runstein, R. E. (2018). *Modern recording techniques* (Ninth Edition). New York: Routledge.

Ide, N. & Pustejovsky, J. (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability for language technology. *Proceedings. of the Second International Conference on Global Interoperability for Language Resources.* Hong Kong, China.

Ide, N., & Romary, L. (2007). Towards International Standards for Language Resources. In: Laila Dybkjær and Holmer Hemsen and Wolfgang Minker. *Evaluation of Text and Speech Systems*, Kluwer Academic Publishers, 263-284 (hal-00650597f). Retrieved on 30 November 2021 from: https://hal.inria.fr/hal-00650597

International Phonetic Association (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Jarmołowicz-Nowikow, E. (2019). *Intencjonalność komunikacyjna gestów wskazujących*. Wydawnictwo Naukowe UAM.
Jarmołowicz-Nowikow, E., & Karpiński, M. (2011). Communicative intentions behind pointing gestures in task-oriented dialogues. *Proceedings of GESPIN*. Bielefeld.
Jassem, W. (1999). English Stress, Accent and Intonation Revisited, *Speech and Language Technology (3)*, 33-50, Poznań.
Karpiński, M. (2007). The intonational realization of requests in Polish task-oriented dialogues. In *International Conference on Text, Speech and Dialogue*, 556-563. Springer, Berlin, Heidelberg.
Karpiński, M. (2013). Acoustic features of filled pauses in Polish task-oriented dialogues. *Archives of Acoustics, 38(1)*, 63-73.
Karpiński, M. (2014). The sounds of language, In: Nau N., Hornsby M., Karpiński M., Klessa K., Wicherkiewicz T., Wójtowicz R. (Eds.), *Book of Knowledge of Languages in Danger.* Online at: http://languagesindanger.eu/book-of-knowledge/
Karpiński, M., Czoska, A., Jarmołowicz-Nowikow, E., Juszczyk, K., & Klessa, K. (2018). Aspects of gestural alignment in task-oriented dialogues. *Cognitive Studies | Études cognitives, 2018(18)*. https://doi.org/10.11649/cs.1640
Karpiński, M., Klessa, K., & Czoska, A. (2014). Local and global convergence in the temporal domain in Polish task-oriented dialogue. *Proceedings of Speech Prosody 2014*, 743-747, DOI: 10.21437/SpeechProsody.2014-137
Karpiński, M., & Jarmołowicz-Nowikow, E. (2010). Prosodic and Gestural Features of Phrase-internal Disfluencies in Polish Spontaneous Utterances, *Proceedings of Speech Prosody 2010*, Chicago.
Karpiński, M., & Klessa, K. (2018). Methods, tools and techniques for multimodal analysis of accommodation in intercultural communication. *Computational Methods in Science and Technology, 24(1)*, 29-41.
Kibrik, A. E. (2017). *The methodology of field investigations in linguistics*. De Gruyter Mouton.
Klessa, K., Karpiński, M., Wagner, A. (2013). Annotation Pro – a new software tool for annotation of linguistic and paralinguistic features. In D. Hirst & B. Bigi (Eds.) Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence, 51-54.
Klessa, K., & Gibbon, D. (2014). Annotation Pro + TGA: automation of speech timing analysis, *Proceedings of the 9th Language Resources and Evaluation Conference*, Reykjavik, Iceland. ISBN 978-2-9517408-8-4.
Klessa, K., & Karpiński, M. (2018). Speaking style variation in laboratory speech: A perception study. In *Proceedings of the 9th International Conference on Speech Prosody 2018,* Poznań, 517-521.

Klessa K., Karpiński M., & Czoska A. (2015). Design, structure, and preliminary analyses of a speech corpus of infant directed speech (IDS) and adult directed speech (ADS). In Kloekhorst A., Kohlberger M. (Eds.) *Proceedings of the 48th Annual Meeting of the Societas Linguistica Europaea. Book of Abstracts*, Leiden, 188-189.

Klessa, K., & Koržinek, D. (2019). Annotation Pro+ CLARIN-PL Align: automatic segmentation and transcription module for desktop uses. In *Proceedings of 2nd Language & Technology Conference*. Poznań.

Klessa, K., Nau, N., Orlovs, O. (2017). Timing patterns variability in Latgalian read speech. In: Abrahamsen, J. E., Koreman, J., & Dommelen, W. A. (Eds.) *Nordic prosody: Proceedings of the XIIth Conference, Trondheim 2016.*

Klessa, K., Wagner, A., Oleśkowicz-Popiel, M., & Karpiński, M. (2013). Paralingua – a new speech corpus for the studies of paralinguistic features. *Procedia-Social and Behavioral Sciences (95)*, 48-58.

Klessa, K., & Wicherkiewicz, T. (2015). Design and Implementation of an Online Database for Endangered Languages: Multilingual Legacy of Poland. In *Input a Word, Analyse the World: Selected Approaches to Corpus Linguistics*. In: Almeida, F.A., Barrera, I.O., Toledo, E.Q. & Cuervo, M.S. (Eds.), Newcastle upon Tyne: Cambridge Scholars Publishing, ISBN (10) 1-4438-8513-4.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech*, *41*(3-4), 295-321.

Koržinek, D., Marasek, K., Brocki, L., & Wołk, K. (2017). Polish read speech corpus for speech tools and services. In *Selected papers from the CLARIN Annual Conference, 2016, Aix-en-Provence, 26–28.10.2016, CLARIN Common Language Resources and Technology Infrastructure (136),* 54–62, Linköping University Electronic Press.

Kung, S. S., & Sherzer, J. (2013). The archive of the indigenous languages of Latin America: An overview. Oral tradition, 28(2).

Labov, W. (1972). Some principles of linguistic methodology. *Language in society*, *1*(1), 97-120.

Labov, W. (1984). Field methods of the project on linguistic change and variation. *Language in use: readings in sociolinguistics*, ed. by John Baugh and Joel Sherzer, 28–53.

Labov, W., Ash, S., & Boberg, C. (2008). Sampling and field methods. In *The Atlas of North American English,* 21-35. De Gruyter Mouton.

Laukka, P. (2004). *Vocal expression of emotion: discrete-emotions and dimensional accounts*. PhD Thesis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Science, Department of Psychology, Uppsala University.

Laver, J. (1994). *Principles of phonetics*. Cambridge University Press.
Lee, A., Bessell, N., Van den Heuvel, H., Saalasti, S., Klessa, K., Müller, N., & Ball, M.J. (2021). The latest development of the DELAD project for sharing corpora of disordered speech. In *Clinical Linguistics & Phonetics (35)*. https://doi.org/10.1080/02699206.2021.1913514
Lehmann, C. (2004). Data in linguistics. *The Linguistic Review 21(3/4)*, 275-310. Retrieved on 30 November 2021 from: https://christianlehmann.eu/publ/lehmann_data_in_linguistics.pdf
Lehmberg, T., Rehm, G., Witt, A., & Zimmermann, F. (2008). Digital text collections, linguistic research data, and mashups: notes on the legal situation. *Library Trends*, *57*(1), 52-71.
Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 44–53.
Lorenc, A. (2019 – 2020). Collection "Polish Cued Speech Corpus of Hearing-Impaired Children". The Language Archive, Retrieved on 20 December 2021 from: https://hdl.handle.net/1839/dbcd8568-d17d-4861-94bb-aa553e943399.
Low, E. L., Grabe, E., Nolan, F. (2001). Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech 43 (4),* 377-401.
Machač, P., & Skarnitzl, R. (2009). *Principles of phonetic segmentation*. Epocha.
MacWhinney, B. (2007). The Talkbank project. In Creating and digitizing language corpora (pp. 163-180). Palgrave Macmillan, London.
Makarova, V., & Petrushin, V. A. (2003). The Map Task Corpus of Spoken Russian. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.
Mallinson, C. (2018). Ethics in linguistic research. *Research methods in linguistics*, 57-84.
Mann, S. (2011). A critical review of qualitative interviews in applied linguistics. *Applied Linguistics*, *32*(1), 6-24.
Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44(2), 314-324. doi:10.3758/s13428-011-0168-7
McNeill, D. (1985). So you think gestures are nonverbal? Psychological review, 92(3), 350. Retrieved on 28 November 2021 from: http://www.cogsci.ucsd.edu/~nunez/COGS160/McNeill_PS.pdf
Mihajlovic, M., & Todorovic, D. (2011). Loudness normalization. In *2011 19th Telecommunications Forum (TELFOR) Proceedings of IEEE,* 1111-1114.
Mueller, S. T., & Piper, B. J. (2014). The Psychology Experiment Building Language (PEBL) and PEBL Test Battery. *Journal of Neuroscience Methods*, 222, 250-259. doi: 10.1016/j.jneumeth.2013.10.024

Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I., & Evans, N. (2019). The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding. *Proceedings of. Interspeech 2019*, 3695-3699, doi: 10.21437/Interspeech.2019-2647

Novak, J. R., Minematsu, N., & Hirose, K. (2012). WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *Proceedings of the 10th Int. Workshop on Finite State Methods and Natural Language Processing,* 45-49.

Oğuz, H., Kiliç, M. A., & Şafak, M. A. (2011). Comparison of results in two acoustic analysis programs: Praat and MDVP. *Turkish Journal of Medical Sciences*, *41*(5), 835-841.

Pawera, N. (2010). *Practical Recording 1: Microphones.* London: SMT. ISBN 978-0-85712-245-2

Podesva, R. J., & Sharma, D. (Eds.). (2014). *Research methods in linguistics*. Cambridge University Press.

Podesva, R. J., & Zsiga, E. (2013). Sound recordings: acoustic and articulatory data. *Research methods in linguistics*, 169-194.

Pol M., Walkowiak T., Piasecki M. (2018). Towards CLARIN-PL LTC Digital Research Platform for: Depositing, Processing, Analyzing and Visualizing Language Data. In: Kabashkin I., Yatskiv I., Prentkovskis O. (Eds) Reliability and Statistics in Transportation and Communication. RelStat 2017. Lecture Notes in Networks and Systems, vol 36. Springer, Cham. https://doi.org/10.1007/978-3-319-74454-4_47

Poldy, C. A. (2001) Headphones. In: John Borwick (Ed.) *Laudspeaker and Headphone Handbook*, pp. 585-686, Oxford: Focal Press.

Port, R. (2008). All is prosody: Phones and phonemes are the ghosts of letters. In *Proceedings of the 4th Internal Conference on Speech Prosody,* 7-16.

Raineri, S., & Debras, C. (2019). Corpora and Representativeness: Where to go from now? *CogniTextes. Revue de l'Association française de linguistique cognitive*, Volume 19.

Rayburn, R. (2012). *Eargle's Microphone Book: From Mono to Stereo to Surround – a Guide to Microphone Design and Application*. Elsevier – Focal Press.

Reichel, U. D., & Kisler, T. (2014). Language-independent grapheme-phoneme conversion and word stress assignment as a web service. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2014*, 42-49.

Réveillac, J. M. (2017). *Musical Sound Effects: Analog and Digital Sound Processing*. John Wiley & Sons.

Rice, K. (2012). Ethical issues in linguistic fieldwork. In *The Oxford handbook of linguistic fieldwork*. Oxford: OUP.

Roederer, J. G. (2008). Sound Waves, Acoustic Energy, and the Perception of Loudness. In *The Physics and Psychophysics of Music,* 76-112. Springer, New York, NY.

Rumsey, F., & McCormick, T. (2006). *Sound and Recording: An Introduction.* Amsterdam: Elsevier – Focal Press.

Sankoff, D. (2008). Problems of representativeness. In *Sociolinguistics,* 998-1002. De Gruyter Mouton.

Schneider, W., Eschman, A., and Zuccolotto, A. (2012). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools, Inc.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2013). Paralinguistics in speech and language—State-of-the-art and the challenge. *Computer Speech & Language, 27(1)*, 4-39.

Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J. R., ... & Weninger, F. (2015). The INTERSPEECH 2015 computational paralinguistics challenge: nativeness, Parkinson's & eating condition. In *16th Annual Conference of the International Speech Communication Association*, Dresden.

Shenoi, B.A. (2006). *Introduction to digital signal processing and filter design*. John Wiley and Sons. ISBN 978-0-471-46482-2.

Steffen-Batogowa, M. (1975). *Automatyzacja transkrypcji fonematycznej tekstów polskich.* Państwowe Wydawnictwo Naukowe.

Stoian-Irimie, D., & Irimie, D. S. (2017). Digital Audio Restoration in Ethnomusicological Research. *Information and Communication Technology in Musical Field*, *8*(2), 63-70.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics, 21*, 121-137.

Śledziński, D. (2007). *Fonetyczno-akustyczna analiza struktury sylaby w języku polskim na potrzeby technologii mowy.* Unpublished PhD Thesis, Adam Mickiewicz University, Poznan, Poland.

Talmy, S. (2010). Qualitative interviews in applied linguistics: From research instrument to social practice. *Annual Review of Applied Linguistics*, *30*, 128-148.

Trochymiuk A., 2008, Wymowa dzieci niesłyszących. Analiza audytywna i akustyczna (Eng. Pronunciation of hearing-impaired children. Auditive and acoustic analysis), In: *Komunikacja językowa i jej zaburzenia (22)*, Lublin: Wydawnictwo UMCS.

Vogel, A. P., & Morgan, A. T. (2009). Factors affecting the quality of sound recording for speech and voice analysis. *International journal of speech-language pathology*, *11*(6), 431-437.

W3Schools: Introduction to XML. Retrieved on 30 November 2021 from: https://www.w3schools.com/xml/xml_whatis.asp

Walker, J. F., & Archibald, L. M. (2006). Articulation rate in preschool children: a 3-year longitudinal study. *International Journal of Language & Communication Disorders*, *41*(5), 541-565.

Walker, J. F., Archibald, L. M., Cherniak, S. R., & Fish, V. G. (1992). Articulation rate in 3-and 5-year-old children. *Journal of Speech, Language, and Hearing Research*, *35*(1), 4-13.

Weibel, S. L., & Koch, T. (2000). The Dublin core metadata initiative. *D-lib magazine, 6(12)*, 1082-9873.

Wells, J. C. (1997). SAMPA computer readable phonetic alphabet. In Gibbon, D., Moore, R. and Winski, R. (Eds.), *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter. Part IV, section B.

Wharton, T. (2003). Interjections, language, and the 'showing/saying' continuum. *Pragmatics & Cognition, 11(1)*, 39-91, http://ftp.phon.ucl.ac.uk/home/PUB/WPL/00papers/wharton.pdf

Wharton, T. (2009). *Pragmatics and non-verbal communication.* Cambridge University Press.

Wiley, D. (Accessed 2021) *The "Open" in Open Content and Open Educational Resources written and published freely under a Creative Commons Attribution 4.0 license*. Retrieved on 30 December 2021 from: http://opencontent.org/definition/

Winkworth, A. L., Davis, P. J., Adams, R. D., & Ellis, E. (1995). Breathing patterns during spontaneous speech. *Journal of Speech, Language, and Hearing Research, 38(1)*, 124-144.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of the 5th Language Resources and Evaluation Conference*, Genoa, Italy, 1556-1559.

Yu, J., Gibbon, D., & Klessa, K. (2014). Computational annotation-mining of syllable durations in speech varieties. In *Proceedings of 7th Speech Prosody Conference*, 20-23.

APPENDIX 1:

SOFTWARE TOOLS AND ONLINE RESOURCES

Please note that we promote freely available software. As always, it is important to give credit to the authors. The form of acknowledgment they expect is often explicitly described on the software web page, in accompanying materials, in help files, or in the "About the program" section in the menu. Some of the software tools listed below are also cited above in the reference list.

Annotation Pro
https://annotationpro.org
A piece of software dedicated mostly to speech annotation, but offering some analytic functions. Among its special features is semi-continuous annotation, where users click on custom graphics and the pointer co-ordinates are read and saved as an annotation. The workspace feature is very convenient when working with larger collections of signals. Another interesting task supported by Annotation Pro is preparing simple experimental procedures. Finally, it has a slot for plug-ins (in C#) which can immensely extend its possibilities.

Audacity
https://www.audacityteam.org/
Free audio editing software capable of multitrack recording. Besides basic editing functions, it offers a rich library of advanced plugins. While it is oriented towards general sound and music editing, field linguists will find it quite useful. Basic editing is extended with a range of advanced filtering and processing functions.
Audacity® software is copyright © 1999–2021 Audacity Team. The name Audacity® is a registered trademark.

ELAN
https://archive.mpi.nl/tla/elan
A popular program used mostly for video annotation. It can work with more than one video file simultaneously (plus corresponding audio files). It has a wide range of options that support manual annotation, as well as some analytic functions.

Praat
http://praat.org
A legendary program for "doing phonetics by computer" used by phoneticians worldwide. Although the user interface is sometimes criticized and perceived as archaic, most users agree that after some time they find it quite ergonomic. Besides a huge range of analytic functions, there are some surprising additions, like articulatory speech synthesis and artificial neural networks.

BAS
https://www.bas.uni-muenchen.de/Bas/BasHomeeng.html
The Bavarian Archive for Speech Signals (BAS) hosted by the University of Munich aims to make speech resources for contemporary spoken German, as well as tools for the processing of digitized speech, available to research and speech technology communities. The available tools, which include G2P services, are available not only for German but also for several other languages.

SPPAS
http://www.sppas.org/
SPPAS is a tool for automatic annotation and analysis of speech, and for the conversion of annotated files to a wide range of formats. SPPAS automatically produces annotations from a recorded speech sound and its orthographic transcription and/or from a video. It also estimates statistical distributions, and supports annotation mining, file management, and visualization of annotations. Available free of charge, with open source code.

CLARIN-PL Mowa
https://mowa.clarin-pl.eu/
Among others, this website provides services for speech recognition, speech alignment, separation of parts of the recording containing speech from others, recognition of speakers, keyword detection and phonetic translation. By default, the tools work for Polish language data. These tools can be used on their own or as part of a larger process to create audio corpora for research. More tools dedicated to the collection and analysis of speech and language resources are offered at https://clarin-pl.eu/.

Illustrations of the IPA
https://www.cambridge.org/core/journals/journal-of-the-international-phonetic-association/illustrations-of-the-ipa
Illustrations of the IPA are concise accounts of the phonetic structure of different languages using the International Phonetic Alphabet (IPA), accompanied

by audio recordings. Typically, each description also includes a transcript and a recording of the fable *The North Wind and the Sun*. A selection of illustrations has been made freely available for download.

SAMPA
https://www.phon.ucl.ac.uk/home/sampa/
The SAMPA (Speech Assessment Methods Phonetic Alphabet) website, including specification of the conversion between IPA and SAMPA as well as SAMPA specifications for a range of languages.

APPENDIX 2:

FURTHER READING

The scope and size of our book is limited. If you need to learn more, there are plenty of excellent publications that go beyond what we discuss here, in terms of methods and techniques, technology, or linguistic background and particular research contexts. Below, we list a small subset of sources which we have not cited, but which we consider informative and relevant to sound recording and processing, or linguistic data sets in general.

Linguistic/phonetic fieldwork – general

Bird, S. (2011). Phonetic Fieldwork in the Pacific Northwest. In *Proceedings of 17th Congress of Phonetic Sciences,* Hong Kong, 76-79.

Bowern, C. (2015). *Linguistic fieldwork: A practical guide*. Springer.

Chelliah, Shobhana L., and Jules Willem. *Handbook of descriptive linguistic fieldwork*. Springer Science & Business Media, 2010.

Chelliah, S. (2013). Fieldwork for language description. *Research Methods in Linguistics*, 51-73.

Dunham, J. R. W. (2014). *The Online Linguistic Database: software for linguistic fieldwork* (Doctoral dissertation, University of British Columbia).

Gordon, M. (2003). Collecting phonetic data on endangered languages. In *Proceedings of 15th International Congress of Phonetic Sciences,* Barcelona, 207-10.

Ladefoged, P. (2003). Phonetic fieldwork. In *Proceedings of 15th International Congress of Phonetic Sciences,* Barcelona, 203-206.

Ladefoged, P. (1993). Linguistic phonetic fieldwork: a practical guide. *UCLA Working Papers in Phonetics*, *84*, 1-25.

Sakel, J., & Everett, D. L. (2012). *Linguistic fieldwork: A student guide*. Cambridge University Press.

Audio/sound recording and processing technology

Bartlett, B., & Bartlett, J. (2017). *Practical recording techniques: The step-by-step approach to professional audio recording*. New York, NY: Routledge.

Eargle, John (2006). *Handbook of Recording Engineering*. Springer.

Everest, F. A., & Pohlmann, K. C. (2013). *Handbook of sound studio construction on a budget*. New York: McGraw-Hill.

Gibson, B. (2011). *Instrument & vocal recording*. Milwaukee, WI: Hal Leonard.

Gold, B., Morgan, N., Ellis, D., & Bourlard, H. (2011). *Speech and audio signal processing: Processing and perception of speech and music*. Oxford: Wiley.

Greene, P. D., & Porcello, T. (2010). *Wired for Sound: Engineering and Technologies in Sonic Cultures*. Middletown: Wesleyan University Press.
Howard, D. M., & Murphy, D. (2008). *Voice science, acoustics and recording*. San Diego, CA: Plural Pub.
Hynes, P. F. (2008). *Sound engineer*. Ann Arbor, Mich: Cherry Lake Pub.
Lane, C., Carlyle, A., & Creative Research in Sound Arts Practice (Organization) (2018). *In the field: The art of field recording,* Uniformbooks.
Morton, D. (2004). *Sound recording: The life story of a technology*. Greenwood Publishing Group.
Müller, M. (2021). *Fundamentals of Music Processing: Using Python and Jupyter Notebooks*. Springer.
Pelton, G. E. (1993). *Voice processing*. New York: McGraw-Hill.
Shea, M. (2005). *Studio recording procedures. How to record any instrument.* New York: McGraw-Hill.
Whitaker, J. C., Benson, K. B., Toole, F. E., & Shaw, E. A. G. (2002). *Standard handbook of audio engineering*, McGraw-Hill Education.
Woram, J. M. (1992). *Sound recording handbook*. Carmel, Ind: SAMS.

ACKNOWLEDGEMENTS

This work was supported by the projects:

– *COLING: Minority Languages, Major Opportunities. Collaborative Research, Community Engagement and Innovative Educational Tools*, MSCA RISE Horizon 2020, Grant Agreement ID: 778384.
– *Research work financed from the funds for science granted of the Ministry of Science of the Republic of Poland for the implementation of an international project co-financed under contract No. 4089/H2020/2018/2 in the years 2018-2023.*
– *MuMoStance: Multimodal Stancetaking: Expressive Movement and Affective Stance. Political Debates in German Bundestag and Polish Sejm*, project funded by the National Science Centre, Poland, ID: 2018/31/G/HS2/03633.
– *Digital Research Infrastructure for the Humanities and Arts DARIAH-PL*, funded from the Intelligent Development Operational Programme, Polish National Centre for Research and Development, ID: POIR.04.02.00-00-D006/20.
– The music score photo included on the front cover shows a fragment of the collection of Dr. Jim Mazurkiewicz. The photograph was taken by the authors during their fieldwork mission in Chappell Hill, Texas, while visiting the Texan Polonia and friends. Thank you!

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
In case of specific questions regarding the use of the book's contents
please contact the authors at {maciej.karpinski | katarzyna.klessa} @ amu.edu.pl

Maciej Karpiński is a phonetician and psycholinguist doing research in the perception and pragmatic aspects of speech prosody, paralinguistic features of speech, multimodal communication, and links between speech and music.
Katarzyna Klessa specializes in empirical, corpus-based linguistics, with a focus on experimental phonetics and investigation of linguistic and paralinguistic features in human communication.
Since 1999, the authors have collaborated on many research projects dedicated to the study of spoken language and multimodal communication in various contexts and from different perspectives, including speech technology, language development, language documentation, and interaction in task-oriented dialogues. At present they both work at the Institute of Applied Linguistics at Adam Mickiewicz University in Poznań.

The book is conceived as a practical, concise guide to the collection, processing and management of linguistic field data. It also contains a certain amount of methodological considerations relating to selected areas of linguistics, the particular demands of speech recordings, and the analysis, storage and sharing of speech material. […] *The authors take into account such factors as the profile of the explored community, legal regulations, and good practices. They share the skills, experience and observations gathered over twenty years with the entire community of interested researchers.* […] *The book is highly recommended for students of linguistics as well as lecturers offering classes in linguistics (including phonetics). Experienced field linguists will also find here discussion on many salient issues, as well as practical hints.*

Dr. habil. Anita Lorenc, UW Professor

ISBN 978-83-66666-89-4

DOI 10.48226/978-83-66666-89-4