The Production of Speech Corpora

Florian Schiel, Christoph Draxler

Angela Baumann, Tania Ellbogen, Alexander Steffen

Version 2.5 : March 21, 2012[1]

[1]This document is prone to frequent updates. You may check *www.bas.uni-muenchen.de/Forschung/BITS/TP1/Cookbook* for the latest version.

Contents

Part I

General

Chapter 1

Introduction

1.1 Preface

BITS (BAS Infrastructures for Technical Speech Processing[1]) is a 100% publicly funded project to improve the spoken language processing infrastructure for German, with a particular focus on spoken German.

One of the main deliverables of BITS is this cookbook-like document describing the production of speech corpora. In this document, the term *speech corpora* refers to collections of digital recordings of speech together with annotation, meta data, and documentation. Speech corpora are the prime source of data for basic and applied research in the area of spoken language communication, and for technology development in the area of *Spoken Language Processing* (SLP), e.g. *Automatic Speech Recognition* (ASR), *Text to Speech* (TTS) or *Speech Synthesis*, or *Speaker Verification* etc.

This cookbook provides prospective users in the scientific community and in engineering with advice on how to produce re-usable, high-quality and consistent speech corpora for their respective needs. Furthermore, it gives an overview of the best practice in this field and presents exemplary role models for some standard cases.

The motivation for the cookbook was the following observation:

> Very often large efforts and huge amounts of money are spent on bad speech corpora, i.e. corpora that serve one particular purpose only and were never meant to be *shared*. These corpora cannot be re-used for other than the originally intended

[1] www.bas.uni-muenchen.de/Forschung/BITS

> purpose and they are difficult to update or to maintain. As a consequence, they totally neglect their potential commercial and research value.

The BAS *Bavarian Archive for Speech Signals*, located at the *University of Munich*[2] has often been asked to add a corpus to its catalogue only to find that the corpus is not usable for any other than for the original purpose. In most cases this is primarily due to the fact that this corpus was poorly specified and that its production process was not monitored properly.

1.2 Intended audience

This document presents guidelines for the best practice in the production of speech corpora. It may be used as an introductory reading for newcomers to the field, or as a reference and check list for the experienced scientist or engineer.

The document does not cover the basic knowledge about *Digital Speech Processing* nor does it go into details of highly specialized topics like the SLP applications mentioned above.

Furthermore, this cookbook does not cover validation techniques for speech corpora. Please refer to [11] for a detailed discussion of this topic.

Finally, although a speech corpus very often contains not only the acoustic signals but also other measurable time signals derived from the process of speaking, throughout this book we will only treat the recording of speech in its various aspects. An exhaustive description of all possible signal recordings with regard to the act of speaking would be far beyond the scope of this cookbook.

The cookbook is organized in such way that a prospective producer of a new speech corpus may follow it like a recipe. In most cases (and if the reader is already familiar with the basic aspects of speech corpora) it will be sufficient to check out the summary check lists at the end of each chapter of part II or summarized in appendix A. Since this cookbook describes speech corpus production step by step, many topics are discussed in more than one chapter, e.g. the recruitment of speakers: it is essential to plan the means by which speakers are recruited and to estimate the costs in the planing phase (chapter Corpus Specifications), then to prepare the recruitment (chapter Preparation of Collection) and finally to perform the recruitment during the collection phase (chapter Collection).

[2] www.bas.uni-muenchen.de/Bas

For a systematic and non-chronological description of speech corpus production refer to the excellent summary in the EAGLES Handbook ([2]), chapters 3 and 4.

1.3 Overview

The cookbook consists of three parts: the first part *General* contains topics of general interest in the context of speech corpora that are better discussed outside the context of the practical cookbook. The fast reader might skip this part and go directly to the second part *Speech Corpus Production* which lists the major steps of a typical corpus production in chronological order. The main phases described there are:

- **Specification**,
- **Preparation of Collection**,
- **Collection**, in most cases overlapped by
- **Post-processing**,
- **Annotation**,
- **Documentation**, and optionally
- **Validation**.

Throughout Part II you will find check lists at the end of the major chapters. They are intended to be used during actual speech corpus production. Check points marked with a single star (*) denote compulsory steps (minimal requirements); check points marked with more than one star denote additional and recommended working steps that will increase the (re-)usability and value of the resulting speech corpus, but also require a greater effort in terms of time and money. The working steps themselves are abbreviated to mere key words. If you are not familiar with the meaning of a working step listed on a check list, please refer to the page number(s) in brackets after the keyword to find the passage(s) with a detailed description of the topic. For example:

> *Specification*
> ...
> ◯ * Define number of sessions (p. 35)
> ◯ * Define number of prompts / recording time (p. 35)
> ◯ ** Define distribution of sex (p. 35)

◯ *** Define distribution of age (p. 36)
◯ *** Define distribution of dialects / place of living / place of education (p. 36)
◯ * Define sampling rate (p. 35)
◯ * Define bits per sample (p. 35)
◯ * Define microphone(s) (p. 35)
◯ * Define acoustical environment (p. 39)
...

In this example all but the third to fifth check box are required for a corpus specification. Not all corpora require a defined distribution of gender; the same is true for age and origin of the speakers. Such a defined distribution will increase the usability of the corpus but will at the same time make the recruiting process more costly. All check lists are collected in appendix A in a format suitable for copying.

Finally, Part III *Examples* contains three prototypical speech corpus examples (WebCommand, SpeechDat-II German, and SmartKom) together with their key specifications and a list of references.

1.4 Terms and Definitions

The following list defines some technical terms deemed important in the context of this cookbook

- Speech Corpus = physical time signals, in most cases sound pressure or other measurable time signals recorded from the act of speaking[3], together with an associated set of annotations, meta data and documentation stored on a digital medium.

- Validation = the (formal) check of a speech corpus with regard to its pre-defined specifications.

- Evaluation = a qualitative assessment of a corpus with regard to its usability in a certain task or development scenario.

[3]Aside from the speech signal these time signals may be: laryngographic signal, electropalatographic signal, coordinate parameters derived from EMA (Electro Magnetic Articulography), X-ray movie (cineradiography), coordinate parameters derived from X-ray micro beam, air flow, nuclear magnetic resonance imaging, ultrasound imaging etc. In this cookbook we will not give any specific instructions on how to use special recording hardware for the listed signals, because this would be far beyond the scope of this book.

- Specification = the fixed technical description of a speech corpus with regards to all of its features (including annotations, meta data and documentation).
- (File) Format = standardized or specified format of digital signal and symbolic (annotations, meta data) data.
- Annotation = discrete (categorized) description associated with a physical signal (coding). Usually consists of a closed set of symbols and a scheme to link these symbols to either points in time or segments in time.
- Domain = topics of verbal communication or the situation in which a verbal communication takes place.
- Prompt = speech item (word, phrase or sentence) presented to a speaker. A *prompt list* or *prompt corpus* is a collection of prompts that define the *spoken content* of the corpus.
- Spoken Content = what was spoken in a speech corpus.
- Meta data = data about data. In this book the term meta data is restricted to three types: *recording protocols*, *comments* and *speaker profiles*.
- Codes = categorized data entries, in contrast to free text. If for instance the meta data parameter *place of birth* is restricted to the German states and the category 'other', then it is a code. A free comment about recording success is no code and therefore not machine readable.

1.5 Acknowledgments

The writing of this book was made possible under a grant of the German Ministry for Education and Sciences (BMB+F grant number 01 IV B01) within the *BITS project*.

Angela Baumann contributed to the example tables in part III, Tania Ellbogen did most of the research for the chapter *Annotation* and Alexander Steffen was responsible for most of the logistics for this book project, did the interviews with external sources and helped with the overall structure.

Aside from the staff members of the *BITS project* and the *Bavarian Archive for Speech Signals (BAS)* we would like to thank the following colleagues for their valuable contributions: Henk van den Heuvel (SPEX), Klaus Jänsch and Phil Hoole (IPSK).

1.6 Disclaimer

The contents of this document represent the joint knowledge of a group of experts to the field of SLP corpora production. It does not claim to cover all known methods and procedures in this field. The authors do not accept any responsibility for actions caused by others that follow the recommendations of this document.

This document may be copied and distributed to third parties for free (no commercial exploitation allowed) on condition that the document is complete and the copyrights are clearly stated.

Chapter 2

Legal Aspects, Contracts

This chapter does not contain legal advice that will enable you to do without your legal department or a specialized lawyer after reading it. This is impossible because this cookbook is intended for international usage and written by technicians, not lawyers.

This chapter is intended to give you some hints as to where in the process of speech corpus production you might or might not think about the legal situation (depending on your special situation, the country you are located and what you are going to do with the corpus). We do not give any practical solutions here but want to make you more sensitive to the legal implications of what you do when producing and distributing a language resource.

In the context of speech corpora – from design to dissemination – there are some legal relationships you should be aware of. The following chapters give some hints about these relationships.

If you or your institution are a member of the ELRA (see below), you might consider accepting their offer for free legal advice:

> "The ELDA team and a set of cooperating experts offer legal and contractual assistance to the ELRA members. This could be useful when you are negotiating for a resource with a producer or if you need information on contractual or legal matters." (*from the ELRA Web Page*[1])

[1] www.icp.grenet.fr/ELRA/org/reasons.php3

2.1 Copyrights, Intellectual Properties

Copyrights and Intellectual Property are well known in the context of written or multi-media content. Software is a special case already, but within the last decades most judicial systems around the world caught up with the special problems related to software. Now, a speech corpus is another thing and can be treated neither like traditional written content nor like software. Usually the first thing to do is to declare your Copyrights clearly in the documentation of the speech corpus (documentation). For instance:

> ©Copyright 2002, 2003 by Bavarian Archive for Speech Signals, Ludwig-Maximilians-Universität München, Germany, D-80538 München, Geschwister-Scholl-Platz 1

In addition to the ©symbol you should use the word 'Copyright' or the phrase "This corpus is the intellectual property of...". Be sure to use the official (legal) name of your institution. For instance in the above example the "Bavarian Archive..." is not an official name; therefore we added the official name of the university where the BAS is located. If you share the copyrights (see below) then you should list all partners after the copyright. If you add or extend to your speech corpus later, add a new year to the copyright line.

To declare a copyright is one thing and anybody can do it. Another thing is to protect it against abuse. Consult your legal advisor about that.

2.2 Speaker and Producer

The next relationship to consider is that between the producer of a speech corpus and the speakers. A speaker is recorded for the purpose that his/her voice is used for scientific investigations or the development of new technology. Basically, every speaker (if he is no public figure) has the right to decide whether recordings of his voice, knowingly or unknowingly, are stored on a media and distributed to others. Therefore it is of paramount importance for the producer of the corpus to clarify the legal situation with his speakers before he starts the recording.

In most countries it is sufficient to advise the speaker about the purpose of the recording and what will happen with the data in the future (*data protection, anonymity*) and then let him sign a short declaration in which he waives his rights to the recorded data and explicitly agrees to the intended usage of the data. This might be not sufficient in your country; please contact a legal advisor about this.

It might be also an good idea to include in this declaration information about the anonymity granted for the speaker. If you are planing to produce more than one speech corpus, you may also ask the speaker on the signed declaration whether he/she would like to participate in future recordings and whether the producer is allowed to store the speaker's contact information for this purpose.

Legal problems may arise in the case of Wizard of Oz (WOZ) recordings (see also section 4.5.4, p. 51) where the speakers might not be aware that they are being recorded.

For details see [2], pp. 143 - 145.

2.3 Client and Contractor

If you are producing the corpus for a client, you will most likely have a contract signed with your client before you start the project. Such a contract should as a minimum define

- the intellectual property rights / copyrights of the speech corpus (belongs to the client, belongs to the producer, belongs to both parties),
- the rights of usage (who? for what purposes? on what time scale? etc.),
- the right of distribution to third parties, and
- the royalties a third party has to pay for the usage of the corpus and how these royalties are distributed among the copyright holders.

This sounds rather easy but be aware: things get complicated very quickly if there are more than one client or more than one producer, if there are legal departments involved or if there are government agencies among the funding institutions. Be prepared to start your project before all the legal problems are solved, because if you have to wait, then your corpus will be probably outdated before you start producing it.

There are other, more practical items to be defined in your contract that should be mentioned here:

- The corpus specification (usually in the technical annex of your contract).
- The time schedule and milestones of corpus production.
- The validation procedure (together with the definition of tolerance measures).

2.4 Copyright Holder and User

Once your speech corpus has been finished and you have checked out the possibilities of distributing it to others, the next legal relationship to consider is that of the copyright holder(s) and the user(s) of the resource. As with software the copyright holders of language resources want to protect themselves against unauthorized copying. There are no technical ways to protect a speech corpus against unauthorized copying; it would hinder the usage of the data too much. On the other hand it is almost impossible to sue a user for copying speech data: in most cases your 'customers' will not even be in your country and proof of abuse will be hard to establish. The best thing you can do is that

- you let the users sign a license agreement that clearly states the rights of usage and clearly forfeits any rights to copy or re-distribute the data, or
- you insert the conditions of usage in your corpus documentation and say that the user, by using the speech data, automatically accepts the conditions stated there.

Both methods are probably legally unsafe. However, the user of a speech corpus is in most cases not a private person but either a company or a scientific institution. It is not very likely that these types of 'customers' will explicitly commit fraud by re-distributing your speech corpus.

2.5 Data Protection

Another important issue is the protection of user data. There are two things to consider here: the speech data and the meta data about the speakers.

The speech data themselves are usually no subject of special concern about data protection once the speaker has agreed to waive his/her rights to the recorded data[2]. However, this might not be the case for biometric databases. A biometric speech corpus is the special case of a speech corpus designed with the aim of developing and/or evaluating systems of voice authentication. In some cases the data provided by the speaker might be abused to break into future security systems based on the new technology[3]. Although this is rather unlikely, we recommend to take extra care in these

[2] Aside of course from the natural concern that you would not like your data to be destroyed or stolen by intruders in your computer system!

[3] And – ironically as it is – the speakers of a biometric speech corpus might be the most vulnerable ones to be broken into depending on the used technology.

cases that the mapping between personal speaker data and the speaker ID[4] within the corpus is inaccessible for everybody including former staff of the speech corpus production project.

Meta data about the speakers, that is personal data like home address, telephone numbers, email etc. are always to be protected and in most countries subject to special laws. Please contact your legal advisors about how to properly store and protect these kind of data, if you decide to collect them.

2.6 Third Party Distribution

You might also consider giving your speech corpus to a distribution agency and avoid most of the problems regarding archiving and distribution. In many cases this option becomes interesting after you or your client have exclusively used the speech corpus for a certain period (usually one to three years). Then you might achieve some return on investment (ROI) by giving a distribution license to one of the following agencies.

2.6.1 ELDA

ELRA's[5] mission is to promote language resources for the Human Language Technology (HLT) sector, and to evaluate language engineering technologies. ELDA is the distribution agency of the ELRA that keeps a large catalogue of speech resources within the European languages. ELDA is not only concerned with speech corpora but with language resources in general (including written text corpora, lexica and terminology databases). ELDA will most likely distribute your speech resource in commission and handle all the legal stuff for you. From time to time ELDA will select some language resources from their catalogue and let SPEX, a validation center in Nijmegen[6], Netherlands, perform a formal validation of the resource. Since ELDA gives a discount to the members of the ELRA, they will most likely ask you to grant discounts as well. However, you should consider this carefully because the member fees of ELRA will not directly benefit you. A better method (for you) would be to increase the royalties for non-ELRA-members.

[4] See also section 3.2.1, p. 29

[5] European Language Resources Association, www.icp.grenet.fr/ELRA/home.html

[6] www.spex.nl

2.6.2 LDC

The Linguistic Data Consortium (LDC)[7] is the largest distribution center for language resources in the US. LDC is located at the University of Pennsylvania (UPenn) and is also organized like an association. Membership fees are considerable higher than ELRAs[8] and the membership model is somewhat different: LDC members not only get a discount on the data, they may use them for commercial developments (non-members may not!) and they receive one copy of all corpora released in their payed membership year for free and older releases at a moderate price.

LDC is mostly funded by the National Institute for Standards and Technology (NIST) and very active in producing their own corpora.

LDC might consider distributing your speech corpus on an exclusive basis. However, they require an exclusive Intellectual Property Rights agreement with you (no other parties have the right to distribute; including yourself).

2.6.3 BAS

The Bavarian Archive for Speech Signal (BAS) is located at the University of Munich, Germany[9]

BAS only produces and archives German speech corpora and pronunciation lexica. BAS will archive, validate and distribute your (German) speech corpus on a non-profit basis[10]. All BAS speech resources listed in its catalogue are also listed in the ELDA catalogue because BAS has a broker agreement with ELDA.

2.7 Sharing Model

Speech corpora productions range from EUR 20.000 for a small monolanguage read speech corpus to several millions of EUR for a large multilanguage, multi-modal WOZ corpus. In almost all cases it makes sense to *share* these corpora.

- Small corpora are often highly innovative – sharing them after a period of exclusive use generates revenue for the owner without compromising his competitive advantage.

[7] www.ldc.upenn.edu

[8] For example, for commercial organizations the yearly ELRA fee is EUR 1.500 while the yearly LDC fee is $ 20.000.

[9] www.bas.uni-muenchen.de/Bas

[10] This does not mean that you are not earning royalties for your corpus, but that BAS does not want to make profit by distributing your corpus.

- Large corpora are often too expensive to produce for a single institution – a common specification, a distributed collection effort, and a one-to-one exchange of corpus data helps to reduce the cost for each partner.

- In general, the value of a corpus multiplies with the number of contexts (e.g. languages, recording environments, etc.) for which it is available.

For the production of a shared corpus, the obvious organizational form is collaboration. This means that partners form a consortium with the aim of creating a shared speech corpus, e.g. a multi-language corpus. Each partner is responsible for a part of the corpus, e.g. his language, and in the end all corpora are exchanged freely within the consortium. Of course a very careful corpus design and strict monitoring by an independent partner outside the consortium are indispensable conditions so that the deal works out satisfactory for all partners.

SpeechDat (M), SpeechDat (II) and SpeechDat Car were the first large corpus productions based on this sharing model; others might follow. See www.speechdat.org for details about the SpeechDat projects.

Chapter 3

Meta Data

The term *meta data* for speech recordings refers not to the recorded speech data itself, but to data *about* these recorded data. The emphasis here lies on the term *data* because meta data does not include documentation of a speech corpus. Meta data consists of categorized, machine-readable data that may be used to classify the speech data contained in the corpus.

Consequently, meta data consists of *codes* (in opposition to *free text*) except for free comments. When you specify your meta data for a speech corpus, it is therefore important not only to specify the type but also the set of possible values.

3.1 Importance of Meta Data

New speech corpora are constantly being produced and becoming accessible to scientists and developers, and the diversity of speech corpora is growing quickly. As a consequence, it becomes more and more difficult for the user to decide which corpus is optimal for his/her work. It is usually not possible to access the speech data itself to check this out, because speech data constitute an expensive resource. But it should always be possible to access the meta data, since this is a formal description of the underlying speech corpus and in itself it is of little commercial value. Therefore, meta data play an important role in the planing phase and for the acquisition of speech corpora.

Unfortunately, many speech corpora in the past were produced under a different paradigm: in most cases the goal was to produce data to be used in a given application, as quickly and as inexpensively as possible. No emphasis was placed on meta data; it was – if at all – considered to be a part of the documentation. Consequently meta data is often not parsable,

not structured and incomplete. A corpus thus quickly becomes virtually useless in terms of re-usability, simply because after a short while there is no longer anybody around who knows the exact properties of the corpus and the circumstances under which it was created.

Until recently no standard existed for the representation of meta data in a formal way. The ISLE Metadata Initiative (IMDI)[1] project has started to define schemata and principles for represention of meta data. The aim is to use meta data browsers to search online for relevant data in a distributed catalogue of speech and language resources[2]. Since there is the hope that in still ongoing projects (e.g. the planned EU project INTERA) more and more speech resources will be added to the IMDI standard and can be browsed over the Internet, it is probably a good idea to include carefully designed meta data files in a speech corpus.[3][4]

In general, the term *meta data* refers to many types of information about the more general category *language resources* from which speech corpora are only a sub-category. However, in the context of speech corpora meta data can be restricted to three main types: *recording protocols*, *speaker profiles* and diverse *comments*. What this means in detail will be outlined in the following sections.

3.2 Recording protocol

Every recording has to have a so called recording protocol in which all important information about the actual recording is logged. To be machine-readable its form has to be standardized (parsable), optimally in XML (see section 4.7.4, p. 59). If the speech corpus contains only recordings under exactly the same conditions, only one recording protocol for the complete corpus is necessary.

3.2.1 Minimal requirements

If the effort for providing meta data shall be reduced to a minimum, five minimal requirements are indispensable to get a useful recording protocol.

[1] www.mpi.nl/ISLE/index.html

[2] In this case the term 'speech data' is not restricted to speech corpora like described in this cookbook. It also refers to text corpora, terminology databases and lexica.

[3] For information about meta data file formats see section 4.7.4 (p. 59).

[4] Other meta data initiatives are Dublin Core, which defines a very small set of descriptors for language resources, MPEG-7 which is an attempt to define a classification system for any type of content of relevance to the home entertainment industry, and OLAC (Open Language Archive) .

These 'at-least-data' are the *session ID*, the *speaker ID*, the *date of recording*, the *environmental conditions* and the *technical recording conditions*.

Session ID

The session ID identifies one particular recording within the speech corpus. It often consists of characters that give broad categorical information about the recording (for instance language, sex of speaker, type of recording setup, domain etc.) and a number. The session ID is then often used in file names within the corpus to denote data belonging to the same recording session.

E.g. a session ID from the Verbmobil II project

G001AC where 'G' = German, 'A' = domain, 'C' = head set

Ideally a session ID should be created automatically to avoid duplicate or malformed IDs.

Speaker ID

The speaker IDs are a identification code replacing the speaker's real name to ensure his/her anonymity. Three capital letters are recommended. The mapping of speaker IDs to real names must not be published with the corpus. If the legal situation in your country allows, we recommend storing the mapping in a safe place. This makes it easier to avoid double speakers in future extensions of your speech corpus (see also chapter 2). The speaker ID will also be used in the speaker profile as well as in signal file headers and comments.

Date of recording

Every recording protocol has to contain the exact date and time of recording. This can be done automatically or manually by the experimenter.

Environmental conditions

Parameters to describe the environmental conditions are *room acoustics*, the sources of *noise*[5] and the presence of *cross talk* of other speakers.[6] Usually only persistent noise or cross talk are described in a recording protocol.

[5]Note that background noise might be played back artificially during the recording and in that case will be easy to describe.

[6]Noise events and cross talk may be subject to annotation techniques (see chapter 8).

- Room acoustics
  As a minimum the description of the room acoustics should give information about the place of recording, in particular whether the recording was in the field, under studio or studio-like conditions or other special room types (*echo canceled studio, studio, quiet office, office with printer/telephone, office with 1/2/5/10 employees, quiet living room (furniture, open/closed windows), living room with TV/PC running, living room with kids playing, standing car, running car, running car in the City, running car on a freeway (velocities, windows up/down, radio on/off, wipers on/off), phone booth on street, phone booth in shopping mall etc.*)

- Sources of noise/background noise
  State the sources of noise or background noise that can be found in the recorded signals: *noise produced by the host computer/sound card, spinning disks, ventilation, noise caused by cellular phones, by fluorescent lights, 50/60 Hz hum, typical office/home/street noise, machines etc.*

  The protocol also should contain a differentiation as to whether the background noise is real or artificially produced by play-back.

- Cross talk
  If other voices are part of the background noise, this has to be noted by a yes/no schema.

3.2.2 Technical recording conditions

For the description of technical recording conditions information about the used microphones, recording device, sampling frequency, bits per sample, coding, the speaker's distance to the microphone, prompting, durations and volume are necessary.

- Microphone
  The manufacturer's name, the type and kind of the microphone(s), for example, if it is a close-talk, a directional or a headset microphone. E.g.

  Ears-free headset Beyerdynamik NEM 192

- Recording device
  The manufacturer's name, the type and kind of the equipment(s) used for the recording.

E.g.

> *Intel Pentium III host with on-board ACI87 sound controller chip, pre-amplifier Beyerdynamik MV 100 set to +20dB*

- Technical specifications of *recorded*[7] signals. See section 4.7 for a more detailed description of the items:
  - Sampling frequency
  - Sample type and width
  - Number of recording channels
- Placement and distance of microphone(s)
  Usually the placement and distance of microphones do not change during the recordings or from session to session. If they do, you must log this in the recording protocol.

  It is a good idea to identify the microphones prior to a recording, e.g. by tapping on each microphone in a predefined sequence. This is especially true if the microphones can be moved or their cables be detached.

3.2.3 Other useful parameters

Aside from the compulsory values described above you may add other data about the recording that might be of interest:

- Name or ID of the recording supervisor
- Details about the recorded domain(s)
- Details about instruction to speaker(s)
- Duration of the recording session
- Type of prompting (paper, face-to-face, screen, voice)
- Emotional speech yes/no
- Details about acoustics: reverberation, S/N ratio etc.
- Supervisor present yes/no

[7] This does not necessarily match the specifications of the signals in the final speech corpus because signals may be altered in the post-processing (chapter 7). For instance very often signals are recorded with 48kHz sampling frequency and then filtered and down-sampled to a lower sampling frequency in the post-processing.

- Interpreter present yes/no
- WOZ: details about 'virtual machine' (see section 4.5.4 (p. 51)
- Type of speech (read, non-prompted, spontaneous, ...)
- Free comments

Most probably you will have to define other, more specific coded recording parameters depending on your special needs.

3.2.4 Example: Verbmobil II

In the following example taken from the Verbmobil II corpus a number of compulsory parameters are not included. For instance the description of the acoustical environment was the same for all recordings and therefore described in the documentation of the corpus. Although this is legitimate in this case we would strongly recommend including such redundant information into the recording protocol to facilitate the browsing of corpora contents as described above.

```
dialogue_name   m144a
recording_date  990421
scenario_date   990421
recording_site  UHH
scenario_id     a
no_speakers     3
speaker1_id     QYX
speaker2_id     HAS
speaker3_id     HCB
speaker1_language       e0
speaker2_language       e1,g0
speaker3_language       g0
speaker1_recmed_spec    h
speaker2_recmed_spec    h
speaker3_recmed_spec    h
speaker1_micbrand       beyer_dynamic_nem_194
speaker2_micbrand       beyer_dynamic_nem_194
speaker3_micbrand       beyer_dynamic_nem_194
```

3.3 Speaker Profiles

The characteristic features of each speaker should be collected in the speaker profile set of the corpus. This may be either a file for each speaker (recommended) or a table summarizing the features per speaker in one line or

column. To be machine-readable the speaker profiles have to be standardized (parsable), optimally in XML (see section 4.7.4, p. 59).

If you are producing speech corpora on a regular basis, you might consider including speaker profiles of your corpora into a database system. Keep the mapping from speaker names to IDs separate from this database (see also chapter 2 for a discussion of data protection).

3.3.1 Minimal requirements

Obligatory speaker information are *ID, sex* and *date of birth*[8].

3.3.2 Other useful parameters

You might also consider the following data to be part of your speaker profiles:

- Mother tongue of speaker
- Second languages of speaker
- Mother tongue of parents
- Second languages of parents
- Pathologies
- Dentures
- Piercings
- Place of elementary school
- Dialect region (difficult)
- Dialect (even more difficult)
- Level of education
- Level of proficiency for a certain task
- Profession (very hard to code)
- Height (including measuring unit)

[8]Do not use the age of the speaker at the time of recording, because you might record the same speaker in a different corpus/release later and want to re-use the speaker profile information.

- Weight (including measuring unit)
- Left/right handed, ambivalent
- Smoker/non smoker
- Stutter
- Hearing status
- Free comments

Most probably you will have to define other, more specific coded speaker characteristics depending on your special needs.

3.3.3 Example: SmartKom

```
id      AMY
sex     m
date_of_birth   660326
own_native_language     g
native_language_father  g
native_language_mother  g
primary_school  Berlin
dialect Berlin
profession      Schauspieler
height  186cm
weight  77kg
smoker  n
right_left_handed       r
```

3.4 Comments

The definition for 'comments' in this context is *all extra information which does not fit into the categories of recording protocol or speaker profiles.* This means that comments often contain information about events / features / observations that were not anticipated by the designer of the corpus. As such they are in most cases very valuable; so there should be a place or procedure to capture comments of speakers / experimenter / labeler etc. in an ordered and safe fashion. Comments are not machine-readable like other meta data. Therefore it is debatable whether they belong to meta data at all. However, for practical reasons we list them in this chapter because it is very easy to insert a free text field entry into a recording protocol file or

speaker profile. Likewise you may add such comment fields into labeler and transcription files.

Comments should be kept in their original version with original wording. Summaries are also possible, but it should be recognizable whether the comment on hand is a summary version or the original version. Beyond that it should be apparent whether the comments have been collected systematically (e.g. in form of a questionnaire) or coincidentally (e.g a subject expressed something about the recording without being asked explicitly). Often system errors have just been detected by speakers' comments.
Comments should be kept with the distributed speech corpus so that they are accessible by prospective users. It is a good idea to keep them in a form (e.g. plain text files) that might be searched for keywords.

Most common are comments about the speaker/speakers behavior:

> How does the speaker approach the 'virtual machine'?
> Has the subject shown emotions?
> What exactly was the gesture?
> ...

Other comments might stem from the experimenter, the labeler, the post-processing or even an external validation group.

Finally, all comments collected during corpus production may be a good source for the documentation of the speech corpus (see chapter 10).

Part II

Speech Corpus Production

This part of the cookbook describes the entire process of speech corpus production in a more or less chronological manner. Figure 3.1 shows the major steps of the process and their relation on a time axis progressing from top to bottom. As you can see, some steps have a strict order because they rely on results or data produced in the previous step, while others may be carried out in parallel. For example, it does not make sense to start with the creation of the pronunciation dictionary before the annotation is finished, because you need a basic transcription to create the dictionary. On the other hand, in many corpus productions collection, post-processing and annotation run in parallel to save time.

Also shown in figure 3.1 is the ideal concept of external validations at least at two points in time by an independent validation institution. Although in most cases insufficient funding prevents such a design, you should at least do an in-house validation then.

All the shown tasks will be discussed in the following chapters in detail. At the end of each chapter you will find a useful check list as a help for your individual speech corpus production.

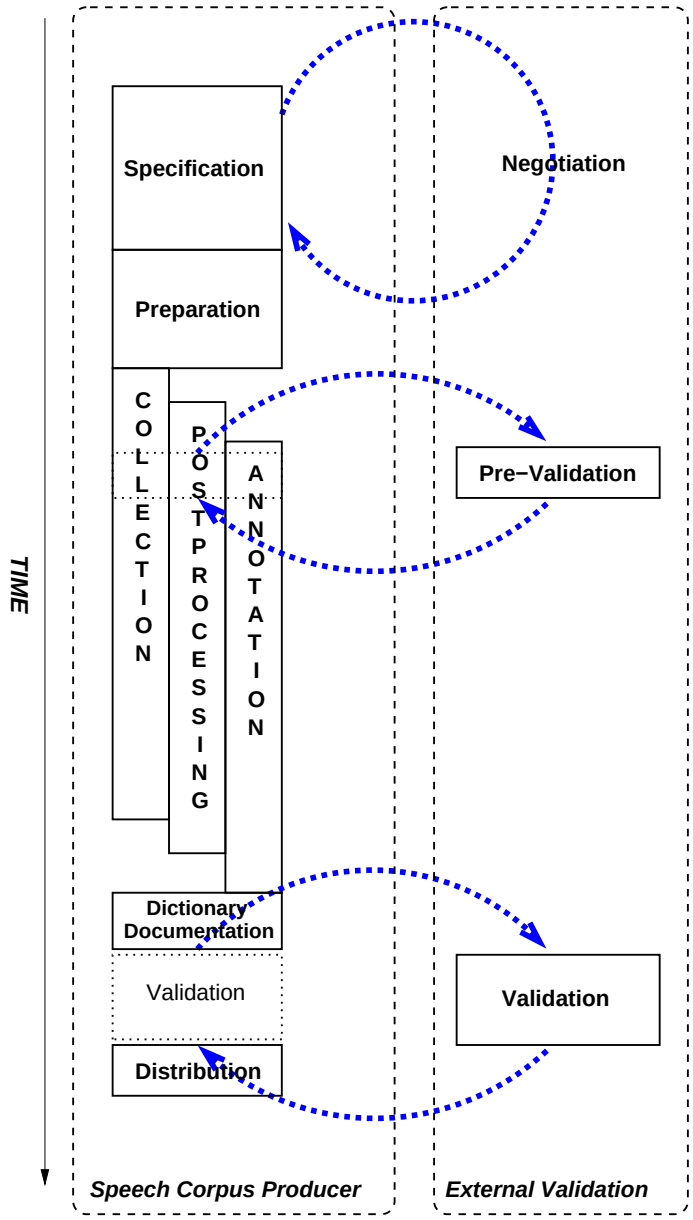


Figure 3.1: Typical schedule of a speech corpus production

Chapter 4

Corpus Specification

As with all projects that require great efforts in terms of workload and money it is absolutely essential to start a speech corpus production with a detailed specification of all desired features, the procedures, the monitoring of the process and the final validation. If you are acting as a contractor, this is probably the phase of the project where you will have the most contact with your client. It is very important to fix all specifications in written form (mostly in the form of a technical annex to your contract) and that your client sign this annex and all later amendments.

The high costs of speech corpus production can be optimally exploited by specifying as many diverse features into one speech collection as possible. For example in a telephone based corpus with the primary aim to recognize digits and numbers the overall costs will not dramatically increase with some additional non-prompted or even spontaneous recordings within the same recording sessions. However, the re-usability of the corpus will be much higher than with a corpus that only contains read digits and numbers.

The following sections give an overview about the basic requirements of any speech corpus specification. There may be additional things to cover in the specs depending on the special nature of your corpus.

In this chapter, the following terms will be used frequently:

- signal data: binary digitized audio data, e.g. WAVE files
- text: free form plain text data, often also referred to as ASCII data, e.g. ISO 8859-1
- markup text: text data containing marker symbols from a closed vocabulary with a given syntax, e.g. XML or HTML text

- formatted text: text with typographic formatting instructions, often in binary format, e.g. Microsoft Word or PDF documents

Markup text is the most flexible type of text because it can be read by both machines and humans, because it enforces minimal consistency constraints, and because it is platform and software independent.

4.1 Speaker Profiles

A speech corpus consists of recordings of humans speaking. Therefore the first things to specify are the characteristics and distributions of these speakers. It is of great importance that the speaker characteristics are documented as elaborately as possible. Although these details may not seem interesting at the time the speakers are recorded, their importance inevitably emerges later. In this case it is often difficult or impossible to recollect the data. Moreover, a well documented speech corpus may also be used for other research purposes, e.g. sociological research. Useful descriptors and criteria are (in order of their importance):

- Distribution of sex; in most cases *50:50*.
- Distribution of age; for example:
  - *Above 16 and under 50*
  - *Equal distribution over the following bins: 12-22, 23-30, 31-40, 41-55*
  - *Under 12*
- Mother tongue; although most corpora imply native speakers of a certain language, it is wise to mention it in the specs. It is also recommended to specify the maximum percentage of non-native speakers, e.g.

  Corpus language: German
  Maximum percentage of non-native speakers: 5%
- Dialectal distribution. There might be the case that a corpus should cover a certain distribution of a number of classified dialects of a language. In general it is very difficult to control the dialectal affiliation of speakers. Most speakers have a very rigid preconception of what dialect (if any!) they are speaking. However, even experts very often do not agree on certain dialectal features and it is therefore very hard to validate features like *10% of the corpus speakers are speaking Bavarian*. Here are some practical recommendations:

  - Specify a recruitment by the factor *place of Elementary School* instead of *dialectal class*. In most cases speakers will keep the dialect they acquired during the period of elementary school. Since most dialectal maps to not match other more familiar geographical areas, try to find a mapping from dialectal regions to states, districts, cities etc. that speakers are familiar with. State this procedure in the specs.
  - Specify a post-recording classification of dialect. This requires an expert in dialects and some time (more costs).
  - Specify a recruitment using 'local media' like local newspapers, local radio stations etc.

- Education / Proficiency / Profession. Some speech corpora require certain social factors like certain proficiencies (*computer expert, computer laymen*), a minimum level or a distribution of different levels of education (*Elementary School, High School, College, University*) or even speakers of a certain profession (*Radiologist, News Announcer*). Be sure that you only specify such characteristics, if you are absolutely positive about the recruiting process.

Other possible factors may be: *pathologies, foreign accents, speech rate, uncooperative speakers (forensic) etc.*

You may also specify here which meta data (see chapter 3) about speakers will added to the corpus.

4.2 Number of Speakers

The number of speakers is one of the most important characteristics of a spoken language corpus. Speech corpora can be roughly divided into the following three classes ([2], pp. 107 - 109):

1. Speech corpora with 1 to 5 speakers are often used in the development of speech synthesis systems or for basic research e.g. where invasive measurements must be made.

2. Speech corpora with about 5 to 50 speakers are often used in experimental factorial research. In general, the number of speakers and the number of repetitions of the speech phenomena that are investigated should be large enough for a meaningful statistical processing if factorial experimental designs are planned.

3. Speech corpora with more than 50 speakers are necessary to adequately train and test speech recognition or speaker verification systems.

Note that a small number of speakers does not necessarily mean a small corpus!

4.3 Contents

The spoken content of a speech corpus is the second major feature that determines the possible usage of the resource. Of course, this feature is not totally orthogonal to other specifications, for instance the speaking style. Basically, there are four main approaches defining the spoken content of a corpus: by vocabulary, by domain, by task or by phonological distribution. These might be applied in a mixed manner in some cases.

4.3.1 Vocabulary

Probably the simplest way to specify the spoken content is by vocabulary. It is more or less derived automatically from the intended usage of the corpus.

For instance, if the corpus will be used to train a speech recognizer on 11 German digits[1] and three command words, then the content definition most likely will require an equal distribution for all 14 items of the vocabulary and their repetitions per speaker, e.g.

> *14 words spoken by 500 speakers with 10 repetitions equals 70000 tokens*

4.3.2 Domain

Another method for controling the contents of a speech corpus collection is by domain. Domain in this context means the topic or field of topics or the situation in which a verbal communication takes place.

The domain could be for instance:

- *Weather*
- *Restaurants in Heidelberg*
- *Speeches in the House of Parliament*

[1] German has two word forms for the digit '2': *zwei* and *zwo*.

- *Fairy Tales*
- *Last nights's TV program*

Although the exact vocabulary cannot be determined by this method, it is a good method for achieving a rather closed vocabulary without restricting the speakers too much.

4.3.3 Task

By instructing the speaker to solve a certain task (either together with one or more human dialog partner(s) or with a 'virtual machine' in a WOZ experiment) the contents of a speech corpus can be reduced to a few hundred words without the problem that speakers feel restricted by the situation. Again the instruction of the speakers more or less defines the size of the resulting vocabulary.

Typical tasks in speech data collections might be

- *Schedule a business meeting*
- *Travel planning*
- *Purchase equipment*
- *Program your VCR*

4.3.4 Phonological Distribution

In some cases – very often in the scientific context or in combination with speech synthesis – the contents of a speech corpus have to be specified not in term of vocabulary but in terms of phonological units, like phonemes, syllables, morphemes.

For instance, a general purpose speech recognition system will require a minimum of repetitions of every possible phoneme in various contexts by each speaker.

Or a corpus for concatenative speech synthesis will require every diphone combination uttered from the same speaker in a minimum of 20 different left and right contexts.

4.4 Speaking Style

Speaking style is another key feature that defines the possible uses of the speech corpus. For instance a corpus containing spontaneous or non-prompted speech will not be useful for a dictation task.

Unfortunately many speech corpora contain only one speaking style and are therefore restricted in their re-use for different applications. This is a pity considering the fact that the recruitment and recording of speakers is the most expensive part of a corpus production. Therefore we strongly recommend specifying at least two different speaking styles for a corpus production. The following list gives an overview of the main speaking styles with rising complexity.

Please keep in mind that the chosen speaking style will interfere with other specifications like the recording setup, the speaker profiles etc.

4.4.1 Read Speech

Most speech corpora contain read speech, either for practical reasons because eliciting non-read speech is more difficult or simply because the intended application or investigation requires read speech. Read speech can be recorded by using so called *prompt sheets* or by displaying text on a graphical output device.

Dictation speech is a special case of read speech: the speakers are asked to read a text as in a dictation task. Exact instructions must be specified how special cases like acronyms and numbers have to be spelled consistently.

4.4.2 Answering Speech

Answering speech covers all recordings that are prompted by a question. These questions can be designed in way that they can be answered only by selecting from a given set of closed vocabulary options. For instance a banking system asks for the credit card number of a client, yes/no questions like *Are you female?*. Or they can be designed to be answered by fee text, e.g. *What did you have for breakfast?* Note that the quality of speech differs considerably from a read text and from spontaneous speech as being used in a dialogue.

4.4.3 Command / Control Speech

Command and control speech is used by speakers in a scenario where they are asked to control a device with a set of known voice commands, in most cases within a Wizard-of-Oz experiment.

4.4.4 Descriptive Speech

Descriptive speech can be elicited by showing a picture, a graph or a movie to the speaker and asking for a description of the shown items. Descriptive

speech is more spontaneous than read, command or answering speech, but can be kept easily within a certain subject thus restricting the vocabulary.

4.4.5 Non-prompted Speech

Non-prompted speech covers all speaking styles that do not use any written text that will be reproduced word-by-word but is not fully spontaneous, that is without any restrictions. For instance the dialog between a pilot and an airport tower is not based on any written text but has to follow certain rules (only one person speaks at any given time) that restrict the speech of both partners.

4.4.6 Spontaneous Speech

Real spontaneous speech can only be recorded in a face-to-face dialog or a very elaborate Wizard-of-Oz setting. The speaker has no restrictions on his speech aside from a topic or a task given by the supervisor.

4.4.7 Neutral vs. Emotional

For some speech corpora it may be essential that the speech either contains or does not contain emotional parts. Eliciting real emotional speech is very difficult (in general, it is only feasible with WOZ) and – in some cases – legally problematic.

4.5 Recording Setup

Before we specify the technical features of the recordings it is important to define an adequate recording setup for the speech corpus production. Basically the recording setup defines the acoustical characteristics of the resulting corpus and therefore also the usability of the data for certain applications or investigations. One can distinguish between open vs. secret recordings. People who know that they are being recorded change their speech behavior. On the other hand, secret recordings impose an ethical problem. Also, there is the risk of spending much time and effort for nothing, if the speakers later do not give their permission on using the recordings. Therefore you should use secret recordings only when there is no alternative. A good method to elicit very natural and spontaneous speech is to occupy the speakers with a task that requires some cognitive activity. People forget that they are being recorded very soon and you have the advantage that

you can choose your equipment for maximal quality and not for expensive secretiveness.

Furthermore the recording setup has an impact on the recruitment of speakers: it is much less expensive to recruit speakers for a telephone recording than in a studio recording (travel costs etc.)

In the recording setup the following general features are specified:

- Acoustical environment. Best not specified in technical terms like *reverberation, signal-to-noise ratio etc.*, but rather in a description of the location itself: *echo canceled studio, studio, quiet office, office with printer/telephone, office with 1/2/5/10 employees, quiet living room (furniture, open/closed windows), living room with TV/PC running, living room with kids playing, standing car, running car, running car in the City, running car on a freeway (velocities, windows up/down, radio on/off, wipers on/off), phone booth on street, phone booth in shopping mall etc.*

- The 'script'. The 'script' defines how the speaker acts in the recording environment. In most cases the only thing specified here is that *the speaker follows instructions while not changing position.* In some cases the 'script' defines actions of the speaker parallel to the recording: *the speaker drives a car, the speakers moves in the living room, the speaker points to certain objects while speaking about them, the speaker uses a phone etc.*

  The script may also define the order of recording prompts and has therefore an impact on the speech characteristics itself. Consider for instance a recording script that presents short utterances in groups of six each. The speaker will read these groups from paper or from a screen and most likely the grouping will influence his/her prosody significantly, for instance by lowering the pitch in the last utterance of each group. To avoid this effect you may overlap the utterances of the groups so that the last item in each group can also be found at the beginning or within another group or use filler phrases.

  Finally, it is recommended that the script contains a training phase before the recordings start and possibly some breaks during the recording script. The speaker gets accustomed to the recording situation in the training phase and any adaptive effects are not represented in the corpus[2]. Frequent breaks in the script allow the speaker to relax and maybe even drink some water to prevent a hoarse voice.

[2]Of course this makes only sense if your not interested in these adaptive effects!

- Controlled background noise. Some corpora require a defined background noise (type, level). This can be only achieved in a studio environment.

- Type, number, position and distance of microphones. Whenever possible we recommend using more than one microphone in a corpus production. Using only high-quality (and high-cost) microphones might increase the quality of your recordings, but not necessarily the usability. Therefore, it is advisable to also use at least one low-cost microphone as it might be used in a product.

Sketches of the recording setups as well as the intended instructions to the speakers can be added to the specifications to clarify the setup.

In the following sections four basic recording setups are discussed; of course mixtures of these are possible and are frequently used to further widen the re-usability of the corpus.

4.5.1 Telephone Recording

Speech recordings over the telephone network are inexpensive and easy to perform. Most ISDN PC cards nowadays allow the setup of an automated speech server that steers the calling speaker through a recording session and the recruitment of speakers is reduced to offering an incentive to a prospective and maybe even remote speaker group. However, the telephone recording setup has some disadvantages, too: the recording quality is restricted to 8 kHz sampling frequency and 12 bits per sample (compressed to 8 bits), there are unavoidable technical disturbances, it is difficult to control/validate the acoustical environment, the kind and distance of the microphone and the recording setup is restricted to read, answered and dialogue speech.

The following additional features may be specified in a telephone recording setup:

- *Fixed telephone network vs. cellular analog phone vs. cellular digital phone*

- *Home vs. office vs. public place vs. running vehicle*

- *Public phone booth vs. private phone*

- *Background noise, cross talk*

- *Hand-held vs. hand-free*

4.5.2 On-site Recording

On-site recordings cover all speech recordings that are performed at one or a limited number of recording sites. They have the advantage that the quality of the recorded signals can be controlled without restrictions. For instance you may use a wide range of low cost to high quality microphones in an on-site recording session.

Basically an on-site recording setup may have all possible setup features as listed before. Aside from that you should distinguish between

- supervised recordings, where a human supervisor is present and may monitor the recording on-line and interfere in case of errors (repeat single recordings), and
- un-supervised recordings, where the speaker follows an automated procedure and errors cannot be corrected on-line.

The latter is more cost-effective because the same supervisor may conduct up to three recordings in different rooms in parallel and errors will be marked in the annotation phase after the collection. However, if all speech items of a recording session are 100% essential, the first method has to be followed.[3]

On-site recordings require more manpower than telephone recordings (for scheduling and supervision). Also, in most cases there are only a few recording rooms available (while in a telephone recording many calls can be handled in parallel) and therefore more time should be allotted to the collection phase. On-site recordings in different locations require careful planning and training of the different crews to avoid recording site specific differences in the data. Be sure that exactly the same hardware is used in all locations. Also the monitoring should be performed by one central institution only.

4.5.3 Field Recording

Recording setups in the field cover all speech recordings performed in the 'real world'.[4] The great advantage is – of course – that all environmental features and in most cases even the speaker profiles match exactly the needs of a certain application. However, the costs are much higher and the reusability is low, because corpora of this kind are usually highly specialized.

Very often field recordings are time-critical in that sense that the location and the speakers are not available all the time. It is worthwhile performing a rehearsal a few weeks before the recording phase to make sure that the

[3] Please note that the annotation phase is in most cases necessary anyway!

[4] Therefore some authors call them 'real world recordings'.

recording devices and all procedures work without problems. If you produce the speech corpus as a contractor, we recommend that at least on the first day of the field recording a representative of your client is present.

4.5.4 Wizard-of-Oz

The term Wizard-of-Oz recordings is used for recordings where the behavior of an application or system (e.g. computer-based spoken language systems) is simulated in such a way that the speaker believes he or she is interacting with the real system. In fact the system behavior is controlled by one or more so called human 'wizards'.

The great advantage of this method is that the behavior of the speaker is very close to that of future users of the intended application. Furthermore, different design aspects of the application can be 'tested' beforehand and data may be analyzed to model user reactions in the application more successfully.

Depending on the effort that is put into the recording setup the acoustical environment can be matched very closely to that of a real-world situation. Therefore the data collected in WOZ technique is usually the best you can get for a complex application.

On the other hand WOZ recordings require a much higher effort in costs and man power: The setting must be so convincing that naive users do not suspect they are being tricked, the recording itself requires as a minimum two persons (one supervisor and one wizard), and finally because the speaker is 'steered' by the simulated system, WOZ requires a lot more training of the persons who do the recordings.

WOZ recordings may be designed in many different settings depending on your needs; therefore it is difficult to give detailed instructions on how to specify them. Also, many problems cannot be foreseen because WOZ recordings are definitely not standardized recordings. The best we can advise you here is that you stay as unspecific about the WOZ technique in the specs as possible. Try to concentrate on the overall intention – for instance that the users must not be aware of the simulation, that the setup matches the real situation as well as possible, and so on – but do not give any hard facts. On the other hand try to get as much information into the specifications as possible about the intended application. This is the basis you have to work on; if you do not exactly know how the 'virtual machine' has to work, you're lost. This is very important if you produce the speech corpus as a contractor.

See the section 5.2.4 (p. 76) for some more technical hints for the WOZ technique. In the section 15 (p. 177) you will find as an example a rather

complicated WOZ recording setup used in the German SmartKom project.

4.6 Annotation

The annotation of a speech corpus refers to all symbolic information that is related to the speech signal, e.g. orthographic transcripts, phonemic transcripts, all kinds of segmentations. See chapter 8, pp. 101, for detailed discussion of annotation within the context of speech corpora.

Since in most cases some kind of annotation is an integral part of the speech corpus, you should define the contents of the desired annotation in the specification. Also it might be a good idea to define the procedures to achieve an annotation as well as the quality control of the annotation beforehand (see validation procedures, chapter 11 and section 6.2).

4.7 Technical Specifications

The technical specifications define the formal properties of the corpus data. Basically all signals and symbolic data must be specified in this section. This section gives an overview about possible categories and values in a standard speech corpus. Please be aware that this list may be extended by other categories and values if you are using special recording devices (other than acoustic signals, e.g. video, laryngograph signals, electromagnetic glottography etc).

See also section 7 for an overview of different data type conversions.

4.7.1 Sampling Rate

The Shannon Theorem or Nyquist Law require that the sampling rate is higher than twice the maximum frequency in the digitized signal. Since speech is more or less located below 8kHz most speech corpora have a sampling rate of 16kHz minimum. Exceptions are telephone recording where the bandwidth is technically reduced to 300Hz - 3300Hz and usually a sampling rate of 8kHz is used. Since the audio CD standard was introduced with 44,1kHz also the dividers 22,05 kHz and 11,025 kHz are used because some audio devices do not process other sampling rates than these. This is also the reason why we recommend avoiding 'exotic' sampling rates, whenever possible.
Laryngograph signals are usually sampled at the same frequency as the speech signal. Because of their low bandwidth speech movement data (e.g.

EMA[5]) can be sampled at about 200 Hz.

- *Telephone Recording: 8kHz*
- *On-site or field recording: 16kHz, 22,05kHz*
- *Laryngograph signal: minimum 16kHz*
- *EMA signals: 200Hz*

4.7.2 Sample Type and Width

The sample type defines the format of a single sampling value; the width of a sample defines the number of bits required to represent the value on a storage medium. Both are of course dependent. Typical values are:

- *Telephone Recording: ALAW (World) or ULAW (US), 8 bits*
  Be aware that sometimes the bit order may be reversed. Decompresses ALAW roughly correspond to 13 bits linear PCM; decompressed ULAW 14 bits PCM.
- *PCM (linear), usually 16 bits, either*
  - *Signed* (values from -32768 to +32767) or
  - *Unsigned* (values from 0 to 65535)
- *ADPCM, 8 bits*
  ADPCM is a form of sound compression that has a good compromise between good sound quality and fast encoding/decoding time. It is used for telephone sound compression and places were full fidelity is not as important. When uncompressed it has roughly the precision of 16-bit PCM audio. Popular version of ADPCM include G.726, MS ADPCM, and IMA ADPCM. (from the sox man page)
- *GSM*
  GSM is a standard used for telephone sound compression in European countries and it is gaining popularity because of its quality. It usually is CPU intensive to work with GSM audio data. (from the sox man page)

If you use more than 1 byte per sample, you also have to define the machine format:

- *Big Endian (Motorola Processors): most significant byte first (10)*
- *Small Endian (Intel Processors): least significant byte first (01)*

[5]EMA = Electro-Magnetic Articulography

4.7.3 Number of Channels, Interleave

Most likely you will have more than one microphone in your setup.[6] Storing several channels in a single file is called interleaving or multiplexing (e.g. stereo audio tracks on a CD contain the samples of the left and right channel in an alternating sequence). However, interleaved signal files may in some cases more difficult to process. Hence, it is advisable to store every signal channel in a file of its own.
In case your recording device (or recording software) delivers multi-channel data you will need to split the signal files. See the post-processing section 7 for details on how to do this.

4.7.4 File Formats

The file format defines in which formal framework the specified data are embedded. Since a speech corpus always contains signals, symbolic data (annotations), meta data and – in most cases – a dictionary, we will describe those separately.

Signal File Formats

There exist quite a number of more or less standardized signal file formats. In this document we will concentrate on the most common formats in speech processing.
In most cases a signal file format consists of a so called *header*, which contains information on the signal, e.g. sampling frequency, sample type and width, machine format, number of channels etc.), and a *body* which contains the digitized signal samples.

- RAW data
  The simplest format: no header, only body with the digitized signal. Disadvantage: you have to get the necessary specifications of the signal from elsewhere. SAM[7] uses for instance raw signal files and stores the signal information in a separate label file with the same base name and a different file name extension.
  Some corpora add an extension that 'defines' the specs of the contained data. For instance:
  - *.dea, .al .la* : ALAW, 8 bits
  - *.deu, .ul .lu* : ULAW, 8 bits

[6] If not, think about it: It does not increase the efforts significantly, but will increase the value of your corpus.
[7] See for instance [2], Part IV, C.

  - *.raw, .pcm* : everything possible

- NIST SPHERE[8]
  The NIST SPHERE format was defined by the Speech Group at the National Institute for Standards and Technology, USA. It consists of a readable header in plain text (7 bit US ASCII) followed by the signal data in binary form. Because of the simple but nevertheless extendable format it is widely used in the speech science community and in many speech corpora. Most scientific tools may recognize NIST SPHERE automatically; other commercial tools may not. Big advantage: since the header information is in plain text, it is very easy to extract and insert values there (this is often a problem with binary headers). Big disadvantage: modifying the header requires modifying the entire file.
  Common filename extensions: *.nis* or *.nist*

- WAVE, RIFF[9]
  The WAVE file format is a subset of Microsoft's RIFF specification for the storage of multimedia files. A RIFF file starts out with a file header followed by a sequence of data chunks. Advantage: Most Windows based tools understand (only) this format. Disadvantage: binary header is not easy to manipulate and to read.
  Extensions: *.wav*

- SHORTEN[10]
  Shorten is not a format but a compression algorithm developed by Tony Robinson. It uses the redundancy of about 50% in speech signals to compress the data accordingly. The header is preserved if it is a standard header known by shorten or if you tell the algorithm how long the header part is.
  Compressed speech files were a big hype in the late eighties but then storage media became so cheap that most people no longer see why it is necessary to go through all the hassle. Also, we found in the SpeechDat project that compression by *gzip* reaches almost the same reduction as shorten and has the advantage that most platforms can decompress data without any additional software installed.
  However, if you think it might be a good idea to use compressed files and you want to use shorten, please inform Tony Robinson about it.
  Extensions: *.shn*

[8] www.nist.gov/speech
[9] ccrma-www.stanford.edu/CCRMA/Courses/422/projects/WaveFormat
[10] www.hornig.net/shorten.html

Annotation File Formats

Annotations are symbolic data associated with the recorded signals of the speech corpus. See chapter 8 for a discussion of the different principles, label categories, methods and tools.

The format in which these data are stored has to be included into the corpus specification. Since no widely accepted standard exists and since very often speech corpora contain a new form of annotation that has never been applied before, many speech corpora contain proprietary formats that were defined only for that special occasion. Some of these formats have become commonly accepted and have been re-used in other collections.

In this section we list some more or less standardized file formats for annotation data and outline their respective properties. These properties can be described by the following criteria:

- platform independence
- self-describing
- plain text vs. markup text
- availability of tools

Where appropriate, the annotation formats will be described in these terms in the following sections.

- SAM Format[11]
  The European SAM project defined a combined signal and labeling format for data collected in the SAM projects. The signal files contain only the raw digitized data while the description files contain the signal specification data, meta data and labeling data (annotations). SAM description files are widely used, e.g. in the SpeechDat telephone speech corpora.

  Pros:

  - Description files and signal files are separate and can be stored and distributed independently.
  - Easy to extend by new labels
  - Does not require any special software.
  - Works on all hardware platforms.

[11] www.icp.inpg.fr/Relator/standsam.html or [2], Part IV, C.

Cons:

- Basically a free-form text format – very weak syntactic constraints
- Does not include flat or hierarchical linking of different annotation tiers.

Recommendation: Use SAM for simple speech corpora, like read or prompted speech by a single speaker.

- EAF (Eudico Annotation Format)[12]
  EAF (Eudico Annotation Format) is an XML and Unicode based annotation format. It supports both time-aligned and symbolic relations between annotation tiers. New tiers can be added easily.

  Pros:

  - Description files and signal files are separated and can be stored and distributed independently.
  - XML: Easy to parse and easy to manipulate.
  - Does not require any special software.
  - Works on all hardware platforms.
  - Open format: definition of new tier types very easy.

  Cons:

  - No label types for events between words (yet).
  - No tiers with both symbolic and time links.
  - EAF events are always time-consuming (no points in time).

  Recommendation: *No experience at BAS with this format yet. Probably the most powerful format.*

- BPF (BAS Partitur Format)[13]
  The generic annotation format used at the Bavarian Archive for Speech Signals (BAS). BPF is quite similar to the old SAM format but has no fixed syntax and semantics for the annotation tiers. A BPF file consists of a SAM compatible header part to store signal specs and meta data and an unlimited number of tier blocks that contain the annotations to the signal. Reference between signal and annotation

[12] www.mpi.nl/DOBES/tools/Eudico-Annotation-Tool.pdf
[13] www.bas.uni-muenchen.de/Bas/BasFormatseng.html

is done via the physical time scale (direct reference) or via so called symbolic links between annotation tiers (indirect reference). The format is open in that new tiers may be added to the file following some basic rules (5 basic annotation types). BPF uses the very basic UNIX concept of filters and lines of code which makes it very easy to create, manipulate and search BPF files on all platforms that provide some basic scripting language.

Pros:

- Description files and signal files are separated and can be stored and distributed independently.
- UNIX filter concept: Easy to parse and easy to manipulate.
- Does not require any special software.
- Works on all hardware platforms.
- Open format: definition of new tier types very easy.
- Integration of other modalities than speech.
- Overlapped speech, points in time and non-speech events.

Cons:

- No hierarchical structuring of tiers.
- No intelligent search structure.
- No specialized software to manipulate or visualize BPF files.

Recommendation: Easy to use and easy to transform into other formats.[14]

- ESPS
  A very simple and widely used label file format for simple event or segment labeling developed by ENTROPICS and used by their signal processing toolbox X-WAVES. Unfortunately ENTROPICS was bought by Microsoft and the support and distribution of X-WAVES stopped in 1999. Nevertheless, many labeling or database tools still use this format, for instance the EMU system (see [4]).

  Recommendation: Comparable to SAM but without the header part.

[14] At BAS there exists a public domain tool *par2ags.pl* to transform BPF into Bird's annotation graph file format (XML).

- AGS (Annotation Graphs)
  Annotation graphs are a general and very powerful concept for representing all kind of symbolic information related to a speech signal[15]. Although the annotation graph *per se* is not a file format, Bird and his colleagues developed an XML based file format AGS that may be used to store all different kinds of annotation graphs.

Meta Data File Formats

For meta data no commonly accepted file format has been proposed yet. Generally, an XML and Unicode based markup text format is recommended.

IMDI[16], Dublin Core, OLAC and other meta data initiatives have all proposed XML applications, but so far none of them has found wide acceptance.

Lexicon Format

Depending on the complexity of your speech corpus you might add the specification of a lexicon covering the entire corpus or parts of it. Again there are no widely accepted standards about a file format for lexica or dictionaries.

In most cases speech corpora come with just a simple three-column list giving for each spoken word form the orthographic representation, the word count and the most likely pronunciation of this word form. This seems quite straightforward, but is clearly not sufficient for languages for which there is no standard orthography, or the orthography cannot be established unambiguously from speech, or in which orthography is not necessarily based on words. Here are some hints:

- Orthography: whenever feasible use Unicode.
- Pronunciation: whenever feasible use SAMPA or X-SAMPA[17].
- Clearly specify what is meant by 'most likely' or 'canonical' pronunciation and how you will produce them.[18]
- Specify whether there might be more than one possible pronunciation of the same word form in the lexicon

[15] See section 8.2 for details.
[16] www.mpi.nl/ISLE/
[17] www.phon.ucl.ac.uk/home/sampa/home.htm or [2], Part IV, B.
[18] There is nothing like "the pronunciation of a word". Your lexicon will always contain word forms where the most likely pronunciation is debatable. For a more detailed discussion of dictionary contents please refer to chapter 9.

- Use a simple plain text list or an XML markup text as the file format (everybody is happy with that because it can easily be imported into any kind of database system).

4.8 Corpus Structure

The corpus structure specification defines the internal structure of the final corpus, the file naming (terminology) and the distribution media. If you're planing a long-term data collection, you'll also define the release packages of your speech corpus here.

Structure specification is of paramount importance if you're working in a consortium of producing partners. If you're the only producer involved, it's up to you and your potential client whether you specify the following or not.

4.8.1 Structure

As mentioned before, it is a good idea to keep signal data files and annotation data separately. The reason for this is that very often users will need only access to the symbolic data of your speech corpus. Furthermore, the annotation part is much more likely subject to updates than the signal data. Therefore it's better to have them separated for an easier maintenance of the corpus.

Small corpora will have the following typical structure in the root of the distribution media:

- DATA : contains all signal files
- ANNOT : contains all annotation files
- META : contains all meta data files
- DOC : contains the documentation
- LEX : contains the lexica (if any)
- TOOLS : contains software to access signal, annotation and lexicon data

Larger corpora (>5 GB) might distribute the DATA part on other media but the basic structure remains the same.

Within the DATA and ANNOT directories organize the files in a way to avoid very large (approx. > 256) numbers of directory entries, and try to

provide a natural order to the prospective user. Depending on the aims of your speech corpus this order of subdirectories may be:

- male / female
- recording sessions
- speakers
- different acoustical environments
- languages
- dialect classes
- speech types (read, non-prompted, ...)
- technical setups (telephone, on-site, ...)
- in ANNOT: different annotation types

4.8.2 File Naming Conventions

The file naming conventions (or nomenclature) define the allowed file and directory names within your speech corpus. A very common approach is to use content-based file names, an alternative approach is to use generated file names.

Content-based file names are constructed from features of the corpus, e.g. language code, speaker gender, type of speech, etc. Content-based names allow access to specific fragments of the corpus simply by filtering file names. Of course the information encoded in the file name must be meaningful and easy to extract. One problem with content-based file names is the platform- or medium dependent length restriction of file names, e.g. 8.3 for ISO 9960 CDs. Another problem is that there is often no natural hierarchical structure in a speech corpus: is it better to organize the recordings by recording location and then by gender, or the other way round?

Generated file names are usually created automatically, e.g. as sequence numbers. Generated file names can easily be organized in hierarchies, e.g. *BLOCKxx/SESxxyy* with xx and yy numbers from 00 to 99. To retrieve fragments of a corpus, a separate document is necessary listing the contents of a signal file.

Some operating systems and programming languages are case sensitive, some are not; some apply their own rules for capitalization, others do not. Sometimes case changes when data is copied to another medium, sometimes

it does not. The lesson here is: do not define a nomenclature that is case sensitive.

Here is an example from the German Verbmobil II corpus:

```
Dialog names are coded as follows:

1st character:
      <lang> [g,e,j,m,n] recorded language
      g(erman), e(nglish), j(apanese), m(ultilingual), n(oise)

2nd to 4th character:
      dialogue number i.e.\ 001

5th character:
      scenario
      a(main), b(information desk), c(remote maintenance),
      d(VM1), n(noise)

Turn names consist of the dialog name (char 1-5) and the following:

6th character:
      technical definition of recording
      c(lose), r(oom), t(elephone)

7th character:

      detailed description of recording means (microphone)
      telephone:
      m(obile), p(hone,analog), w(ireless), d(ect)

    close:
      h(eadset), n(eckband microphone), c(lip microphone)

    room:
      r(room)

8th character:
      channel coding
      [1..n]
9th character: '_'

10th - 12th character:
      turn number starting with '000'

13th character: '_'
```

```
14th - 16th character:
      <sp_id> speaker ID


The extensions code the contents of the file:
       .nis  NIST file
       .trl  transliteration
       .spr  speaker protocol file
       .rpr  recording session protocol
       .par  symbolic information in "partitur" format

Each recording consists of a set of files like the following:

Type                            Name                    Location

signals                         <turn>.nis              data/<dialog>/
recording session protocol      <dialog>.rpr            data/<dialog>/
speaker protocol file           <lang>_<sp_id>          spr/
transliteration                 <dialog>.trl            trl/
Bas Partitur Files (BPF)        <turn>.par              par/<dialog>/
```

In the above example the *dialog name* is used as a structural element to sort files into groups, while the *turn name* is the prefix to signal and annotation files.

4.8.3 Distribution Media

Specify here on which media you will distribute the speech corpus to partners, your client or other parties. Basically you have the choice between CDROM (650MB), DVD-R/RW (4,7GB), DVD+R/RW (4,7GB), DVD-RAM (5,2GB), tapes (up to 500 GB), removable drives like ZIP (250MB), JAZ (2 GB), or hard disks (up to 160GB and growing).

Refer to chapter 12 for a more detailed discussion of the different media.

You might also specify here how many copies will be produced and who is going to cover the costs for the raw storage media. Also keep in mind that special devices might be needed to produce the distribution media and therefore this should be accounted for in the funding scheme.

4.9 Release Plan / Validation Procedures

The specification (or the contract with your client) should contain a rough schedule which marks the major steps of the production and milestones or

delivery dates.

If you plan to perform a pre-validation and/or final validation (highly recommended), you should also include the details and schedule for the validation procedures here (see sections 6.2, p. 84 and 11, p. 135 for more information regarding the validation procedures).

In a long-term speech corpus production (one to several years) you will probably not wait until the very end of the project but rather distribute parts of the corpus in releases. In that case the specifications should contain a mile stone plan that defines when and what data will be released (most likely together with the validation procedures for these releases). Also you might define in the specifications an update plan for the already released data because in most cases you or your partners/clients/validation center will find errors in the released parts of the corpus that have to be fixed. These updates are often forgotten in the specification and therefore often no funding is available for error updates.

4.10 Meta Data

The sense and significance of meta data have already been discussed before (see chapter 3, p. 27). It is advisable to include at least the contents – perhaps also the formats – of the meta data files of your speech corpus into the specification because this might be a point of interest for your partners or client.

Basically your specified meta data should reflect more or less the *variable parameters* in your specifications of the speaker profiles, of the recording setup and the technical specifications. *Constant parameters* do not need to be included in the meta data because they will already be written into the corpus documentation. However, it might be a good idea to include even such redundant information into the meta data as well with respect to the possibility to browse speech corpora as mentioned before.

4.11 Documentation

Although not very common, the documentation procedures might also be specified before-hand. This is probably a good idea in large projects with several producing partners working on a common goal. Partners could use a distributed pre-formatted documentation template and fill it in accordingly. This might facilitate the validation of the documentation as well.

See section 6.1 for some hints on what should be documented in a speech corpus and how.

Check List Corpus Specifications

◯ Speaker Profiles * (p. 42)
◯ Number of Speakers * (p. 43)
◯ Spoken Content * (p. 44)
◯ Speaking Style * (p. 45)

Recording Setup General *(p. 47)*

- ◯ Acoustical Environment **
- ◯ 'Script' *
- ◯ Background Noise **
- ◯ Microphones *
- ◯ Sketch **

Recording Setup Telephone Recording *(p. 49)*

- ◯ Distribution of telephone type (fixed, cellular ...) ***
- ◯ Public phone booth vs. private phone ***
- ◯ Hand-held vs. hand-free ***

Recording Setup On-site Recording *(p. 50)*

- ◯ Supervised vs. non-supervised *

Field Recording *(p. 50)*

- ◯ Schedule a rehearsal **

Wizard-of-Oz Recording *(p. 51)*

- ◯ Specification of 'virtual machine' *

◯ Sampling rates * (p. 52)
◯ Sample Type and Width * (p. 53)
◯ Signal File Formats * (p. 54)
◯ Annotation File Formats * (p. 56)
◯ Annotation Contents and Procedures * (p. 52)
◯ Meta Data File Formats ** (p. 59)
◯ Meta Data Contents *** (p. 64)

Chapter 5

Preparation of collection

After you (and your client) have agreed on the specifications of the speech corpus you will need some time to prepare the collection phase. Do not underestimate the time required for this preparatory phase: very often it takes longer to prepare a data collection than to actually record the data.

5.1 Instructions and Prompting

All speakers participating in a speech corpus recording need some kind of instruction before the recording starts. The instructions may range from a very reduced instruction set in psychologically inspired experiments or Wizard-of-Oz recordings to very detailed and strict instructions for an unsupervised collection over the telephone network. Although possible, we strongly discourage you from relying only on a verbal instruction given by the supervisor or experimenter before the recordings. To improve the consistency of the recordings always use a written instruction, or use pre-recorded instructions.

Make the instructions as simple and as unambiguous as possible. Don't overload them with background information but give a brief outline of what the collected data is going to be used for. Outline the contents, the speech style, the recording technique and the estimated length of the recording.

In many speech collections you will prompt the speaker to produce specific utterances. This can be done acoustically (resulting in mimicked speech styles) or on paper or a monitor (resulting in read speech). Except for special purposes where a certain prosody, loudness or emotional speech style should be elicited we strongly recommend using written prompt material. You may also use direct questions — again acoustical or written — to elicit

certain utterances (answering speech), ask for descriptions of an image or a video movie (descriptive speech) or give directions for a monologue or dialogue (non-prompted speech).

In any case formulate your prompts so that they are un-ambiguous with regard to phonemic form and stress. If you are prompting with questions, restrict the number of possible answers:
"What is your favorite dish? (name one)"

Technically the prompting can be done in several different ways:

- Let the speaker read the prompts or questions from paper[1].
- Display the orthographic form of the utterance or the question on a screen.
- Display of images or video movies.
- Playback of pre-recorded acoustical prompts or questions.
- Direct questions from the experimenter (interview).
- Combinations of these.

Depending on the recording setup you may use an automated process that controls the recordings within one recording session. For telephone recordings (see below) using a telephone server this is a must, but also in most other types of recording setups we recommend using some script language to control the recording times, the signal file names, the questionnaires for the speaker etc.

Here are some practical hints for the setup of your recording procedure:

- If you are using a paper version and an automated process simultaneously (like in a telephone recording), be sure that they are consistent. Nothing confuses a speaker more than inconsistent instructions.
- Use some dummy recordings at the beginning of the session and possibly also at the end. You may announce them explicitly as training prompts.
- Carefully design the order of prompts. Avoid sequences of similar utterances, e.g. many sequences of numbers. This may result in a droning speech style.
- Use a beep or a visual marker (for instance a 'red light') to indicate when a recording begins or ends.

[1] Often this may result in un-wanted background noise like paper rustle, page turning etc.

- Give feedback to the speaker. For instance:
  "You have already finished more than half of the recordings!"

Test your procedure, including instructions set, prompt material and automated recording program on naive speakers before starting the collection.

5.2 Recording Techniques

5.2.1 Telephone Recordings

For telephone recordings you need:

- An ISDN telephone account.
- Hardware that allows you to handle and record phone calls (nowadays this will be a ISDN interface of some kind).
- A software library or DLL to access your hardware.
- A control program that allows you to model the recording session, normally a simple chain of played back instructions from the server and recordings of the speech of the calling speaker.
- Speech prompts recorded from a clear and easy to listen to voice (you might need a studio-like environment for that or you can order them from a supplier).
- A good 'beep' for the prompting (you get very good beeps from the Internet).
- Finally, the 'script' itself, describing the session.

Unfortunately there are no public-domain ready-to-use software packages for the setup of an ISDN speech recording server available. There is of course the possibility to buy a professional VoiceXML engine plus ISDN hardware, but in most cases the investment is not justified. If you are lucky enough to own a VoiceXML engine, simply design your recording session in a VoiceXML document and run it through your hardware. Here are some useful hints if you are going to design your own server:

- If you are planning a single corpus recording, do not try and develop a complete VoiceXML machine. Although there are some very powerful tools for the handling of XML available (especially for JAVA), the effort is probably too high.

- Most manufacturers of low-cost ISDN cards provide libraries for the API to their hardware. In most cases these APIs are compatible with the *Common ISDN API (CAPI)*. Also, they might provide some demo applications for their cards that can easily be adapted to your needs.
- Prompt the start of the recording by playing a short 'beep' sound file.
- A silence detector during the recording can be used to avoid empty recordings.
- Most ISDN interfaces will allow you to detect DTMF tones sent from the calling phone. Make use of this capability to give the caller better control during the recording session. For instance the caller might
  - skip already known instructions
  - call a help message
  - repeat bad recordings
- A speech detector may be used to shorten the overall session time by adjusting the individual recording times to the actual length of the input. However, a speech detector will not work reliably in all situations, e.g. loud environment noise, or technical noise which is common in mobile phone connections. If you plan to use an automatic speech detector, try to keep its configuration simple with at most 2 to 3 parameters to adjust[2]. Then add at least half a second before and after the detected speech to your recorded sound file.
- If you have to use a fixed recording interval for each speech item, try to find a speaker that is extremely slow and test your system. Every recording prompt needs to be adjusted to an individual length. If you simply set a very long fixed recording length for all recordings, the speakers will have to wait a (subjectively) very long time between prompts and will start to make other noise or even utter to themselves.
- The raw data provided by the telephone company and recorded by your ISDN card will be either in ULAW (US, Asia) or ALAW (EU)[3].
- Low cost ISDN cards do not provide 'echo-canceling'[4]. Thus the prompt beep might be audible in the recording.

[2]Typically a threshold and two timing parameters for speech and silence: When the signal is higher than the threshold for more than T1, it's speech starting; when the signal stays under the threshold for longer that T2, it's speech end

[3]MIME types: audio/x-alaw-basic or audio/x-ulaw-basic

[4]'Echo' in this context means that the signal sent to an analog telephone will be heard with a certain time delay in the channel coming from the analog telephone

To design your recording session you will need a number of pre-recorded sound files for instructions, greetings, help messages, prompting etc. Here are some hints for the production of these sound files:

- If you do not have the proper equipment and a studio, consider ordering these sound files from a professional studio. Mention that you will need the sound files for playback over the phone, so they might add some compression to the signals which makes the speech much more understandable.
- Use a voice that is clear and easy to understand for your pre-recorded prompts and instructions.
- Carefully remove any DC component from the prompt sound files; they might cause a strong clicking noise when played in your telephone server.
- Adjust your recording level so as to avoid clipping the sound file. Clippings[5] in the original signal tend to be much more audible in the compressed form.

Finally, here are some design hints for the 'script' itself:

- At the begin of the session you should clearly explain what is going to happen and what is the purpose of the data collection. If possible insert a mechanism that allows the caller to skip these messages by pressing a button.
- Insert informational voice messages into your script, e.g.

  You have now completed more than half of the recording.
  ...
  Please remember to speak only after the beep.

- Clarify the legal aspect of the speech recording. For instance it is a good idea to include a sound file like the following at the very beginning of the script:

  The recordings of your voice during this call will be used for the development of future speech recognition techniques. For this purpose your voice recordings will be distributed anonymously to scientists and developers. If you do not agree to that, please hang up now.

[5]Samples that have the maximum value of your sample format, for instance +32767 in 16 bit

5.2.2 On-site Recordings

If you do on-site recordings in a professional studio environment, you'll most likely also have the trained staff to do the recordings. Then you may skip this section entirely.

If you, however, plan to set up your own recording hardware, you might find the following hints useful. A good thing to prevent unsuccessful recording sessions are check lists. Provide a check list for the recording supervisor that has to be run through before every recording session and in which all settings and procedures are listed.

Acoustical Environment

Be sure to follow the specifications regarding the acoustical environment (see 47). Do not alter the furniture (not even the position) during the collection phase unless it is specified to do so. Furniture – especially carpets, sofas and drapes – may alter the reverberation of the room significantly. [6] Document the setting by some pictures from different angles.

Microphones

Do not alter the position of your microphones; include points in your check list for the control of the position of clip-on microphones and headsets (distance and position to mouth). If you are using head set microphones, be sure that the cable does not touch the microphone arm or other hard surfaces when the speaker moves. The sound will be transmitted to the microphone.

Also add a check point that batteries of microphones are checked before every recording session. Always keep fresh batteries available.

Never change the type of microphone during a recording phase. If you have to (e.g. because a microphone is broken), document the change in the recording protocol and documentation.

Amplifier and Level

High quality microphones very often need an amplifier for the signal before the signal is fed into the recording device. Make sure that the amplifier has the proper input type (symmetric, asymmetric) and possibly the right phantom voltage. Also, the output type must match the input type of the recording device. For instance, most high quality microphones are symmetric, but most A/D cards for Intel platforms have an asymmetric input.

[6] You may mark the position of furniture by taping markers on the floor.

Symmetric microphone lines have the advantage that electromagnetic noise (usually in form of 50/60 Hz hum) is not induced and therefore you may safely use long microphone lines. If you intend to use high quality, symmetric studio microphones in combination with a computer as the recording device, consider using a semi-professional mixer between microphones and computer. Most mixers have symmetric inputs and (at least some) asymmetric outputs that may be connected to standard sound cards. If your budget allows the additional costs you may also use a digital mixer and a digital sound card that can be connected directly to the digital output of the mixer. This has also the advantage that you can record up to 16 channels in parallel into one computer.

Whatever microphones you use (low or high quality) try to set them up in way that the average speech level will be within the lower 40% of your dynamic range. For example: Using 16 bit samples the average speech level should not exceed $0.4 * 32767 = \pm 13106$. Or set the maximum sound pressure to - 12 dB. Test this with a speaker with a very loud voice. Add the settings of the amplifier to your check list; do not change the settings during one recording session, and if you must alter it, document this in your recording protocol. Also add a check point that batteries of amplifiers are checked before every recording session. Always keep fresh batteries available. If you use AC power amplifiers, make sure there is no 50/60Hz hum on your signals. Beware: it is not sufficient to simply listen to the recordings to check the signal quality because some D/A cards or headphones might not reproduce low frequencies; it is much better to check a sonagram or calculate some spectra.

Recording Device

Most likely you will not start with an analog recording device. Even for field recordings under extremely difficult conditions you will find affordable digital and portable recording devices. Basically you may use any kind of digital recording device; however, we recommend using a hardware that lets you store the signal on a storage medium on-line without an extra copying step whenever possible to avoid unnecessary work load. Here are some basic hardware options you have for on-site recordings:

- Record into a workstation
  Use a high quality A/D card to avoid disturbance from the host computer. You may improve the results by changing the slot and find the best position of your card with the lowest noise. In most cases standard cards will not do the job because they provide only two channels. But there are quite a number of affordable 4- and 8-channel

cards available.

If the recording host has to be located in the room where the recording takes place, try to get an acoustically insulated chassis and silent, low spinning hard drives.

- Record into a laptop
  Not a good idea. Most laptops produce severe noise on the A/D input. Test the A/D input carefully before buying the laptop.[7]

  If you want to use a laptop, record via a USB audio device (e.g. Griffin Technology's iMic). An alternative without laptop is to use a standalone digital recording device (e.g. Nomad Jukebox). Such a device can record up to 4 hours of uncompressed audio data onto its internal hard disk and then allows a rapid transfer of data to the computer via USB or FireWire.

- Record on a DAT tape recorder
  Many people out there still do that because the quality is outstanding. Also, you'll automatically have a backup medium (the tape), if something happens to your recorded data later (and as Murphy says: "What can go wrong will go wrong!").
  On the other hand you then have to copy your digitized data from DAT to your computer. Most DAT players have a digital port (either electronic or optical) that can be connected to special cards (e.g. TripleDAT). Then you have your data in interleaved blocks and with 48 kHz (or 44.1 kHz) sampling frequency. You will most likely then filter them digitally and downsample to a suitable sampling frequency for speech.
  All this takes time and man power from your project budget.

For larger setups (more than 4 microphones and possibly several feedback channels for simulated background noise) you might consider a semi-professional music mixer to handle all the amplifying stuff.

Recording Software

If you use a computer to record, you will need some software to access and control the A/D card. There are plenty of applications for recording sound from your sound card; probably the manufacturer of your sound card will provide one as well. A good tool for testing the quality of your setup is *Praat* ([3]) because it allows you to record and to check the spectrum and sonagram for noise and hum. Of course you may also buy a professional

[7] In most cases you can hear the spinning up of the hard drive quite clearly.

or semi-professional recording software (often in combination with a 4- or 8-channel card).

Some of the commercial recording software packages will also allow you to filter and downsample signals after the recording and change them into different signal file formats.

5.2.3 Field Recordings

Most of the things said in the previous section are also true for recordings in the field. Here are some additional practical hints:

- Most of your devices will be battery-powered. Be sure that your recording team always keeps fresh batteries ready. Test the devices before each recording (check list).

- In larger installations you might have AC power available. Power lines outdoors very often have bad grounding that may cause 50/60Hz hum on your signals. Do not perform the pretest in the lab; do it in the woods.

- Even worse are emergency power supplies; they might produce all kinds of noise in your signals. Look very carefully (spectrum) at your signals after the pre-test. Use AF filters in the power cords to your recording devices.

- Switch off all cellular phones within a perimeter of at least 20m around your recording site. They produce audible interferences with sound devices at irregular intervals.

- If you plan recordings in the running vehicle, you may have all sorts of problems getting a computer running reliably. Standard PCs cannot be powered by DC 12V; you will need an industrial version of the Intel PC. Alternatively you may use a laptop for prompting only[8] and do the recording with a digital recording device independently.

- Be prepared to cope with nasty weather conditions; some microphones react allergically to humidity.

- Plan to backup recorded data immediately after each recording. Use built-in CD-R drives or a second DAT recorder to copy your signals. Store the recording media in a dry and cool place.

[8] For most laptops adaptors for DC 12Vi are available.

- Plan plenty of time for the setup and testing. Then double the time in your schedule. It's better to have an extra coffee break than to ruin recordings.

5.2.4 Wizard-of-Oz Recordings

WOZ recordings may be designed in many different settings depending on your needs or the features of the simulated application ('virtual machine'); therefore it is difficult to give detailed instructions on how to set up a WOZ recording site. Please also refer to the section 5.2.2 for general hints and tips as they probably apply for WOZ recordings as well. Here are some general hints specially for the WOZ technique that might be useful:

- Do not use a room with a one-way mirror. This technique is so well-known by now that almost every speaker will become suspicious. If you need to watch or video tape the speaker, use small cameras instead, and if you cannot hide them successfully, simply add them to the 'virtual machine'[9]
- Take care that the location of the wizards (the 'control room') is acoustically insulated from the recording room, so that the wizard can produce live speech output and/or communicate (possibly via an intercom) with the recording supervisor.
- Since the microphones are further away from the recording device in a WOZ setting than in a usual setting, make sure that you don't get a hum on your signals. If you experience disturbances, try to set up the microphone amplifier in the recording room and not in the control room.
- In most WOZ recordings you will need a synthetic speech output of the 'virtual machine' instead of written feedback. According to our experience it is next to impossible for a wizard to type absolutely error free. On the other hand it is quite easy to distort the human voice – for instance by using sound devices from electronic music stores – and to train the wizards to speak in a very controlled manner. The only weak points in this setting are laughing or coughing (no machine laughs or coughs!).
- Do a lot of training sessions with cooperative, non-naive users and ask them what went wrong. Since WOZ recordings are very expensive

[9]The application that is simulated by the WOZ experiment. For example if you need a camera in a simple command and control recording, tell the speakers that the machine uses a camera to detect the point in time when he/she starts speaking.

because of the manpower you need, it is better to do some extra off-line training sessions before you recruit naive speakers.

- Clarify the legal aspects of recordings in which the recorded person is unaware of the situation. For instance you should ask your legal department whether it is ok *not* to tell the speakers *immediately* after the recording that it was a fake. Because if you *have* to tell them the truth right after the recording, very soon you will not be able to find naive speakers in the vicinity any more.

- Simulating very complex ('intelligent') systems is difficult because the wizard constantly has to decide whether a user input is still within the capabilities of the 'virtual machine'. A good way to handle this is to prepare *task flow maps*. Task flow maps are like a semantic map that shows the wizard which possible ways through a dialog with the 'virtual machine' are allowed and which are not. Since the wizard has to think about it (time delay) it is also very helpful to have some pre-recorded standard answer ready that can be played by just pushing a button[10]. For instance there should be a general "Sorry-I-did-not-understand-you" button and probably others that fit exactly to the weak points in the 'virtual machine' setup.

- Keep the recording session short and keep the speakers occupied. In long sessions it is more difficult to maintain the illusion that the speaker talks to a machine. Bored speakers start to look around or think about clever ways to trick the system.[11]

This is by no means a complete list of hints. Many problems you'll find in the pre-tests might be solved by a better or more specific instruction to the speakers as well.

5.3 Questionnaires and Forms

You should prepare a number of forms for the collection phase. As a minimum you should provide:

[10] This also increases the 'machine-likeness' of the simulation because these pre-recorded answers sound exactly the same every time.

[11] In the SmartKom WOZ experiments we asked speakers to solve a simple task with the help of the 'virtual machine'. However, one of the speakers finished the task very quickly and spent the rest of the recording time testing the system with a kind of *von Neumann* test: he repeatedly asked the system to meet him at the cinema and maybe later to have dinner together. Fortunately, our wizard kept a straight face (straight voice) and kept on hitting the button saying *Sorry Sir, I did not understand. Could you please state your question again?* again and again and again...

- Speaker questionnaire for speaker meta data
- Form for the recording protocol
- Statement about the transfer of Intellectual Property Rights to be signed by the speaker
- Receipt form for the incentive of the speaker

Furthermore, you might think of other questionnaires about the recording situation itself (especially in WOZ recordings) and forms for the documentation, if you work with other partners at various recording sites. Questionnaire and forms may be either on paper or in electronic form (see also section 6.1, p. 83). If you are using an automated recording process, you may include the questionnaires in that.

5.4 Legal Aspects

Be prepared to face all kinds of legal problems during the collection as listed in chapter 2 (p. 19). Consult your legal advisors to get information on how to formulate forms and statements you will need in your relation to the speakers. Take care that signed documents are kept in a safe place. If you pay incentives to your speakers let them confirm that they received the money. Include necessary advisory steps into your check lists for the experimenter.

5.5 Check Lists

All manual actions taking place during the actual recording should be fixed in check lists. For instance you might use a check list for the technical setup of the recording devices, another check list for the recording procedure itself and a third one for all the activities after the last recording e.g.: backups, cleanup of disk space etc.

Design your check lists before the pre-test (see next section) and let other members of your staff test them for consistency and comprehensibility. Keep copies of the check lists for the final documentation of the speech corpus.

5.6 Pre-test

The pre-test is not part of the speech corpus nor part of the pre-validation, if you plan to do one. The pre-test is the only way to eliminate all the bugs

that are still in your procedures, your check lists, your software and in your staff.[12]

Never — I repeat — never skip the pre-test of your recording setup and equipment. Do it with cooperative speakers and do it way ahead of the scheduled start of the collection. We've never experienced a speech corpus production that worked without any kind of trouble right from the beginning.

Do everything exactly as you would do it in the real collection. Keep a log file to record all actions you are taking. Let your pre-trained staff test the prepared check lists and software. Don't let the developers of the software do it!

Shoot plenty of pictures for the documentation; later in the collection phase you will probably not have the time and will make the speakers nervous.

Do spot checks on the signals right after the recording. Send them together with your log files to your client / partners for comments.

If you're preparing a field recording, do the pre-test under exactly the same environmental conditions as planned for the collection; if possible on the same location.

5.7 Planning of Recruitment

The last thing to do before going into the collection phase is the preparation of the speaker recruitment.

Again, the recruiting technique you'll use depends on the kind of speech corpus you're producing and on how much funding you have. You will find general hints from our long experiences with different corpus productions in section 6.6 (p. 88). Here is some advice as to why you might need to think about the recruitment way ahead of the start of the collection:

- If you have plenty of funding, you might consider to out-source the problem to an advertising or market research agency. Some agencies keep extensive databases that may be used to find speakers of your desired profile and mail them directly. Expect costs of about EUR 20-40 per successfully recruited speaker (without the incentive). We mention that here already because these agencies usually need some time to prepare (2 months).

- If you do the recruitment yourself, assign one person from your staff to it. This person should then take care during the collection phase that

[12] Believe us: the bugs are there!

enough speakers of the needed profiles are scheduled for the recording slots. Give that person some time to get familiar with the problem (1 month).

- Usually it's a good idea to start the recruitment at least a month ahead of time. Your best asset in the recruitment business are the recruited speakers themselves. Offer them more incentives for each new recruited speaker (snow ball systems) and they will go for it like starved squirrels for the roasted peanuts. But to get a significant mass this takes some time. Therefore it is good to start early and do a lot of pre-scheduling.

Please also refer to the chapter 2 (p. 19) for legal advice in the matter of storing speaker information.

It is a good idea to collect data about one's speakers in a database [2], pp. 138:

> "Speakers should be thought of as a primary and very valuable resource in speech recordings. It is therefore advisable to build a speaker database which contains for each speaker
>
> - a unique speaker id,
> - administration data (name, address, telephone),
> - personal information (place and date of birth, languages, education, etc.),
> - physiological data (sex, size, weight, etc.),
> - speaker history (list of recordings, etc.),
> - remarks.
>
> The recruitment of speakers should have two goals: provide a sufficient number of speakers for a given speech data collection, and provide sufficient information about the speakers which can be used to build or extend a speaker database."

Check List Preparation of Collection

◯ Instructions * (p. 67)
◯ Prompt List ** (p. 67)
◯ Automated Recording Procedure *** (p. 67)
◯ Test of Instructions, Prompts, Procedure * (p. 67)

Recording Techniques Telephone (p. 69)

◯ ISDN Account *
◯ ISDN Hardware + DLL (CAPI) *
◯ Control Program *
◯ Speech Prompts + Beep (Check for DC and clippings) *
◯ The 'script' *
◯ Silence detector ***
◯ Speech Detector ***
◯ Adjust / test recording intervals / detectors *
◯ Check for 'echos' *

Recording Techniques On-site, Field + WOZ (p. 72)

◯ Acoustical Environment *
◯ Microphones *
◯ Amplifiers, set levels *
◯ Recording Devices *
◯ Recording Software *

Recording Techniques Field (p. 75)

◯ Batteries *
◯ Check AC Grounding + Power Supplies *
◯ Banish Cellular Phones *
◯ Recording Devices *
◯ Be prepared for bad weather *
◯ Daily Backup *

Recording Techniques WOZ (p. 76)

◯ Observation technique (no mirrors) **
◯ Acoustically insulated recording and control rooms *
◯ Simulate Synthetic Speech Output *

- ◯ Clarify Special Legal Aspects *
- ◯ Task Flow Maps *

◯ Legal Aspects * (p. 78)
◯ Prepare Doc Forms and Questionnaires * (p. 77)
◯ Prepare Check Lists * (p. 78)
◯ Pre-test * (p. 78)
◯ Plan Recruitment * (p. 79)

Chapter 6

Collection

This chapter describes the activities at the core of every speech corpus production: the actual recording of the speech signals. Most of the technical and and other advice you've already found in the previous chapter. In the following we summarize some more practical hints that might be useful during the actual collection phase. Basically these hints can be structured into *Ongoing Documentation, Pre-Validation, Quality Control, Data Pipelining* and *Recruitment*. The order of the sections in this chapter are not meant to be chronological. You should read them all before you start the collection.

6.1 Ongoing Documentation, Logging

Ongoing Documentation or Logging is of paramount importance to ensure the later usability of the speech corpus. All processes of the data collection must be documented in such a way that the user of the speech corpus understands all aspects that might be of importance for the later usage of the data.

There are basically two ways to do the logging during the speech data collection: on paper or online.

Logging on paper is easy and can be performed everywhere without computer hardware. However, in most cases the written data must be transferred into machine readable form later which means additional costs. It is much better to perform online logging, either by using a customized editor or into a database system via a Web server.

Practically all modern database systems allow the access and input of data via a Web interface. The advantage of this method is that different data from different processes can be easily linked together. For instance you

might use the same database system for the scheduling of your recording sessions and to input the required meta data about recordings and speakers. Care has to be taken that the basic rules of data protection are observed. See also section 2.5 (p. 22).

The following list gives the obligatory data to be logged (marked with one *) and other possible data of interest logged during the collection phase:

- Recording Protocol *
  These data are the basis for the meta data files about each individual recording session or recording procedure that have to be included in the final speech corpus. Follow your specifications about your recording protocol (section 4.10 (p. 64)) or refer to section 3.2 (p. 28) for a basic discussion of meta data.

- Speaker Protocol *
  These data are the basis for the meta data files about each individual speaker participating in your speech corpus production. Follow your specifications about your speaker profiles (section 4.10 (p. 64) or refer to section 3.2 (p. 28) for a basic discussion of meta data.

  Both – recording and speaker protocol – should contain codes and free text comments as discussed in section 3.4 (p. 34).

- Comments of Speakers

- Questionnaires

- Statistical Data
  For instance, how many recorded words in unsupervised recordings, S/N ratios, other technical conditions, covered dialects or other required specifications (languages, locations, sex, age groups etc.)

If you are working on a large data collection with many staff members or project partners at different locations, you might also think of an automated Web information system, where interested parties can monitor the progress of the collection and react to certain developments[1].

6.2 Pre-Validation

In a large speech corpus collection it is highly recommended that you perform a pre-validation after a small amount of collected data, preferable

[1] For instance to strengthen the recruitment efforts in areas that are not covered yet.

conducted by an external validation center. Do not confuse the terms pre-test (section 5.6) and pre-validation: the pre-test is only concerned with the testing of the technical setup and procedures; the pre-validation deals with real data that will be part of the resulting speech corpus.

The optimal model for a pre-validation is that after a pre-defined number of recorded speakers the speech signals, the annotations, the meta data and documentation files are transferred to an external validation center which will perform a formal validation of the data. The collection awaits the results of this validation, then reacts to found errors or other recommendations and then continues.

In practice there will often be some restrictions on this ideal situation: in most cases the annotation files won't be ready after such a short collection time, and the same is probably true for documentation and meta data files. Nevertheless, at least the speech signals should be validated against their specifications.

6.3 Quality Control

Aside from the pre-validation an ongoing quality control is necessary to detect systematic errors or deviations from the specifications early. There are basically two different forms of quality control: the control of the recording process itself and the control of the recorded contents (also known as *monitoring*).

6.3.1 Monitoring

A monitored speech recording is a technique where the speaker is required to follow a strict 'script' (i.e. reads a text) and a supervisor is present to ensure that the speaker does not deviate from the given 'script'. If this is done systematically over the whole corpus you might skip the annotation part of your speech corpus production entirely[2].

The following is partly taken from [2], p. 129:

[2]In fact this has been done in BAS corpus productions but we do not recommend it. The reason for this is that the costs for a online monitoring is often higher than a annotation after the recording. Furthermore, a post-recording annotation may find errors that go by unnoticed even in a monitored recording and additional characteristics in the speech signal may be annotated that the supervisor was not able to detect (for instance a disturbance of the signal caused by a malfunctioning recording device). Finally, we think that a speech corpus of monitored speech utterances is in most cases not what the users of speech corpora really need: The scientist or developer of a speech application has to cope with errors in the spoken language input. Therefore they should not be omitted from the corpus but rather labeled. Only corpora for speech synthesis might justify a monitored recording technique.

> "Monitoring is the task of controlling and modifying technical and phonetic characteristics on-line, i.e. during the course of a recording. Validation relates to an off-line (or post hoc) technical or phonetic evaluation of the material recorded. Two on-line monitoring paradigms can be distinguished: one in which any deviation or error is signaled to the experimenter only, and another one in which also the speaker is informed that a particular error has occurred. Some characteristics of recorded speech can only be evaluated after the recording has taken place. In the technical domain such characteristics are the signal-to-noise ratio for the whole material, and an analysis of noises that were recorded with the speech."

6.3.2 Control of Recording Process

The easiest way to ensure the compliance with the specifications (recording setup, technical specifications) is to include spot tests into the check lists of the experimenter / supervisor and to make sure that they are observed. Spot tests may concern the speech data itself, the required questionnaires and data forms and the backup procedures (see also section 5.5, p. 78). Spot tests may be automated to some extent, but if you do so, we recommend adding some random component to make sure you're not checking the same data again and again.

6.4 Security

As already mentioned earlier you'll need to set up sufficient procedures to protect your recorded data.

6.4.1 Security against Theft

Although unlikely, you should consider the possibility that data is ripped off your recording equipment while your staff are not present. This may be a severe problem if you work for a client because often you will find articles in standard contracts that explicitly make you – the producer – responsible for ensuring the security of the recorded data. Include safety regulations in your start and end check lists. Use lockable computers or hard drives and security equipment to chain your equipment in unlocked rooms or unsafe locations.

6.4.2 Security against Data Loss

Primarily, your concern should be the usage of daily backup procedures to protect against hard disk failures. Add them to your check lists and make sure that sufficient backup media are available. Provide a safe place to store them, preferably not at the recording site. As mentioned before, good backup media are CD-Rs or tapes.

As a second concern protect your data against vandalism. Use passwords and disconnect your equipment from the network when not used. Make sure that no unauthorized personnel may access the equipment because most computer systems are vulnerable if the intruder has direct access to the system.

6.5 Data Logistics

Logistics of data concerns all problems regarding the data transfer and data storage during the collection phase. Depending on your corpus specifications speech data will accumulate on your recording devices quickly and will have to be transfered to a safe location and/or to post-processing and annotation.

6.5.1 Storage

In most cases this is trivial and can be solved by using a file server (with backup utility) and corresponding procedures that ensure that the data are transfered to this server right after each recording. Try to minimize times where data is stored on hard disks without any backup. If you want to be absolutely safe against data loss, use a parallel DAT recorder running during the recording sessions.

Also calculate the maximum amount of disk space needed to store the data of a single recording session and make sure that on your recording device space is always left for at least two or three additional sessions. Think of suitable methods to prevent other users from installing or copying data to your recording devices; you might also insert a check point into the check list of the experimenter to check for available disk space before each recording.

6.5.2 Data Pipelining

Speech corpus collections are usually not a strictly linear process as depicted in this cookbook. Therefore it is most likely that you will start to process or even annotate your recorded data while the collection is still in progress.

The term Data Pipelining refers to the logistical problem of ensuring that the required data are at the required location at the right time.

In large projects where many post-processing steps and annotation procedures are necessary and where these processes might be conducted in parallel by different working groups, this problem can be the hardest to tackle. One aid for avoiding costly idle times is to design a dynamic data flow chart where staff members can see online what data are available and what data are to be processed next and even what data are to expected in the near future. One practical way to realize this might be a Web interface generated by a database which all processing steps are logged into. A single working group may check out data for annotation in the database, and later on, after finishing the job, mark the data as ready. If this is done systematically and consistently, it is easy for the management to detect bottle necks or idle times early enough to react accordingly.[3]

The concerns about storage and safety mentioned above also apply to the whole pipeline, of course. Very often you will find that idle times are not caused by too slow or too fast working groups but by missing resources like disk space. Always be prepared to store the data of up to 10 recording days 'on the side' because there is a problem in the data pipeline that has to be solved. If you do not provide this, the whole pipeline might come to a stop, which might cost you a lot of money.

6.6 Recruitment

If you don't need to do the recruiting yourself, you're lucky[4]. Or – as mentioned earlier – you might out-source the recruitment to an external agency. In most cases however you will have the problem of getting speakers of the right kind, at the right place and at the right time. Again, the recruiting technique you'll use depends on the kind of speech corpus you're producing and on how much funding you have. In the following you will find some useful hints for your recruitment during the collection[5].

6.6.1 Basic Recruiting Techniques

- Your best assets in the recruitment business are the recruited speakers themselves. Offer them additional incentives for every new recruited speaker (snow ball system) and they will go for it like starved squirrels

[3] See for instance [8] for a description about data pipelining in the SmartKom project.
[4] For instance, if you work in a company and will recruit the own employees as speakers.
[5] Please also note the sections about recruitment in chapter 5, p. 79, and chapter 2, p. 20.

for roasted peanuts. But to get a significant mass this takes some time. Therefore it is good to start early and do a lot of scheduling.

- If your speech corpus requires the same speaker to be recorded more than once, do not pay the incentive before the last recording session. Give feedback to speakers, especially if they are recorded via telephone lines, to keep them motivated. That way you'll minimize the number of incomplete recording sets.

- Some remarks about different advertising methods:

  - Newspaper, TV or radio ads:
    Not very effective and usually rather expensive.

  - News paper articles or small radio/TV stories about the collection:
    Very effective! The problem is to find somebody who is interested in the project and will do a story about it[6]. Try the popular science sections.

    Great advantage of this method: you may select news papers or radio/TV stations based onyour desired speaker profiles. For instance if you're looking for young female speakers, you might contact young fashion magazines; if you're looking for speakers with a certain dialect, you might try a local radio station in that area.

  - Internet:
    Absolutely not effective. Informative Web pages about the project and Web forms (to register as a speaker) may be of help with the recruitment, but they are not sufficient to get to the speakers.

  - Associations / schools / colleges / clubs:
    If you have special speaker profiles to fulfill, use your imagination, a good search engine or the yellow pages to identify institutions for this particular group. For instance you might contact the local Association of Turkish Tennis Players, if you need native speakers of Turkish. Or contact the local high-schools and offer them a free tour through your lab and record them in small groups while the rest plays Tetris on your transcriber workstations.

[6] If you're working for a non-profit organization this is much easier than if you work for a company.

6.6.2 Incentives

In any case you'll have to pay the speakers an incentive; money is still the most effective incentive but not easy to distributed by mail (in case of remote speakers like in a telephone recording). Think of valuable things that are easy to be mailed, for instance telephone cards, cash cards for Internet purchases etc.

The value of your incentives can roughly be estimated from the time the speaker has to spend for the recording: 2 minutes equals 1 EUR is a good rule of thumb (not counting the time to get to the recording location; just the pure time spent there).

Incentives for telephone recordings are usually at a lower rate because the speaker can participate from wherever he wants to: 3 minutes equals 1 EUR.

As a minimum incentive we consider the sum of 5 EUR[7].

[7] All numbers taken from the year 2002.

Check List Collection

Set up Logging Procedures *(p. 83)*

- ◯ for the recording protocol *
- ◯ for speaker meta data *
- ◯ for speaker comments **
- ◯ for questionnaires **
- ◯ statistical data ***

◯ Organize Pre-validation ** (p. 84)
◯ Set up Procedures for Quality Control * (p. 85)
◯ Check for Security * (p. 86)
◯ Provide enough Storage * (p. 87)
◯ Organize Data Pipelining * (p. 87)
◯ Choose your Recruiting Technique * (p. 88)
◯ Define Incentive and their Distribution * (p. 90)

Chapter 7

Post-processing

Post-processing includes all processing steps from the recorded raw signal data to the final distributed corpus. The following processing steps might not all be necessary in your corpus collection; however, some of them are (marked with a *): *file transfer from recording device to computer, file name assignment*, filtering, cutting, synchronization, re-sampling, format conversion*, special conversion for annotation* and *automatic error detection**. Please note that some of these processing steps may be applied after or between the annotation steps described in the next chapter depending on the structure of your data pipelining (see section 6.5.2, p. 87).

We deem this chapter to be quite relevant for the prospective producer of a new speech corpus, because the costs and man power needed for post-processing is often neglected or at least grossly under-estimated. Please review this chapter carefully before you calculate the overall costs of your corpus production and take into account all the necessary post-processing steps for your individual corpus production.

Although the order of the processing steps is in principle arbitrary[1], the most effective order is given in the following description.

7.1 File Transfer

As discussed in section 5.2.2 you might use recording devices that are not computers and will store the digitized signal data on media that cannot be read directly from your post-processing machine. If this is the case in your speech corpus production, the very first step is of course the transfer

[1] Except for filtering before down-sampling!

to a proper signal file. Set up check procedures to ensure that no data loss may happen during this transfer. Since the transfer techniques from non-computer devices to a computer is often designed without automatic hand-shake and size-verification procedures, you might loose data without any error messages from your devices.

7.2 File Name Assignment

The first thing to do after a recording session is to assign the correct filenames to the signal files[2]. Although it is possible to use an internal terminology during the post-processing, we do not recommend it to avoid unnecessary confusion. Therefore you should include an item in the check list for the experimenter that he/she correctly classifies the recorded raw data immediately after the recording and gives them the required names.

It's a good practice to use the prefix of a file name for your terminology and the suffix to mark the type of data file. For instance, if your recording device delivers raw signal files without header with 48kHz sampling rate, you might use the suffix *.raw48* for these raw data files.

Since the assignment of file names is often done manually, it is wise to add a small parser in your automatic error detection routines (see section 7.8) to find wrong or impossible file names according to your specified terminology.

7.3 Editing

Depending to your special recording setup it might be necessary to splice the relevant data segments from a larger raw data recording[3]. For instance you might use a DAT recorder to record dialogues in the running car and you have no means to automatically start and stop your recording device. Then you will end up with a signal file containing the complete recording session including instructions and remarks from the experimenter or the participating speakers that you do not want to be part of the final corpus.

To cut out the relevant signal segments and split them into correctly named single files you may either use automatic procedures or do it manually using a sound editor or a combination of both.

Fully automatic editing procedures require either a good silence detector or – in case that there is background noise present in your signal – a good

[2]Correct with regard to your specified terminology, see section 4.8.2, p. 61.

[3]This process is sometimes also referred to as *segmentation* but we prefer the term *editing* to distinguish it from the segmentation of speech into linguistically units.

speech / non-speech detector, and in addition you have to know when and how many pauses are to be expected in the speech recording. Such a technique might work reliably if your recording contains single utterances where no between-speech-silence occurs. If you are recording sentences, turns of a dialogue or even free conversation, this technique will most certainly fail. In any case we recommend that you verify the resulting cut signal files in your annotation step (see next chapter) for editing errors. Alternatively you might use a semi-automatic procedure that detects codes in your raw signal file to get the editing information. For instance in a telephone recording of a dialogue between two parties, we asked the speakers to press a certain button on their DTMF phone before starting to speak[4]. The whole session was recorded into a single channel raw file[5] and later in post-processing this file was automatically cut into the turns of the dialogue ([9]). The DTMF codes might also be created automatically by a computer which controls the text prompting as done successfully in the SpeechDat Car project ([12]).

In most cases however you will need to cut out relevant signals manually. It depends on your individual corpus design which segments are a good choice. In our practice we encountered the use of whole dialogues, turns, sentences, dialog acts, phrases, words or even single phonemes. Also keep in mind that the physical editing of your raw signal files might also be avoided by providing only the segmental information – as done for instance in the Verbmobil II corpus collection. In Verbmobil II the whole dialogue between two partners was recorded in several synchronized channels and only the begin and end of each turn was marked in an annotation file, so that partners might cut out automatically relevant speech segments stemming from one speaker ([10]).

To physically edit signals by hand you can use any sound editor or probably as a best choice the *Praat* program ([3]).

7.4 Filtering

In some cases you might need to filter your raw data. Very often this will be necessary when you intend to down-sample your raw data (see following section). But it might also be the case that you encounter a constant disturbance in your recorded signal data that cannot be avoided in the recording. Typically these are 50/60Hz hums that might be filtered very effectively using a digital high pass or notch filter. You may design your own filter or

[4] Of course in this case no real spontaneous conversation is possible, because the partner has always to wait for the other partner to finish.

[5] Using a simple conference call and an ISDN card.

use tools like *genfilt* from the public domain *SFS* software package[6]. Use a different suffix for the filtered signal files to avoid confusion and double filtering of the same data. Refer also to [13] for a more detailed discussion of techniques to improve the quality of your recorded signals.

7.5 Re-sampling

Very often you will find the situation that your recording device does not record with the desired sampling rate as specified for the final corpus. A typical case is the recording with a DAT recorder which usually allows only either 48kHz or 44.1kHz sampling rates. These high sampling rates are required for HiFi recordings but not for speech where a maximal sampling rate of 22.05kHz is sufficient.

To save space in the final distribution the signals have to be down-sampled. Prior to down-sampling you have to be sure that the recorded signals do not contain any frequencies higher than half of the intended sampling frequency after down-sampling[7]. You may either take care of that in the recording process itself or filter the raw data using a low pass filter before down-sampling.

There might also be the case that you have to re-sample your data to higher sampling frequencies, for instance to meet special requirements of your partners, an annotation tool or your client. In this case no information is added to the signal and therefore no filtering is necessary.

Re-sampling can be done using public domain tools like *sox*[8]. Some professional tools automatically filter the signals before re-sampling. Check their respective manual to be sure. Be aware that re-sampling in most cases causes a degradation of quality of the raw signal. It depends on the algorithm used and the sample format of the data how good the quality of re-sampled data will be. In most cases we have found that *sox* delivers sufficient quality. More importantly, *sox* explicitly states how rate conversion is performed – most other applications and tools do not disclose this information.

7.6 Format Conversion

Most likely you will have to convert your final signal files into a standard format as given in your corpus specification (see section 4.7.4, p. 54 for

[6] www.phon.ucl.ac.uk/resource/sfs/
[7] Nyquist or Shannon Theorem
[8] www.spies.com/Sox/

a detailed description of the most common speech file formats). If your target file format is a standard, you will probably use the general speech file conversion tool *sox* mentioned above. *Sox* has the great advantage that it is a command line tool and therefore easy to include in your scripts. Again we recommend that you include some simple check procedures into your check lists to ensure that no data loss has occurred in the conversion.

If you want to use a compression technique like *shorten* or *gzip*, you should place the format conversion at the end of your data pipeline because your annotation tools most certainly will not work with compressed input data (see section 4.7.4 for a discussion of compression techniques).

7.7 Special Conversion for Annotation

In some corpus productions you will use special annotation tools that require a speech file format other than the specified format. For instance in the SmartKom corpus production the annotation of speech, facial expression and gestures was done with a variety of annotation tools that each require a different input format. A considerable effort was necessary just to provide the right formats for all annotation groups.

If possible, choose your annotation tools so as to avoid such unnecessary conversions.

7.8 Automatic Error Detection

Automatic error detection denotes all check procedures that may be carried out automatically. There is no certain place in the chain of the post-processing where automatic error checks are most recommended. Whenever possible include automatic checks after each processing step, especially after steps that require manual work.

Here are some possible checks that can be carried out automatically:

- Check for empty files
- Check for correct length
- Check for correct terminology
- Check for speech contained in the file or only silence
- Check for required number of files per recording session
- Check for total disk space of a recording session

- Check for S/N ratio
- Check for consistent header content

An easy way to implement such automatic checks is the usage of a UNIX shell. Most checks are simple UNIX commands or can be easily implemented in a script language like AWK, perl or BASH. If you are using MS platforms[9], you can use the public domain *Cygwin* applications that allow you to run a UNIX-like environment on your PC[10]. You may check [11], chapter 5 for a collection of small examples scripts.

In a large data pipeline it may happen that checks are omitted and nobody notices until later. A good way to avoid this is a central log file or database where all checker programs may enter their results together with date and the checked data.

[9] Our most sincere regrets.
[10] www.cygwin.com/

Check List Post-processing

In this check list the processing steps that might not *be obligatory are marked with **.*

Chapter 8

Annotation

The annotation of a speech corpus denotes all symbolic[1] information that is directly related to the speech signal, either via the physical time scale, in which case we speak of a *segmentation and labeling*, or via some semantic content of the speech signal, in which case we speak of a *transcription or tagging*.

For an in-depth discussion about annotation of speech corpora refer to [2], pp. 146 - 161.

This chapter will give a short introduction to annotation types as well as practical hints on contents, procedures and annotation tools in the context of speech corpora.

8.1 Types of Annotation

The following list of annotations is taken from the documentation of the BAS Partitur Format[2] and will give you an idea of what different types of annotation might be used and what has already been done so far. Pure transcriptions or tagging are marked with an (T), while segmentations and labellings are marked with an (S):

- Orthographic transcript (T)
- Canonical pronunciation (citation form) (T)
- Broad phonemic/phonetic segmentation and labeling (S)

[1] Symbolic in the sense of discrete or categorical.
[2] www.bas.uni-muenchen.de/Bas/BasFormatseng.html

- Word segmentation (S)
- Dialog act labeling (T)
- Syntactic-prosodic labeling (T)
- Prosodic labeling and segmentation in Tobi (S)
- Phonetic segmentation and labeling in IPA (S)
- Noises: articulatory and technical (S)
- Segmentation or tagging of cross talk (T/S)
- Parts-of-Speech (T)
- Syntax trees (T)
- Translations (T)
- Turn segmentation (S)
- Prosodic labeling of accents and boundary types (S)
- User state segmentation and labeling (S)
- Meta-linguistic events: breathing, laughing, cough, hesitations. (S)
- Discourse events: false starts, stutter, repeats etc. (T)
- Glottal pulses (S)

Note that a transcript contains no information about the time relation of its contents aside from the fact that usually a chunk of speech is associated to a chunk of transcript. For example, if the corpus is structured in paragraphs of read text, then each signal file stores the speech of one paragraph while the associated transcription file stores the transcript of what was said in the signal file, but there is no fine-grain time information about when each individual word starts and ends within the signal file.

A segmentation requires either

- a point in time or
- a starting time and ending time or
- a starting time and duration

of the labelled category. For example, in a phonemic segmentation and labelling each segment will consist of the phoneme category (coded for instance in SAM-PA), the begin of the phonemic segment and the duration:

```
IPA:   1.2758934 0.097867  e:
```

8.2 Data Model

Liberman and Bird[3] claim that all types of annotations as described above may technically be represented in form of a directed, acyclic graph where the nodes of the graph may (but do not need to) represent points in time while the arcs between the nodes represent labels or tags. This data model is very useful if you intend to write your own annotation software. At the LDC[4] you will find a number of publicly available software tools and libraries (Annotation Graph Tool Kit, AGTK) that might be used for that purpose and that are all based on the data model of Liberman and Bird. Since the internal data model is independent of the file format, you may use these tools and libraries for all kinds of processing tools that deal with the input, output or transformation of annotation data. Please refer to the documentation on the above-mentioned web-site as well as to [1].

8.3 Orthographic Transcription

The most basic type of annotation that makes a collection of speech recordings into a speech corpus is some kind of orthographic transcription. This can range from a simple chain of words per recording item (based for instance on the script that was used during the recording) to an extensive labeling of several different semantic layers[5]. The choice about what is to be included in the transcript is dependent on the type of speech corpus and the intended usage. For example, a corpus of read speech items over the telephone network with the aim to train automatic speech recognition algorithms does not need any elaborated labeling of discourse events. A corpus containing dialogue speech between two or more persons that is subject to scientific investigations will require much more effort.

8.3.1 General Rules for Transcription

- Follow the 'natural segmentation' of the corpus into the individual signal files and create one transcription file or one line in a table or one entry in a database per signal file.[6]

[3] See for instance [1].

[4] agtk.sourceforge.net/

[5] In some cases the latter is called a *transliteration* to distinguish it from a simple orthographic transcript. Beware: some authors do not even use the term *transcript* for the orthographic representation at all, because they reserve this term for the phonemic or phonetic representation.

[6] If your corpus is not segmented into signals files of the size of an utterance or less, consider incorporating such a segmentation into the transcription. For example in the

- Use a standard spelling and character coding.
- Use capital letters only according to your spelling rules; not at the beginning of sentences.
- If you use punctuation marks, always separate them from the last word by a white space; in most cases it is even better to omit punctuation completely.
- Do not use any white space characters in any other meaning than to separate items in the transcript. For instance do not use a format where a certain number of blanks is required to mark the beginning of a turn. This will lead to severe problems in the parser.
- Do not allow any digits in the transcript but represent spoken digits, cardinals or numbers as their written names, e.g. '456' as 'four hundred fifty six', '6th' as 'sixth' or '72.5' as 'seventy two point five'.
- Use a format that is brief[7] and readable. Unfortunately, formats that are easy to parse, like XML do not meet this requirement. Therefore, you might consider using an intermediate format for the transcription work and transform this format later into something like XML.

8.3.2 Possible Transcript Items

The following table gives a rough overview about possible labels and tags contained in the orthographic transcript[8]. You may review the following tags and decide which of them might be useful for your special needs. In the third column you see an example of how the items may be tagged. Of course you may also use an XML-style tagging instead.

Assume for the following list of tags that a dialogue between two or more speakers is transcribed turn by turn by listening to the signals.

Verbmobil II speech corpus the first edition consisted of signal files of approx. 10 minutes length that contained the speech of one dialogue partner over a whole dialogue. In the transcript this long recording was segmented into dialogue turns and numbered throughout the dialogue starting with '000'. Furthermore, the transcribers were asked to markup the begin and end of each turn on the time scale resulting in a rough segmentation of the recording which simplifies the later usage of the corpus considerably.

[7] That is: does not need a lot of redundant typing.

[8] Most of these are more thoroughly explained in section 15.2.

Item	Remarks	Example
Lexical unit	Standardized spelling/character coding. Define a lexical unit: words, interjections, neologisms? Lexical units are usually not tagged.	`station`
Spelling	Spelling of a word or abbreviation letter by letter.	`$U $S $A`
Acronyms	Official substitutes for words or phrases, spelled like a word	`OPEC`
Proper names	All names that cannot be translated into another language: People's names, street names, restaurants etc.	`~Peter` `~Marine+World`

Numbers	Numerals, combinations of numbers and ordinal numbers. All number written as words	`#three`
		`#twenty`
		`#hundred`

Neologism	Word that has been made up by the speaker	`*deliverator`
Foreign Words	Words that are from another language and have not been officially adopted by the main language	`<*IT>saluti`

Off-Talk	person is speaking to himself or herself and not to the partner(s) of the dialogue	`what<OOT>`
		`do<OOT>`
		`I<OOT>`

Read Off-Talk	Off-talk caused by reading aloud	`seven<ROT>`
Command Words	Words to operate a dialogue system	`!KEYComputer`
Lengthening	Markup of sounds within an item that are lengthened	`so<L>rry`
Garbage	words completely or partly incomprehensible	`<%> three%`
Truncation	Item is truncated for several reasons (technical, stutter etc.)	`so the que=` `by hel= <*T>`
Interruption	Items may be interrupted for several reasons: pauses, breathing, hesitations etc.	`trans_ <A>` `_lation`
Missing signal	Missing parts of the signal for technical reasons have to be marked in the transcript.	*see 15.2*

Unusual Pronunciation	Slang, dialect, contractions, assimilations or mispronunciations. May either be marked in orthographic or phonetic transcript. It is important to keep the 'correct' orthographic form to allow lexical mapping.	`no <!1 nope> going to <!2 gonna>`
Repetition	Stutter of parts or complete items.	`like +/to/+ to see`
False start	Breaking off an utterance and starting a new one.	`-/this evening/- tomorrow`
Breathing	Clearly audible breathing	`<A>`
Filled Pauses	Pauses filled with vocalization or nasalization or a combination or other articulatory noises with the same intention	`<uh> <hm> <uhm> <hes>`
Empty Pause	Temporary unfilled gaps in the speech. Usually not marked at the beginning or the end of a recording.	`<P>`
Articulatory Noise	Noise produced by the articulatory system of a speaker but no filled pause.	`<noise> <cough> <laugh> <smack> ...`
Other noise	Noise caused by background events, by touching the microphone, by the recording equipment etc.	`<#> <#microtouch> <#knock> <#hum> ...`
Cross talk	Overlapping speech caused by speakers interrupting each other. If information is needed about who is interrupting whom and where, this can be rather complicated.	*see 15.2*
Overlay	Overlay of noise or crosstalk may be marked by using a bracket system. Recommendation: tag each overlayed word individually.	`here <:<#> you:> are`
Prosody	Prosodic events like emphasis, main and secondary phrase accent and boundaries may be marked up in the transcript.	`[PA] [NA] [B3] [B9]`

8.3.3 Transcription Example

```
w253_hfd_001_AEW: hallo [PA] [B3 fall] . <#> <"ahm> [B2] ich wollt'
 fragen [NA] [B2] , was heute abend [NA] im Fernsehen [PA] kommt [B3
 fall] .

w253_hfw_002_SMA: hallo . <P> <#> was kann ich f"ur Sie tun ?

w253_hfd_003_AEW: <"ah> [B2] ich w"urde ganz gern [NA] das
 Abendprogramm [PA] wissen [B3 fall] .

w253_hfw_004_SMA: wenn ich Ihnen einen Tip geben darf , <P> <#> heute
 kommt ~Der+Bulle+von+T"olz auf ~Sat-Eins um #zwanzig Uhr #f"unfzehn .

w253_hfd_005_AEW: -/und wa=/- [B9] <"ah> [NA] [B2] gibt es heute [NA]
 abend eine *Sportshow [PA] [B3 cont] ? <P> zum Beispiel [NA] Fu"sball
 [PA] [B3 rise] ?
```

8.3.4 Transcription Method

As anyone can imagine a transcriber can make many mistakes, especially in complex transcription formats. To simplify the transcription process and to end up with a formally correct transcription some measures have to be taken:

- Train the transcribers extensively and test their performance from time to time on bench-mark examples. Use a fixed manual for the transcription and stick to it throughout the work on the corpus.
- Use a text editor that allows 'hot keys' for marker strings and a simple online parse of the input and that displays the various marker types in different colors; for instance use `xemacs` or `WWWTranscribe` (see next section).
- Use a simple re-play tool to allow the transcriber to listen to the sound channels quickly and easily. In longer recording files (more than 5 sec) the tool should allow you to mark and re-play parts of the signal as with a sound editor. Be sure to use a tool that is not capable of modifying the signal, or protect your signal files by setting their rights to read-only.
- In complex transcriptions use a structured process:
  - On the first level produce a simple transcript with only the lexical items together with their immediate markers (*numbers, names,*

spellings, neologisms, foreign words, hard to identify, truncations). This *base transcription* may also be used for a first rough usage of the recorded speech data, in cases where a partner or client is not willing to wait for the final transcription.

  - On the second level let a different group of labelers add the more complex markers (*off-talk, lengthening, interruptions, comments on pronunciation, repetition/correction, false starts, breathing, filled and empty pauses, noises, crosstalk, superpositions, prosody*).
  - Finally pass the transcriptions through a correction level where all data are reviewed by a small number of experts (preferably by one person only).

- Use technical verification techniques after the final or after each processing level:
  - Extract lexical items and compare them to a 'valid word list' to detect typos or inconsistent spellings.
  - Run the data through a formal parser to detect syntactical errors.

8.3.5 Existing Transcription Formats

The design of an individual transcription format, the training of a transcriber group and the setup of the transcription process takes a lot of time and effort. Therefore you might also consider using an already existing format:

- SpeechDat : see [15][9]
- Verbmobil :
  www.bas.uni-muenchen.de/
  Forschung/Verbmobil/VMTrlex2d.html
- SmartKom :
  www.bas.uni-muenchen.de/
  Forschung/SmartKom/Konengl/engltrans/engltrans.html
- MATE : Deliverable 2.1
  www.ims.uni-stuttgart.de/projekte/mate/mdag/

or even to out-source the whole transcription process to an institution that is specialized in that task. The latter is very advisable if you are working on a foreign language and do not have any native speakers on your staff.

[9] www.speechdat.org/speechdat/deliverables/public/SD132V24.PDF

8.3.6 Transcription Tools

We cannot give here an exhaustive list of available software tools that allow different types of transcripts to be done. In general we can say that for the transcript you don't need any special hard- or software other than what is available on most personal computers.

Depending on the signal format used in the corpus you will need a simple tool to replay parts of the signal, optimally in conjunction with a rough signal display where the transcriber may mark parts of the signal and listen to them via headphones. If you are working on multi-channel transcripts (for instance the dialogue between two persons recorded with two separate close microphones), it will help to identify overlapping speech by displaying both signals synchronously on the screen.

To enter the text use a simple text editor that does not perform any automatic formatting functions. As mentioned above, in complex transcription schemes you may consider using an editor with built-in parsing and/or high-lighting capabilities e.g. `xemacs`). You may speed up the transcription process by using hot-key functions to insert marker strings, to automatically transform digits and numbers into strings etc. To reduce the number of typos in the transcript the editor might use a built-in spelling test that keeps an adaptable list of words.

Please also refer to section 8.6.1 at the end of this chapter for the WWW-Transcribe tool that might be useful if you intend to work with a distributed transcriber group over the Internet.

8.4 Tagging

Tagging refers to the markup of categorical classes on the words or larger chunks of the speech signal. Tagging does not require a direct relation to the physical time scale, but usually its labels or tags refer to the transcript. Examples:

- A parts-of-speech (POS) tagging assigns a syntactic category to each word of the transcript.
- A dialog act labeling assigns a dialog act class from a closed set to groups of words from the transcript.
- A tree-bank tagging assigns the leaves of a syntactic tree to the consecutive chain of words in the transcript.

The relation of the tags to the words or larger chunks may either be expressed by repeating the transcript in the tagging or by giving pointers

(usually word numbers) to the transcript. The latter method has the advantage that typos or other errors in the transcript may be pruned without affecting the tagging given that the order of words in the transcript remains the same.

For example in the BAS Partitur Format (BPF)[10] the transcript and dialog act labeling of a dialog turn could be represented as follows:

```
ORT:  0  good
ORT:  1  morning
ORT:  2  have
ORT:  3  we
ORT:  4  met
ORT:  5  before
DIA:  0,1  GREETING_AB
DIA:  2,3,4,5   QUERY_AB
```

As you can see the transcript assigns a unique number to every word which than may be used in different tagging (and segmentations) as a pointer to words.

Taggings are produced manually or automatically. In case of manual tagging the same measures have to be taken as in the case of complex transcripts to ensure consistent and reproducible results (see section 8.3.4).

8.5 Segmentation and Labeling

In contrast to the transcript or other tagging that do not directly refer to the speech signal via the physical time scale a segmentation always contains a combination of time information and categorical content. We distinguish here between *segments* vs. *points-in-time* as well as between *manual* vs. *automatic* segmentation and labeling

8.5.1 Segments vs. Points-in-Time

Speech events may either cover a certain time span, a segment, or happen at a certain point in time. Segmental events are for instance: phonetic features (for instance *voiced*), phones, syllables, morphs, words, dialog acts, dialogue turns, while events that have only a single point in time might be: glottal pulses, bursts, energy peaks or valleys, fundamental frequency peaks or lows, voice onsets, accents, syllable nuclei.

In most speech corpora you will encounter segmentations in turns, dialog acts or words, on a much smaller scale also segmentations in phones and

[10] www.bas.uni-muenchen.de/Bas/BasFormatseng.html

prosodic categories. As a rule of thumb we can say that the effort for segmentation and labeling increases dramatically and inversely proportional to the size of the labeled units[11]

8.5.2 Manual Segmentation

> "Manual segmentation refers to the process whereby an expert transcriber segments and labels a speech file by hand, referring only to the spectrogram and/or waveform. [...] The manual method is believed to be more accurate. Also, the use of a human transcriber ensures that the segment boundaries and labels (at least at the narrow phonetic level) are perceptually valid. However, there is a need for explicit segmentation criteria to ensure both inter- and intra-transcriber consistency, together with (ideally) some form of checking procedure. Sets of guidelines for manual segmentation have been developed by various projects. One such is Hieronymus et al. (1990), which uses the four levels of underlying phonemic, broad phonetic, narrow phonetic and acoustic. It also retains the same base phonemic symbol even at the acoustic level, to facilitate the automatic determination of boundaries at the phonetic level once the boundaries at the acoustic level have been determined. One should not expect more than 90% agreement between experts." *(From [2], p. 152.)*

The basic principles that were listed in section 8.3.4 apply also for the manual segmentation and labeling. You should focus on a consistent training of the segmenters and labelers to maximize inter-labeler agreement (see also section 8.7.

8.5.3 Automatic and Semi-automatic Segmentation

> "Automatic segmentation refers to the process whereby segment boundaries are assigned automatically by a program. This will probably be an HMM-based speech recognizer that has been given the correct symbol string as input. The output boundaries may not be entirely accurate, especially if the training data was sparse. Semi-automatic segmentation refers to the process whereby this automatic segmentation is followed by manual

[11] For instance in the Verbmobil projects we found that the time to segment a dialogue into turns my be achieved in 5 times real-time while a phonemic segmentation required 800 times real-time.

> checking and editing of the segment boundaries.
> This form of segmenting is motivated by the need to segment very large databases for the purpose of training ever more comprehensive recognizers. Manual segmentation is extremely costly in time and effort, and automatic segmentation, if sufficiently accurate, could provide a shortcut. However, it is still necessary for the researcher to derive the correct symbol string to input to the autosegmenter. This may be derived automatically from an orthographic transcription, in which case it will not always correspond to the particular utterance unless manually checked and edited. The amount of inaccuracy that is acceptable will depend on the uses to which the database is to be put, and its overall size." *(From [2], p. 153.)*

At the time of writing[12] there are only a few fully automatic methods known that yield usable results. These are

- Segmentation into words, if the word chain is known and the speech is not very spontaneous[13].
- Markup of prosodic events according to a reduced Tobi set[14]
- Time-alignment of a chain of phonemes using Hidden Markov Modeling[15].
- Segmentation and labeling into phonemic units by MAUS[16] requiring the word chain and a statistical rule set about pronunciation.
- The 'elitist approach' developed by Steve Greenberg. Yields a stream of articulatory features that may be combined into phoneme categories[17].

All these automatic methods do not achieve the same performance as a human segmenter and labeler. However, for some applications and investigations they might be sufficient. Lately automatic segmentation into phonemic units as well as automatic prosodic tagging became rather important in

[12] Oct 2002

[13] For instance the XWaves Aligner or the HTK package, a public domain software package developed by the University of Cambridge, htk.eng.cam.ac.uk/

[14] See for instance work that has been done at the IMS Stuttgart, www.ims.uni-stuttgart.de.

[15] Again XWaves Aligner or HTK.

[16] See for instance [5] or www.bas.uni-muenchen.de/Forschung/Verbmobil/VM14.1eng.html

[17] See www.icsi.berkely.edu/~steveng

the field of speech synthesis by unit selection, because this method requires large quantities of reliably segmented and labeled speech units from one speaker.

8.5.4 Annotation Methods

The scope of this cookbook does not allow us to cover all possible segmentation and labeling procedures. The development of a segmentation scheme requires probably even more time and effort than the transcription scheme discussed in the previous sections. We therefore strongly recommend that you select an already existing scheme and follow the recommendations given there. The following (incomplete) list of projects that involved segmentation and labeling schemes might give you some directions:

- *Kiel Corpus of Read/Spontaneous Speech*
  In this project a moderate amount of read speech from the PhonDat corpus and non-prompted speech from the Verbmobil I corpus[18] was segmented and labeled into phonetic/phonemic units together with a selection of prosodic markers. The formats used in this project are called `S1` and `S2` and were developed in the *PhonDat* projects. Since this format is very hard to parse, we do not recommend using this format for segmentation and labeling.
  See *www.ipds.uni-kiel.de/forschung/kielcorpus.en.html* for details.

- *BAS Verbmobil I*
  In the Verbmobil I corpus distributed by BAS several segmentations and labellings are contained: phonetic/phonemic manual segmentation, phonetic/phonemic automatic segmentation using the MAUS method, prosodic segmentation and labeling in GTobi, word segmentation. See *www.bas.uni-muenchen.de/Bas/BasKorporaeng.html#VMI* for details.

- *Segmentation of the Switch-board Corpus*
  Parts of the Switch-board Corpus[19] have been segmented and labeled into syllable units by Steve Greenberg and his group. For details see *www.icsi.berkeley.edu/real/stp/*

[18] www.bas.uni-muenchen.de/Bas/BasKorporaeng.html
[19] A corpus of telephone dialogue recordings available at LDC, www.ldc.upenn.edu

8.6 Manual Annotation Tools

Since transcriptions, tagging and segmentation of speech data tend to be rather idiosyncratic, there are not many general purpose annotation tools available. In many projects special tools have been developed, or existing and publicly available tools have been adapted for special needs. Also, the production of speech corpora is a rather infrequent enterprise and therefore only very few commercial tools are developed and distributed in this area.

For these reasons in this section we will only describe two general purpose tools that are publicly available and may be of interest for the producer of a speech corpus involving manual annotation: `WWWTranscribe` (mainly used for transcription of large speech data bases) and `Praat` (often used for segmentation and labeling)

8.6.1 WWWTranscribe

WWWTranscribe is a tool for the annotation of audio signals via the WWW. It features an oscillogram display of the speech signal, audio output, editing buttons that simplify the task of annotating the signal, and a formal consistency checker for the annotations. WWWTranscribe was developed at the Bavarian Archive for Speech Signals (BAS)[20] within the SpeechDat project. Currently[21], it supports orthographic transcriptions according to the SpeechDat guidelines; other annotation systems can be added simply by extending the annotation object class hierarchy.
WWWTranscribe is implemented in Java using only the standard JDK classes to guarantee platform independence.

In WWWTranscribe, the transcriber logs in and enters the ID of the session to be transcribed. A session consists of a number of recordings, each containing a single utterance corresponding to a prompt in the interview. Once a recording is selected, the transcription page is displayed. It contains a single output button with a speaker icon, a signal display, transcription and comment text fields, an assessment menu, and save and clear buttons (see figure 8.1). A click on the speaker button outputs the speech signal as sound. For read items, the original text of the prompt sheet is displayed in the transcription field, for spontaneous speech this field is initially empty. Any text in the transcription field can be edited. The buttons below the transcription field perform some basic conversion tasks on the text in the transcription field, e.g.:

[20] Contact the author Chr. Draxler, draxler@bas.uni-muenchen.de, for more information regarding WWWTranscribe.
[21] Oct 2002.

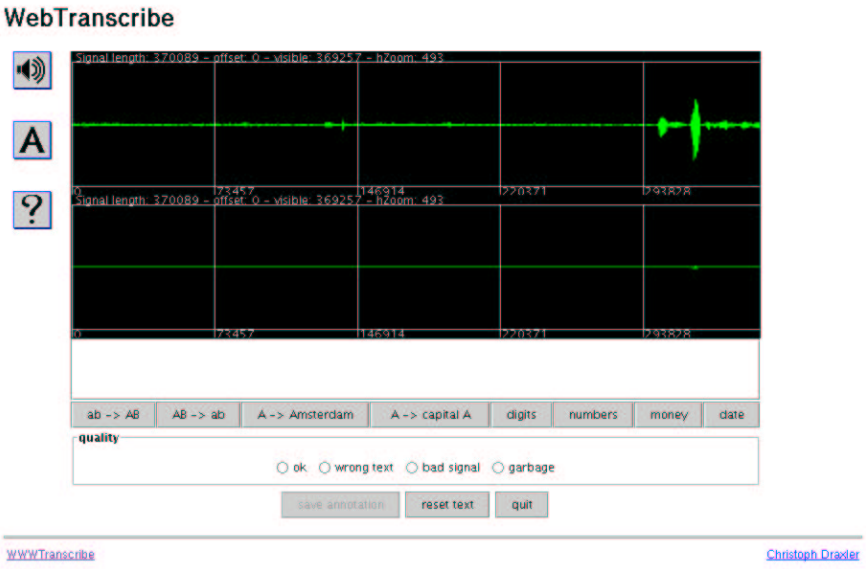


Figure 8.1: Transcription page of WWWTranscribe

- text to lower or upper case
- digit sequences to orthographic digit or number strings
- money amounts and date expressions to orthographic strings

The assessment popup-menu allows the transcriber to select general noise markers. Comments on the recording, e.g. on the quality of the speech or the signal, may be entered into the comment field. The save button saves the transcription to the file system at the server site in the SpeechDat SAM database exchange format.

WWWTranscribe performs an automatic consistency check on the annotation text so that only formally valid annotations are entered into the annotation database.

At the BAS WWWTranscribe has been successfully used for a wide range of transcription, tagging, validation and evaluation tasks. WWWTranscribe is currently being packaged for public distribution[22].

[22]See www.bas.uni-muenchen.de/Forschung/BITS for updated information about the

8.6.2 Praat

Praat is a widely used general purpose tool to analyze and manipulate digital speech data. It was (and still is) developed by Paul Boersma and David Weenink at the Institute of Phonetic Sciences, University of Amsterdam[23]. Although the first aim of Praat was to give students and scientists of Phonetics a handy tool for manipulating speech data and for creating stimuli for perception experiments, the Praat tool very quickly evolved into a general purpose speech tool that may be used for segmentation and labeling as discussed above.

Features

Praat allows multi-tier labeling, labeling of synchronized multi-channel signals, it may be used to label segments or points in time, and it contains a vast number of analysis tools and algorithms that may be of interest in your special case.

Praat also contains a script language, numerical tools for optimization, manipulation tools, statistics, a graphical processor to render results into a graphical format, a speech synthesis component and learning algorithms.

For a detailed and up-to-date description of Praat as well as to contact the authors please refer to www.praat.org.

Segmentation and Labeling

Praat is a tool, not a labeling system. It allows you to define as many labeling tiers as you want but the content of these layers is up to you (see section 'Annotation Methods' above). Praat reads most standard speech signal files[24] and write the results of the segmentation and labeling into a proprietary format called 'TextGrid' which is fortunately a readable plain text file.

Usability

Praat runs on most computer platforms[25] and is free available for educational and non-commercial usage. Contact the authors, if you intend to use Praat in a company or commercial project.

---

availability of WWWTranscribe.

[23] www.praat.org/

[24] including the NIST SPHERE format, which is rather seldom

[25] We have tested Windows, Linux and Macintosh.

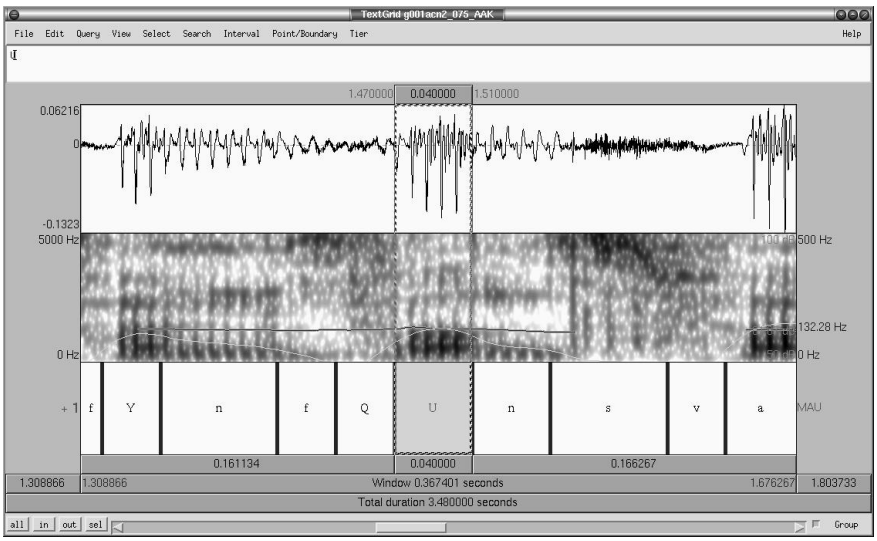


Figure 8.2: Annotation window in Praat

We have used Praat successfully since 2001 in many scientific and speech corpus projects. Praat is also used in the large Dutch and Flemish speech corpus collection CGN as the official tagging and segmentation tool[26].

8.7 Internal Validation

As already mentioned transcriptions and other annotations should undergo a final correction step, preferably performed by one single person, to ensure a good and consistent quality of the annotation. In this correction step all errors found should be logged into a file and used to improve the training of the labeler group. Also, this logging may be used to measure improvements in the labeling procedure.

As a further way to evaluate the quality of your annotation teams you might measure the inter-labeler agreement of the final results. In most cases this can be done automatically by using some automatic alignment method as being used in automatic speech recognition to compare the results of a recognizer with a reference transcript.[27] To estimate the inter-labeler

[26] See lands.let.kun.nl/cgn/ehome.htm

[27] For instance the tool HResults from the HTK package htk.eng.cam.ac.uk

agreement you need a representative sample of data to be annotated by two or preferably more labelers independently. Then the results of these labelers or labeler groups are matched against each other and the average coverage between groups or labelers is calculated.

You will find that the inter-labeler (and intra-labeler) agreement gets worse with the decrease of segment length. That is, the segmentation of dialog acts will be much more consistent than the segmentation of phonetic units. If you are planning to measure the inter- or intra-labeler agreements of segmentations, you'll have to evaluate the labels and the time information independently. However, they are dependent in a way, because, if a label is missing, the time information of the adjacent segments will be distorted. There exists no widely accepted measure for inter- or intra-labeler agreement. You may find some hints in the PhD thesis of A. Kipp ([6]).

Typical values for inter-labeler agreements are

- Orthographic transcript, read speech: 99%
- Orthographic transcript, spontaneous speech: 97%
- Syllable labeling, spontaneous telephone speech: 80%
- Phonemic labeling, spontaneous studio speech: 94%
- Phonemic boundaries within a window of +/- 20 msec, read speech: 95%
- Phonemic boundaries within a window of +/- 20 msec, spontaneous speech: 85%
- Prosodic tagging of accents and boundaries, spontaneous studio speech: 66%

Check List Annotation

◯ Select/define annotations * (p. 101)
◯ Integrate annotations into the data pipeline * (p. 87)
Always produce some kind of orthographic transcription:
◯ Define/select the orthographic transcription * (pp. 103, 108)
◯ Set up the transcription rules/method * (p. 107)
◯ Define the delivery format of the transcript * (p. 108)
◯ Choose/program the tools for transcription * (pp. 114, 109)
◯ Train the group of transcribers *
◯ Set up check procedures for the transcription * (p. 107)
◯ Test for inter-transcriber agreement *** (p. 117)

For each other annotation type, tagging (p. 109) or segmentation (p. 110):
◯ Define the annotation contents and rules *
◯ Define the delivery format of the annotation *
◯ Choose/program/test the tools for annotation *
◯ Train the labelers *
◯ Set up check procedures *
◯ Test for inter-labeler agreement *** (p. 117)

Chapter 9

Pronunciation Dictionary

The pronunciation dictionary lists the most likely pronunciation or citation form of all words that are contained in the speech corpus. A pronunciation dictionary cannot be considered as a 'must' but the usability of the corpus – especially for ASR – will increase considerable if you provide this information together with the corpus. As in other parts of this cook book the choice of how to encode the pronunciation of your corpus can range from very simple and achievable with automatic procedures to very complex and time-consuming. In any case be prepared to reserve some man power for the creation of a dictionary, because in most cases there is manual work involved.

9.1 File Format

There is no widely accepted standard format to code pronunciation dictionaries, although such standards are emerging right now. However, in most cases dictionaries formats are an 'over-kill' for the simple purpose of pronunciation because they are more concerned with other linguistic data, e.g. syntactic, morphologic, semantic.

If you've decided to include a dictionary in your corpus and have not already specified a file format for it, please refer to section 4.7.4 for some basic hints.

9.2 Pronunciation Encoding

As mentioned earlier there are a number of coding schemes for phonetic or phonological units available. Probably the most universal and widely spread coding is the SAM Phonetic Alphabet (SAMPA and XSAMPA) as defined by Wells[1]. SAMPA and XSAMPA provide phonological codings in 7 Bit ASCII for a large number of European and other languages.

Apart from the actual coding scheme, you have to decide about the contents as well. The minimum content as described in section 4.7.4 will be a simple table containing a consistent orthographic representation and a *most likely* or *canonical* pronunciation. Since the latter is not a well defined term for most languages, please make sure that you come up with a definition that may successfully be used in the creation of the dictionary. In some cases you may refer to a standard dictionary[2] or even better to a standardized rule set of pronunciation[3]. If this is not possible, provide a minimal rule set for problematic cases to be used by your staff during the work on pronunciation. Include this rule set into your documentation of the corpus.

If you are working on a German speech corpus, you may use the BAS rule set for manual transcription as given in appendix C.

9.3 Lexical Encoding

The same is basically true for the orthographic representation: in some languages there exists no standardized form for spelling[4]. The orthographic representation forms the lexical access to the dictionary and is therefore quite important for the usability. Try to be pragmatic about it, but at the same time ensure that the used spellings are consistent throughout the dictionary. Make sure that the spellings you are using in the dictionary match those being used in your annotation or transcription files[5]. If you're using a standard dictionary, name it and the edition you're using in the documentation.

---

[1] www.phon.ucl.ac.uk/home/sampa/home.htm or citeEagles1997, Part IV, C.

[2] For instance for German the official 'Ausspracheduden'; for American English the 'Webster'; for British English the 'Oxford Dictionary'.

[3] For instance for Spanish where there is a formal relationship between orthographic and phonemic form.

[4] However, for most languages there exist at least one dictionary that is widely accepted to be a standard. For instance in German the official 'Duden'; for American English the 'Webster'; for British English the 'Oxford Dictionary'.

[5] This includes the consistent usage of special characters like the German Umlauts.

9.4 Additional Contents

You may add more information to your dictionary such as:

- word count in the corpus or parts of the corpus (especially in parts that contain conversational speech).
- other likely pronunciation variants aside from the canonical pronunciation.
- pronunciation variants found in your corpus together with their respective word count[6] based on a phonetic segmentation.
- Syllable/morph/compound word boundaries.
- Syntactic information.
- Primary and secondary word accents.
- Additional inflections of word forms, e.g. if the word 'goes' is in your corpus, you might extend the dictionary by the forms 'go', 'went' and 'gone'.

Do not forget to describe these contents accordingly in the final corpus documentation. Make sure that these contents are parsable.

9.5 Examples

9.5.1 Simple List – Verbmobil

The following example is taken from the German part of the Verbmobil pronunciation list that was created together with the speech corpus. Note that it contains no blanks and the separation between orthographic and phonemic column is a single tab sign. The orthographic representation is encoded in LaTeX; the pronunciation in an extended German SAM-PA alphabet. Primary word stress is marked with a ' before the vowel or diphthong that carries the stress; secondary stress is marked with a ". Compound boundaries are marked with a #. Stress is only marked in words with more than one syllable containing vowels other than schwa (schwa syllables are never stressed).

[6] Sometime referred to as *empirical pronunciation variants*.

```
"Ubernachtungen  Qy:b6n'axtUN@n
"Ubernachtungskosten     Qy:b6n'axtUNs#k"Ost@n
"Ubernachtungsm"oglichk  Qy:b6n'axtUNs#m"2:klICk
"Ubernachtungsm"oglichkeit      Qy:b6n'axtUNs#m"2:klICkaIt
"Ubernachtungsm"oglichkeiten    Qy:b6n'axtUNs#m"2:klICkaIt@n
"Uberschneidung  Qy:b6Sn'aIdUN
"Uberschneidungen        Qy:b6Sn'aIdUN@n
"Uberschneidungsprobleme        Qy:b6Sn'aIdUNs#pro:bl"e:m@
"Uberschneidungspunkt    Qy:b6Sn'aIdUNs#p"UNkt
"Uberschneidungstermine  Qy:b6Sn'aIdUNs#tE6m"i:n@
"Ubersetzungshilfe       Qy:b6z'EtsUNs#h"Ilf@
```

9.5.2 Simple List – The HTK Standard

In HTK[7] a pronunciation dictionary contains one entry in each line. The first column contains the word ID (usually a standard orthographic spelling) followed by an optional number denoting an a-posteriori probability that this pronunciation variant occurs given the fact that the word has occurred. In the remainder of the line the linguistic units are listed (separated by white space) that code the corresponding pronunciation of the word entry. As indicated by the possible a-posteriori probability a HTK dictionary may contain more than one entry per word ID. If no a-posteriori probabilities are given, these variants are considered to be equally probable; otherwise the given probabilities should be sum up to 1 for all entries belonging to the same word ID.

Note that this standard is based on a ASR system and does not define the orthographic nor the phonemic coding scheme. Therefore it is not sufficiant to say in the documentation that the dictionary is in the HTK format; you have to document your coding schemes as well.

```
going   0.856   g @ U I N
going   0.144   g @ U I n
```

9.5.3 Enriched Dictionary – PHONOLEX

The PHONOLEX dictionary is not part of a speech corpus but a general purpose pronunciation list to be used for German speech applications. The PHONOLEX format allows several dictionary sources to be combined into a common format by adding a source marker to each entry. Consequently,

[7] HTK = Hidden Markov Model Toolkit : a public domain software package developed by the University of Cambridge, htk.eng.cam.ac.uk/

there may be more than one entry for the same lexical encoding with different contents depending where the entry stems from.

PHONOLEX is basically a list of canonical pronunciations (most of them in German SAM-PA) but entries may be extended by syntactic, semantic or other markers as well as by lists of empirically found pronunciation variants together with their source and word count.

In the following example the first entry stems from a list of fully inflected words created by the University of Saarbrücken (OR marker), the syntactic word class is `nomen` (CL marker) and the pronunciation was created automatically using the text-to-phoneme system `P-TRA` (TP marker). This entry contains no empirical pronunciations because it is based on a generative linguistic system.

The remaining three entries all stem from the Verbmobil project (OR marker) and have therefore no syntactic word class and the canonical pronunciation has been created manually (TP marker). Empirical pronunciations found in the corpus using the automatic segmentation method `MAUS` are listed below the canonical pronunciation.

```
"Ubernachtungsgeldes
CL:nom  OR:sb   TP:ptra
Qy:b6naxtUNsgEld@s
*
"Ubernachtungskosten
OR:vm   TP:manu
Qy:b6n'axtUNs#k"Ost@n
y:b6naxtUNskOst@n       1       VM      MAUS
y:b6naxtUNskOsn 1       VM      MAUS
*
"Ubernachtungsm"oglichk
OR:vm   TP:manu
Qy:b6n'axtUNs#m"2:klICk
y:b6naxtUsm2:klICk      1       VM      MAUS
*
"Ubernachtungsm"oglichkeit
OR:vm   TP:manu
Qy:b6n'axtUNs#m"2:klICkaIt
y:b6naxtUNsm2:klICkaIt  1       VM      MAUS
y:b6naxtUNsm2:kICkaIt   1       VM      MAUS
*
```

Check List Pronunciation Dictionary

To create the dictionary you will most likely proceed through parts of the following procedures (depending on what resources you have):
◯ Define the orthographic representation for your corpus and transliterate your data or render your text material accordingly *
◯ Create a complete list of unique words. Watch out for capital letters at the beginning of sentences[8] *
◯ Define the desired contents of each entry in your dictionary *
◯ Use automatic procedures to create as much content as possible such as: look-up existing dictionaries, text-to-phoneme converters, part-of-speech taggers, etc. (pass 1) **
◯ Verify the contents of pass 1 and/or create information manually from scratch and produce a corrected version of the dictionary (pass 2) *
◯ If possible, let this be done by one person for the complete dictionary **
◯ Repeat the last step by a second person for the complete dictionary (pass 3) **
◯ Automatically find the differences between pass 1 and pass 2 or between pass 1 and pass 3 where pass 2 and pass 3 are not consistent and discuss these inconsistencies with a group of experts to come up with the final version of the dictionary **
◯ Repeat the last four steps for all content types that need manual labeling/verification *
◯ Use a simple parser to ensure a proper coding of the final dictionary. Especially look out for inconsistent usage of blanks and tab signs. You may also check for homophones and homographs and check whether they are really valid for your language.

Sources for existing pronunciation dictionaries may be the ELDA[9], the LDC[10] or the BAS[11].

[8] A proper transliteration should not contain any of these!
[9] www.icp.grenet.fr/ELRA/home.html
[10] www.ldc.upenn.edu
[11] www.bas.uni-muenchen.de/Bas

Chapter 10

Documentation

The documentation of a speech corpus summarizes all relevant information regarding the production and the intended usage of the corpus. It does not contain meta data or any kind of symbolic data directly related to the speech signals (annotations). The documentation consists of descriptive text (preferably in English), figures and optionally pictures.

However, the distinction between documentation and meta data in the above definition is often a fuzzy one: in many speech corpora data that are essentially meta data can be found in the corpus documentation and for a simple reason: In most cases these 'meta data' are constant for the entire speech corpus and therefore not listed in every speaker profile or recording protocol. Furthermore, some authors define meta data in a much broader way than it is usually done in practice: For instance they also include parameters that describe the speech corpus (usually given in the corpus specifications) such as number of speakers, number of recorded items, technical specifications etc. To include all these data into the meta data set makes sense but only if there are standardized ways to access these data. Since these techniques are emerging just now, in this cookbook we follow the traditional way and include these data under the label 'documentation'.

We have mentioned documentation two times earlier in this cookbook. In the chapter Corpus Specification we listed it as a possible item to be specified beforehand, mainly in larger projects with many producing partners (see section 4.11, p. 64). In the chapter Collection we gave a few hints about what and how to log relevant information during the collection process (see section 6.1, p. 83). In this chapter we will merely give an overview of what we deem to be essential parts of any speech corpus documentation. As usual this will not be an extensive listing because we cannot foresee the special

needs of future speech corpus productions.

In summary the corpus documentation consists of the following parts:

- Introduction (= executive summary)
- Copyrights and disclaimers
- Version number and date
- List of documentation files
- Corpus description
  - Numbers (speakers, recording etc.)
  - Structure
  - Contents
  - Terminology (file naming)
  - Technical specifications of signal files
  - Other parts of the corpus: dictionary, translations etc.
- Recruitment
  - Speaker profiles
  - Recruitment technique
  - Legal aspects
- Recording
  - Setup
  - Script
  - Technique
  - Log file
- Post-processing
- Annotation
  - Contents
  - Procedure
  - File formats
- Meta data

  - Contents
  - File formats

- Spoken texts, prompt files
- Original corpus specification
- Validation reports
- Publications, internal reports
- Comments
- Corpus history
- Known errors

In the following sections we will document each of the above listed items.

10.1 Starting Document

Usually the documentation consists of a 'starting document' that gives a first overview of the corpus and all documentation files and points to other documents in the distribution or even to web sites.

Documentation files may be plain text files, Postscript, PDF or even be maintained on the Web[1] but there should always be a copy of all Web files in the documentation on the distribution medium. Place all documentation files in a common directory called `DOC` or in subdirectories of this directory.

For example: most of the BAS speech corpora contain annotation files in the BAS Partitur Format (BPF). The BPF is documented on a Web page[2], since there are frequent changes when new tiers are added to the format. The documentation of the BAS speech corpora therefore contains an URL of this documentation but also contains a copy of the Web documentation at the time of media production.

The 'starting document' should be named `README` or `REPORT` and contain at least the Introduction, the Copyright, the Version Number and Edition Date, the List of Documentation Files and the Corpus Description:

The **Introduction** describes the main features and the intended usage of the corpus in one paragraph. This information may later be used in catalogues etc. For example:

[1] Avoid formats that require non public domain software to read, such as Word, StarOffice, etc.

[2] www.bas.uni-muenchen.de/Bas/BasFormatseng.html

```
This is the documentation for the WEBCOMMAND database created
in Jun - Aug 2002 as a subcontract to Siemens Company.
WEBCOMMAND contains recording sessions of 48 native speakers
of France and Great Britain. All speakers read a list of 130
prompts from a screen. They are recorded with two microphones:
a high quality headset and a high quality microphone fixed to
a 'webpad' held on the lap.
```

Clearly state the **Copyright and Disclaimers** immediately after the introduction (see chapter 2). Be sure to make absolutely clear who may use the corpus for what purposes and who is eligible to distribute the data. Then give the **Version Number** of the speech corpus in this distribution, the **Date of edition** and the **Date of the last update**.

The **List of Documentation Files** is simply a commented directory listing of the documentation directory.

The **Corpus Description** should contain all information about the corpus in its present state: Numbers about size, speakers and recordings, distributions about certain speaker characteristics and special recording conditions, contents of the speech recordings, the structure of the corpus on the distribution media, the distribution media itself and the usage of these, the technical specifications of the signal files, file formats, terminology, other parts of the corpus such as dictionary, translation files etc.

The speech corpus may be identified in a more formal way by using standardized entries in the head of the corpus description as in the following example:

- Corpus ID PD1
- Corpus full title PhonDat I : Di-phon balanced speech corpus
- Corpus short title PhonDat I
- Corpus creator University of Bonn, Bochum, Kiel, München
- Corpus version 3.0
- Corpus edition BAS
- Corpus status finished
- Corpus date 01/01/1993
- Corpus update frequency not specified

- Corpus date updated 07/19/2001
- Corpus publisher BAS
- Corpus publisher type University
- Corpus size 3.2 GByte

10.2 The Core Documentation

The documentation of the **Recruitment** describes the common profiles of the speakers as well as the recruiting method that was used. For example it might be interesting to know whether the speakers were paid for their job or not. Were they paid more for a successful job? Were there any other sources of motivation? Also legal aspects might be listed here, e.g. the rights of usage of the data.

The documentation of the **Recording** and the **Post-processing** is basically a repetition of the corresponding part in the corpus specifications with the slight but important difference that here the *real* recording conditions should be described. If there exists a **Log File** of the production, it should be included here. If possible include **pictures** from the recording setup and recording sites. Draw **diagrams** to illustrate the exact positions of speakers and microphones.

The **Annotation** should be documented for each of the used annotation layers in great detail. Not only the mere contents and file formats should be given but also the exact procedures on how the annotations were produced. For manual annotations there must be a copy of the annotation guide lines included here. Education and training of the labelers should be indicated, tools and their usage described.

If you use any automatic procedures, insert a copy of the source code of your scripts or programs here or give proper reference to public domain software and describe exactly how it was used. Describe the methods of quality control that were applied to the annotations; define the character set that is used in the annotation files as well as tag sets, phonetic alphabets etc.

If you are using XML in the annotation files, give pointers to the corresponding DTDs.

The documentation of the **Meta Data** should contain a precise definition of each entry in the meta data files. Give complete lists of the codes you are using and comment on how the data were gathered. For instance, if an entry in the speaker profile files describes the dialectal variety of a language by naming the state or province of a speaker, you should mention

here how this information was obtained: was it from an interview with the speaker (self-assessment), was is by asking for the place of elementary school or was it from a judgment of one or a group of experts about dialects of that language.
If you are using XML in the meta data files, give pointers to the corresponding DTDs.

10.3 Other Documents

If the corpus contains read speech, the **Prompt Texts** must be given in the corpus documentation. This might be a simple list of spoken items or – if every speaker has spoken a different set – a corpus of prompt files for each speaker.

Include the **Original Corpus Specification** in the documentation. It might contain important information for the user that you are not aware of. Also, it might be useful for colleagues that plan to work on similar speech corpora as you do.

The **Validation Reports** – be they external or internal – are an important part of your documentation. They might be the basis for any prospective user of your corpus to decided whether this corpus fullfils his/her requirements.

If there already exist any **Publications** with regard to the speech corpus, ask the authors to include a copy in the documentation, or list the references to them. These publications might give the user valuable insights into how the corpus may be used and what are certain characteristics of the speech data.

Any **Comments** of project partners, funding organization or users might be listed in the documentation as well. Be sure to ask the authors of these comments for their permission.

The **Corpus History** is basically a chronological list of the changes to the corpus after completion. It should name all changes of the version of the speech corpus together with the date and with what was altered in the corpus and where the updated files might be downloaded from.

Finally, since no speech corpus is absolutely error-free – there should be a **List of Known Errors** that have not been and probably will not be fixed for various reasons (for instance when a recording file is corrupt, but the speaker is not available any more so the recording cannot be repeated).

Check List Documentation

'Starting Document'

- ◯ Introduction *
- ◯ Copy rights, disclaimers *
- ◯ Version number and dates *
- ◯ List of documentation files *
- ◯ Corpus description *

Core Documentation

- ◯ Recruitment *
- ◯ Recording *
- ◯ Post-processing *
- ◯ Annotation *
- ◯ Meta data *

Other Documents

- ◯ Spoken texts, prompts *
- ◯ Original corpus specification **
- ◯ Validation reports **
- ◯ Publications, internal reports ***
- ◯ Comments ***
- ◯ Corpus history *
- ◯ Known errors *

Chapter 11

Validation

The term validation has been mentioned throughout this document several times. If you intend to validate a speech corpus that you do not produce yourself (for instance if you act as an external validation institution or if you want to verify the contents of an ordered speech corpus), you should refer to the excellent overview of Henk van den Heuvel ([16]) or to the document 'Validation of Speech Corpora' ([11]) edited by the authors[1].

In this section we give some basic hints about validation as far as they are relevant within a speech corpus production. The main points are:

- internal vs. external validation
- when to validate
- what to validate

11.1 In-house vs. External

In-house validation refers basically to quality control during or after production of the speech corpus and carried out by members of your institution. It is definitely more economical than an external validation which requires quite an effort of time, money and manpower.

However, we do recommend using an external validation whenever possible, because in-house validations tend to be not very effective. The reason for this is similar to the well-known fact that for instance a programmer that is looking for a bug in his code cannot 'see' the error, because he is involved

[1] See www.bas.uni-muenchen.de/Forschung/BITS/TP2/Cookbook/ for a downloadable version of this document.

far too deeply in the process. An external observer however often simply points to a quite obvious and simple solution. That's why programmers tend to blab about their programming problems a lot.

The same is true with errors in a speech corpus production process. Therefore it is vital to perform external validations as often as possible.

Very often there is no funding available for external validations or – even worse – there is nobody to be found who might act as an external validation institution. If the speech corpus is produced for a client or a partner institution, the obvious solution is to make that partner or client act as the external validation. However, if you do that, make sure that in your contract there are precise guidelines to be found on what has to be validated when and how (see the following sections). If you have no client or partner in your constellation there remain a few institutions who might be willing to act as a validation institution for your corpus production: SPEX in Nijmegen, Netherlands[2], BAS in Munich, Germany[3] and University labs that are working on Phonetics and/or Computer Linguistics[4].

11.2 When to validate

When is the best time to validate the results of the speech corpus production? This depends on the size, the time scale and the intended usage of your speech corpus. Smaller corpora productions that take less than 3 months to be finished will require only the pre-validation and the final validation. Larger corpora might have separate validations of individual releases.

11.2.1 Pre-Validation

For details about the pre-validation please also refer to section 6.2, pp. 84. The pre-validation should take place after a small sample of speakers has been recorded, their data have been post-processed and annotated. The collection process should wait for the results of the pre-validation and implement any necessary corrections to the processes before continuing with the recordings (see figure 3.1, pp. 40).

[2] www.spex.nl

[3] www.bas.uni-muenchen.de/Bas

[4] See www.ims.uni-stuttgart.de/phonetic/joerg/worldwide/lingphon.html for some links to such institutions.

11.2.2 Release Validation

Release validations should take place at defined milestones of the speech corpus production. To improve the internal consistency within the individual releases, it is a good idea to wait for the results of a release validation before starting to collect data for the next release. Therefore, the validation times should be included in the overall time schedule of the project. Also your contract should specify how to deal with errors found. If there is, for instance, a systematic error throughout the first three releases, should they be corrected or not? Is funding available for such updates?

11.2.3 Final Validation

The final validation takes place after the speech corpus production is declared to be completed. Take care that the funding structure and the overall schedule allows some funding and time for corrections after the final validation (updates). It is very unlikely that a final validation will come up with no errors at all.

11.3 What to validate

Basically every item included in the specifications of the speech corpus may be subject to validations. What will be validated and what will be regarded as an error has to be included in the contract or the corpus specifications. Typically, the following parts of the speech corpus are validated:

- *Documentation:* consistency, completeness, structure
- *Meta data:* completeness, parsability, contents (samples)
- *Signal data:* completeness, technical quality, acoustical quality (samples), contents (samples)
- *Annotation data:* completeness, parsability, contents (samples)
- *Media:* readability (on different computer platforms)
- *Dictionary:* completeness, quality (samples)

Then there is always the distinction between qualitative and quantitative validation results. A quantitative result might be for instance that more than 5% of the signal data are clipped. Validations of this type will usually be carried out automatically (sometime referred to as 'formal' validation). If the documentation contains a description of the signal file format that is

unclear or inconsistent, this would result in a qualitative validation result. Validations of this type require manual work, and are often carried out only on a randomly selected sample from the corpus.

Finally, there has to be an agreement on what is treated as an error and what is deemed to be within the tolerance measures. For instance, if the specification demands a 50/50% gender distribution throughout the corpus, there also has to be given a tolerance percentage +/- X%[5]

To get a better idea about what parts of a speech corpus are validated with what procedures or measures please refer to [11] or to the SpeechDat example in part III.

11.4 Validation Reports

The results of all validations should be documented in reports and included into the final corpus documentation. It is considered to be a distinctive quality feature of a speech corpus to contain such validation reports.

11.5 Example

The following example is fictitious and will therefore not contain all possible items to be validated. It should merely give you an idea how a specification for corpus validation might look:

The speech corpus for the following validation example consists of un-supervised telephone recordings by 1000 speakers with the orthographic transcript as annotation.

Quantitative (Formal) Validation Procedure:
Check for 134 recording items per speaker.
Check for empty signal files.
Check for signals files with clippings; must always be less than 5%.
Check for S/N; must be more than 15 dB.
Check for correct terminology for all data files according to specs.
Check for one annotation file per signal file.
Check if annotation files are parsable.
Check for complete and parsable speaker profile per speaker.
Check for complete and parsable recording protocol per recording.

[5]Also, in this special example it must be stated in the specifications whether the distribution is with regards to the speaker numbers or with regards to the amount of material recorded by the speakers.

Check for 50/50% gender distribution +/- 5%.
Check for age distribution in two groups 18 - 32 and 32 - 64; both groups have 50% +/- 5%.
Check for parsability and completeness of dictionary.

Qualitative Validation Procedure:
Check documentation for completeness and consistency.
Check 5% randomly selected annotation files by independent manual transliteration and cross check results; 3% word errors (including insertions and deletions) allowed.
Check 10% randomly selected entries from dictionary for correct pronunciation; 2% phonemic errors (including insertions and deletions) are allowed.

Check List Validation

In this check list the processing steps that might not *be obligatory are marked with **.*

◯ Decide between in-house or external * (p. 135)
◯ Schedule pre-validation ** (p. 136)
◯ Schedule release validation ** (p. 136)
◯ Schedule final validation * (p. 136)
◯ Define validation content * (p. 137)
◯ Validation reports into documentation * (p. 138)

Chapter 12

Distribution

The final stage of the speech corpus production is the production of distributable media. The major points to be considered here are:

- Which media to choose and how to produce them
- Compression and Compatibility problems
- Signal data vs. symbolic data
- Safety, verify procedures and version control
- Larger edition vs. Burn-on-Demand
- On-line distribution

12.1 Media Production

The most common and accepted medium still is the CDROM because every computer platform has a drive for that. DVD are getting more popular quickly because the consumer industry causes the prices for drives to drop rapidly.

For speech corpora larger than 5GB it is probably best to press the physical signals on DVDs and the annotations (which are usually much smaller) on a separate CD-ROM.

Burners for CD-R and DVD are reasonably priced and should be included in the budget of the speech corpus production. Choose quality drives to avoid unnecessary drop-outs. Standard computer networks are

fast enough that you may set up a burner station on one host and store the master images on your file server[1].

Alternatively you might out-source the CDROM or DVD production to a company which produces 'real' (that is pressed) CDs or DVDs. However, this might only be economical if you plan at least 100 copies to be produced (see also section 12.5).

Another thing to consider here is the so-called file system (FS) you will install on your media. For CDROM there is basically one widely accepted FS called ISO9660 but it comes with many different extensions. We recommend using the Rockridge extension (for UNIX systems) and the Joliet extension (for MS systems). Both allow you to use longer file names than the basic ISO9669 which is restricted to the old DOS 8.3 convention. Both can be installed in parallel and do not interfere with each other. Please note that the addition of such extensions may increase the total size of your data significantly, if your corpus contains a lot of files.

On DVD you may use ISO9660 as well or preferably the new UDF file system that overcomes most shortcomings of ISO9660: is allows long file names, a larger number of files, etc. Most computer platforms (tested on MS, Macintosh and Linux) detect ISO9660 and UDF automatically.

The use of very many small files (typically your annotation files) with less than 4kB will increase the net size of your data as well, because most FS will reserve a minimum block size per file (typically 4kB, 8kB or 16kB). Take this into consideration when dividing up your corpus over several media volumes. The best way to test the actual size of a volume is to produce an image file and check whether it fits on the medium.

If your corpus exceeds several 100GB you'll probably consider distributing either on special tapes (which makes it harder for the users to access the data) or on inexpensive IDE hard disks. In the latter case the FS on the hard disk should be VFAT that can be mounted by all computer platforms.[2]

If you're using HD drives as distribution media, you might install a swappable IDE drive slot on one of your hosts; this makes it quite easy to change drives without opening the case.

---

[1] 2002: We've had good experiences with the following constellations:
- CD-R 8x, Linux, 100 MBit network
- DVD-R 2x, Macintosh, local
- DVD+RW 2x, NT, Linux server, 100 MBit network

[2] Although VFAT is defined for a maximum size of 125GB we found that older Linux kernels (< 2.4.18) will only handle partitions of a maximum size of 65GB. You can circumvent this problem by adding more than one partition to the hard drive.

12.2 Compression / Compatibility

By using standard compression algorithms you may reduce the total amount of speech data to 55-65% depending on the technical specifications of your speech signals[3]. There also exist special compression algorithms for speech[4], but we found that they do not yield significantly better compression rates than standard algorithms (like `gzip, zip`) when used in 'loss-less' mode[5].

However, we do not recommend using compression at all. Working with uncompressed data directly from the distribution medium is much more convenient, while on the other hand the reduction of costs do not justify the additional effort. Furthermore, by using compression on your distribution media you'll increase the probability of software incompatibilities on the user side.

If you're using a well established standard medium like CD-R, it's very unlikely that you'll run into hardware compatibility problems. On the other hand, large tapes and magneto-optical media may require special hardware to read.

Avoid special hardware whenever possible to avoid trouble with extinct or not supported hardware in the future. The author has seen cases where a valuable speech corpus could not be loaded any more, because it was produced on special DEC magneto-optical disks. The data were actually lost because of that fact.

12.3 Signal / Symbolic Data

As mentioned before it is a good concept to separate signal and symbolic data in the corpus structure and henceforth also in the distribution. In medium-sized or large speech corpora you might not be able to keep all the data online at all times. However, you will very likely need to access the annotation data of the whole corpus. Therefore it makes sense to store these data either on a separate medium or to copy them on to all volumes of the speech corpus.

For example in the SmartKom corpus on every single DVD-R volume you will find the complete set of annotation files for the total corpus. In the WebCommand corpus these data are stored on a separate CDROM[6].

[3] In general high sampling rates tend to contain more redundancy than lower sampling rates and are therefore easier to compress.

[4] For instance Tony Robinsons `shorten`.

[5] Never use a non loss-less compression algorithm on your speech data. Don't even think about it!

[6] The latter has of course the disadvantage that a mixed set of media have to be

Another advantage of keeping the symbolic information separate, is the fact that these data are much more likely subject to updates than the signal files. Since the symbolic data occupy usually less than 1% of the total corpus size, these updates can be easily distributed to users via a download from a FTP server.

12.4 Safety / Verify / Versions

The produced and validated speech corpus is a very valuable resource. Therefore all care should be taken against all kinds of possible data loss. Never rely on one storage medium alone and never keep all storage media in the same location. For instance the archive data at BAS are stored in three (four) different ways: on a file server with an independent backup system, on CD-R or DVD-R shut away in a safe place, and on a Tivoli storage system in another part of town. During and possibly also after the production phase there will be changes to your data. Be sure to set up reliable procedures to distribute all these changes to all of your data locations.

Always use verify procedures to ensure that data transfers were performed successfully, especially when you transfer over the network. Use `diff -r` on UNIX systems to detect differences between your target and source data quickly. Make use of build-in verify procedures in CD-R or DVD-R burner software. To be 100% sure mount the ready medium on a different host and run an additional verify to the source data.

It's recommended to use a version control or at least manually set version numbers on your data that are increased after each update or change. Every speech corpus documentation must contain a change log where all changes are documented together with the corresponding version number. We recommend a two-part version number X.Y where X is increased only after major changes that imply that for instance software which uses the corpus has to be adapted, while Y is increased for error corrections (updates) only.

If possible, set up a mailing list of all users of the speech corpus and inform them about version changes automatically.

12.5 Larger Edition vs. Burn-on-Demand

If you're using CD-R or DVD-R as a distribution medium, there are basically two ways to produce them: the production of a large series of identical copies (traditional edition) or to produce needed copies on demand.

---

distributed.

To produce a larger edition has the advantages that

- the production may be out-sourced
- the production price per copy is lower
- hardware incompatibilities are very unlikely

This method is therefore clearly recommended if you do not intend to maintain and update the corpus on a regular basis.

To burn on demand means that always the latest version of the data will be copied on to the medium. The advantages here are

- you do not need to decide beforehand how many copies to invest in
- updates are possible
- you may easily switch to new media types

For example, the speech corpora at BAS are always distributed in their newest version, because BAS is actively working on most of its speech resources. Using this concept it was possible that for instance the RVG1 corpus which initially consisted of 32 CD-Rs will now be distributed on 5 DVD-Rs as well.

On the other hand the SpeechDat corpora produced by BAS and distributed by ELDA were produced in a larger edition because nobody is responsible for maintaining these corpora any more.

12.6 On-line Distribution

Smaller speech corpora may also be distributed on-line, for instance by a password protected FTP server. Using an appropriate database system it might even be possible to distribute parts or excerpts of a speech corpus. For instance a prospective user might only be interested in the female speech of a large corpus, or even more specificly, only in certain spoken words that might be indexed via a word segmentation of the corpus.

Distribution servers of this kind do already exist for special scientific speech data and are usually free to use. They require a considerable effort to set up and maintain.

For speech resources that are not absolutely freely available there are still many practical and legal problems to solve.

We recommend allowing the free download of the meta data and perhaps also of the annotation data of a speech corpus. Meta data are essential for prospective users to help them decide whether a speech resource meets

their special needs. Annotation data are in most cases of not much use for commercial users without the corresponding signal data, but they might be of academic interest.

Check List Distribution

In this check list the processing steps that might not *be obligatory are marked with **.*

◯ Select media * (p. 141)
◯ Compression? * (p. 143)
◯ Store symbolic data separate ** (p. 143)
◯ Safety/verify procedures * (p. 144)
◯ Print, burn-on-demand or online? * (p. 144)

Part III

Examples

The third part of this cookbook describes the specifications of three prototypical speech corpora: *WebCommand, SpeechDat* and *SmartKom*.

WebCommand is an example for a low-cost small-size corpus production, *SpeechDat* describes the specs of an international and commercial speech corpus production in the field of telephony, and finally *SmartKom* is a good example for a complex scientific corpus collection of multi-modal data including speech data.

	WebCommand	SpeechDat	Smartkom
Content	Commands	Diverse	Dialogue
Language	English/French	13 European	German
Speaker	40	5000	400
Type	Read	Read	Spontaneous
Signal	Online	Telephone	Online
Channels	2	1	9
Environment	Office	Field	Studio
Size	9 GB	30 GB	25 GB
Annotation	SpeechDat	SpeechDat	SK Transliteration

The examples are non-fictitious and by no means meant as role models for an ideal corpus specification. The descriptions were taken from the real corpus contents and missing or badly designed contents are commented on accordingly.

To make the link to the remaining contents of this cookbook easier and to simplify comparisons between the different corpora styles the main description of each corpus is structured in a table more or less according to chapter 4 of this cookbook.

Chapter 13

WebCommand

13.1 Corpus Specification of WebCommand

WebCommand is a speech corpus for the development and validation of speech recognition algorithms for British English and French. The target application is a portable full-size touch screen controlled by voice commands, a so-called 'Web Pad'. This device is intended primarily for communication, i.e. video phone, email and Internet access.

The pre-validation and the final validation have been done by the producer itself, although we recommend asking a third independent institution for both. However, this might be justified because of the relatively small size of the corpus and the very constrained budget of the client.

In the following, the corpus specification of WebCommand will be presented in the manner of a check list. The elements of this check list have already been discussed in this order in chapter 4. If elements are not applicable for WebCommand, they're marked with a 'n.a.'.

Speaker Profiles	Speakers are native speakers of British English or French and at least 18 years old. Gender distribution is 50:50, all dialects allowed, education level not specified
Number of Speakers	At least 40 speakers had to be recorded, 20 for British English and 20 for French. The number of male and female speakers had to be preferably equal in every language.
Contents:	The contents of the corpus were specified by the client in form of a plain text command list. The text corpus was fixed – that is all speakers recorded in one recording room spoke the same corpus of 135 command words. There are in total four text corpora: one for each of the two recording environments (see below) in the languages British English and French.
- Vocabulary	English: 163 words; French: 188 words
- Domain	Control commands and names
- Task	No task specified
- Phonologic Distribution	No distribution specified
Speaking Style:	
- Read Speech	+
- Answering Speech	–
- Command/Control Speech	–
- Non Prompted Speech	–
- Spontaneous Speech	–
- Neutral/Emotional	–
Recording Setup:	On-site Recording
- Acoustical environment	Each speaker is to be recorded on-site in two different recording rooms P and S on different days. The acoustical background consisted only of the hum of the recording device which was a regular Macintosh Desktop PC approx. 50 cm from the head of the speaker. The PCs were rated to be rather silent.
- Script	Speakers read prompts from the CRT display in their native language

- Background noise	no artificial background noise specified
- Microphones	The speaker wears an ear-free headset Beyerdynamik NEM 192; a second Beyerdynamik MCE 10 is mounted on the upper left corner of a dummy laptop case that the user holds with both hands on his/her lap to simulate free speaking.
Technical Specifications:	
- Sampling Rate	22050 Hz
- Sample Type and Width	Sample Type: linear, not compressed.
- Number of Channels	Two channels recording: left channel: Beyerdynamik NEM 192; right channel:Beyerdynamik MCE 10.
- Signal File Format	File format: WAV stereo (RIFF)
- Annotation File Format	SAM annotation files according to SpeechDat specifications and a summarized annotation table for each recording block.
- Meta Data File Format	Table SPEAKER.TBL gives a mapping of 4-digit speaker id to sex, age and mother tongue. Table SESSION.TBL contains a mapping of 4-digit session id to speaker id, place of recording, microphone types, channel mapping, environment. The file SUMMARY.TXT contains the SpeechDat compliant summary of recordings: for each recording session all individual recordings are listed in the line. If a recording is missing, a '-' is listed instead of the three-digit prompt number.
- Lexicon Format	Two-column plain text file: orthography and pronunciation coded in SAM-PA

Corpus Structure: - Structure	Recordings are stored in separate subdirectories for each combination of recording environment and language. The corpus contains 47 complete sessions (130 recordings per session). Care is taken that each speaker is recorded in complete sessions in each of the two recording rooms. Additional incomplete recording sessions are collected in the directories NOT_USED_FR (4 sessions) and NOT_USED_EN (7 sessions) respectively. Signal data are stored on DVD; a separate CDROM contains documentation, annotation files and pronunciation dictionaries.
- Terminology	Session names are coded as `SES#***` where `#` codes the combination of environment and language and `***` encodes the session number, e.g. SES6013 is the 13th recording session of a French speaker in room P. A mapping from speaker IDs to sessions, as well as the speaker profile can be found in the file `SESSION.TBL`. A recording file name is encoded as `Q1#***YYYY.WAV` where `YYYY` denotes the number of the text prompt (000-129) e.g. Q16013051.WAV contains the two microphone signals in a WAV stereo file of the 52nd prompt of the 13th recording session of French speakers in room P. The channel assignment for the microphones is stored in the file `SESSION.TBL`.
- Distribution Media	The corpus consists of two DVD-5 with a total size of 7.5 GByte plus a CD-ROM with the label files and documentation. On one DVD the data of the British speakers are stored; on the second DVD the data of the French speakers.

Release Plan	06.05.02 : Start of project, delivery of the prompts for both languages by ordering company. 01.07.02 : Database British English will be delivered to ordering company. 15.07.02 : Database British English will be delivered to ordering company. The client agrees that the corpus is offered to third parties via the national catalogue of the BAS and the international catalogue of the European Language Resource Association (ELRA) after a blocking period of one year. If the ELDA acts as a broker to deliver the corpus to a third party, ELDA earns a commission of 20% of the agreed royalties. A discount for research and for members of the ELRA is not provided.
Documentation	`REPORT.TXT:` main documentation including copyrights, history and error log (see section 13.4 for a complete listing) `SAMEXPORT.TXT:` summary of annotation `SESSION.TBL:` recording protocol: mapping of 4-digit session id to speaker id, place of recording, date of recording, microphone types, channel mapping, environment `SPEAKER.TBL:` speaker protocol: mapping of 4-digit speaker id to sex, age and mother tongue Documentation of SpeechDat annotation guidelines and format and pictures from the recording setup

13.2 Meta Data of WebCommand

13.2.1 Recording Protocol

In case of WebCommand it would have been too costly to create a separate recording protocol. The recording protocol for WebCommand is reduced to a table with one line per recording; general conditions are part of the documentation. The following list of minimal requirements for the recording

protocol as given in section 3.2 contains a + if the content is given (optionally followed by a citation from the documentation) or a - if the content is not given.

Session ID	+ `SES` + 4 digit number set automatically
Speaker ID	+ 4 digit number set automatically
Date of recording	+
Environmental conditions	+ Room Acoustics: "The acoustical environment of both rooms is quiet office environment." + Sources of Noise/Background Noise: "There is only one computer (Mac desktop mounted in front of the speaker). No other noise sources." + Cross Talk: "No other noise sources."
Technical recording conditions	+ (see section 13.4, part 'Recording situation')
Microphones	+ Head set: Beyerdynamik NEM 192, left channel; Web-pad mic: Beyerdynamik MCE 10, right channel
Recording device	+ The signal of the microphones is amplified by a Beyerdynamik MV 100 amplifier: headset mic + 20 dB, web-pad mic + 20 dB and then connected to the standard Mic input of the recording Mac.
Technical specifications of recorded signals	+ Sampling rate: 22050 Hz; Bits per sample: 16; Length per prompt: 5.7 sec
Placement and distance to microphone(s)	+ The speaker wears a ear-free headset Beyerdynamik NEM 192; the second mic is a Beyerdynamik MCE 10 mounted on the upper left corner of a dummy laptop case that the user holds with both hands on his/her lap.
Name or ID of the recording	+
Details about the recorded domain(s)	n.a.
Details about instruction to speaker(s)	+
Duration of the recording session	+

Type of Prompting (paper, face-to-face, screen, voice)	+ screen
Emotional speech yes/no	+ no
Details about acoustics (reverberation, S/N ratio etc.)	–
Supervisor present	+ no (just for the first 3 training prompts)
Interpreter present	n.a.
WOZ: details about 'virtual machine'	n.a.
Type of speech	+ read
Free comments	–

13.2.2 Speaker Profiles

The speaker profiles were collected on paper forms and then transfered into a web-based data-base. Meta data were then distributed in form of a plain text table.

The following table of recommended meta data about speakers (refer to section 3.3 for details) contains a + if the information was collected and a – if not.

Speaker ID	+
Sex	+
Date of birth	+
Mother tongue	+
Second languages of speaker	–
Mother tongue of parents	–
Second languages of parents	–
Pathologies	–
Dentures	–
Piercings	–
Place of elementary school	+
Dialect region	+
Dialect	–
Level of education	–
Level of proficiency for a certain task	–
Profession	–

Height	–
Weight	–
Left/right handed, ambivalent	–
Smoker/non smoker	–
Stutter	–
Free comments	–

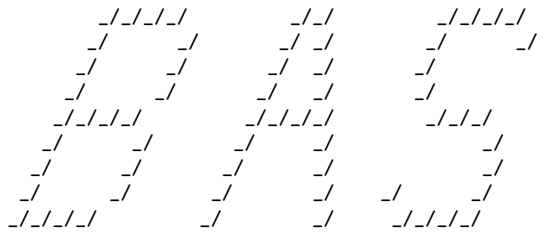
13.3 Comments to WebCommand

Two major errors occurred during the production of the WebCommand speech corpus that caused the production costs to be about 25% higher than estimated. Both errors might have been avoided if the production had adhered strictly to the standards recommended in this cookbook.

The first error was a very reduced in-house pre-validation after the setup of the recording hardware and procedures. Because of this sloppy procedure it went unnoticed that the same prompts were displayed in both recording rooms. According to the specification however, the prompts had to differ. This error was not even detected in the final validation but was discovered by the client *after* the first delivery. It is very likely that a proper pre-validation by an external partner (or the client) might have detected this logistical error much earlier thus saving a considerable amount of manpower and money (see section 6.2 for details about pre-validation).

The second error was of a technical nature. Although the pre-validation was performed on both recording setups in the two recording rooms and showed valid recordings, it turned out in the *final* validation that in one of the recording rooms the input selection was later unintentionally set to another channel and all recordings were in fact empty. A number of speakers had to be contacted again and asked for an additional recording session to fill up the missing data. This error could easily have been detected very early if quality control in form of random tests on the recorded data had been performed (for details about quality control refer to section 6.3).

13.4 WebCommand Documentation



BAVARIAN ARCHIVE FOR SPEECH SIGNALS

University of Munich, Institut of Phonetics
Schellingstr. 3/II, 80799 Munich, Germany
bas@bas.uni-muenchen.de

WEBCOMMAND 1.1 - on-site recordings for webpad voice control

This is the documentation for the WEBCOMMAND database created in
Jun - Aug 2002 as a subcontract to Siemens Company.

WEBCOMMAND contains recording sessions of native speakers of
France and Great Britain. All speakers read a list of 130 prompts from
a screen. They are recorded with two microphones: a high quality headset
and a high quality microphone fixed to a 'webpad' hold on the lap.

------------------- Contents of this file ------------------------

```
   Recording situation
   Naming conventions
   Signal file formats
   Transcription and error markers
   Annotation format
   Known errors
   History


----------------- DVD directory structure --------------------------

The corpus consists of two DVD-5 with a total size of 7.5 GByte plus a
CD-ROM with the label files and documentation ('DOCCDROM').

On one DVD (WebCommand_EN, #1) the british speakers are stored; on the
second DVD (WebCommand_FR, #2) the french speakers.

Recordings are situated in the 'BLOCK' directories:

BLOCK40  :  british, room P, 26 sessions
BLOCK50  :  british, room S, 26 sessions
BLOCK60  :  french, room P, 21 sessions
BLOCK70  :  french, room S, 22 sessions

The corpus contains 47 complete sessions (130 recordings per session).
Care is taken that each speaker is recorded in complete sessions
in each of the two recording rooms.
Additional incomplete recording sessions (speakers did not record a second
session, or corrupted sessions) are collected in the directories NOT_USED_FR
(4 sessions) and NOT_USED_EN (7 sessions) respectively.

The CDROM 'DOCCDROM' contains additional documents about the
corpus recording and annotation as well as pronunciation dictionaries:

PRON_FR.LEX     : Pronunciation dictionary, SAM-PA, french
PRON_EN.LEX     : Pronunciation dictionary, SAM-PA, english
TRANSCRP.PDF    : description of rules and conventions of SpeechDat
                  transcription (German)
TRANSCRP_EN.PDF : description of rules and conventions of SpeechDat
                  transcription (English)
PICS/           : Pictures of the recording setup
BLOCK##/        : SAM annotation files to recording block ##
REPORT.TXT      : this file
SAMEXPORT.TXT   : condensed summary of all SAM label files in one table
SUMMARY.TXT     : SpeechDat conform summary of recordings: foreach recording
                  session all individual recordings are listed in one line.
                  If a recording is missing, a '-' is listed instead of the
                  three-digit prompt number.
SPEAKER.TBL     : mapping of 4-digit speaker id to sex, age and mother tongue
SESSION.TBL     : mapping of 4-digit session id to speaker id, place of
                  recording, date of recording, microphone types, channel
                  mapping, environment
```

```
----------------- Recording Situation  -------------------------

Each speaker (complete sessions only!) was recorded in two different
recording rooms P and S on different days. Each session consists
of 130 prompts as given in the prompt lists doc/PROMPTS*.
The speaker wears a ear-free headset Beyerdynamik NEM 192; the second mic
is a Beyerdynamik MCE 10 mounted on the upper left corner of a dummy
laptop case that the user holds with both hands on his/her lap.

The recording setup is documented with photos in the directory PICS.

During the recording the user does not have to use the keyboard or the
mouse. The acoustical environment of both rooms is quiet office environment.
There is only one computer (Mac desktop mounted in front of the speaker);
no other noise sources. The signal of the microphones is amplified by a
Beyerdynamik MV 100 amplifier: headset mic + 20 dB, webpad mic + 20 dB
and then connected to the standard Mic input of the recording Mac.
Each session starts with a short instruction of the speaker, then the
microphones are mounted by the supervisor and a short training session
(not recorded) of 5 prompts is performed. Then the supervisor leaves the
room for the rset of the session. The prompting and recording runs
automatically; for each prompt a fixed time slot of 5.7 sec was recorded.
The timing is controlled by a 'red light' control: a red light indicates
not to speak, the yellow light indicates to get ready and then together
with the green light the prompt is displayed and the speaker reads from
the sreen. After the fixed recording time the red light comes again and
the cycle starts anew.


Recording specs:

Minimum speakers per language                  20
Minimum speakers per sex                       20
Recording sessions per speaker                 2
Prompts per session:                           130 (000-129)
Length per prompt:                             5.7 sec
Sampling rate:                                 22050 Hz
Bits per sample:                               16
File format:                                   WAV stereo
Head set:              Beyerdynamik NEM 192, left channel
Webpad mic:            Beyerdynamik MCE 10,  right channel
Amplifier:             Beyerdynamik MV 100, set to +20dB, LF Cut off


---------------------- Naming conventions ----------------------


Session names are coded as follows:
```

```
SES####    where #### denotes the session number


Session numbers starting with '4' : british speaker, room P
Session numbers starting with '5' : british speaker, room S
Session numbers starting with '6' : french speaker, room P
Session numbers starting with '7' : french speaker, room S

e.g. SES6013 is the 13th recording session of a french speaker in
room P.

A mapping from speaker IDs to sessions, as well as the speaker profile
can be found in the file TABLE/SESSION.TBL

Each recording file is named as follows:

Q1####%%%.WAV     where: #### denotes the session number
                         %%%  denotes the prompt number (000-129)


e.g. Q16013051.WAV contains the two microphone signals in a WAV stereo
file of the 52nd prompt of the 13th recording session of french speakers
in room P. The channel assignment for the microphones is stored in the
file TABLE/SESSION.TBL



------------------------- Signal file formats ----------------------

All recording files are stored in WAV standard format.
See specs aboce for details.

-------------------------------------------------------------------------

Transcription and error markers

All recordings were annotated according to SpeechDat conventions.
See the document doc/TRANSCRP.PDF for details about this.

The transcription files (SAM label format) are stored
on a separate CD-ROM in a file system hierarchy that mirrors
that of the signal files, i.e.\ BLOCKxx/SESxxxx.

The same information is also stored in a semicolon delimited text file
SAMEXPORT.TXT.

The SAM label names are the following (this is also the field
order of SAMEXPORT.TXt):

LHD     SAM Header specification
DBN     database name
SES     session number
```

```
CMT     comment
SRC     name of signal source file
DIR     directory path of signal file
CCD     corpus code of signal file
BEG     begin recording
END     end recording (in samples)
REP     recording place
RED recording date
RET recording time
CMT     comment
SAM     sample rate
SNB     sample number of bytes
SFB     byte order
QNT     quantization
NCH     number of channels
CMT     comment
SCD     speaker code
SEX     speaker gender
AGE     speaker age
ACC     speaker accent
CMT     comment
MIP     microphone position
MIT     microphone type
ENV     environment
CMT     comment
LBD     label file body
LBR     prompt text
LBO     transcription of utterance
ELF     end of label file

e.g.

LHD: SAM 6.0
DBN: Siemens WebCommand Database
SES: 6005
CMT: *** Recording data ***
SRC: Q16005004.WAV
DIR: BLOCK60/SES6005
CCD: 004
BEG: 0
END: 126064
REP: University of Munich, Phonetics Institute
RED: 04.07.2002
RET: 13:54:42
CMT: *** Signal data ***
SAM: 22054
SNB: 2
SBF: lo_hi
QNT: PCM
NCH: 2
CMT: *** Speaker data ***
SCD: 1005
```

```
SEX: F
AGE: 23
ACC: FR
CMT: *** Environment data ***
MIP: HEADSET=RIGHT, WEBPAD=LEFT
MIT: HEADSET=BEYERDYNAMIC_NEM_192,WEBPAD=BEYERDYNAMIC_MCE_10
ENV: P-ROOM
CMT: *** Label file body ***
LBD:
LBR: 0,126064,,,,appeler Nicolas Moulin
LBO: 0,63032,126064,appeler Nicolas Moulin
ELF:

---------------------------------------------------------------------------

Known errors


Remark: The subdirectories NOT_USED_* contain sessions that are incomplete,
either because speakers were not recorded a second time, or because signal
files were corrupted.

---------------------------------------------------------------------------

History

01.06.02 : start of recording
20.07.02 : start of validation
01.08.02 : end of recording
08.08.02 : end of validation
09.08.02 : delivery date 1.0
19.08.02 : delivery date 1.1 (update of DOCCDROM only)
```

Chapter 14

SpeechDat II German

14.1 Corpus Specification

SpeechDat-II is an EU-funded project to create telephone speech databases for the development of speech recognizers and speaker verification for voice-driven applications and tele-services. The main motivation for SpeechDat-II was to

- collect comparable databases in all major European languages for both the fixed and the mobile telephone networks,
- establish a standard for telephone speech data collections by publishing all database specifications,
- exchange the databases within the project, and
- make the databases available to the general public after a given blocking period.

In SpeechDat-II, competitors on the market collaborate to share the effort of creating a database, and then individually exploit these databases to develop competing applications, devices and services.

SpeechDat-II is a successor to the pilot project SpeechDat-M, and it has been succeeded by a number of further projects, e.g. SpeechDat-E for the East European languages, SpeechDat-Car for data collection in mobile environments, OrienTel for the Mediterranean languages, and numerous similar projects in all parts of the world.

The German SpeechDat-II data collections were performed by the BAS at Munich university under a subcontract to Siemens for the fixed telephone network, and Vocalis for the mobile telephone network.

In the following, the corpus specification of the fixed network German SpeechDat II data collection will be presented in the manner of a check list. The elements of this check list have already been discussed in this order in chapter 4. If elements are not applicable for SpeechDat, they're marked with 'n.a.'.

Speaker Profiles	Primarily native speakers of German; gender distribution 50:50% with a tolerance of +/- 5%, three age classes (16–30, 31–45, 46 and older): each of them a minimum of 20%; for the dialectal distribution Germany is divided into 11 regions corresponding to the larger federal states with a number of speakers proportional to their population; education level not specified
Number of Speakers	5000
Contents:	
- Vocabulary	Digits, numbers, date and time expressions, simple application words and phrases, spellings, person, company and geographic names, phonetically rich words and sentences
- Domain	not specified
- Task	not specified
- Phonologic Distribution	applied only to phonetically rich words and sentences
Speaking Style:	
- Read Speech	+
- Answering Speech	+
- Command/Control Speech	–
- Non Prompted Speech	+
- Spontaneous Speech	–
- Neutral/Emotional	–
Recording Setup:	Telephone Recording
- Acoustical environment	3 environments specified: office, home, telephone booth: minimum of telephone booth 2% of recordings
- Script	Prompt sheet and guided dialog by telephone server
- Background noise	natural, dependent on environment

- Microphones	not specified, but classification between rotary and DTMF phones required
Technical Specifications:	
- Sampling Rate	8000 Hz
- Sample Type and Width	ALAW, 8 bit
- Number of Channels	1
- Signal File Format	RAW header-less data
- Annotation File Format	SAM
- Meta Data File Format	Tab delimited ISO-8859 text files
- Lexicon Format	Tab delimited ISO-8859 text file, pronunciation coded in SAM-PA
Corpus Structure:	
- Structure	Hierarchical file structure according to recording sessions
- Terminology	signal file names encode recording session and prompt item
- Distribution Media	CD-R
Release Plan	SpeechDat II is to be available through ELRA after a 12 month blocking period after the end of the project
Validation	External pre-validation after 10 recordings; external final validation of the entire data base
Documentation	Specifications publicly available, recording logs and final validation report included in the distribution

14.2 Meta Data of SpeechDat

14.2.1 Recording Protocol

The SpeechDat-II partners were free to choose their recording equipment. Most partners used the ADA software developed at the Polytechnical University of Catalonia in Barcelona, or proprietary software. For the German data collection, proprietary software was used. The recording platform was a standard PC running Windows NT connected to an ISDN prime rate interface (30 channels). Some recording meta data was registered directly by the recording software, e.g. date and time of call, session ID, etc., other data was added later during the preparation of a recording session for transcription, e.g. prompt sheet number, region and network of call, etc.

The following list of minimal requirements for the recording protocol

as given in section 3.2 contain either a + or - if the content is given or a citation from the documentation files.

Session ID	+
Speaker ID	+
Date of recording	+
Environmental conditions	Room Acoustics: + class of 3 different environments Sources of Noise/Background Noise: – Cross Talk: –
Technical recording conditions	(in documentation)
Microphones	(in documentation)
Recording device	(in documentation)
Technical specifications of recorded signals	(in documentation)
Placement and distance to microphone(s)	–
Name or ID of the recording	+
Details about the recorded domain(s)	n.a.
Details about instruction to speaker(s)	–
Duration of the recording session	+
Type of Prompting (paper, face-to-face, screen, voice)	paper (in documentation)
Emotional speech yes/no	–
Details about acoustics (reverberation, S/N ratio etc.)	–
Supervisor present	n.a.
Interpreter present	n.a.
WOZ: details about 'virtual machine'	n.a.
Type of speech	read (in documentation)
Free comments	–

14.2.2 Speaker Profiles

The speaker data was established during the preparation of a recording session for transcription. A given set of individual recordings containing information about the speaker was listened to and the data was entered into the database containing the speaker meta data, e.g. speaker gender, accent, age, etc.

The following table of recommended meta data about speakers (refer to section 3.3 for details) contains a + if the information was collected and a – if not.

Speaker ID	+
Sex	+
Date of birth	+ (age)
Mother tongue	– (German implied but not verified)
Second languages of speaker	–
Mother tongue of parents	–
Second languages of parents	–
Pathologies	–
Dentures	–
Piercings	–
Place of elementary school	+
Dialect region	+
Dialect	–
Level of education	–
Level of proficiency for a certain task	–
Profession	–
Height	–
Weight	–
Left/right handed, ambivalent	–
Smoker/non smoker	–
Stutter	–
Free comments	–
Comments	–

14.3 Comments to SpeechDat

In the project proposal, the planned duration for SpeechDat-II was 24 months. In reality, however, the project took more than 36 months! The main reasons for this significant delay were threefold:

- the specification phase took much longer than expected,
- speaker recruitment was much slower than anticipated, and
- the final validation uncovered grave shortcomings in some databases which had to be corrected.

The size and the heterogeneous composition of the project consortium, consisting of a variety of industrial and academic partners, made the specification of a common subset of items a very tedious task. The requirements of application developers are quite different from those of service providers or of academia. The question of database exchange value was difficult to solve: is a database of 500 Luxemburg German speakers equal in value to a 5000 speaker database of standard German? How to incorporate late entrants into the consortium – all these questions had to be solved, and they had to be solved by consensus because the project plan did not foresee sanctions for non-cooperative partners.

Speaker recruitment turned out to be the single most critical issue. None of the project partners had experience with such a large speech database collection. Project partners with a good geographic and demographic coverage among their employees found it relatively easy to motivate their employees to participate – examples are national telecom companies. Professional market research companies in general were not used because of the high cost – e.g. in Germany they asked for more money than was available for the entire German data collection – and the lack of a guarantee that they would provide the requested number of speakers.

Most SpeechDat-II databases were ready for validation at about the same time. This imposed a heavy workload on the validation agency; originally it was planned to deliver the databases in sequence so that their validation could proceed with a constant effort over a longer period of time. During the validation grave errors were found in some databases. These errors had to be corrected, either by recording additional material, re-annotation or re-creation of lexica. In some cases, not all errors could be corrected and the database had to undergo an acceptance vote. The most important lesson learned here was that there should be at least three validations: a formal validation of all prompt material prior to any recordings, an early validation of the first few recordings prior to the main recording phase, and

a final validation. For very large databases, an intermediate validation is very useful.

SpeechDat has effectively set the standard for many successor projects. It is a show case for the collaboration of academia and industry, and it has proved that direct market competitors can effectively share the effort creating resources while at the same time keeping up the competition for the development of devices, applications and services.

14.4 Specification Documents

All SpeechDat-II specifications are publicly available on the SpeechDat web site www.speechdat.org. These documents include a description of the overall project goals, the language specific requirements, the database contents and the database exchange formats.

The following document is the README-file for the German fixed telephone network database. It outlines the contents and the structure of the database. The DESIGN.DOC document listed in the README file gives a detailed description of the real contents of the database, and VALREP.DOC is the final validation report of the validation agency. The documentation is contained on every one of the 17 CD-ROMs on which the database is distributed.

```
                    GERMAN SPEECHDAT(II) FDB4000
                         CD-ROM COLLECTION

                             Version 2.0
                        Copyright(C) 1999 by
                         SIEMENS AG, Munich


Compiled by: Chr. Draxler
             Department of Phonetics and Speech Communication
             University of Munich
             Schellingstr. 3/II
             D 80799 Munich

             +49/89/2866 9968
             +49/89/280 0362 fax

             draxler@phonetik.uni-muenchen.de


The German SpeechDat(II) FDB4000 consists of 4000 calls stored on
17 CD-ROMs in the final SpeechDat(II) database exchange format
as defined in deliverable SD 1.3.1 V.4.3:
```

```
CD-ROM Structure
----------------
/-- DISK.ID
/-- README.TXT
/-- COPYRIGH.TXT
/-- FIXED1DE -- +- DOC-----+-- DESIGN.{DOC | PDF | PS}
                |          +-- ISO88591.{PDF | PS}
                |          +-- SAMPALEX.{PDF | PS}
                |          +-- SAMPSTAT.TXT
                |          +-- SUMMARY.TXT
                |          +-- TRANSCRP.{PDF | PS}
                |          +-- VALREP20.TXT
                |
+- INDEX---+-- A1TRNDE.SES
                |          +-- A1TSTDE.SES
                |          +-- CONTENTS.LST
                |
                +- PROMPT--+-- SHEET.{PDF | PS}
                |
+- SOURCE--+-- CC_PIN.TXT
                |          +-- DEFTSTDE.PL
                |
+- TABLE---+-- LEXICON.TBL
                |          +-- SESSION.TBL
                |          +-- SPEAKER.TBL
                |
+- BLOCKyy-+                (with yy=[10..58])
           +-- SESyyzz --+  (with zz=[00..99])
                       + -- A1yyzzcc.DEA (signal file)
                       + -- A1yyzzcc.DEO (SAM label file)
                       (cc = corpus code)

The BLOCK directories contain the actual recordings.
Each call is written to a SES directory, where the 4-digit number
in the directory name identifies the session uniquely. The signal
and label files are held in the session directory; for each
signal file (extension .DEA) there is the corresponding SAM label
file (extension .DEO).

Note: file name extension mappings:

.DOC     Microsoft Word 6
.PDF     Adobe Portable Document Format
.PS      Adobe PostScript
.TXT     DOS-formatted ISO 8859-1
.PL      perl script
.TBL     tab-delimited ISO 8859-1 table file
.DEA     8 KHz 8 bit alaw encoded raw signal file
.DEO     ISO 8859-1 encoded SAM label file
```

```
The following directories contain documentation and related
information:

DOC    : DESIGN.{DOC|PDF|PS} Contents description of the
                                        German FDB4000
         ISO88591.{PDF|PS}   ISO8859-1 (ISO Latin) code table
         SAMPALEX.{PFD|PS}   German SAM-PA table
         SAMPSTAT.TXT        SNR values
         SD131V43.DOC        Database Exchange Format
                                        Specification
         SD132V24.DOC        Orthographic and Transcription
                                        Conventions
         SUMMARY.TXT         German FDB1000 summary file
 TRANSCRP.{PDF|PS}   the validation and transcription handbook
         VALREP.TXT          validation report by SPEX with
                                        responses by
                             U-Munich


INDEX  : A1TRNDE.SES  training set file
         A1TSTDE.SES  testing set file
         CONTENTS.LST contents of the database

                      The order of fields in the table is

                      VOL DIR SRC CCD CRP SCD SEX AGE ACC LBO

                      and the fields are separated by tabs.


PROMPT : contains a Portable Document Format and PostScript file

         SHEET.{PDF|PS}    prompt sheet layout in the form it was
                                       distributed to speakers

SOURCE : contains the follwing DOS formatted ISO 8859-1 files

         CC_PIN.TXT   150 16-digit credit card numbers and
                                        150 6-digit PIN codes

         DEFTSTDE.PL  perl script to define training and test sets
                      for the German FDB 4000


TABLE  : contains the following DOS-formatted ISO 8859-1 files

         LEXICON.TBL  the lexicon file with the following
                      tab-delimited fields

                      ORTHOGRAPHY FREQUENCY SAM-PRONUNCIATION

         SPEAKER.TBL  the speaker information file with the following
```

```
                tab separated fields

                SES AGE SEX ACC

SESSION.TBL     the session information file with the following
                tab separated fields

                SES RED RET AGE SEX ACC REG ENV

                this file is used to generate the training and
                test set files A1trnDE.ses and A1tstDE.ses
```

Chapter 15

SmartKom

> "The SmartKom multi-modal corpus was produced in the years 1999 - 2003 at the Bavarian Archive for Speech Signals (BAS) located at the University of Munich (LMU). The corpus was 100% funded by the German Ministry for Education and Science and is therefore freely available for all kinds of usage except re-distribution to third parties.
>
> The primary aim of the corpus was the empirical study of Human - Computer interaction (HCI) in a number of different tasks (domains) and technical setups (scenarios)."
>
> *(from the corpus documentation)*

In the SmartKom data collection subjects were recorded while using a self-explanatory, user adaptive man-machine interface (MMI). The MMI is simulated using a Wizard-of-Oz setup (WOZ, see section 4.5.4) and interprets speech and gesture input and analyses the facial expression of the user. The total corpus consists of a number of speech channels, four video channels, the output of a graphic tablet or finger point detector and a separate multi-modal biometric data collection. The resulting video data and multi-channel recorded spontaneous speech data serve as a basis for research and development of speech recognition, gesture recognition and the user model of SmartKom.

In the following only the speech part of the WOZ data collection is described.

15.1 Corpus Specification

The SmartKom recordings are carried out in three different technical setups (Public, Home, Mobil) and in an open number of task domains which do not overlap between technical setups. Most of the following specifications were defined at a special workshop organized by the group that produced the final corpus. Attendees of this workshop were all partners of the SmartKom consortium.

In the following, the corpus specification of the total SmartKom speech data collection (all technical setups, all task domains) will be presented in the manner of a check list. The elements of this check list have already been discussed in this order in chapter 4. If elements are not applicable for SmartKom, they're marked with 'n.a.'. The following check list for SmartKom covers only the recorded WOZ speech data without considering the biometric speech corpus.

Speaker Profiles	Primarily native speakers of German; gender distribution 50:50%; age ranging from 15 to 60 years; dialectal distribution not specified; education level not specified
Number of Speakers	open, depending on effort and funding; if feasible: equal proportions of speakers recorded in different technical setups and different task domains
Contents:	
- Vocabulary	free speech, no restrictions to vocabulary
- Domain	depending on the implemented task domains in the SmartKom prototype; at the writing of the specifications only a few task domains were defined: cinema guide, electronic program guide (EPG), VCR control, touristic information, navigation (by foot and by car), restaurant guide, office tasks
- Task	depending on the selected domain; each recording consisted of one primary and one secondary task, e.g. primary task: to find a cinema for tonight in Heidelberg, secondary task: to find a restaurant for dinner after the cinema
- Phonologic Distribution	not specified

Speaking Style:	
- Read Speech	–
- Answering Speech	+
- Command/Control Speech	+
- Non Prompted Speech	+
- Spontaneous Speech	+
- Neutral/Emotional	+
Recording Setup:	Wizard-of-Oz Recording
- Acoustical environment	normal office, reverberation time dampened by curtains, furniture and acoustical absorbers on walls and ceiling
- Script	subjects are told to assess the performance of a new prototype for a market study; no further explanations about the functionality of the system; description of the task to be solved; experimenter leaves room after introduction to the task; each subject is recorded in two sessions on the same day with a brief interruption between sessions
- Background noise	playback noise on two channels (back and front) recorded in different environments depending on technical setup
- Microphones	1 directional microphone Sennheiser ME66/K6 on top of front camera (approx. 60 cm from mouth), microphone array of 4 Sennheiser ME104 situated at the upper end of the display area, 1 headset Sennheiser ME104 or stereo clip-on Sennheiser ME104
Technical Specifications:	
- Sampling Rate	48000 Hz
- Sample Type and Width	PCM, 16 bit
- Number of Channels	9/10 (6/7 microphones, voice output, background noise back/front)
- Signal File Format	Microsoft WAVE
- Annotation File Format	SmartKom Transliteration, BAS Partitur Format (BPF)
- Meta Data File Format	XML (DDTs provided)
- Lexicon Format	Tab delimited 7-Bit ASCII text file, pronunciation coded in extended German SAM-PA

Corpus Structure:	
- Structure	Hierarchical file structure according to recording
- Terminology	Signal file names encode corpus type, recording session, technical setup, primary task and channel
- Distribution Media	DVD-R (5GB); each recording session is stored on one DVD
Release Plan	Data are released to partners as they become ready; a final integrated release is planned at the end of the project through BAS
Validation	On-going validation of current releases by partners; external final validation of the entire data base by BAS
Documentation	Not specified

15.2 Transcription

The SmartKom transcription format[1] is an overlay of very different information layers to the speech signal. From a technical point of view these layers would be better represented in separated layers coded in XML. However, we found that it is much more time consuming to produce 7 different layers of information than one complex transcript. If the format is valid and parsable, you may separate the layers later automatically[2]. Furthermore, the complex transcript is easier to read because the time relations are more obvious.

Assume for the following list of tags that a dialogue between a machine and a human being is transcribed turn by turn by listening to the signals.

- Lexical units: Lexical units are written in a standardized spelling and character coding. Furthermore, a definition of lexical units is needed, e.g. words, interjections, reduced forms of words, etc.
  In the SmartKom format the spelling is defined by the German Duden and a list of neologisms and foreign words (to keep the spelling of these consistent), the coding is LaTeX and the character set is 7 Bit ASCII. The lexical unit comprises only normal words and interjections[3].

[1] For a detailed description of the SmartKom transliteration format refer to www.bas.uni-muenchen.de/Forschungsprojekte/SmartKom

[2] which is partly done in the BAS Partitur Format of Verbmobil or SmartKom.

[3] That is: everything that is not marked in any way is either a normal word or an interjection. All other cases are tagged individually.

- Spelling: The spelling label is used when the subject spells a name letter by letter, e.g. for referring to the orthography or in abbreviations like 'USA'. The letters are always uppercase and separated by a comma or a dash, the latter mostly in abbreviations, e.g.
  `my name is Smith, $S , $M , $I , $T , $H .`

- Acronyms: Acronyms are official substitutes for particular words. The label only has to be placed once, at the beginning of the acronym. Acronyms must be pronounced like a word, e.g.
  `&OPEC`

- Proper names: All words are marked as proper names that can't be translated into another language; this includes surnames and first names of people, names of streets, hotels and restaurants, company names, names of institutions, local places, national holidays etc. Words are not labeled as proper names if they do not only appear in one language and thus can be translated e.g. names of international holidays, names of countries and continents, the names of the seven seas etc. If the proper name consists of several words that in regular orthography are separated by spaces, they will be linked by a '+' sign between each part of the name. For instance:
  `~Peter ~Marine+World ~Zur+Blauen+Traube`

- Numbers: Numbers are numerals, combinations of numbers and ordinal numbers. Two-digit numbers are labeled as one word. All numbers are written as words, e.g.
  `#three #twentytwo #first`

- Neologisms: 'Neologism' is a term referring to a word that has been made up by the speaker and does not appear in a regular dictionary. It could be slang or a slip of the tongue. e.g.
  `*forrowed`

- Foreign Words: Foreign words are words that stem from another language than that used by the speaker in that dialogue. In these cases an international language code marker is attached to the beginning of the word, e.g.
  `<*IT>Milano`

- Off-Talk: Especially in Wizard-of-Oz recordings you can find Off-Talk, i.e. when a person is speaking to himself or herself and not to the partner of the dialogue or the machine. You distinguish between 'Read Off-Talk' (ROT; the person is reading something aloud) and 'Other

Off-Talk' (OOT; any other speech which does not belong to the dialogue). For example:

```
now<OOT> what<OOT> do<OOT> we<OOT> have<OOT> here<OOT>
<hm> ~Arabic+Nights<ROT> can you give me
```

- Command Words: Command words are words that speakers use to operate the system by means of meta language, e.g.

  ```
  !KEYSmartakus
  ```

- Lengthening: Markup of sounds within or at the end of a lexical unit that are lengthened. It may also be used for pre-final lengthening, with plosives that have a particulary long closure phase and in the event of an aspiration phase that is stronger or longer than normally. The label is directly added to the letter representing the sound affected, e.g.

  ```
  giv<Z>e so<Z>rry
  ```

- Not or hardly identifiable words: This label can be used if it is impossible to understand a part of what has been said within the recording. Words that are not identifiable can either be completely incomprehensible or may be partially understood but not with certainty. The SmartKom format uses the label `<%>` in place of a non-understandable word, and a trailing `%` if we can understand a word partially but not well enough to identify it without any doubt, e.g.

  ```
  enough%  I have <%> enough
  ```

- Truncated Words: Truncated words occur when the speaker has begun to articulate a word but doesn't finish it. In other words, the item is terminated at a point where some of the component sounds have already been produced, while the rest has been cut off before being articulated. The equal sign is used here as the label; it is also placed during a series of stutters where parts of a word are repeated but the word as a whole is still not completely pronounced, e.g.

  ```
  the +/que=/+ question is could you hel= <*T>
  ```

- Articulatory Interruptions: Lexical items can be interrupted by various phenomena such as pauses, breathing, hesitations, slips of the tongue, mispronunciations etc. Such events can be marked up by adding a underscore followed by a blank space at the point of interruption. Then we insert the interrupting element and finally conclude with the remaining part of the interrupted word which is preceded by another blank space and underscore, e.g.

  ```
  this e_ <A> _vening
  ```

- Technical Interruptions: Technical interruptions are caused by a temporarily broken or missing section of the audio signal, something that might happen due to technical or other errors. There are four distinguishable types of technical interruption:

  1. `<T_>` is used when the beginning of a utterance is missing. In this case it is attached to the beginning of the first lexical item occurring, again without a blank space and regardless of whether this item seems to be complete or fragmental.
  2. `<*T>` is used when larger parts of an utterance are missing. It's a substitute for the missing speech.
  3. `<*T>t` is used when the end of an utterance is missing.
  4. `<_T>` is used when the last part of a word is cut off. In this case the label is attached to the end of the last word.

- Comments on pronunciation: The pronunciation comment indicates that the subject uses an unusual pronunciation (like foreign accent or dialect, word contractions, assimilations or mispronunciations). Thus, pronunciation comments show the deviation between actual pronunciation and the most likely form. In the case of contractions the number of the contracted words is given after the exclamation mark of the label. The label follows the lexical item, separated by a blank space, e.g.

```
no <!1 nope>  haben wir <!2 hamma>
```

- Repetition or Correction: There's a tendency in spontaneous speech to stutter and also to correct such disfluencies. The brackets `+/.../+` are used when the speaker repeats a word or a phrase or when he substitutes a new word for the one he started with, but continues with the same word class, e.g.

```
I would like +/to/+ to see
```

- False Start: A false start is characterized by the subject beginning an utterance, breaking it off before completion and continuing the utterance with an entirely new thought. The label is placed in the same way as the repetition/correction label, e.g.

```
-/this evening/- tomorrow I will
```

- Breathing: Clearly audible breathing, inhalation or exhalation, often occurs at prosodic or syntactic boundaries. In the transcript only breathing that can be heard well has to be marked. If the punctuation mark and the breathing label collide, the punctuation mark is put first,

e.g.
`please show me <A> the way . <A>`

- Filled Pauses: In spontaneous speech filled pauses are defined as pauses that are filled with some vocalization (or nasalization). A filled pause may occur when a speaker thinks about something. The speaker actually interrupts his speech while continuing his articulation. This articulation is however neither a word nor part of a word and should thus not be treated as such. As a consequence a punctuation mark cannot follow a filled pause, it has to come first. Nevertheless a filled pause can make a turn of its own. In SmartKom transcripts the four labels `<"ah>` (vocalic), `<"ahm>` (vocalic/nasalized), `<hm>` (nasalized) and `<h"as>` (others) are used. English adaptations four these four markers `<uh>` , `<uhm>` , `<hm>` and `<hes>` are also allowed.

- Empty Pause: Empty pauses can be defined as temporary, unfilled gaps in speech. They can be overlayed by cross talk, but cannot overlay actively. Just as with the filled pause labels punctuation marks always come first. Empty pauses at the beginning or at the end of a turn are not transcribed, e.g.
`could you please <P> tell me`

- Human Noises: Speakers also produce sounds that have no real meaning, such as laughing, coughing, swallowing etc. These are all labeled as `<Noise>` or `<Ger"ausch>` (German for noise). If one of these noises occurs for a long period of time, without being interrupted (a speaker laughing for example), a single label will be sufficient. As usual, punctuation marks come first.

- Technical Noises: Noises that can't be attributed to the speaker are technical noises. These might be caused by the recording instruments, by dropping things or by people moving around while recording, e.g.
`hello <#> !KEYSmartakus`

- Cross talk: Cross talk occurs when the subject and the system (or two subjects) speak at the same time or when noises occur while the subject speaks. From the point of view of the subject a cross talk may be either passive or active, depending on whether the speaker is the one who has been interrupted or the one who has interrupted. In either case the labeling indicates both the turn components passively affected by and the turn components actively affecting the interference. It's quite usual that within a dialogue there are several speaker interferences. This is why interferences are numbered consecutively,

e.g.

```
A: I'd like1@ to1@
B: @1please @1give me
A: here you can2@ see2@
B: @2that's @2right
```

- Superposition of noise: Any part of an utterance may be superimposed by one or more noises that are either background noises or noises produced by a speaker. If a noise appears during a word, brackets are used to embrace both, the noise and the word, e.g.

```
I <:<Ger"ausch> will:> take   here <:<#> you:> are
```

- Prosodic events: It is quite possible to mark up prosodic events in a transcript. In SmartKom, primary and secondary accent of the utterance as well as boundaries are marked in square brackets after the corresponding word item (see the following example transcript).

15.3 Transcription Example

By using all the above markers, a transcription can become rather complex. The transcribers have to be trained carefully in the conventions in order to make a transcription like the above a feasible task. In the following you can see a transcribed dialogue from the SmartKom project.

```
; DVD:
; Version: 1.0
; Dialog: w253_hf
; ENC: TEX
; zuletzt bearbeitet am: 28.05.02
; VPK: AEW
; ATMO: Wohnung
; Offtalk: wenig
; Erst:ulim , Pros+Korr: pet , Korr: babala
; PROS:
; Tonqualit"at:
;
w253_hfd_001_AEW: hallo [PA] [B3 fall] . <#> <"ahm> [B2] ich wollt'
 fragen [NA] [B2] , was heute abend [NA] im Fernsehen [PA] kommt [B3
 fall] .

w253_hfw_002_SMA: hallo . <P> <#> was kann ich f"ur Sie tun ?

w253_hfd_003_AEW: <"ah> [B2] ich w"urde ganz gern [NA] das
```

```
 Abendprogramm [PA] wissen [B3 fall] .

w253_hfw_004_SMA: wenn ich Ihnen einen Tip geben darf , <P> <#> heute
 kommt ~Der+Bulle+von+T"olz auf ~Sat-Eins um #zwanzig Uhr #f"unfzehn .

w253_hfd_005_AEW: -/und wa=/- [B9] <"ah> [NA] [B2] gibt es heute [NA]
 abend eine *Sportshow [PA] [B3 cont] ? <P> zum Beispiel [NA] Fu"sball
 [PA] [B3 rise] ?

w253_hfw_006_SMA: <#> hierf"ur erfolgt die Ausgabe auf dem Display .

w253_hfd_007_AEW: mhm , <#> gut <;sehr leise> . <PP> <%> <P> <hm> [NA]
 [B3 fall] <P> <hm>1@> <Ger"ausch> <*T>t

w253_hfw_008_SMA: @1f"ur diese Dienste m"ussen Sie sich identifizieren
 . <#> schreiben Sie Ihren Namen oder geben Sie das Pa"swort an .

w253_hfd_009_AEW: ~Schornsteinfegermeister<ROT> [B3 cont] . <A> <P>
 h"a<OOT> [NA] [B3 cont] ? <P> <#> *boa<OOT> [PA] [B3 fall] . <P> <#>
 <:<#> Hilfe<Z>:> [PA] [B3 cont] . <P> hallo [PA] [B3 rise] ? <#> <P>
 <"ah> [NA] [B2] wie <:<#> bedient:> [PA] man das [B3 cont] ?
 <Ger"ausch>

w253_hfw_010_SMA: ich habe Sie nicht verstanden .

w253_hfd_011_AEW: wie bedient <!1 bendiet> [PA] man dieses Ger"at [PA]
 [B3 fall] ? <#> <P> <hm> <P> <Ger"ausch> gut<OOT> .

w253_hfw_012_SMA: einen Augenblick , bitte . <#> <PP>2@> <#> einen
 Augenblick , <#> bitte . <PP> <#> einen Augenblick , bitte .

w253_hfd_013_AEW: @2und <!1 un'> @2jetz' ?
;gefl"ustert

w253_hfd_014_AEW: oh [PA] [B3 cont] .

w253_hfw_015_SMA: ich habe nichts gefunden .

w253_hfd_016_AEW: <"ahm> [NA] [B2] <P> oka<Z>y [PA] [B3 cont] . was
 gibt es heute abend [PA] auf ~Pro-Sieben [NA] [B3 rise] ?

w253_hfw_017_SMA: bitte suchen Sie Ihr Programm nach einer der
 Kategorien aus .

w253_hfd_018_AEW: <:<#> oh:> je [NA] [B3 fall] .
```

```
w253_hfw_019_SMA: bitte w"ahlen Sie einen Zeitraum .

w253_hfd_020_AEW: <h"as> [B2] <P> <Ger"ausch> ich h"atte [PA] ganz
 gern<Z> [B2] den Zeitpunkt <%> [B9] <*T>t

w253_hfw_021_SMA: bitte w"ahlen Sie eine oder mehrere Sparten .

w253_hfd_022_AEW: hallo da% [B3 cont] .
; sehr leise

w253_hfw_023_SMA: bitte w"ahlen Sie einen oder mehrere Sender .

w253_hfw_024_SMA: <#> einen Augenblick , <#> bitte .

w253_hfd_025_AEW: kein% Problem [PA] [B3 fall] .

w253_hfw_026_SMA: einen Augenblick , <P> bitte .

;EOF
```

15.4 Meta Data

15.4.1 Recording Protocol

The following list of minimal requirements for the recording protocol as given in section 3.2 contains a + if the content is given (optionally followed by a citation from the documentation) or a - if the content is not given.

Session ID	+
Speaker ID	+
Date of recording	+
Environmental conditions	Room Acoustics: (in documentation) Sources of Noise/Background Noise: + Cross Talk: n.a. (because always the case)
Technical recording conditions	(in documentation)
Microphones	+
Recording device	(in documentation)
Technical specifications of recorded signals	(in documentation)
Placement and distance to microphone(s)	(in documentation)

Name or ID of the recording	+
Details about the recorded domain(s)	+
Details about instruction to speaker(s)	(in documentation)
Duration of the recording session	–
Type of Prompting (paper, face-to-face, screen, voice)	n.a.
Emotional speech yes/no	+
Details about acoustics (reverberation, S/N ratio etc.)	–
Supervisor present	(in documentation)
Interpreter present	n.a.
WOZ: details about 'virtual machine'	(in documentation)
Type of speech	n.a.
Free comments	+

15.4.2 Speaker Profiles

The following table of recommended meta data about speakers (refer to section 3.3 for details) contains a + if the information was collected and a – if not.

Speaker ID	+
Sex	+
Date of birth	+
Mother tongue	+
Second languages of speaker	+
Mother tongue of parents	+
Second languages of parents	–
Pathologies	+
Dentures	–
Piercings	+
Place of elementary school	–
Dialect region	+
Dialect	+

Level of education	+
Level of proficiency for a certain task	–
Profession	+
Height	+
Weight	+
Left/right handed, ambivalent	+
Smoker/non smoker	+
Stutter	–
Free comments	+

The following sections show two examples of SmartKom meta data: a recording protocol and a speaker profile.

15.4.3 SmartKom Recording Protocol

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<!DOCTYPE RPR SYSTEM "rpr.dtd">
<RPR>

  <HEAD>
    <version number="2" subnumber="0" />
  </HEAD>

  <DVD no="40" />

  <SPEAKER id="ABS" />
  <SESSION-PARAMETERS>
    <session_id value="w090_pk" />
    <recorded_domains>
      <domain_planned use-case="Kino" />
      <domain_recorded use-case="Kino" />
      <domain_recorded use-case="Restaurant" />
    </recorded_domains>
    <atmosphere place="telephonebox" number="1" volume="normal" />
    <background pattern="Kubismus1" />
    <pen mode="finger" />
    <emotions evoked="no" />
    <content_variation version="1" />
    <recording_date year="2001" month="01" day="31" />
    <recording_location value="LMU" />
    <recording_setup wizard="wizard1" />
    <experimenter name="Christine_Enzinger" />
    <wizard_speech_output name="Sebastian Weberbeck"
    distortion="normal distortion" />
    <wizard_navigation name="Katerina Louka" />
    <session_sequence_no position="2" />
```

```
</SESSION-PARAMETERS>

<DATA-TRACKS>
  <VIDEO>
    <Data fieldname="(m)_mimic" device="Sony DSR-PD 100 AP" present="yes" />
    <Data fieldname="(l)_lateral" device="Sony DSR PD 1" present="yes" />
    <Data fieldname="(i)_SIVIT-video" device="SIVIT1" present="yes" />
    <Data fieldname="(o)_beamer-output" device="1024x768" present="yes" />
    <Data fieldname="(g)_four_fold_view" device="unknown" present="yes" />
  </VIDEO>

  <GESTURE>
    <GData fieldname="(k)_SIVIT-coordinates" present="no" />
    <GData fieldname="(t)_graphic-tableau-coordinates" present="no" />
  </GESTURE>

  <AUDIO>
    <Data fieldname="(d)_directional_mic"
    device="Sennheiser ME 66 / K6" present="yes" />
    <Data fieldname="(1)_array-mic_1"
    device="Sennheiser ME 104" present="yes" />
    <Data fieldname="(2)_array-mic_2"
    device="Sennheiser ME 104" present="yes" />
    <Data fieldname="(3)_array-mic_3"
    device="Sennheiser ME 104" present="yes" />
    <Data fieldname="(4)_array-mic_4"
    device="Sennheiser ME 104" present="yes" />
    <Data fieldname="(a)_clip-mic_1"
    device="Sennheiser ME 104" present="no" />
    <Data fieldname="(b)_clip-mic_2"
    device="Sennheiser ME 104" present="yes" />
    <Data fieldname="(h)_headset-mic"
    device="Sennheiser ME 104" present="yes" />
  </AUDIO>

  <ENVIRONMENT faked="unknown">
    <EData fieldname="(w)_wizard-mic" present="yes" />
    <EData fieldname="(p)_atmosphere-front" present="yes" />
    <EData fieldname="(q)_atmosphere-back" present="yes" />
  </ENVIRONMENT>

  <ANNOTATIONS>
    <ANNOTATION fieldname="trl" present="yes" />
    <ANNOTATION fieldname="marker" present="yes" />
    <ANNOTATION fieldname="gesture-labels" present="yes" />
    <ANNOTATION fieldname="userstate_labels_trp" present="yes" />
    <ANNOTATION fieldname="userstate_labels_ush" present="yes" />
    <ANNOTATION fieldname="userstate_labels_usm" present="yes" />
  </ANNOTATIONS>

  <QUICKTIME>
    <quicktime_frame present="yes">
```

```
    </QUICKTIME>

  </DATA-TRACKS>

  <BEHAVIOUR-OUTCOMES>
    <Emotions outcome="none">
      <anger outcome="none" />
      <joy outcome="none" />
      <surprise outcome="none" />
      <uncertainty outcome="none" />
    </Emotions>

    <personal_relation outcome="none" />
    <gesture outcome="none" />
  </BEHAVIOUR-OUTCOMES>

  <COMMENTS>
    <COMMENTS-ON-SPEAKER-BEHAVIOUR>
etwas herumexperimentiert: wollte ins Schwimmbad.
    </COMMENTS-ON-SPEAKER-BEHAVIOUR>

    <COMMENTS-TRL>
    </COMMENTS-TRL>

    <COMMENTS-GES>
    </COMMENTS-GES>

    <COMMENTS-USH>
    </COMMENTS-USH>

    <COMMENTS-USM>
    </COMMENTS-USM>

    <COMMENTS-TRP>
    </COMMENTS-TRP>

    <OTHER-COMMENTS>

flschlicherweise Aufnahme mit Headset-Mikro.
Ist also auf den Videos zu sehen.
    </OTHER-COMMENTS>
  </COMMENTS>

</RPR>
```

15.4.4 SmartKom Speaker Profile

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<!DOCTYPE SPR SYSTEM "spr.dtd">
<SPR>

  <HEAD>
    <version number="2" subnumber="0" />
  </HEAD>

  <SPEAKER>
    <Personal_Data speaker-id="ABI" sex="F" date_of_birth="19660922"
      height="175 m cm" weight="72 kg" handed="right" />

    <School degree_state="ST (Sachsen-Anhalt)"
    degree="(Fach-)Hochschulreife" profession="Sekretrin" />

    <Languages mothertongue="DEU" mothertongue_mother="DEU"
      mothertongue_father="DEU" dialect="E2 (Oberschsisch)" bilingual="no">
      <foreign_languages>
        <language value="ENG" />
        <language value="LAT" />
        <language value="RUS" />
      </foreign_languages>
      <bilingual_languages>
        <language value="none" />
      </bilingual_languages>
    </Languages>

    <Culture german_nationality="yes" cultural_environment="unknown" />

    <Experience speech_singing_training="no"
      computer_experience="yes" speech_dialogue_experience="no" />

  </SPEAKER>

  <RECORDING-SPECIFIC>
     <glasses exists="no" />
     <smoker exists="no" />
     <beard exists="no" />
     <piercing exists="no" />
     <jewels exists="no" />
  </RECORDING-SPECIFIC>

  <COMMENTS>

  </COMMENTS>

</SPR>
```

15.5 Comments on SmartKom

The SmartKom Speech Corpus is a special case of a scientific corpus production. Because the outcome of the total project cannot be defined in detail at the beginning, specifications for the corpus production tend to be inaccurate and open. However, this may also be considered to be an advantage because that way the corpus production can be adapted to the needs of the project partners.

There are 3 major problems with this kind of corpus production:

1. Logically, the corpus production should start ahead in time before the rest of the partners start their work. That way the necessary data will be available when needed and not at the end of the total project. However in most cases this is not possible because of the funding structure and because it is almost impossible to define the exact data type needed beforehand.

2. A data collection that adapts to the progress of a scientific project tends to yield many different and inconsistent data types. For example, if during the project an evaluation of special modules is needed and the data collection provides very specialized data for this purpose, these data might not easily be integrated into a monolithic corpus. Care has to be taken that all differing data types are documented in great detail to ensure the future re-usage of the corpus.

3. In most cases the funding for a scientific corpus production ends at the same time as the scientific work. This is a problem because data will be produced up to the very last minute and will not be properly integrated into the corpus. The solution is to arrange for a third party outside of the project that will take care of the corpus after the scientific project has ended. This institution must be funded independently from the project and must take the responsibility for the data for a longer time span. In the case of SmartKom the BAS took over the data after the SmartKom project was finished.

Bibliography

[1] Steven Bird and Mark Liberman (2001): A formal framework for linguistic annotation, Speech Communication 33(1,2), pp 23-60.

[2] Dafydd Gibbon, Roger Moore, Richard Winski, eds. (1997). Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin New York.

[3] Paul Boersma and David Weenink: The PRAAT System, www.praat.org/

[4] Steve Cassidy: The EMU System, www.shlrc.mq.edu.au/emu/

[5] Kipp, Andreas & Wesenick, Barbara & Schiel, Florian (1997): Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech. Proceedings of the EUROSPEECH 1997, Rhodes, Greece, pp. 1023–1026

[6] Kipp, Andreas (1998): Automatische Segmentierung und Etikettierung von Spontansprache. Thesis, Technical University of Munich.

[7] J. Sotschek (1984): Sätze für Sprachgütemessungen und ihre phonologische Anpassung an die Deutsche Sprache. Tagungsband DAGA: Fortschritte der Akustik, pp. 873-876.

[8] Türk, U. (2001): The Technical Processing in SmartKom Data Collection: A Case Study. Proceedings of EUROSPEECH Scandinavia, Aalborg, 2001, pp. 1541-1544.

[9] F. Schiel (2002): The TAXI Corpus, www.bas.uni-muenchen.de/Bas/BasTAXIeng.html

[10] F. Schiel (2000): The Verbmobil Corpus, www.bas.uni-muenchen.de/Bas/BasKoporaeng.html#VMI

[11] F. Schiel, Chr. Draxler (2003): The Validation of Speech Corpora, to appear.

[12] Chr. Draxler: The SpeechDat Car Projects, www.speechdat.org/SP-CAR/

[13] H.R. Pfitzinger (2000): Removing Hum from Spoken Language Recources. Proc. ICSLP 2000, vol. III, pp. 618-621. Beijing.

[14] J. Llisteri (1994): Prosody Encoding Survey, Multext – LRE Project 62–050.

[15] Francesco Senia, Jeroen van Velden (1997): Specification of orthographic transcription and lexicon conventions; LRE-4001 - SD1.3.2; 1997.

[16] H. van den Heuvel, L. Bouves, E. Sanders (2000): Validation of Content and Quality of Existing SLR: Overview and Methodology. ELRA/9901/VAL-1 Deliverable 1.1, Jan 2000.

Appendix A

Check Lists – Summary

- **Corpus ID:**
- **Corpus full title:**
- **Corpus short titel:**
- **Corpus creator:**

Specification

◯ Hand-held vs. hand-free ***

Recording Setup On-site Recording *(p. 50)*

◯ Supervised vs. non-supervised *

Field Recording *(p. 50)*

◯ Schedule a rehearsal **

Wizard-of-Oz Recording *(p. 51)*

◯ Specification of 'virtual machine' *

◯ Sampling rates * (p. 52)
◯ Sample Type and Width * (p. 53)
◯ Signal File Formats * (p. 54)
◯ Annotation File Formats * (p. 56)
◯ Annotation Contents and Procedures * (p. 52)
◯ Meta Data File Formats ** (p. 59)
◯ Meta Data Contents *** (p. 64)
◯ Lexicon Format * (p. 59)
◯ Corpus Structure * (p. 60)
◯ Terminology * (p. 61)
◯ Distribution Media * (p. 63)
◯ Release Plan ** (p. 63)
◯ Documentation *** (p. 64)

Preparation of Collection

◯ Instructions * (p. 67)
◯ Prompt List ** (p. 67)
◯ Automated Recording Procedure *** (p. 67)
◯ Test of Instructions, Prompts, Procedure * (p. 67)

Recording Techniques Telephone (p. 69)

◯ ISDN Account *
◯ ISDN Hardware + DLL (CAPI) *

◯ Control Program *
◯ Speech Prompts + Beep (Check for DC and clippings) *
◯ The 'script' *
◯ Silence detector ***
◯ Speech Detector ***
◯ Adjust / test recording intervals / detectors *
◯ Check for 'echos' *

Recording Techniques On-site, Field + WOZ (p. 72)

◯ Acoustical Environment *
◯ Microphones *
◯ Amplifiers, set levels *
◯ Recording Devices *
◯ Recording Software *

Recording Techniques Field (p. 75)

◯ Batteries *
◯ Check AC Grounding + Power Supplies *
◯ Banish Cellular Phones *
◯ Recording Devices *
◯ Be prepared for bad weather *
◯ Daily Backup *

Recording Techniques WOZ (p. 76)

◯ Observation technique (no mirrors) **
◯ Acoustically insulated recording and control rooms *
◯ Simulate Synthetic Speech Output *
◯ Clarify Special Legal Aspects *
◯ Task Flow Maps *

◯ Legal Aspects * (p. 78)
◯ Prepare Doc Forms and Questionnaires * (p. 77)
◯ Prepare Check Lists * (p. 78)
◯ Pre-test * (p. 78)
◯ Plan Recruitment * (p. 79)

Collection

Set up Logging Procedures *(p. 83)*

- ◯ for the recording protocol *
- ◯ for speaker meta data *
- ◯ for speaker comments **
- ◯ for questionnaires **
- ◯ statistical data ***

◯ Organize Pre-validation ** (p. 84)
◯ Set up Procedures for Quality Control * (p. 85)
◯ Check for Security * (p. 86)
◯ Provide enough Storage * (p. 87)
◯ Organize Data Pipelining * (p. 87)
◯ Choose your Recruiting Technique * (p. 88)
◯ Define Incentive and their Distribution * (p. 90)

Post-processing

In this check list the processing steps that might not *be obligatory are marked with **.*

◯ File Transfer from Recording Device to Computer ** (p. 93)
◯ File Name Assignment According Terminology * (p. 94)
◯ Define Suffices for Different Processing Steps * (p. 94)
◯ Cutting ** (p. 94)
◯ Filtering ** (p. 95)
◯ Re-sampling ** (p. 96)
◯ Format Conversion * (p. 96)
◯ Special Format Conversions for Annotation ** (p. 97)
◯ Automatic Error Checks * (p. 97)

Annotation

◯ Select/define annotations * (p. 101)
◯ Integrate annotations into the data pipeline * (p. 87)
Always produce some kind of orthographic transcription:

◯ Define/select the orthographic transcription * (pp. 103, 108)
◯ Set up the transcription rules/method * (p. 107)
◯ Define the delivery format of the transcript * (p. 108)
◯ Choose/program the tools for transcription * (pp. 114, 109)
◯ Train the group of transcribers *
◯ Set up check procedures for the transcription * (p. 107)
◯ Test for inter-transcriber agreement *** (p. 117)

For each other annotation type, tagging (p. 109) or segmentation (p. 110):
◯ Define the annotation contents and rules *
◯ Define the delivery format of the annotation *
◯ Choose/program/test the tools for annotation *
◯ Train the labelers *
◯ Set up check procedures *
◯ Test for inter-labeler agreement *** (p. 117)

Dictionary

To create the dictionary you will most likely proceed through parts of the following procedures (depending on what resources you have):
◯ Define the orthographic representation for your corpus and transliterate your data or render your text material accordingly *
◯ Create a complete list of unique words. Watch out for capital letters at the beginning of sentences[1] *
◯ Define the desired contents of each entry in your dictionary *
◯ Use automatic procedures to create as much content as possible such as: look-up existing dictionaries, text-to-phoneme converters, part-of-speech taggers, etc. (pass 1) **
◯ Verify the contents of pass 1 and/or create information manually from scratch and produce a corrected version of the dictionary (pass 2) *
◯ If possible, let this be done by one person for the complete dictionary **
◯ Repeat the last step by a second person for the complete dictionary (pass 3) **
◯ Automatically find the differences between pass 1 and pass 2 or between pass 1 and pass 3 where pass 2 and pass 3 are not consistent and discuss these inconsistencies with a group of experts to come up with the final version of the dictionary **
◯ Repeat the last four steps for all content types that need manual label-

[1] A proper transliteration should not contain any of these!

ing/verification *
◯ Use a simple parser to ensure a proper coding of the final dictionary. Especially look out for inconsistent usage of blanks and tab signs. You may also check for homophones and homographs and check whether they are really valid for your language.

Sources for existing pronunciation dictionaries may be the ELDA[2], the LDC[3] or the BAS[4].

Validation

In this check list the processing steps that might not *be obligatory are marked with **.*

◯ Decide between in-house or external * (p. 135)
◯ Schedule pre-validation ** (p. 136)
◯ Schedule release validation ** (p. 136)
◯ Schedule final validation * (p. 136)
◯ Define validation content * (p. 137)
◯ Validation reports into documentation * (p. 138)

Distribution

In this check list the processing steps that might not *be obligatory are marked with **.*

◯ Select media * (p. 141)
◯ Compression? * (p. 143)
◯ Store symbolic data separate ** (p. 143)
◯ Safety/verify procedures * (p. 144)
◯ Print, burn-on-demand or online? * (p. 144)

[2] www.icp.grenet.fr/ELRA/home.html
[3] www.ldc.upenn.edu
[4] www.bas.uni-muenchen.de/Bas

Appendix B

Web References – Summary

In the following, you'll find all Web references that occur in the main text for easier lookup (in their order of appearance).

URL	Topic/Content
www.bas.uni-muenchen.de/Forschung/BITS/TP1/Cookbook	The newest version of this book
www.bas.uni-muenchen.de/Forschung/BITS	BITS Home
www.bas.uni-muenchen.de/Bas	BAS Home
www.icp.grenet.fr/ELRA/org/reasons.php3	ELRA: legal assistance
www.icp.grenet.fr/ELRA/home.html	ELRA Home
www.spex.nl	SPEX Home
www.ldc.upenn.edu	LDC Home
www.speechdat.org	SpeechDat Home
www.mpi.nl/ISLE/	ISLE Project
www.nist.gov/speech	Speech Group at NIST, NIST software
ccrma-www.stanford.edu/CCRMA/Courses/422/ /projects/WaveFormat	WAVE format
www.bas.uni-muenchen.de/Bas/BasFormatseng.html	BAS file formats
www.hornig.net/shorten.html	Shorten compression
www.icp.inpg.fr/Relator/standsam.html	SAM standards

www.mpi.nl/DOBES/tools/Eudico-Annotation-Tool.pdf	EUDICO file format
www.phon.ucl.ac.uk/home/sampa/home.htm	SAM Phonetic Alphabeths
www.phon.ucl.ac.uk/resource/sfs/	SFS software
www.spies.com/Sox/	SOX software
www.cygwin.com/	CYGWIN Shell for Windows
www.speechdat.org/speechdat/deliverables/ /public/SD132V24.PDF	SpeechDat transcript
www.bas.uni-muenchen.de/Forschung/Verbmobil/ /VMTrlex2d.html	Verbmobil transcript
www.bas.uni-muenchen.de/Forschung/SmartKom/Konengl/ /engltrans/engltrans.html	SmartKom transcript
www.ims.uni-stuttgart.de/projekte/mate/mdag/	MATE transcript
www.bas.uni-muenchen.de/Forschung/ /Verbmobil/VM14.1eng.html	MAUS automatic labelling
www.icsi.berkely.edu/~steveng	Elicit Segmentation
www.ipds.uni-kiel.de/forschung/kielcorpus.en.html	Kiel Corpus
www.icsi.berkeley.edu/real/stp/	Switchboard corpus segmenation
www.praat.org/	PRAAT software
www.bas.uni-muenchen.de/Bas/ /BasGermanPronunciation	BAS ruleset for German Pron.

Appendix C

BAS – Rules of Transcription

The following is a copy of the 'Transcription Conventions for Canonical German' as being used at the Bavarian Archive for speech signals[1].

C.1 Aims and Objectives

The main objective of the transcription of a word list is consistency. A word list has to be logical in itself and also in comparison to other word lists. How difficult it is to keep long lists consistent can be seen from the standard works on German pronunciation like the "DUDEN - Das Aussprachewörterbuch", for example. This and the fact that we now and then aim at a somewhat more phonetic transcription than that provided by the canonic forms of the DUDEN makes it necessary to suggest some modifications and own conventions. Nevertheless, the DUDEN should be considered the decisive reference for all transcriptions.

The transcription conventions are structured such that different word lists from different sources can be edited with different amounts of effort (depending on the requirements). The conventions that can be found in the chapter 'Basic Transcription' hold for all types of word lists. All other conventions only have to be paid attention to if the information they regulate is relevant to the respective transcription task.

[1] See www.bas.uni-muenchen.de/Bas/BasGermanPronunciation/ for an updated version of this document

To give an account of the type of transcription made, a simple check list like the following will be enough. It should be infact always attached to the transcribed list. The example shows a check list for the VERBMOBIL lexicon:

Transcription according to	(URL of this document)
Primary accent	+
Secondary accent	+
Morpheme markers	-
Compound markers	+
Function word markers	+

C.2 Basic Transcription

In order to ensure easy handling and implementation of the transcribed lists, all transcriptions should be done in SAMPA. See also: http://www.phon.ucl.ac.uk/home/sampa/german.htm

There is a difference in the respresentation of the glottal stop, though: While SAMPA uses the symbol /?/, we will use /Q/, as this is the easiest way to avoid overlaps with meta symbols and punctuation marks.

C.2.1 Vowels

One of the biggest problems of the basic transcription is the correct account of vowel quantities and qualities. It is advisable to use the DUDEN when in doubt, even if the respective word is unusal or not of German origin: Firstly, the DUDEN contains a lot more than one would expect initially. Secondly, it is possible to derive more complex units from simpler ones.

Example: Arcor can be derived from:

Arco	Q'arko
+ Chor	ko:6
=	Q'arko:6

With vowel length it also has to be taken into consideration that a shift of stress may well result in the shortening of a vowel. Transcription mistakes often occur within a word family.

Example:	Telefon	telef'o:n
	telefonieren	telefon'i:r@n

But see also:

Pulli	p'Uli
Pullis	p'Ulis

In the usage of the glottal stop we suggest some deviation from the conventions of the DUDEN. While the DUDEN doesn't indicate the glottal stop in the beginning of words, and not consistently within words, we've decided to leave out the glottal stop in only one case: If the vowel occurs in a position where it is usually never stressed.

Example:	und	Q'Unt
	Aimee-und-Jaguar	QEm"e:Untj'a:gua:r
	ab	Q'ap
	unabh"angig	Q'UnQaph"ENIC

Also different to what we find in the DUDEN should be the handling of non-syllabic /i/. Instead of diacritics we suggest the usage of the symbol /j/, which is much more in accordance with the phonetic reality.

Example:	Funktion	fUNktsj'o:n
	Kom"odien	kom"2:dj@n

C.2.2 Vocalised r

The so-called 'Lehrerschwa' is also used in the DUDEN, though again not completely consistently. We modify in so far as we lay down that /6/ has to be used whenever /r/ would follow immediately after a vowel within a syllable. An exception to this rule is after /a/, here remains /r/.

Example:	6	Lehrer	l'e:r6
	i:6	Bier	b'i:6
	I6	Schirm	S'I6m
	y:6	Tür	t'y:6
	Y6	N"urnberg	n'Y6nb"E6k
	e:6	der	d'e:6
	E6	verkehrt	fE6k'e:6t
	E:6	w"ar	v'E:6
	2:6	BurgerKing	b'2:6g6k"IN
	96	W"orter	v'96t6
	u:6	Uhr	Q'u:6
	U6	durch	d'U6C
	o:6	vor	f'o:6
	O6	Information	QInfO6matsj'o:n
But:		Bar	ba:r
		Mark	mark
		mehrere	m'e:r@r@
		w"are	v'E:r@
		Brauerei	braU@r'aI

As the counterexamples show /r/ has to be used instead of /6/ as soon as /r/ forms the onset of the next syllable ('wär' vs 'wäre', 'mehr' vs 'mehrere', 'Brauer' vs 'Brauerei').

C.2.3 Consonants

The first reference for the transcription of consonants is again the DUDEN. Special attention has to be paid to the German 'auslautverhärtung' and to the account of the various German r qualities. The term 'auslautverhärtung' refers to the devoicing of voiced fricatives and plosives in the coda. It is a phenomenon that is typical for German and very rare in other languages. As we try to represent the pronunciation of a German native speaker, the 'auslautverhärtung' should also appear in the transcription of foreign words, even if there would not be any 'auslautverhärtung' in the native pronunciation (see also 'Foreign Words'). The DUDEN does not indicate the 'auslautverhärtung' in foreign words.

Example:	Abend	Q'a:b@nt
	Subkultur	s'UpkUltu:6
	BigBrother	bIkbr'aD6

The various German r qualities must not be represented with any other symbol but /r/. Even though some German pronunciation dictionaries prefer using /R/, this symbol should be strictly avoided here.

A common transcription mistake can be found on syllable borders, when the same consonant occurs in the coda of the first syllable as well as in the onset of the second syllable. Here it is actually necessary to use the same symbol twice.

Example:	Autobahnnummern	Q'aUtoba:nn"Um6n
	heraussuchen	hEr'aUsz"u:x@n

Assimilation processes do not have be taken into consideration when transcribing. The boundary between really common and quite uncommon forms would be too fuzzy, consistency would be hard to guarantee.

C.2.4 Reductions

Reduction processes, like the syllabification of /m/, /n/ and /l/ under elision of schwa, do not have to be transcribed. Otherwise we would also have to indicate assimilation processes, which could easily lead to inconsistentencies (see also 'Consonants').

Example:	allem	Q'al@m
	daneben	dan'e:b@n
	Einzelheiten	Q'aInts@lhaIt@n

C.2.5 Foreign Words

Foreign words should be transcribed with a certain adaption to German pronunciation habits. As there are hardly any conventions for this type of transcription (the DUDEN only provides the 'original' pronunciations), we can only give some rough guidelines here. It is generally advisable to avoid exaggerations, transcriptions should represent a reasonably talented speaker (so no 'Mock German English', please!). The native pronunciation remains the reference.

- Voiced plosives, fricatives and affricates are substituted by their voiceless opponents when occurring in the coda ('auslautverhärtung').

Example:	big	bIg	(Eng)
		bIk	(Ger)
	Deneuve	d@n'2:v	(Fre)
		d@n'2:f	(Ger)

- /s/, on the other hand, is produced voiced when occuring in the onset of a syllable.

  Example: Dolby-Surround d"Qlbis@r'aUnd
  d"Olbiz9r'aUnt

Vowel qualities have to be mainly modified in view of the German pronunciation of English words:

- English distinguishes between more closed /e/ and more open /{/. For the German pronunciation, /E/ should be used in both cases.

  Example: Brenda br'end@
  br'Enda
  BigDaddy bIgd'{di
  bIkd'Edi

- English /V/ has to be replaced with German /a/. Those two qualities are not very different, anyway, and using the German symbol helps reducing the number of additional phonemes.

  Example: brother br'VD@
  br'aD6

- English /Q/ has to be replaced with German /O/ (Note: /Q/ does not stand for glottal stop here, but for a open back rounded vowel quality)

  Example: McDonalds m@kd'Qn@ldz
  m@kd'On@lts

- English final /@/, often respresented by the graphemes ¡-er¿ or ¡-a¿, is replaced with /6/ or /a/ depending on the orthography.

  Example: brother br'VD@
  br'aD6
  Brenda br'end@
  br'Enda

- English /@/ can be replaced with German /9/ when occurring in an unstressed position.

  Example: Dolby-Surround d"Qlbis@r'aUnd
  d"Olbiz9r'aUnt

- English further back /A:/ has to be replaced with German /a:/.

  Example: Hugh Grant hju:grA:nt
  hj"u:gr'a:nt

- The English Diphthong /eI/ has to be replaced with the German Monophthong /e:/.

  Example: Take-Away t'eIk@v"eI
  t'e:kEv"e:

- English /3:/ has to be replaced with German /2:/.

  Example: Worst-Case w3:stkeIs
  v2:stke:s

  This example shows also, that the typical English /w/ has to be replaced with /v/ for German pronunciations.

The French nasalisation of vowels can be indicated by adding a tilde after the relevant vowel.

Example: Restaurant rEstor'a:

English 'th', which has to be transcribed as /T/ and /D/ respectively, as well as the voiced fricative /Z/ and the voiced affricate /dZ/ (English, French, Italian, ...) should not be changed - that is unless they occur in the coda and are subject to the 'auslautverhärtung'.

Example:	brother	br'aD6
	Regie	reZ'i:
	Giardino	dZard'i:no

When transcribing words from less common languages it is advisable to look at the orthography in order to decide on what could be the most likely German pronunciation.

Example:	Tarragona	tarraG'ona	(span)
		tarag'ona	(dt)

C.2.6 List of All Symbols

Consonants

p	b	t	d	k	g		
f	v	s	S	C	j	x	h
m	n	N	l	r			

Affricates

pf ts tS

Glottal Stop

Q

Vowels

I		E	a	O	U	Y	9	6
i:	e:	E:	a:	o:	u:	y:	2:	

Dipthongs

aI	i:6	y:6	e:6	2:6	a:6	u:6	o:6
aU	I6	Y6	E6	96	a6	U6	O6
OY			E:6				

Foreign Symbols

T D Z dZ

C.3 Accents

Normally there shouldn't be more than one primary accent (') per entry, the DUDEN can always be used as a reference. Secondary accents (") are not indicated in the DUDEN. They occur mainly in phrases, compounds and prefix verbs.

Example:	Aimee-und-Jaguar	QEm"e:Untj'a:gua:r
	"Offnungszeiten	Q'9fnUNsts"aIt@n
	abbuchen	Q'apb"u:x@n

C.4 Morpheme Markers (+)

Morpheme markers (+) are placed in between morphemes, i.e. in between the "meaningful units of language that cannot be further divided".

Example: Fernsehturm f'E6n+ze:+t"U6m

A meaningful unit can also be a unit that indicates nothing but a grammtical function.

Example: Filme f'ilm+@

As morpheme markers can only be placed within a transcribed unit there will be no overlaps with the function word markers (see also Function Word Markers).

C.5 Compound Markers (#)

Compound markers (#) should only be used if two or more transparent, or in simplified terms 'meaningful', units can be separated. A meaningful unit in that sense is a content word that keeps its meaning even though a second unit is added.

Example:	Fernsehturm	f'E6nze:#t"U6m
	dunkelblau	d'UNk@l#bl"aU
	staubsaugen	St'aUb#s"aUg@n

Not marked, however, is:

Example: Bahnhof b'a:nh"o:f

This is due to the fact that a 'Fernsehturm' is still a 'Turm', but a 'Bahnhof' is not a 'Hof' anymore.

Parallel to that not marked are constructions with prefixes and suffixes, as in these cases only one 'meaningful' unit is involved.

Example:	vorlesen	f'o:6l"e:z@n
	Sicherheit	zIC6h'aIt

C.6 Function Word Markers (+)

As a function word are mainly those words marked that cannot be inflected, i.e. all pronouns, articles, prepositions, conjunctions and adverbs; in addition all forms of the so-called copula verbs 'haben' (to have) and 'sein' (to be).

Not marked should be adverbs that are at the same time basic forms of adjectives.

Example:

über	Q'y:b6+
abends	Q'a:b@nts+
bin	b'In+
den	d'e:n+
etwas	Q'Etvas+
ihre	Q'i:r@+
irgendwann	Q'I6g@ntv'an+
normalerweise	nO6m'a:l6vaIz@+
sondern	z'Ond6n+
und	Q'Unt+
werden	v'e:6d@n+

But:

kurz	kU6ts
täglich	t'E:klIC

As here the marker (+) comes after the transcribed unit there will be no overlaps with the morpheme markers (see also Morpheme Markers).