

SOUTHERN AFRICAN LANGUAGE CORPORA SERIES

General Editors

JENS ALLWOOD, RUSANDRÉ HENDRIKSE & MMEMEZI MFUSI

Advisory Editorial Board

D Prinsloo (Pretoria), J Rammala (Polokwane); B van Rooy (Potchefstroom); G van Huysteen
(Potchefstroom)

Volume 1

Jens Allwood and Rusandr  Hendrikse

Guidelines for Developing Spoken Language Corpora
Pretoria, 2005

GUIDELINES FOR DEVELOPING SPOKEN LANGUAGE CORPORA

JENS ALLWOOD

University of Gothenburg

RUSANDRÉ HENDRIKSE

University of South Africa

COLLABORATING AUTHORS

Ellen Breitholtz (University of Gothenburg)

Leif Grönqvist (University of Gothenburg)

Magnus Gunnarsson (University of Gothenburg)

Mmemezi Mfusi (University of South Africa)

Nozibele Nomdebevana (University of South Africa)

Hildegard van Zweel (University of South Africa)

DEPARTMENT OF LINGUISTICS

UNISA, PRETORIA

2005

Table of Contents

CHAPTER 1: THE SPOKEN LANGUAGE CORPUS PROJECT

| | |
|----------------------------------------------------|---|
| INTRODUCTION | 1 |
| ELECTRONIC NETWORK | 1 |
| A CORPUS LINGUISTIC APPROACH..... | 2 |
| WHY SPOKEN LANGUAGE? | 2 |
| WHAT IS INCLUDED IN THE CORPUS? | 4 |
| A LIST OF COMMON SOCIAL ACTIVITIES..... | 4 |
| CORPUS-DRIVEN AND CORPUS-BASED APPROACHES | 5 |
| AN EXAMPLE OF A KEY WORD IN CONCORDANCE LINES..... | 5 |
| AN EXCERPT FROM AN ANNOTATED CORPUS | 6 |

CHAPTER 2: RECORDING SPEECH

| | |
|------------------------------------------------------------|----|
| INTRODUCTION | 8 |
| DATA COLLECTION METHODS | 8 |
| DATA STORAGE | 9 |
| DATA RECORDING..... | 9 |
| INFORMATION ABOUT A RECORDED SPOKEN LANGUAGE ACTIVITY..... | 11 |
| WHAT TO RECORD – ACTIVITY TYPES..... | 13 |
| WHAT TO DO WITH THE RECORDED MATERIAL | 14 |
| SPECIFICATIONS FOR A TRANSCRIPTION COMPUTER..... | 14 |
| EXAMPLE OF THE RECORDING REGISTER..... | 16 |

CHAPTER 3: TRANSCRIPTION STANDARD

| | |
|----------------------------------------------------------------------|----|
| INTRODUCTION | 17 |
| THE PARTS OF A TRANSCRIPTION | 18 |
| HEADER CONVENTIONS..... | 18 |
| HEADER LINES | 18 |
| LIST OF INSTITUTIONAL CODES TO BE USED IN THE TRANSCRIPTION IDs..... | 20 |
| LIST OF LANGUAGE CODES TO BE USED IN THE TRANSCRIPTION IDs | 20 |

| | |
|--------------------------------------------------------------------------------------------------|----|
| THE TRANSCRIPTION BODY | 24 |
| GENERAL TRANSCRIPTION PRINCIPLES AND RULES | 24 |
| 1. TRANSCRIPTION TEXT FORMAT | 24 |
| 2. PUNCTUATION | 25 |
| 3. NUMERALS | 25 |
| 4. TRANSCRIPTION AUTHENTICITY | 26 |
| 5. LANGUAGE MIX | 28 |
| 6. CONTRIBUTIONS AND OVERLAP | 29 |
| 7. LINE FORMATS OF CONTRIBUTIONS | 30 |
| 8. WORDS | 30 |
| 9. NUMERALS | 32 |
| CONVENTIONS USED IN THE TRANSCRIPTION OF THE BODY | 32 |
| CONTRIBUTION LINES | 33 |
| HOW TO TRANSCRIBE A CONTRIBUTION | 33 |
| 1. LINGUISTIC STRUCTURE-RELATED CONVENTIONS | 33 |
| 2. PROSODY-RELATED CONVENTIONS | 34 |
| 3. DISCOURSE-RELATED CONVENTIONS | 34 |
| GUIDELINES FOR DETERMINING THE LENGTH OF A PAUSE | 35 |
| GUIDELINES FOR TRANSCRIBING DIFFERENT TYPES OF OVERLAPPING | 36 |
| 4. RECORDING-RELATED CONVENTIONS | 37 |
| COMMENT LINES | 38 |
| HOW TO WRITE COMMENT LINES | 38 |
| EXAMPLE OF COMMENT LINES IN A TRANSCRIPTION | 39 |
| COMMENTS ON OVERLAPPING CONTRIBUTIONS | 39 |
| STANDARD COMMENTS | 40 |
| 1. GENERAL | 40 |
| 2. OUTLINE AND ILLUSTRATION OF INSTANTIATIONS OF THE VARIOUS TYPES OF STANDARD COMMENTS | 41 |
| 2.1 COMMENTS ON VOCAL SOUNDS | 41 |
| 2.2 COMMENTS ON PROPERTIES OF SPEECH | 42 |
| 2.3 COMMENTS ON SPECIAL EXPRESSIONS | 42 |
| 2.4 CLARIFICATION COMMENTS | 42 |
| 2.5 COMMENTS ON SPEAKER'S MOOD | 42 |
| 2.6 COMMENTS ON THE PROPERTIES OF THE TRANSCRIPTION | 43 |
| 2.7 COMMENTS ON GESTURES | 43 |
| SECTION LINES | 43 |
| EXAMPLE OF SECTION LINES IN A TRANSCRIPTION | 44 |
| TIME CODING | 44 |
| EXAMPLE OF TIME LINES IN A TRANSCRIPTION | 45 |
| <hr/> CHAPTER 4: QUALITY ASSURANCE PROCEDURES <hr/> | |
| INTRODUCTION | 46 |
| THE AUTOMATIC CHECKING PROCEDURE | 46 |
| HOW TO SET UP AND USE THE AUTOMATIC CHECKING TOOL | 46 |
| TRANSCRIPTION EDITING PROCEDURE | 47 |
| TRANSCRIPTION CHECKING PROCEDURE | 48 |

| | |
|---------------------------------------------------------------------------------------------------|----|
| <hr/> CHAPTER 5: RECORDKEEPING AND ARCHIVING PROCEDURES <hr/> | |
| INTRODUCTION | 49 |
| RECORDING-RELATED RECORDKEEPING..... | 50 |
| HOW TO SET UP AND MAINTAIN A RECORDKEEPING SYSTEM FOR RECORDINGS | 51 |
| TEMPLATE FOR RECORDKEEPING OF RECORDINGS..... | 52 |
| TRANSCRIPTION-RELATED RECORDKEEPING..... | 53 |
| HOW TO SET UP AND MAINTAIN A RECORDKEEPING SYSTEM FOR TRANSCRIPTIONS | 53 |
| TEMPLATE FOR RECORD KEEPING OF TRANSCRIPTIONS | 54 |
| CONSOLIDATED INTERIM PROGRESS REPORTS ON THE CORPUS | 55 |
| HOW TO SET UP A CONSOLIDATED REPORT | 55 |
| TEMPLATE FOR CONSOLIDATED REPORT | 55 |
| ARCHIVING THE VARIOUS CORPUS MATERIALS | 55 |
| MEDIA ARCHIVE | 55 |
| TRANSCRIPTION FILES ARCHIVE | 55 |
| BUILDING A CORPUS | 57 |
| HOW TO OFF-LOAD THE TEXT OF A TRANSCRIPTION IN THE CORPUS FILE..... | 57 |
| <hr/> APPENDIX 1: IDENTIFYING ACTIVITY TYPES <hr/> | |
| A LIST OF POSSIBLE FEATURES FOR CLASSIFYING ACTIVITIES | 58 |
| <hr/> APPENDIX 2: PREPARING THE RECORDED MATERIAL FOR TRANSCRIPTION <hr/> | |
| CONVERSIONS OF DATA MEDIA | 61 |
| HOW TO SET UP AND TRANSCRIBE FROM A VCR/TV | 61 |
| HOW TO CREATE A MEDIA FILE AND TRANSCRIBE DIRECTLY ON A COMPUTER | 62 |
| <hr/> APPENDIX 3: EXERCISES AND GUIDELINES ON CORPUS-RELATED SOFTWARE AND PROCEDURES <hr/> | |
| HOW TO TRANSFER A RECORDING FROM A DV CAMERA TO CD/DVD | 64 |
| HOW TO MANAGE INFORMATION ON TRANSCRIPTION FILES AND FOLDERS | 65 |
| RUN THE GORALLT PROGRAM | 67 |
| CREATE CORPUS | 70 |
| STATISTICS | 70 |
| ANALYZING THE RESULTS | 71 |
| GET BAREGREP | 72 |
| BASIC SEARCHES..... | 72 |
| REGULAR EXPRESSIONS..... | 72 |
| <hr/> APPENDIX 4: CORPUS-RELATED SOFTWARE <hr/> | |
| TURNING VIDEO AND AUDIOMATERIAL INTO COMPUTER FILES | 74 |
| TRANSCRIBING | 74 |
| SEARCH AND RETRIEVAL | 74 |
| CONCORDANCE PROGRAMS | 75 |

APPENDIX 5: HOW TO VISIT THE CORPUS WEBSITE AT UNISA 76

Editors' Notes

The primary aim of the Southern African Spoken Language Corpus Project (SOUTH-TALK) is the development of linguistic resources for all the official languages of Southern African in the form of electronically archived multimodal corpora of spoken language use.

It is an inter-institutional project with Gothenburg University and the University of South Africa as the principal administrative institutions for the project. The following universities are equal partners in the project providing research resources for creating corpora for the specified languages of the areas in which they are located:

University of Limpopo (Turfloop campus): *Sesotho sa Lebowa, Xitsonga, Tshivenda*

The Technical University of Venda: *Tshivenda*

Northwest University (Potchefstroom and Mafikeng campuses): *Setswana, Afrikaans, English*

Free State University (Bloemfontein campus): *Sesotho sa Lebowa*

Umtata University: *isiXhosa*

Fort Hare University: *isiXhosa*

KwaZulu Natal University: *isiZulu*

University of Swaziland: *siSwati*

University of Botswana: *Setswana*







University of Lesotho: *Sesotho sa Lesotho*

University of Port Elizabeth: *isiXhosa*

This publication was originally intended to be just a manual that sets out the principles, methods and techniques underlying the recording, transcribing and archiving of spoken language corpora as well as the utilization of spoken corpora in research applications. In the process of writing it, we realized however that the guidelines should be embedded in a more general framework, and that it should also allude to more general pertinent issues. We felt for example, that the question about the identification of activity types or the question about the status of the notion word in agglutinating languages deserve some discussion even in this manual. The manual thus became much more detailed than we had originally planned, as a result of which many parts of it may not be relevant, say to somebody who is just interested in the transcription or the recording guidelines. Thus, although we would advise everybody participating in the project (be they recorders or transcribers) to read through the whole manual, the manual could also be used in a selective manner. Each chapter deals with a specific aspect of corpus development. Recorders could therefore refer to the guidelines on recording set out in Chapter 2. Transcribers could focus on the transcription guidelines in Chapter 3. Transcription checkers could refer to the guidelines on checking set out in Chapter 4.

In order to facilitate the use of the manual a few icons are used to draw the attention of the user to special types of information in the relevant sections. The icons and their significances are listed below.


ICON KEY


| | |
|-----------------------------------------------------------------------------------|----------------------------------|
|  | Terminology |
|  | Note |
|  | More detail |
|  | Example |
|  | Computer-related information |
|  | Principles, rules and guidelines |

We would like to acknowledge the editorial assistance and advice of Penny Sanderson, Elizabeth Ahlsen and Louie Swanepoel in the preparation of this book. It will be an easy way out to lay the blame with them for any remaining errors. Alas, the editors will have to bear the brunt, although Penny, Elizabeth and Louie made it less likely that there will be too many obvious errors.

Finally, without the financial support of SIDA and the NRF and the infrastructural support of our respective institutions, the University of South Africa and the University of Gothenburg, the spoken language project would never have materialized. We are greatly indebted to these institutions for enabling and facilitating our research endeavours in this project.

The Spoken Language Corpus Project

 **The word corpus (plural corpora) here refers to an extensive computer-based collection of linguistic communicative data**

 **The word multimodal refers to information pertaining to more than one production modality (e.g. speech or gesture) or sensory modality (e.g. auditory and visual) modality**

Introduction

This manual presents the principles and methods that will guide the user in the processes and procedures of creating a multimodal corpus of spoken language for the official languages used in Southern Africa. The corpora for the various languages are multimodal in the sense that they capture communication data from both the auditory (speech) and the visual (gestures and facial expressions) modalities.

In this introductory chapter we give a brief outline of a corpus linguistics approach, the uniqueness of spoken language, the spoken language corpus project and the network of participating institutions and their roles. We also suggest some significant research opportunities that should follow from this project and some benefits that should emanate from research applications based on the various corpora.

Electronic network

In order to coordinate and facilitate the activities of the different project participants, an internet network of all the participating institutions is envisaged. Each participating institution in the project is presumed to have access to a PC and an internet connection. At the centre of such a network would be an internet website which would act as the hub to which the websites of participating institutions would be linked. Each linked website would require a page master who should upload new data on a regular basis and maintain the information on that local website. A standardised template for the participating individuals/institutions will be downloadable from the central website.

Initially the central website which would function as a hub will contain the following:

- A short overview of what the project entails.
- The individuals and institutions involved in the project (with links to those individuals and institutions.)
- A clickable map of Southern Africa which indicates the regions where the languages and the institutions researching them are located.
- Samples of transcriptions of spoken data.
- Links to where resources/tools for capturing data and transcription can be found.

Eventually, the website will have links to an electronic database where other researchers can access the transcribed corpora of the project. The central website will be hosted at the University of South Africa at:

<http://www.unisa.ac.za/corpusproject>

If you are unsure of how to browse the internet, link to this website or how to set up a local website look at the relevant procedures outlined in Appendix 2 at the end of this manual.

A corpus linguistic approach

Creation of corpora is a basic feature of what has become known as “corpus linguistics” which actually is less of a type of linguistic theory, than it is an effort to provide linguistics with an increased empirical basis for theoretical work (Kennedy 1998). Thus, most corpora that have been collected have usually striven to be neutral in regard to particular linguistic theories. We might say in this regard that the goal of corpus linguistics is to provide useful empirical resources for linguistic theories in general. In practice, this goal is perhaps not always attained, since some linguistic approaches downplay the importance of recorded material, claiming that intuitions concerning language use are of equal value to recorded examples of actual language use. Such approaches tend not to make heavy use of corpora even though, in principle, they would probably also benefit from a more corpus oriented approach. Another reason why the goal is not always attained is that corpora, because of their size, invite researchers to employ quantitative (often statistically based) types of analysis. This is not a kind of analysis which is always easy to combine with all types of linguistic theory. However, it should be noted that quantitative or statistical analysis is not required to use a corpus. A corpus can also very well be used to look for single examples in order to support or disconfirm a particular theoretical hypothesis.

Why spoken language?

There are many good reasons why creating a multimodal corpus of spoken language is an important challenge. Here are some of them:

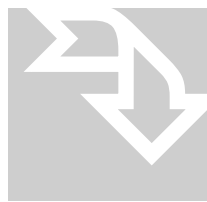
1. Speech is a unique ability of human beings and as such it is intimately tied up with the evolution of the human species. Face-to-face interaction in spoken language combined with gestures allowed sharing of complex information and coordination of complex action and interaction in a way which had survival value. Understanding how spoken language interaction works is therefore a basic challenge for linguistic theory.
2. Another reason for being interested in spoken language is the fact that roughly 4 000 or two-thirds of the 6 000 - 8 000 human languages (Ethnologue) in the world have no written representation. They are purely spoken languages and should for this reason perhaps also be studied as spoken languages, rather than through some expeditiously constructed form of written language. In addition to this, even if a language does have a written code, for many languages in the world, a large proportion of their speakers may still be illiterate or very rarely use the written code of their language. Spoken language thus remains the basic medium of linguistic interaction.
3. Cultural activities, practices and customs go hand in hand with spoken language. As such, spoken language is actually a record of indigenous knowledge systems. A multimodal corpus of spoken language is therefore a natural part of a digital archive of the cultural and communicative practices of a particular group of people.
4. Many practical applications can be derived from an increased understanding of spoken language. For example, consider new possibilities of giving education in spoken language in combination with use of graphics, perhaps employing computers or mobile phones. This would clearly be desirable for illiterate people or for speakers of languages without writing systems, but it would also be desirable in more literate societies. In general, increased understanding of spoken language, based on data in multimodal spoken language corpora, combined with the development of systems for speech production (synthesis) and speech understanding (recognition), might help us to create
 - better programs for education in general
 - better programs for language teaching/learning in particular
 - better programs and learning aids for people with communicative handicaps
 - better speech controlled appliances, such as videos, micro ovens, time-table information etc.

What is included in the corpus?

A common goal in creating a corpus of language use (written or spoken) is to make the corpus representative of as many varieties of language use as is possible. This is not as simple as it may sound since language use can vary in many different ways. There is, for example, regional variation (different dialects), gender variation (male and female speech), age variation (child, adolescent, adult and old age), social class (caste) variation, educational variation and, last but not least, activity variation i.e. variation due to the difference in the social activity for which the language is being used (Allwood 2000; Allwood 2001). In fact, the main reason for choosing the notion of social activity as the basis for selecting material for our corpora is the fact that spoken language can be very different in different social activities. Compare the language used in a church sermon with the language used in a friendly chat, the language used in a patient – doctor consultation with the language used in an auction, the language used in a court of law with the language used in a bargaining session in a market place. Differences between social activities are one of the main reasons for variation in spoken language.

It is virtually impossible to make a single corpus representing all these factors underlying variation in language use. For the purposes of the spoken language corpora in SOUTHTALK (Southern African Spoken Language Corpora Project) we have opted for differences in social activities as the basis for variation in language use. The recordings of speech in this project will therefore be based on differences in social activities. However, we also take into consideration the language use effects of the other factors (such as dialect or gender) where they are noticeable.

Given that we are interested in capturing linguistic variation related to differences in social activity, we have to decide on what activities to record in order to get a good sample of the relevant varieties of language use. Since we are not aware of a general theory or taxonomy of the range of possible human social activities or how frequent the different types of activities are, we can only give a list of activities that are fairly common in many cultures and language communities.



A list of common social activities


formal/friendly conversation
family gathering
party
asking for and giving directions
telephone conversation
taxi driver – passenger talk
bus driver – passenger talk
an interview
bargaining in a market place
shopping (customer – shop assistant talk)
academic seminars
patient – doctor consultation
classroom teaching
formal meeting
auction
court proceedings
parliamentary debates

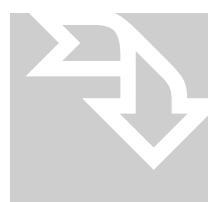
cultural rituals (e.g. initiation, seasonal ceremonies)
religious rituals (e.g. a church sermon or a funeral service)
tribal meetings
chore activities (e.g. washing, cooking, herding)

Even though there is no general theory for the range of possible human activities, it is possible to provide a number of dimensions or features along which human activities can vary. These dimensions may then be used to select activity types for recording in such a way that they are as different as possible from each other, which will give the corpus a better coverage of different types of spoken language use. The list of possible features for the classification of activity types is given in Appendix X. However, since it is also desirable that spoken language use can be compared from one language and culture to another, we recommend that the types of activities given in the list above be recorded in the first place. These activities have been chosen because they are fairly common and have a high degree of autonomy, i.e. they are identifiable as an independent activity.

Corpus-driven and corpus-based approaches

Various types of linguistic and extra-linguistic speech-related information can be retrieved from language corpora (written and spoken) depending on the kinds of information that are represented in the transcription of a recording. Even if a corpus simply contains a collection of written texts or spoken language samples that have been transcribed in the standard orthography of the relevant language very interesting and useful information can be extracted from the corpus. In fact, until fairly recently such “raw corpora” (i.e. corpora containing just texts without any additional annotations) were the norm in corpus linguistic studies. The corpus linguistic approach that works with “raw corpora” became known as the **corpus-driven approach** (cf. Leech 1991). This approach uses a special tool to retrieve information from a corpus, namely a **concordancer**. The application of this tool to retrieve information from a corpus is based on a special method, called KWIC (Key Word In Context) analysis. The researcher names a key word say *run*, and the concordancer then searches the corpus for all the occurrences of *run* as well as the context in which it occurs and expresses the finds in **concordance lines**. Consider a sample of concordance lines that contain the key word *run* below.

 **The word concordance refers to the citation of lines that contain a designated word and a concordancer is a software program that excerpts the relevant citation lines from a text or a corpus**




An example of a key word in concordance lines

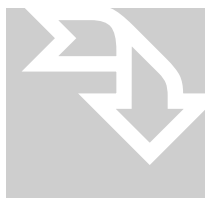
The key word in the concordance lines below is *run*.

Can you run a bath
experience to run a business
you will run the risk
training to run the marathon
trying to run me down
took a run from the

In the corpus-driven approach there is therefore no need for special annotations of the linguistic material in a corpus.

The raw data in a corpus can however also be enriched with various kinds of annotations. These annotations allow far more subtle searches of a corpus. For instance, linguistic features peculiar to female or male speech, linguistic features characteristic of different genres and speech activities, linguistic features characteristic of spoken language such as overlapping as well as various kinds of morphosyntactic facts can be obtained from a corpus as long as these kinds of information are overtly reflected in mark-up of the raw material in a corpus. The mark-up of a corpus is done by means of standardized annotation schemas as well as morphosyntactic **tags** (cf. van Halteren 1999). The corpus linguistic approach that subsumes the annotation of the material in a corpus is known as the **corpus-based** approach (cf. Leech 1991).

 **The word tag refers to the labels assigned to the morphemes and syntactic categories in the utterances contained in a transcription.**



An excerpt from an annotated corpus

The annotations reflect both general and morphosyntactic bits of information.

\$N: <ok>

@ <code switch: *English*>

\$P: <1 and then>1 [wen {a} <<proIIs>>
u<<indIIs>>be<<pastcon>>no<<ass>>m<<n3>>sinde<<n/stem>>]
[u<<indIIs>>m<<oc1>>caphuk<<v/stem>>el<<appl>>e<<perf>>] > <2
but>2 nje<<adv>>
e<<locgen>>ku<<n15>>hamb<<v/stem>>eni<<locsuf>>
kwe<<poss15>>xesha<<n/stem>>ku<<ind15>>za<<futdef>>
{ku} <<inf>>fun<<v/stem>>ek<<neut>> {a} <<basicv>>
u<<subIIs>>m<<oc1>>thand<<v/stem>>e<<subjv>> kuba<<conj>>
ngu<<idc1>>mama<<n/stem>> {wa} <<poss1>>kho<<possproIIs>>
u<<indIIs>>za<<futdef>>
{ku} <<inf>>m<<oc1>>caphuk<<v/stem>>el<<appl>>a<<basicv>>
nje<<adv>> nga<<inscon>>laa<<demIII3>> m<<n3>>zuzu<<n/stem>>

@ <1 code switch: *English*>1

@ <2 code switch: *English*>2

SOUTH TALK follows the corpus-based approach and one of the main objectives of this manual is to present the participants in this project with the standard annotation schema for transcribing spoken language that has been developed in the Department of Linguistics at Gothenburg University (Sweden). The annotation conventions were developed with a view to capturing as many as possible of the peculiarities and characteristics of spoken language.

The corpora in SOUTHTALK will ultimately also be annotated with morphosyntactic tags to reflect morphosyntactic information. In this manual we will not address morphosyntactic tagging since it involves a host of other considerations and issues which are not pertinent to the transcription objectives that we would like to attain in capturing the characteristics of spoken language. The morphosyntactic tagging (both manual and automated) of the corpora in SOUTHTALK will be addressed in the next volume in the Language Corpora Series.

Recording speech

Introduction

The development of spoken corpora, which is the main concern of this project, is characterized by a number of phases. Of these, the recording phase is the most crucial. Without physical samples of recordings there is no data base for corpus linguistic analyses. Since recordings constitute the building blocks in the development of a corpus, the successful execution of other operations such as transcribing of utterances, annotation and grammatical analyses depends on the existence of fairly extensive collections of multimodal communicative linguistic data. This entails the employment of recording equipment such as digital video cameras. Audio recorders will occasionally be used simultaneously with the video cameras in order to improve sound quality or as a back-up. Video recordings are important in order to capture the way we always use our bodies in spoken language. We lose an important source of information about the nature of spoken language if we do not include recordings that allow us to study how body and speech are used together. Another reason to use video recordings is that they give a much better understanding of the context in which spoken language is used. This is essential, since face-to-face spoken language communication depends on activating contextual information to a much greater extent than written language communication.

Data collection methods

Many methods of data collection have been used in linguistics. Some examples are armchair reflection, interviews, questionnaires or experiments. Although compatible with these methods, the main method of data collection in assembling a corpus of interactive spoken language is the observation of authentic interaction through audio and video recording. In the case of a written language corpus, the main method will instead involve collecting specimens of writing. The best solution here will usually be direct use of electronically available printed material, but if this is not available, it can, for example, be made electronically available by a process of scanning. In the worst case it has to be re-typed. Irrespective of whether the corpus contains spoken or written language material, the objective is to get samples of language use which are as natural as possible.

Turning to the specific requirements of collecting data for a spoken language corpus, it is important that the persons who are to be recorded (the speakers) are informed about the recording and its purpose and that they agree to be recorded. This can be done either by asking for permission to record in advance (which is to be preferred) or by giving information about the recording after it has been done and then asking for permission to include it in the corpus. The informants should also always be given the right to delete (partially or fully) what has been recorded. The purpose of assembling a corpus is not normally to record particular persons or specific views. Rather it is to get a window on the language as it is typically spoken. The focus is a general one, namely on how language and communication work, rather than a particular one such as focusing on the content of what particular people say. Thus, as we shall see, it is normal practice to anonymize all persons who contribute to the corpus. Procedures such as these mentioned above are usually not as important in creating a written language corpus, since written language, to a much greater extent than spoken language, is already created for public consumption.

Data storage

A basic step in data collection concerns how the data is stored or registered. Just as there are several methods of data collection, there are several methods of data registration. Some of them are the following: memorization, field notes (or other types of notes), observation protocols, audio and video (analog or digital) recordings as well as electronic texts.

Data collection and data registration/storage are to some extent independent of each other. Thus, the answers given in an interview or an observation of an interaction can, for example, be stored in person's memory, written down on paper or audio/video recorded. Here our focus will be on audio/video recordings since we want to get as close as possible to spoken language as it is actually occurring in a natural situation (in the sense that the recording session is not arranged by the researcher).

Data recording

The following two very important points regarding data recording need to be made:

- *Make sure the data is as ecologically valid as possible.*
What this means is that we should try to record spoken language in typical and naturally occurring activities, not specifically arranged by the researcher for recording. We want our corpus to have ecological validity, i.e. be typical of the environment it represents rather than being the result of an environment created by the researcher.
- *Make sure the data is of the highest possible quality.*
The second requirement is that all recordings should be of a high audio and video quality. This is essential if it is going to be possible to transcribe, analyze and annotate the recordings later.

Unfortunately the second requirement of high quality sometimes comes into conflict with the first requirement of ecological validity. When this happens we should always attempt to meet both requirements as far as possible. In order to do this the following might be helpful:

To be able to fit all participants into the picture, the camera has to be placed at least a few meters from the participants. The LCD screen on the camera should be used to ensure that all the participants fit into the view of the camera. Unfortunately, increasing the distance between the camera and the participants may have an adverse effect on the audio recording quality of the internal microphone of the camera. Therefore an external microphone closer to the speakers is preferable, and even better would be to have one separate microphone for each participant, attached to their clothes. If separate microphones make the conversation less natural, then stick to one external microphone.

Moving the camera to zoom in on the current speaker all the time will make it impossible to study the interaction between the speakers. It will also distract the participants more if the recorder is operating the camera all the time. It is much better to keep the camera fixed on a tripod in one location with a view of all participants.

For the field worker to make a good recording the following guidelines should be followed:

- In order to get the full cooperation of the participants in a speech event that will be recorded, it is necessary to explain the nature and the purpose of the project to them.
- One should also address the possible concerns of the participants about being recorded on camera and how the recorded material will be handled. The participants should be made aware that the content of what is said in the recordings is of secondary importance, since we are interested in how things are said.
- To facilitate the transcription of facial expressions and gestures all participants in a speech activity should be captured simultaneously.
- The camera should focus on all the participants involved in the recorded activity and not shift from speaker to speaker.
- Avoid having any light(s) including direct sunlight in the background as this will have an adverse effect on the picture quality of the video.
- Remain as unobtrusive as possible so that your movements and actions do not distract the participants.
- Finally, make sure that you can hear what everyone is saying and that you can fit in as many of the participants as possible on the monitor of the camera.

Finally, ensuring that there are enough cassettes and charged batteries before embarking on a recording trip cannot be overemphasized.

Information about a recorded spoken language activity

A very important task of the recorder of a spoken language activity is to gather all the relevant information about the activity. This information is of crucial importance when the recording is transcribed as you will see in Chapter 3. For this purpose we have designed a form that must be filled in by the recorder. Relying on one's memory to recall all the relevant information some time after a recording may be disastrous, causing the loss of important information about the activity. Copies of this form must therefore be available during a recording field trip and should be filled out during or immediately after the recording. Following this procedure is not only essential, but may also enable the recorder to clarify relevant issues (such as the anonymity of the participants, omission of parts of the recording, dialectal variations, etc).

The recording form given below specifies all the categories of information that are required in the header of the transcription of a recorded activity as will be explained in Chapter 3. Recorders should however feel free to include under the comment heading other bits of information about the recording or the activity that they may regard as important for some or other reason.

In most cases recorders will probably do recordings on several tapes during a fieldwork trip. In order to ensure that each recording form is linked to the relevant tape containing the recorded activity, the following procedures should be followed:

- Assign a number to the tape that you are going to use to record the current activity. Write this number together with the date and your name on a sticker and paste the sticker on the tape.
- Write the number of the tape, the date and your name also on the tape box.
- Enter the tape number, the date and your name on the recording form in the relevant spaces.
- In the event that more than one activity is recorded on the same tape a separate recording form for each one of the activities should be filled in. The number of the tape containing the various activities should be entered on each one of the recording forms for the respective recorded activities.
- In the event that several tapes have been used to record the same activity, each tape should have a different sequential number reflecting the sequence in which they have been used to record the activity. Since only one recorded activity is involved, there will be only one recording form. The various tape numbers should therefore be entered on the recording form.

- After the recording of an activity, write the name of the activity on the tape box. If more than one activity was recorded on the tape, all of the activities should be listed on the box.

RECORDING SESSION FORM

1. **Number of tape(s) used for the recorded activity:**

2. **Name of Recorder:**

.....

3. **Date of Recording:**

.....

4. **Time of Recording:**

.....

5. **Duration of Recording:**

.....

6. **Locality of Recording:**

.....

7. **Main language used in recorded activity:**

.....

8. **Recorded activity title:**

.....

(E.g. informal conversation; formal meeting; lecture; medical consultation; religious service; shop; bus driver/passenger; court case; group discussion)

9. **Number of participants:**

10. **Profiles of participants (name; approx. age; level of education; gender; dialects, etc.):**

@Participant 1:

.....
.....
.....
.....

@Participant 2:

.....
.....
.....
.....

@Participant 3:

.....
.....
.....
.....

@Participant 4:

.....
.....
.....
.....

11. Comments:

.....
.....
.....
.....
.....
.....

What to record – activity types

In view of the fact that we are interested in recording varied social activities representative of the language concerned, the following provisional list of communicative activities is provided to guide the field worker:

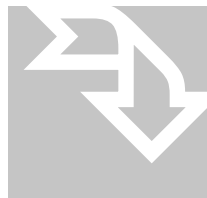
- informal / friendly conversation
- a family gathering
- a party
- requesting or giving directions
- telephone conversation
- taxi driver – passenger talk
- bus driver – passenger talk
- an interview
- bargaining in a market place
- shopping (customer – shop assistant talk)
- academic seminars

patient – doctor consultation
classroom teaching
a formal meeting
auction
court proceedings
parliamentary debates
cultural rituals (e.g. initiation ceremonies)
religious rituals (e.g. a church sermon or a funeral service)
tribal meetings
chore activities (e.g. washing, cooking, herding)

Since this list is not exhaustive, field workers are encouraged to identify other social activities to supplement those mentioned above. In general, however, we recommend that only activities which contain significant portions of spoken language be included. Thus, ceremonies of various types containing mostly dancing or music are not a good choice for recording, since they do not contain the kind of data which the project is primarily aimed at, namely spoken language use.

What to do with the recorded material

Each institution that participates in the project should set aside a special room that should be appropriately furnished with lockable cupboards where the recorded tapes could be stored in the interim before they are transcribed and where they could be archived after they have been transcribed. Preferably, this room should also be equipped with (a) suitable computer(s) on which the transcriptions will be done.



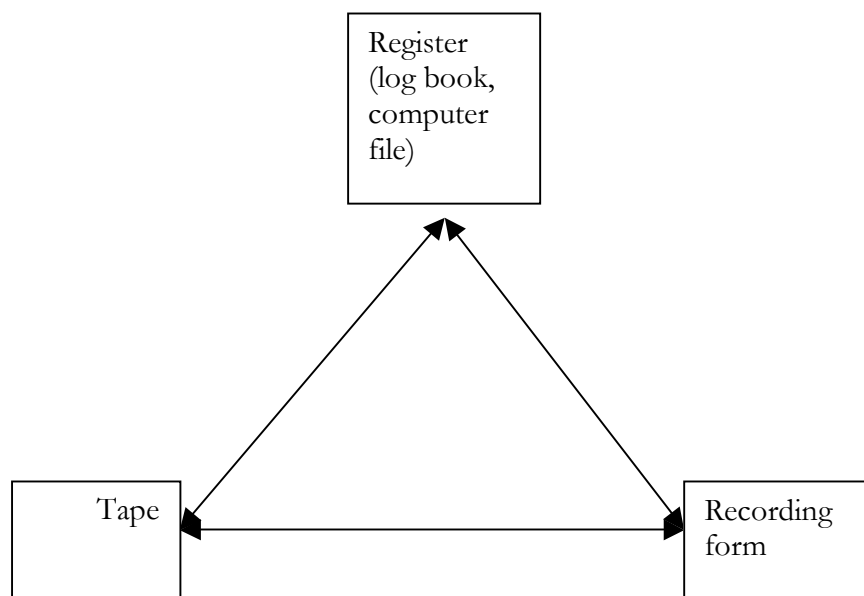
Specifications for a transcription computer

The computer(s) that will be used to do the transcriptions should at least be equipped with the following hardware:

- A Firewire card (enabling the transfer of the digital video material on the digital camera to a video file on the computer).
- A network connection (enabling access to the internet).
- A DVD writer (for backup purposes).
- A reasonably sized hard disk.

To keep track of all the recorded activities, tapes and recording forms a register (both in paper and electronic format) containing the relevant information should be kept. In addition, one should establish some or other link between the various items mentioned above since they will be kept in separate places - the recording

forms will be kept in a file, the tapes will be shelved and the register will be in a log book and in a file on the computer. The diagram below illustrates the nature of this link.



In order to link these three items we need an index that will be common to all three of them. The index should be based on the **tape number**, the **name of the recorder**, and the **date of the recording**. Following the procedures outlined above, these bits of information should have been written on the tape sticker and should have been entered on the recording form, but must now also be entered in the register (log book and computer file). Although the same number may appear on various tapes (i.e. tapes used on different recording occasions and by different recorders) the recorder's name together with the date of the recording will ensure that the index remains unique and can be linked to the recording session form in each case.

The coding procedure that we have described here is an interim measure in order to keep track of the recorded material. Each tape (and its backup) will receive another more permanent unique code (for archival purposes) during the transcription phase. The composition of this code will be explained in Chapter 3.

The register (both the log book and computer file) should contain the following information that can be obtained from the recording form.

- Tape number
- Recorder's name
- Recording date


- Language
- Activity name (If more than one activity was recorded on the tape, they should be listed separately.)
- Recording transcribed or not



Example of the recording register

Note: The same table format should be used for the log book as well as for the register file on a computer. The register should be updated (by a person designated to carry this responsibility at each institution) whenever recorded material is returned to the transcription room.

| Recorder | Date | Tape Nr | Language | Activity | Transcribed |
|----------|----------|---------|----------|------------------------------------|-------------|
| Sekere | 15/06/04 | 1 | Sotho | Formal Meeting | No |
| Sekere | 15/06/04 | 2 | Sotho | 1. Friendly conversation | No |
| | | | | 2. Lecturer-student discussion | Yes |
| Mmemezi | 13/02/05 | 1 | Xhosa | Informal conversation with patient | Yes |
| Nozibele | 16/04/05 | 1 | Xhosa | Interview | Yes |

 **A standard is a set of norms or conventions for a particular purpose**

Transcription Standard

Introduction

Since present day language technology does not allow us to work directly on speech and gestures in audio and video recordings of spoken language activities, we have to approach spoken language interaction by transcribing or writing down what is said and done. For this reason transcriptions are still very important in the corpus linguistics study of spoken language use. In order to use computers to analyze the corpus and in order to find phenomena that are of a similar kind, it is vitally important that all the transcriptions of recorded spoken language samples should conform to a uniform transcription format. In this chapter we describe and illustrate the transcription standard for computer-readable transcriptions of audio and audio-video recordings of spoken language. The standard used in the SOUTHTALK project is based on the standard that has been developed in the Department of Linguistics at Gothenburg University (Allwood 1999; Nivre 1999; Nivre et al. 2004) with appropriate adaptations for the African Languages of Southern Africa.

The goals of the standard are the following:

- it should enable the transcriber to capture as many of the essential features of spoken language interaction as possible;
- it should be reliable and yield consistent transcription results, i.e. it should be fairly easy for different transcribers to obtain the same transcription results of a recorded speech event;
- it should be easy to learn and apply reliably (hence the omission of certain aspects of spoken language);
- it should enable comparison between spoken and written language use (thus the spoken language transcription standard should not unnecessarily be unlike the standard for written language).

The standard is therefore an attempt to create a way of transcribing spoken language which will enable us to capture the most important features of spoken language, while allowing comparison with written language, and at the same time being both easy to learn and easy to apply consistently.

This means that not everything that is important in spoken language is captured in the transcription. In particular, most features related to prosody and intonation are left out, since they are not easy to capture in written language, are not easy to learn to transcribe and usually result in low reliability.

The parts of a transcription

Every transcription consists of two parts in the order shown below:

- Header
- Body

In the following sections the transcription conventions applicable to each one of these two parts of a transcription will be explained and illustrated. First, we consider the header.

Header conventions

The header is always the first part of a transcription and it captures all the relevant information about a recorded linguistic activity as well as various bits of information about the transcription itself. Information about the recorded linguistic activity (e.g. the name of the recorder, the participants, the duration of the recording, etc) can be obtained from the recording session form. The header therefore contains various types of information and each type is represented on a separate information line. Each information line of the header is introduced by the character @ followed by a space and the name of the type of information, followed by a colon, followed by a space and then the relevant information. The structure of a header line can thus be represented by the following template:

@ Type of information: relevant information

In the section that follows we give examples of the various header lines according to the types of information they contain together with explanatory notes on each type of line.

Header lines

@ Recorded activity identification code (ID): F-XV-01-13-01

The recorded ID code is unique to this particular recorded activity in the corpus and it should encode the following bits of information in a sequence of letters and numbers separated by hyphens:

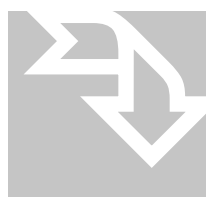
- The institution where the recording was made and where the material is archived. This bit of information is encoded in the first letter of the code, e.g. F (Fort Hare) in the code above.

- The language of the recorded speech activity and the recording medium (video or audio). The letters XV in the code above capture these two bits of information, X representing Xhosa and V representing Video. In the case of an audio Xhosa recording the code would be XA.
- The number of the research project. The digits 01 in the code above show that this is a transcription of material in the first corpus project, namely spoken language corpora. A second project might be on written corpora and would be numbered 02.
- The number of the tape in the archived collection of tapes in a specific medium (audio/video) containing the material that is currently being transcribed. The number 13 shows that this is the 13th tape in the collection. Note. This is **not** the same number as the one that was assigned to this tape by the recorder when it was used to do a recording. Rather it is a unique number of a video or audio tape within the archived collection of video or audio tapes. The transcriber assigns this unique number to the tape when the material on the tape is transcribed.
- If more than one tape has been used to record the same activity, all the tapes receive the same number according to the convention described above, but they are distinguished from each other with a decimal number following the main number. Say, a recorded activity runs over 3 tapes, then all the tapes will receive the same number, e.g. 18, but the first tape will be numbered 18.1, the second 18.2 and the third 18.3.
- The number of the activity recorded on a tape that is currently being transcribed. Each separate activity that has been recorded on the same tape is sequentially numbered and separately transcribed. The last two digits in the code, 01, show that this is either the only activity that has been recorded on this tape, or that it is the first of more than one activity that has been recorded on this tape.

The following template shows the sequence of the various bits of information in the transcription ID code:

| Institution | Language and recording medium Video/Audio | Project nr | Tape nr | Activity nr |
|-------------|----------------------------------------------|------------|---------|-------------|
| | | | | |

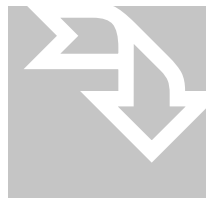
Each one of the participating institutions has a unique code as indicated in the list below:



List of institutional codes to be used in the transcription IDs


| | | |
|-----------------------------------------|---|---|
| Fort Hare | F | |
| Free State University | S | |
| KwaZulu-Natal | | K |
| Northwest University (Mafikeng) | | M |
| Northwest University (Potchefstroom) | P | |
| Technical University of Venda | | V |
| University of South Africa | | U |
| Unitra | T | |
| University of Botswana | B | |
| University of Lesotho | L | |
| University of Swaziland | W | |
| University of Limpopo (Turfloop campus) | | N |
| University of Port Elizabeth | | E |

Similarly, every language in the project has a unique code:



List of language codes to be used in the transcription IDs

| | | |
|---------------------|---|---|
| isiNdebele: | | N |
| Sesotho sa Lebowa: | S | |
| Sesotho sa Lesotho: | L | |
| isiSwati: | I | |
| Xitsonga: | T | |
| Setswana: | W | |
| Chivenda: | V | |
| isiXhosa: | X | |
| isiZulu: | Z | |
| Afrikaans: | A | |
| English: | E | |

 **The plural (s) in brackets indicates that an information line may contain a list of items, separated by commas.**

@ Name of recorder(s): Mvuyisi, Mmemezi

@ Duration (of recorded activity): 01:31:16

The duration of the activity in the format hh:mm:ss (hours:minutes:seconds).

@ Recorded activity date: 2004/03/01

The date of the recording in the format yyyy/mm/dd.

@ Recorded activity type: Formal meeting, staff meeting

The name for an activity type should be indicative of the nature of the activity, e.g. “formal meeting”, “shopping” or “informal conversation”. The list of activity types given in Chapter 2 in the section, What to Record, could be used as a guideline for identifying activity types. Where sub-types of an activity type can be distinguished the sub-type should be also be indicated. For example,

the activity type “formal meeting” may have sub-types such as “committee meeting” and “political meeting”.

@ Recorded activity title: A formal meeting of the staff members of the Dept. of African languages discussing examination results

It is important that the title contains at least a key word that identifies the activity type and correlates it with similar activities, e.g. a telephone conversation, a dialogue between a shop assistant and a client, a conversation between family members, etc.

@ Short name: Teaching staff meeting

A short version of the title to be used in transcription inventory tables, etc. A key word for the activity (e.g. shopping, sermon, lecture, etc.) may be ideally suited for the purposes of creating a short name.

@ Recorded activity location: Fort Hare University

The place where the activity took place together with other relevant locational information, e.g. Department of African Languages.

@ Activity mode: face-to-face

The concept of mode refers to the nature of the communicative interaction as well as the communication medium. To simplify matters, we suggest that only the following terms be used: face-to-face, telephone, face-to-face (TV), face-to-face (radio), one-way (TV), one-way (radio). By “face-to-face (TV)” is intended, for example, a televised discussion where there is interaction both in front of the camera and one-way communication with the viewers. “One-way (TV)” refers to the case where there is only or mostly one-way communication with the viewers, e.g. a news broadcast.

@ Participant: N = F61 (Nomsa)

First name initials are used to identify the participants in a spoken language activity, e.g. N for Nomsa in the information line above. Every participant is also assigned a unique gender-number code which is used throughout the corpus for that individual in all the recorded activities that he/she participated. In the information line above F61 is the unique participant code for Nomsa. Four symbols are used to differentiate the gender of the speaker, namely M (adult male), F (adult female), B (boy), and G (girl). Transcribers must keep a record of these codes in order to assign the correct code to the same individual in the transcription of the different recorded activities in which he or she participated and to assign unique codes to “new participants” in the corpus.

To protect the anonymity of speakers, pseudonyms should generally be created to replace all names which could identify participants. The initial of the pseudonym should then be used in the header to identify the speaker and the full pseudonym where the relevant name is used in the contributions.

Each speaker should be specified on a separate participant information line in the header, e.g.,

@ Participant: N = F61 (Nomsa)

@ Participant: V = B15 (Vuyiso)

If two participants have the same first name or their names have the same initial, the initial of one of the participants can be doubled in order to distinguish the two, e.g.,

@ Participant: N = F61 (Nomsa)

@ Participant: NN = F73 (Nozibele)

Sometimes the participants take part in a speech activity as a group, say an audience listening to a speech or a congregation listening to a sermon. In such cases it would be impractical if not impossible to identify all the participants individually. Yet, it is still important to note their contributions in the speech activity, e.g. overlapping applause, overlapping interjections, etc. Participant groups are also indicated on a separate participant line similar to individual participants, but in such cases we use the letters GR to indicate that the relevant contribution is by a whole group.

The participant line for groups in the header should then be:

@ Participant GR = Group (audience)

or

@ Participant GR = Group (congregation)

@ Tape(s) ID code: F-XV-01-13

This is the name of the tape (including its backup CD or DVD) on which the recording was made. The tape code, excluding the activity code, is identical to the recorded activity ID.

NB! This code must be written by the transcriber on the tape, backup CD or DVD and their covers

If more than one activity was recorded on a tape the recorded activities are numbered according to the sequence on which they occur on the tape and these numbers are written on a sticker on the tape as well as on the cover of the tape in the following manner.

F-XV-01-13

01: *Activity name*

02: *Activity name*

03: *Activity name*

If the recorded activity runs over more than one tape, the same tape code appears on all the tapes, but, as we have explained before, the tapes are distinguished from each other by means of a decimal number added to the main number. E.g. if three tapes were used to record the activity, they should be represented as follows:

@ Tape(s): F-XV-01-13.1-01, F-XV-01-13.2-01, F-XV-01-13.3-01

@ Transcription ID: F-XV-01-13-01-T1.

The transcription identification code is exactly the same as the recorded activity ID, but for the addition of two digits, cf. T1 in the example. These digits indicate that it is the first transcription of the recorded activity. Subsequent transcriptions of the same recorded activity should be reflected accordingly in the transcription ID, e.g. T2, T3, etc.

@ Transcriber(s): Mvuyisi, Nozibele

The name(s) of the person(s) who transcribed the recorded activity. If more than one person was involved in the transcription, all of them should be listed. The place where another transcriber took over the transcription should be indicated in a comment line in the transcription.

@ Transcription date(s): 2004/10/12-25

This is the date on which the transcription was made in the format yyyy/mm/dd. If the transcription was done by different transcribers on different dates, these dates must be listed on the same line separated by commas.

@ Transcription system: Standard Xhosa orthography

The orthography used in the transcription, e.g. South African orthography or Lesotho orthography in the case of a transcription of Sotho in the Lesotho orthography.

@ Electronic checking: Nozibele

The name of the person who did the electronic checking of the transcription (cf. Chapter 4 for quality assurance procedures). This procedure should preferably be followed by the transcriber immediately after the transcription.

@ Editor(s): Mmemezi Mfusi

The name(s) of the person(s) who edited the transcription (cf. Chapter 4 for quality assurance procedures).

@ Checker (s): Ncedile Saule

The name(s) of person(s) who checked the transcription against the recording (cf. Chapter 4 for quality assurance procedures).

@ Section title: Welcome

If different sections can be distinguished in the activity, each section title should be listed on a separate line, e.g.,

@ Section title: Tabling of minutes

@ Section title: Discussion

@ Time coding: None

The duration of each section should be indicated on a time line in the transcription that gives information about the amount of time elapsed from the start of the recorded activity. A time line in the transcription body is started by the special character #, and the time is given in the format hours:minutes:seconds, e.g. 00:13:24.

@ Comment:

Any additional information that the transcriber may want to add about the recorded activity, e.g. the quality of the recording.

The transcription body

The body of a transcription is the second and the most important part of a transcription. It represents in various ways as realistically as possible the communicative actions and utterances of a spoken language activity. The transcription body consists of two kinds of information:

- Contributions (the transcription of the spoken language of the interlocutors)
- Information lines (various bits of information about the contributions that are added by the transcriber)

Before we take you step-by-step through the transcription procedures we need to give general outline of the principles underlying the transcription task.

 **An interlocutor is a participant in a speech event**

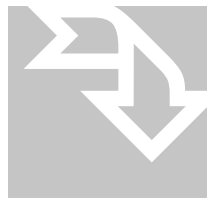


General transcription principles and rules

1. Transcription text format

The transcription text should be **unformatted**. You could do the transcription in a word processor such as Microsoft Word or WordPerfect, but you need to save the file in Plain Text format. This format

makes it easier to move the transcriptions from one computer to another and makes it also easier to analyze the transcriptions.



How to create an unformatted text file

With a blank document open in your word processor do the following:

Step 1. Click on *File*

Step 2. Click on *Save as ...*

Step 3. In the *Save As* window open the options in the *Save as type* bracket

Step 4. Click on *Plain Text*

Step 5. Click on *Save*

The text of the transcription will now be unformatted.

2. Punctuation

Neither punctuation marks (i.e. full stop, question mark, comma, etc) nor capital letters (at the beginning of an utterance or for proper names including place names) may be used in the transcription. Thus, punctuation marks are omitted everywhere in the transcription of the utterances and capital letters are written as normal letters. For example,

yiza kum mmemezi ndifuna ukuthetha nawe

(Notice the absence of the capital letter (Y) in *yiza*, the omission of the comma (,) after *kum*, the absence of the capital letter (M) in the name *mmemezi* and the omission of the exclamation mark (!) after *mmemezi*, the absence of the capital letter (N) in *ndifuna* and the omission of the full stop after *nawe*.)

o kae

(Notice the absence of the capital letter (O) and the omission of the question mark (?) after *kae*.)

uphuma kwazulu_natal

(Notice the absence of the capital letter (U) in *uphuma* and the capital letters (K, Z and N) in *kwazulu_natal*.)

3. Numerals

Numerals are always transcribed with letters, never with numbers. For example, the number 16 will be transcribed the way it is said, namely *sixteen*. This also applies to dates, e.g. the date 3 May 2005 should be transcribed as it is said, namely *the third of may two thousand and five*.

4. Transcription authenticity

It is extremely important that a transcription captures certain important features of spoken language. Spoken language typically contains communicative sounds that would not be treated as standard word forms in written language. Yet, they are communicatively meaningful in that they are used by the interlocutors to indicate (dis)agreement, queries, support, understanding, etc. These sounds are collectively called **feedback** (cf. Allwood, Nivre et al. 1993; Allwood 2000) abbreviated as **fb**. Speakers also use such sounds in the management of their own communication in order to keep the floor while organizing their thoughts or to solicit some response from the other interlocutor(s). These are collectively termed **own communication management** abbreviated as **ocm**. From a spoken language point of view these sounds must therefore be regarded as “words”. Consider some of these spoken language words.



Examples of feedback and own communication management “words”

Feedback:

mh (query)

h' (disagreement)

e: (agreement)

h (request for response or clarification)

Own communication management:

hm hm hm (keeping the floor while organizing thoughts)

There is currently no orthographic standard for the representation of feedback and own communication management “words” and transcribers very often have to devise their own representations. Nevertheless, there is a need for some form of standardization of such representations which is not only applicable to a specific language but which could possibly be applied across languages, specifically the African languages included in this project. For the purpose of standardizing the representation of these expressions, we propose the following guidelines and procedures.



Principles and procedures for standardizing fb and ocm expressions

Principles:

- The transcription of fb and ocm expressions should be as phonetic as possible, using articulation as a cue for the selection of appropriate orthographic symbols. With the exception of the diacritic for the glottal stop (the superscript comma as in *h'*), the use of IPA diacritics should be avoided.
- The transcriptions of fb and ocm expressions should use the written standard forms for these expressions where they exist. However, the duplication of vowels to indicate length typically used in written language forms should not be followed.

- Where new word forms for fb and ocm expressions are created the transcription should as far as possible be based on the standard orthographic symbols for the language in question.
- Once an orthographic representation for a fb or ocm expression is created it should be consistently used in all the transcriptions for the language in question throughout the corpus.

Procedures:

In many instances fb and ocm expressions are very similar in various languages. In order to reflect this similarity and also in order to harmonise and standardise the orthographic representation of these expressions we suggest that the procedures outlined below should be followed:

Following the principles guiding the creation of an orthographic representation of fb and ocm expressions given above, a transcriber “proposes” a possible orthographic representation for a fb or an ocm expression. In the transcription the representation is enclosed in comment brackets, i.e. < >. The transcriber then gives a brief but clear characterisation of the significance of the expressions in a comment line. Consider a few Xhosa examples of such characterisations in comments for the expressions below which would have been transcribed and enclosed in angle brackets in contribution lines:

\$A: ... <rha>
 @ <fb: *expression of disgust particularly with regard to a bad smell*>
 \$A: ... <shu>
 @ <fb: *expression of pain when burning*>
 \$A: ... <heheyi>
 @ <fb: *expression of joy*>
 \$A: ... <yho>
 @ <fb: *expression of surprise*>

The transcribers and the checkers of the transcriptions at each one of the institutions participating in the project should have regular editorial sessions to discuss the proposed orthographic representations of these expressions. Once this “orthographic committee” has reached agreement on the orthographic representations the proposed representations should be send to the project editor at the following address:

Mr JM Mfusi, Dept of Linguistics, PO Box 392, Unisa, Pretoria. 0003

The editor will on a regular basis circulate the standardised orthographic representations for the various expressions for the various languages to the project contact persons at the various institutions.

Another kind of vocalization that is typical in spoken language is sound imitations such as the imitations of animals and objects. These sounds, with the exception of ideophones, should not transcribed but should rather be treated as vocal gestures

and represented in the standard manner applicable to all gestures as will be explained further on.

Spoken language use characteristically shows certain “deviant” forms that would be regarded as ungrammatical or non-standard expressions. Some of these are illustrated below.



Examples of “deviant” expressions in spoken language use

Repetitions: ndiyazi ndiyazi

Incomplete utterances: *Andazi ukuba ...*

Self-interruptions (false starts): *UNozibele mandithi uBuli ...*

Truncation and word merging: ufike ebusuku > ufik' ebusuku

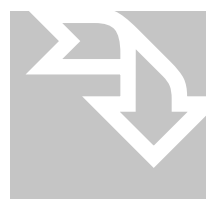
Furthermore, spoken language is typically accompanied by communicatively significant gestures and facial expressions. In order to make a transcription as authentic as possible, conventions have been developed to capture some of the most important spoken language features in the transcription. These conventions will be explained and illustrated when we deal with the actual transcription further on. The important point is that the transcriber is **not an editor**. The transcriber must transcribe what he/she actually hears and sees, not what he/she thinks is grammatically correct or stylistically more appropriate.

5. Language mix

The languages of the world have always influenced one another and have adopted linguistic features and material from each other. Nowadays, particularly with the advances in information technology, spoken language use in many languages shows various degrees of English vocabulary and phraseology. The Southern African languages are no exception to this trend. Since these foreign intrusions are very much part of spoken language use they are transcribed as they appear in the recording, but commented on by the transcriber depending on the degree and nature of the intrusion as we will show you when we deal with the comments in more detail further on.

Language mixture takes on various shapes as the examples enclosed in angle brackets in a transcription of Xhosa show:

njengoba <i aids> < especially> < esouth africa> < istatistics> sesona lizwe
liphuma < if> < bekuwinwa> < iaward> < igold medal> besizoyifumana
< because i think> thina < we are more important> uba < it is> < yireality>
u{ku}ba < i aids> < it kills really then>



A typology of language mix

The following types of language mix are distinguished in this project,

Code switch:

The whole expression (word, phrase, sentence or even one or more sections) comes from another language. The

imported expression is not integrated into the language phonologically or morphologically, for example *we are more important* in the excerpt above.

Code mix:

An expression contains grammatical and lexical elements from two different languages. Normally, the lexical material comes from the foreign language, say English, and the grammatical elements from the base language of the speech event. The respective pronunciations of the languages that are mixed remain more or less intact and the orthographies of both languages are retained in the transcription.

i aids, esouth africa, istatistics, bekuwinwa, iaward, igold medal, yireality

Adoptive:

The lexical material from the foreign language is adapted to the morphological and phonological systems of the base language. In the case of adoptives the standard orthography of the base language is used as far as possible

Uyafonisha ('phone'), *ibhegi* ('bag'), *iyainkonviniensha* ('inconvenience')

Where the foreign language material has been fully indigenized and standardized over a period of time, it is no longer regarded as foreign language matter, for example *isikolo*, *istrato*, *ivenkile*.

Instances of language mix are transcribed according to the orthographic principles outlined above, but they are enclosed in angle brackets in the transcription and commented in comment lines as illustrated below.



Examples of comments on language mix

\$A: njengoba <1 i aids>1 <2 especially>2 <3 esouth africa>3 <4 istatistics>4 sesona lizwe liphuma <5 if>5 <6 bekuwinwa>6

@ <1 code mix: *English* {*aids*}>1

@ <2 code switch: *English*>2

@ <3 code mix: *English* {*South Africa*}>3

@ <4 code mix: *English* {*statistics*}>4

@ <5 code switch: *English*>5

@ <6 adoptive: *English* {*win*}>6

(The conventions used in comment lines will be discussed and illustrated further on.)

6. Contributions and overlap

Contributions are the basic units of the body of a transcription. For the purposes of a transcription a contribution can be defined as the communicative activity (both verbal and non-verbal) of a single interlocutor. That is, a contribution always belongs to a single interlocutor. This means that even if more than one interlocutor says the same thing (or performs the same communicative act simultaneously) their verbal or non-verbal acts are separately and sequentially represented in the transcription.

Spoken language is typically interactive. This means that participants in the interaction in one way or another respond to the speaker holding the floor during or after his/her turn. These feedback responses come in various forms: verbal as well as non-verbal e.g. gestures and facial expressions.

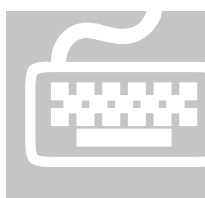
The feedback of the non-turn-taking interlocutors can **overlap** with the speech of the current speaker. (Note that overlaps occur in all kinds of situations when two or more persons are speaking simultaneously, not only in feedback situations) The fact that overlaps coincide with the speech of the turn-taker

presents a problem for the transcriber. This problem has been solved in the Gothenburg transcription standards by means of a simple convention. We will discuss and illustrate this convention in more detail when we take a closer look at the transcription standards. For now, it is important to note the rule that underlies the convention:



The overlap rule

The contribution of a speaker is fully transcribed without any breaks, i.e., line breaks, irrespective of how long it is, irrespective of the length of pauses in the speech of the speaker, and irrespective of any overlapping response(s) (verbal or non-verbal) of the other interlocutor(s). The verbal and non-verbal responses of the other interlocutors in the speech event are treated as separate contributions and are transcribed as such. That is, they are transcribed sequentially on separate contribution lines.



Note that overlaps occur in all kinds of situations when more than one person is speaking simultaneously, not only in feedback situations.

For instance, two participants may start their respective contributions at the same time.

7. Line formats of contributions

The transcription of every contribution should start with the speaker sign \$, followed by the speaker's initial, a colon, a space and the transcription of the contribution, e.g.,

\$N: emnandi yona

8. Words

The transcription of a word must always be continuous. No special symbols other than the letters making up a word may be inserted anywhere in the word. There is one exception to this principle, namely the glottal stop that very often occurs in feedback words. The glottal stop is represented by the apostrophe e.g., h' (indicating disagreement).

In written text words are typically separated by means of surrounding spaces. Although typographical space works reasonably well in marking words orthographically especially in isolating languages, there are many instances where it is not a good indicator of word-like elements. Consider some of these problematic cases illustrated below.

Disjunctive orthographies

In the orthographies of some agglutinating language, such as the Sotho languages, certain prefixes (which are not considered independent words) are also separated from one another and from the lexical stem by typographic space. For example,

O a tsamaya.

Place names

Place names further illustrate the problematic nature of dealing with word status in orthographies. Strictly speaking, a name, irrespective of the number of elements that it is made up of, should be represented orthographically as a single word, but this is rarely the case:

New York, South Africa

Newfoundland [New(ly) found land], *Mpumalanga* [phuma ilanga]

Idioms and fixed phrases

For various reasons idioms are regarded as non-compositional elements. That is, they are treated as unitary expressions even though they are made up of words. However, the words in an idiom lost their independent meanings and together assumed a collective meaning in a manner similar to lexical item.

kick the bucket

amaphuthi ahlath' inye

In addition to idioms there are many expressions that have become so conventionalized that they have also acquired unitary status like lexical items. They are so pervasive in language that we very often overlook the fact fixed phrases.

thank you, excuse me, by the way, of course

kanti ke, ke kaloku ke, into yokuba

These orthographic issues have implications for corpus linguistic studies both in the transcription of the expressions as well as in the subsequent retrieval of various kinds of information for research purposes. On the one hand, the orthographic separation of morphological elements may adversely affect searches that involve word frequency counts of a corpus. Such searches are based on orthographic space and the disjunctively written morphological elements in the Sotho languages will therefore count as words. On the one hand, the spelling of some proper names made up of different words may yield rather distorted results in certain corpus searches. Since capital letters may be not be used in the body of a transcription as was pointed out in the principles and rules of transcription the *new* in *New York* will not be recognized as part of a proper name, but rather as an adjective. Finally, fixed phrases very often function as various unitary word categories such as adverbs, conjunctives and complementizers, which is not reflected in their spelling. A search for

these kinds of expressions will only be possible if they are represented as units in the transcription.

In order to express multi-member expressions as units, the transcriber should use the following convention.



The underscore convention

The various elements of unitary expressions (morphological complexes, place names, idioms and fixed phrases) that are separated by orthographic space in the written standard should be linked by underscore, for example,

o a tsamaya should be transcribed as *o_a_tsamaya*

New York and *South Africa* should be transcribed as *new_york* and *south_africa* respectively

amaphuthi ablah' inye should be transcribed as *amaphuthi_blah_inye*

kanti ke and *ke kaloku ke* should be transcribed as *kanti_ke* and *ke_kaloku_ke* respectively.

9. Numerals

Numerals, wherever they occur (dates, time, etc), should always be transcribed with letters. Consider the transcription of numerals in the examples below:

Quarter past 9 transcribed as: *quarter past nine*

10 Rand 95 cents transcribed as: *ten rand ninety five cents*

I want 3 apples transcribed as: *i want three apples*

Let us now consider in more detail the transcription standard as a whole.

Conventions used in the transcription of the body

The body of a transcription consists of four types of lines

Contribution line contains the transcription of a single interlocutor's contribution. A contribution line is identified by the dollar sign (\$) which is therefore the first symbol of such a line.

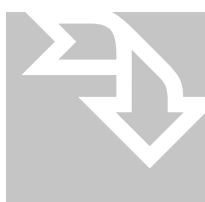
Comment line contains the comment(s) of the transcriber on some aspects of the foregoing transcribed contribution. A comment line is identified by the at sign (@). Comment lines are therefore always preceded by this symbol.

Time line indicates the duration of a selected part (usually a section) of the recorded speech event. The number sign (#) indicates a time line and it is the initial symbol of such a line in the transcription.

Section line presents the title of a delimitable section of the speech event. The paragraph sign (§) introduces a section line.

Contribution lines

In this section we focus on the conventions that apply to the transcription of contribution lines.



How to transcribe a contribution

A contribution is started by a dollar sign (\$), followed by a speaker initial, a colon and a space, e.g. \$A: . Then follows the transcription of all the utterances of the contribution of the interlocutor whose turn it is. The recommended practice is to regard a contribution as completed only when (i) another participant talks without overlap, or (ii) it is clear that the activity or sub-activity is over. A contribution is therefore regarded as a transcription unit irrespective of the number or length of the utterances of a single interlocutor (even if more than one interlocutor make the same utterance simultaneously). It is treated as a continuous whole uninterruptible by other contributions (including overlaps) or other types of information such as information lines.

Four types of conventions are used in the transcription of contributions:

Linguistic structure-related conventions (pertaining to characteristic linguistic peculiarities found in spoken language forms);

Prosody-related conventions (pertaining to certain characteristic sound aspects of utterances);

Discourse-related conventions (pertaining to characteristic illocutionary peculiarities found in interactive spoken language use);

Recording-related conventions (pertaining to quality of speech)

Each of these types will be discussed in turn in sections 1-4 below.

1. Linguistic structure-related conventions

- **The curly bracket convention: { }**

In spoken language adjacent words very often merge as a result of which parts of the preceding and/or following word may disappear. Since these omitted parts of the relevant words are important for the morphosyntactic tagging of the transcription they should be recovered in the transcription in a fashion that would show that they were not actually uttered. The recovered segments are enclosed in curly brackets, but the merged words are not separated by any spaces.

Example transcription:

\$N: ub{e} uthe kum uy{a} edolophini kub{a} ufun{a}
ukuthenga iswekil{e} emhlophe

- **The plus convention: +**

Incomplete words and interrupted words (i.e. when a speaker pauses at some point in a word) are common in spoken language. These phenomena are represented by means of the plus sign at the position where the word is broken off or interrupted without any spaces between the sign and the word.

Example transcription:

\$J: jonga nokhwezi ndiyak+

2. Prosody-related conventions

- **The uppercase convention:**

Emphasized and contrastively stressed utterances are transcribed with uppercase (capital) letters.

Example transcription:

\$J: INENE INENE ndithi kuni
\$K: andicingi TU mna

- **The colon convention:**

The lengthening of a word is indicated by a colon immediately after the word (i.e. without any space between the word and the colon).

Example transcription:

\$V: ungunyana kamfi: wena
N: e:

3. Discourse-related conventions

- **The slashes convention: /, //, ///**

Pauses by a speaker during a contribution are indicated by slashes, a single slash for a short pause, a double slash for a fairly long pause, and three slashes for a significantly long pause.

Example transcription:

\$X: uandile / lo kaxhanga // ngumde ngeentonga / kunjalo nje
uphaphamile /// ndiyaphinda ndithi uphaphamile kulo mzi
wakwangqika nizakuthini/ nizakuthi makahambe /// uno:tshe



Guidelines for determining the length of a pause

A short pause (/) has a duration of the same order or magnitude as a word (given the current speech rate).

A very long pause (//) has a duration of several seconds noticeable as a “gap” in the speech flow.

When in doubt as to the length of a pause, mark the pause as intermediate (/ /).

Silence between two consecutive contributions is treated as a pause and it is either indicated at the end of the first contribution or at the beginning of the second contribution. In some cases, there may be clear indications that the silence belongs to one of the participants rather than the other. If this is not the case, the silence should as a rule be marked as a pause at the end of the first contribution.

Note also that any contribution of another interlocutor during such pause of a speaker who has the floor is regarded as an overlapping contribution and must be treated according to the overlapping convention described below.

- **The overlap square brackets convention: []**

A characteristic feature of most spoken language activities is that interlocutors may speak at the same time as the speaker whose turn it is. This phenomenon is called **overlapping** and occurs most commonly in the form of feedback that non-turn-taking participants give to the speaker although there are many other forms of overlapping as well. In such cases the following rules apply:

- (a) Each contribution is transcribed as an uninterrupted whole.
- (b) The overlapping contributions are placed sequentially in the transcription according to the order in which they occur in the turn-taking contribution.
- (c) The stretch in the turn-taking contribution where an overlap occurs, is enclosed in square brackets with a numerical index on the inside of the left bracket and the same index on the outside of the right bracket. The index on the left bracket is followed by a space and then the overlapped part of the contribution.
- (d) The overlapping contribution is transcribed in its proper sequential position and enclosed in square brackets with the same numerical index as the one used in the overlapped stretch of the turn-taking contribution. Once again, the index on the left bracket is followed by a space and then the overlapping contribution.

Example transcription:

\$N: andifuni kwenda [1 mna]1 ndingekasebenzi [2 laphela elo xesha]2

\$K: [1 ngoba]1

\$Z: [2 ukwenda]2

\$Z: uyaphosisa wethu yinto entle

Note that the contribution of speaker Z, when it is her turn, is transcribed on a separate line from the transcription of her overlapping contribution. Since the overlap relates to the contribution of speaker N, the overlap is transcribed directly after the contribution it relates to even though speaker Z is taking the turn after speaker N.



Guidelines for transcribing different types of overlapping

Several overlaps by one interlocutor within a contribution of the current speaker

It frequently happens that the contribution of the current speaker is overlapped by several overlaps of another participant. In such cases, each overlap, even though it is uttered by the same participant, is transcribed (enclosed in square brackets) on a separate contribution line.

Example transcription:

SZ: mamelani apha [1 kufuneka]1 sizihloniphe [2 inkosi zethu]2
ngokuzinika yonke into eziyifunayo [3 zingalambi]3 noba nithini na ke [4
kodwa lisiko lidala linenkqayi]4
\$M: [1 simamele maan mfondini]1
\$M: [2 ungumni wena]2
\$M: [3 kowu uphambene ke ngoku]3
\$M: [4 hlonipha kwedini usiyeke thina]4

Simultaneous starts

Another case that deserves mention is where two participants speak simultaneously, say after a pause. In such a case, the overlap of the speaker who had the floor prior to the pause is regarded as a continuation of his/her contribution and it should be transcribed as part of that contribution without a break. The other speaker's contribution is regarded as an overlap and should be transcribed as a subsequent contribution.

Example transcription:

\$N: [andifuni kwenda] mna ndingekasebenzi
\$Z: [mamelani ke]

Nested overlaps

This type of overlapping occurs when two or more participants overlap with the contribution of the current speaker in such a way that the overlaps overlap with one another as well. Consider the graphic representation of nested overlaps below where the dotted line represents the contribution of the current speaker.

\$S:[1.....]2.....[1.....]2.....

In the transcription standard such nested overlaps are not allowed. The problem of nested overlaps is resolved by assigning the overlapped part of a contribution that is common to both overlaps to the second one as follows:

\$S:[1.....]1|2.....[2.....]

Example transcription:

\$K: hayi ayikho tu into yo {ku}ba imali ayikho [1 into kunayo nje]1 [2
kukugeza]2 kwaba Bantu ayizange ingabikho imali

\$P: [1 kakade nje]1
\$N: [2 bayageza]2

Chinese boxes

The problem of Chinese boxes obtains when several participants overlap with the contribution of the current speaker at the same point. In such a case the various overlapping overlaps are represented by the same pair of overlap bracket in the contribution but transcribed as separate overlapping contributions of the different participants. Consider the example below.

Example transcription:

\$P: Kwakungathiwanga imali ikhona na [1 injani na le nto]1 inyani ayikho kwaba bantu into endiyibonayo mna abafunikukhupha mali

\$K: [1 yho inene kunzima]1 sizazakuhlupheka

\$T: [1 buza mnt {w} ana] wabantu]1

\$Z: [1 ubuza kubani]1

- **Gesture convention: < >**

In spoken language gestures very often constitute the sole response of interlocutors in a speech event. Thus, a nod or shake of the head, a smile or a raised finger without any verbal utterance may have the same communicative effect as a verbal expression of approval/disapproval or asking permission to ask a question or comment on something said by the turn-taker. Such gestures may therefore be regarded as contributions in their own right. Obviously, it is impossible to transcribe gestures. Instead, they are designated in a transcription by the comment convention, i.e. by a description of the gesture in angle brackets in the contribution line of the interlocutor who made the gesture. The commented contribution is anchored in the contribution of the turn taker where the gesture was made by means of angle brackets with a space between the brackets.

Where the current speaker is making a gesture during his contribution the angle brackets with a blank space between them are placed in the transcription where the gesture occurred as exemplified in the contribution line of speaker P below.

Example transcription

\$P: < > bonakele

@ <gesture: *shaking head*>

4. Recording-related conventions

- **Round brackets convention: ()**

Any stretch of a contribution (word or sequence of words) that the transcriber is uncertain about (normally for reasons of limited audibility), which is therefore approximately transcribed, is enclosed in round brackets.

Example transcription:

\$K: (andazi) +phinde <1 nezikajuly>1 <2 so>2 ndifuna nje ukuba ke ukuba nithini na niyibona njani nina

Note that the string *andazi* is enclosed in round brackets. This indicates that the transcriber is unsure (because of limited audibility) whether the speaker actually said *andazi* or something else. In the context of the rest of the utterance it would seem that the speaker actually said *andazi*.

- **Three dots enclosed in round brackets: (...)**

Any stretch of a contribution (word or sequence of words) that the transcriber cannot transcribe at all (normally for reasons of limited audibility) is represented by three dots enclosed in round brackets.

Example transcription:

\$Z: mandiqale ngabantu abadala ndize ndandule ukuthetha nolutsha kha(...) ndingx(...) ewe (...) hamba uye egxulu

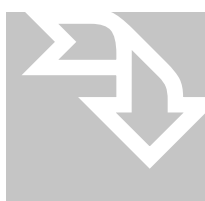
\$F: ndixeleleni bethuna (...)

\$L: (...)

All the instances of (...) in the transcription above represent cases where the transcriber could not make out at all what was said and could therefore not even attempt to make an approximate transcription.

Comment lines

In this section we give an outline of the structure and function of comment line as well as their placement in the transcription body.



How to write comment lines

The purpose of comments in the comment lines is to allow the transcriber to add information pertinent to various aspects of the immediately preceding transcribed contribution line. Typically, the comments are about peculiarities in the language used (e.g. language mix), gestures and names and so on. Although there are

standard types of comments, as we will explain in a subsequent section, transcribers should use their discretion and comment on aspects of a contribution that might be noteworthy.

Comment lines occur immediately after the contribution to which they relate, i.e. before overlapping contribution lines.

Each comment line begins with the special character @ and contains one or more comments, enclosed in angle brackets (< >), and referring back to a comment anchor in the preceding contribution also enclosed in angle brackets. The comment anchor is that part of the preceding contribution marked by surrounding angle brackets that the transcriber wants to comment on.

A comment line consists of a designation of the type of comment followed by a colon. After the colon the specific comment is typed in italics. If more detail could be added to the comment, this additional

information is enclosed in curly brackets. Consider the template of a comment line and the examples below:

```
@ <type: comment {more detail}>
@ <gesture: raising hand>
@ <name: tribal {amaMpondomise}>
```

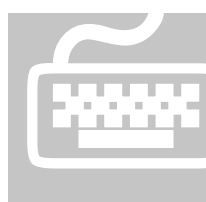
If more than one comment is made on the preceding contribution, the comment anchors in the contribution as well as the corresponding comments are sequentially numbered. The numbers are placed on the inside of the left angle bracket followed by a space and on the outside of the right angle bracket. Each comment linked to a different comment anchor appears on a new comment line.



Example of comment lines in a transcription

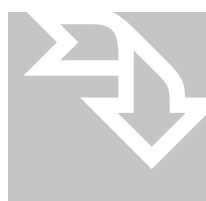
In the examples below the relation between the anchor(s) in a contribution line and one or more corresponding comments in comment lines are illustrated.

```
$A: njengoba <1 iaids>1 <2 especially>2 <3
esouth_africa>3 <4 istics>4 sesona lizwe liphuma <5
if>5 <6 bekuwinwa>6
@ <1 code mix: English {aids}>1
@ <2 code switch: English>2
@ <3 name: country, code mix: English {South Africa}>3
@ <4 code mix: English {statistics}>4
@ <5 code switch: English>5
@ <6 code mix: English {win}>6
```



If there are several comments that relate to the same comment anchor in the preceding contribution, all of them are enclosed within the same pair of angle brackets, but the individual comments are separated from one another by commas. See, for example comment number 3 above. Note also that the original English word is given as more detail within the curly brackets in

comments 1, 3, 4 and 6.



Comments on overlapping contributions

Comments are always anchored to a specific contribution. In a situation where there are comment anchors in parts of a contribution of the current speaker where other speakers overlap, the angle brackets of the comment anchors are placed inside the square brackets indicating the overlaps. Consider the excerpt from a transcription below. Note, in

particular,

- the placement of the angle brackets for the comment anchors inside the overlapped parts of the current contribution;
- the ordering of all the comment lines of the comment anchors in the current contribution before the overlapping contributions;
- the ordering of the comment lines on the overlapping contributions immediately after the overlapping contribution in which they are anchored and before the next overlapping contribution.

Note also that the various overlapping contributions by one speaker have been transcribed on different contribution lines even though these lines immediately follow each other in some instances.

```

$M: eny{e} into kaloku aph{a} <1 edutywa>1 uyazi u{ku}ba kwenzeka
[1 ntoni]1 <2 hafu>2 [2 <3 yegcuwa>3]2 [3<4 ishophisha>4 apha]3 <5
iwillowvale>5 <6 ishophisha>6 [4 apha]4 <7 nehafu>7 <8 yengcobo>8
[5 <9 ishophisha>9 apha]5 <10 nehafu>10 [5<11 yomthatha>11 apha
<12 embhashe>12 nayo <13 ishophisha>13 apha]5
@ <1 name: place {Idutywa}>1
@ <2 adoptive: English {half}>2
@ <3 name: district {iGcuma}>3
@ <4 adoptive: English {shopping}>4
@ <5 name: place, code mix: English {Willowvale}>5
@ <6 adoptive: English {shopping}>6
@ <7 adoptive: English {half}>7
@ <8 name: place {iNgcobo}>8
@ <9 adoptive: English {shopping}>9
@ <10 adoptive: English {half}>10
@ <11 name: place {uMthatha}>11
@ <12 name: river area {uMmbhashe}>12
$F: [1 e:]1
$F: [2 <okay:>]2
@ <code switch: English, fb: expression of understanding>
$F: [3 <o:>]3
@ <fb: expression of understanding>
$F: [4 <allright>]4
@ <code switch: English>
$F: [5 <loo>1 <2 i can say>2 into edala loo nto <3 yisenta>3]5
@ <1 fb: expression of understanding>1
@ <2 code switch: English>2
@ <3 adoptive: English {centre}>3

```

The next section outlines some standardized types of comments.

Standard Comments

1. General

Comments are subjective additions to the transcription of the recording of a speech event and largely dependent on the discretion of the transcriber. However, in the Gothenburg spoken language project some types of comments have been found to be frequently recurring and an attempt has been made to standardize the comments. The following six categories of standardized comments are distinguished in Gothenburg Transcription Standard:

Vocal sounds

Properties of speech

Special expressions

Clarifications

Events and moods

Properties of the recording

To these we have added,

Gesture comments

2. Outline and illustration of instantiations of the various types of standard comments

2.1 Comments on vocal sounds

The comments on vocal sounds can be divided into two groups:

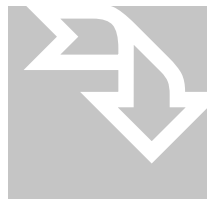
- Vocal sounds made by the speaker in the course of his/her contribution.

In this case the identity of the speaker who makes the vocal sounds is left unspecified in the comment. Note also that the comment anchor in the contribution is empty except for a space between the angle brackets indicating the passing of time. This is not regarded as an overlap as such and is therefore not transcribed as an overlap.

- Vocal sounds made by one or more of the other interlocutors in the course of someone else's contribution.

Since the vocal sounds of the other interlocutors overlap with part of the contribution of the current speaker, the overlap is marked in the usual manner and in the overlapped contribution line a comment anchor is enclosed in angle brackets inside the square brackets designating the overlap. Consider the example below.

\$L: impondo [ndisuka ngapha ke cofimvaba ngoku ndandisagoduka
eastern cape] ja
\$GR: [< >]
@ <vocal sound: *laughter*>



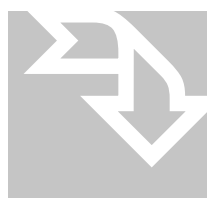
The following comments on vocal sounds are distinguished:

@ <hesitation sound>
@ <laughter>
@ <chuckle>
@ <giggle>
@ <sigh>

@ <puff>
@ <click>
@ <clear throat>
@ <cough>
@ <sneeze>
@ <yawn>
@ <whistle>
@ <snort>

2.2 Comments on properties of speech

Comments on properties of speech always refer to the current speaker. The part of the contribution that is commented on should therefore be enclosed in the comment anchor brackets.



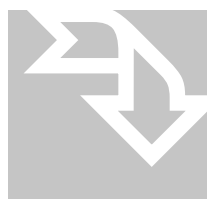
Comments on properties of speech:

@ <quick/slow>
@ <loud / soft>
@ <shouting>
@ <mumbling>
@ <singing>
@ <other language>

2.3 Comments on special expressions

Comments on special expressions relate to two types of expressions:

- Expressions that deviate from the norm, e.g. in pronunciation and in non-standard forms such as foreign language and dialectal, sociolectal or idiolectal expressions.
- Expressions that have a special word form status such as names, abbreviations, etc.

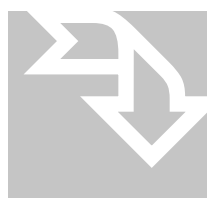


Comments on special expressions:

@ <pronunciation>
@ <code switch>
@ <code mix>
@ <adoptive>
@ <dialect>
@ <sociolect>
@ <idiolect>
@ <name>
@ <abbreviation>
@ <acronym>
@ <letter>
@ <onomatopoetic>

2.4 Clarification comments

Clarifications are comments by the transcriber on the quality of the speech in the contributions that might affect the reliability of the transcription.

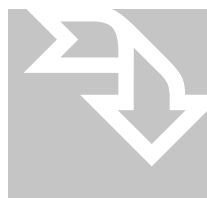


Comments clarifying problems with the recorded material

@ <unclear> (This comment specifies an inaudible part of an otherwise audible part of an expression.)
@ <incomprehensible>
@ <inaudible>

2.5 Comments on speaker's mood

The transcriber can comment on the mood of the speaker

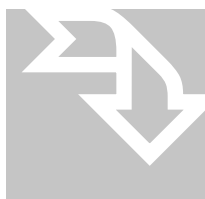


Comments on moods

@ <surprised>
@ <shocked>
@ <amused>
@ <angered>
@ <bored>
@ <sad>
@ <happy>
@ <excited>
@ <in a hurry>

2.6 Comments on the properties of the transcription

These are more technical comments on the transcription process in relation to certain aspects of the recording.



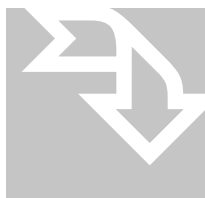
Comments on properties of the transcription

@ <end of tape> (If the recorded activity stretches over more than one tape, this comment could be used to indicate where the switch to the other tape took place in the transcription.)

@ <not transcribed> (If, for some reason, a part of the recorded activity was not transcribed this comment together with the reason(s) for not transcribing it.

2.7 Comments on gestures

These comments pertain to non-vocal gestures that are communicatively significant, i.e. they relate to the speech event.



Comments on non-vocal gestures

@ <gesture: facial {smiling / frowning / winking}>

@ <gesture: body {nodding / shaking the head / raising a hand or finger / pointing a finger}>

Section lines

A speech activity very often has identifiable sections, e.g. a beginning (greetings, etc), various discussion topics (each of which makes up a section), and an ending. If there are such clearly identifiable sections in a recorded speech event, the transcriber should also divide the transcription into correlating sections by means of section lines.

Section lines should conform to the following requirements:

- A section line is indicated with the special character §.
- The transcription body ends with a special section line: § END.
- If no identifiable sectioning is possible, the transcription body should begin with the section line, § START.

- Each section name must be unique within the transcription.



Example of section lines in a transcription

The dotted lines indicate that part of the transcription that has been left out in the example. The section lines have been highlighted for illustration purposes.

§: **About reduced period of Xhosa group visit to Cape town and Eastern Cape**

\$K: ... phinde <1 (ne)zikajuly>1 <2 so>2 ndifuna nje ukuva ke ukuba nithini na niyibona njani nina

....

@ <name: language {isiXhosa}>

#: 00:02:36

§: **About vacant post**

\$N: noba kuthiwa <1 mzo>1 akho sithuba pha <2 kula centre>2 masambe <3 mzo>3

....

#: 0:03:45

§: **Who are going to attend group visit**

\$K: <1 then>1 ke sawubuye sizibeke ke ezinye {e} izinto ndizawujonga kakuhle u {ku} b {a} <2 izikolo>2 zivulwa nini zivalwa nini kwela cala sibone ke nabantwaba zawuhamba apha ke kweli cala lethu umtshana sekudal {a} ezinikezele kangangokuba nango <3 septemba>3 xa <imali> ibikhona ebezawuya ngomso xa bendisithi ndi+

.....

#: 00:04:15

§: **Invitation for recording drama**

\$N: aph {a} ekuhambeni kwethuba mhlawumbi ndizawunicela bethunani njengokuba ninje ninje <1 sistage>1 <2 idrama>2 <3 yeradio>3 ngoba ukufumana imvume e <4 sabc>4 abafundi bethu banikwe ezi sezikhoyo ziphumayo <5 kwasabc>5 kuwe phantsi int {o} endiye ndathi <6 masi sike>6 ndiyenze ngoku u {ku} ba mandibhale kodwa sendibuyibhala <7 then>7 siyenze thina apha <8 ootabs>8 aba

.....

Time coding

Time lines can be used anywhere in a transcription body to give information about the amount of time elapsed from the start of the recorded activity. Usually, it is a good idea to correlate the time lines (if they are used at all) with section boundaries. The following format requirements apply to time lines.

- A time line is started by the special character #.
- The time is given in the format HH:MM:SS (Hours:Minutes:Seconds) e.g. 01:11:42.

If you are using Soundsciber as your transcription tool, you can get the time information directly from the Soundsciber screen on your computer.



Example of time lines in a transcription

The time lines have been highlighted for illustration purposes.

§: About reduced period of Xhosa group visit to Cape town and Eastern Cape

\$K: ... phinde <1 (ne)zikajuly>1 <2 so>2 ndifuna nje ukuva ke ukuba nithini na niyibona njani nina

....

@ <name: language {isiXhosa}>

#: **00:02:36**

§: about vacancy post

\$N: noba kuthiwa <1 mzo>1 akho sithuba pha <2 kula centre>2 masambe <3 mzo>3

#: **0:03:45**

§: Who are going to attend group visit

\$K: <1 then>1 ke sawubuye sizibeke ke eziny{e}izinto ndizawujonga kakuhle u{ku}b{a}<2 izikolo>2 zivulwa nini zivalwa nini kwela cala sibone ke nabantwaba zawuhamba apha ke kweli cala lethu umtshana sekudal{a}ezinikezele kangangokuba nango<3 septemba>3 xa <imali> ibikhona ebezawuya ngomso xa bendisithi ndi+

.....

#: **00:04:15**

§: Invitation for recording drama

\$N: aph {a} ekuhambeni kwethuba mhlawumbi ndizawunicela bethunani njengokuba ninje ninje <1 sistage>1 <2 idrama>2 <3 yeradio>3 ngoba ukufumana imvume e<4 sabc>4 abafundi bethu banikwe ezi sezikhoyo ziphumayo <5 kwasabc>5 kuwe phantsi int{o} endiye ndathi <6 masi sike>6 ndiyenze ngoku u{ku}ba mandibhale kodwa sendibuyibhala <7 then>7siyenze thina apha <8 ootabs>8 aba

Quality assurance procedures

Introduction

Three complementary quality assurance procedures are used in SOUTHTALK. The first procedure checks the correctness of technical aspects of a transcription, e.g. information structures in the header and body, the correct use of line indicator symbols and speaker initials, the opening and closure of certain kinds of brackets, etc. This is an automated checking procedure using a software tool developed at Gothenburg University. The second procedure involves the “proofreading” by a transcription editor who is both knowledgeable in the language of the recorded speech activity as well as in the transcription standard. The third procedure checks the reliability of the transcription against the original audio or audio-video recording of the speech activity.

The automatic checking procedure

This procedure is followed by the transcriber immediately after the completion of the transcription. This procedure simply ensures that the basic technical quality of a transcription conforms to the transcription standard. It is a supplementary form of quality control and the checking of a transcription cannot solely rely on its application.



How to set up and use the automatic checking tool

The checking tool can be found on the internet at the following address

<http://www.ling.gu.se/~leifg/gts>

and should be applied in the following steps (cf. also the exercise on automatic checking in Appendix 3):

Step 1. Open your web browser and go to the above address.

Step 2. Copy your transcription and paste it in the text window.

Step 3. Click “Check”. The results, i.e. the instances of the transcription that do not correspond with the standard, will appear on the screen.

Step 4. Correct the mistakes one by one in your transcription file (save continuously) after a few corrections, copy and paste the transcription again, click “check” and carry on until all mistakes have been corrected.

Transcription editing procedure

An editor who should be conversant in the language of the speech activity that has been transcribed and familiar with the transcription standard scrutinizes the transcription (both header and body). The purpose of the editing is to ensure that

- the linguistically relevant conventions such as underscore, truncations involving the curly bracket convention, incomplete words involving the plus convention have been properly applied in the transcription;
- the sequencing of contributions, overlaps, etc conform to the standard;
- the transcription of foreign code(s), punctuation, numerals, names conform to the standard;
- the comments in the transcription are appropriate and that the conventions used conform to the standard.

After the transcription has been edited, the editor and the transcriber go through the transcription together discussing problematic issues and clarifying moot points. This discussion session is important for several reasons. It will obviously help the transcriber to become more vigilant in subsequent transcriptions. Furthermore, issues relating to spoken language use that have not been properly dealt or which have not been addressed at all in this manual are bound to come up as the transcribed corpus grows. For instance, non-standard comments, the orthographic representation of feedback and own communication management expressions, etc may be of such a nature that transcribers cannot resolve them on their own. It will therefore not only be reassuring for the transcriber to have somebody to discuss these issues with, but it may be in the interest of the project as a whole that these issues are properly noted and dealt with by a project editorial committee. Some of these issues, particularly the standardization of feedback and own communication transcriptions, may be relevant for the other languages in SOUTHTALK as well. In addition, some of these issues may be of such a nature that they should be addressed in future revisions of this manual. We therefore urge the editors of transcriptions to forward these issues to the transcription editor of the project, Mr JM Mfusi at the following address:

J M Mfusi
Dept of Linguistics
P O Box 392
Unisa
Pretoria 0003

Alternatively, editorial issues and comments can be posted to the project website given below:

<http://www.unisa.ac.za/corpusproject>

Transcription checking procedure

The aim of the third checking procedure is to do a quality check of the transcription vis-a-vis the recording of the speech event. The checker obtains a hard copy of the transcription and checks it while watching a playback of the digital video file of the recording on a computer or of a video playback of the recording on a TV. The checking should, among others, ensure that the contributions are assigned to the right interlocutors, that all the contributions have been captured, that all overlaps have been properly noted, that the speech-related linguistic features such as truncations, incomplete words, etc are properly reflected in the transcription. The checker should make appropriate comments about the quality of the transcription on the hard copy of the transcription. These comments should be discussed with the transcriber who should, where necessary, make the relevant improvements to the transcription.

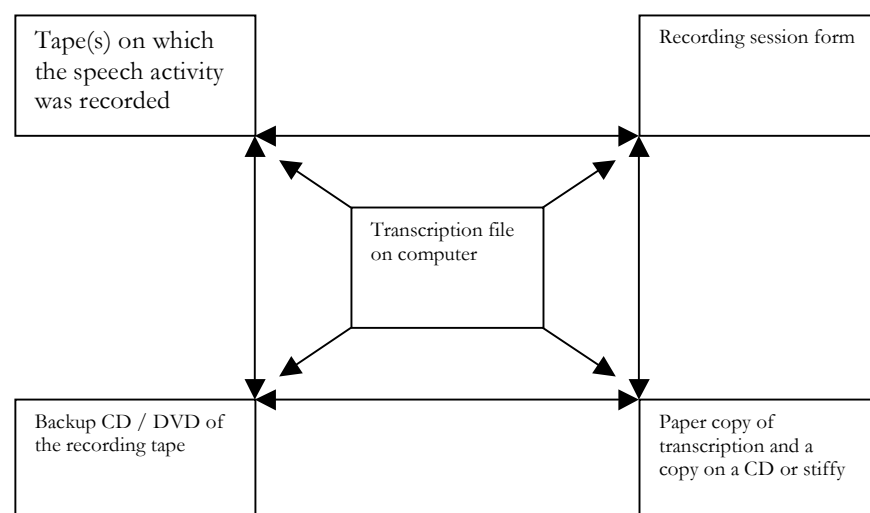
After the checking and improvements of the transcription have been concluded the name of the checker should be added to the transcription header. A hard copy of the transcription should be produced and filed. Backups of the electronic file containing the transcription should be made on a floppy or CD and properly archived in a transcription archive. The backup material must have the same name attached to it as the transcription ID in the header of the transcription. The electronic transcription file is then added to the electronic corpus. Relevant information about the transcription file must be entered in appropriate record files and inventories.

In the next chapter we take a closer look at recordkeeping, inventories and archiving procedures.

Recordkeeping and archiving procedures

Introduction

During the transcription phase the various materials of a specific instance of data collection came together in a system of interlocked items. The transcription of a recorded activity is the link that ties the tape(s) that an activity was recorded on, the recording session form, the video file of the recorded activity and the backup(s) of the recording tape together. This interlocking system of items can be represented in the following diagram.



After the transcription phase the physical link between these various items will be broken as they will be stored in different places – the tapes and their backups in the media archive, the recording session form in a file, the transcription in the integrated electronic corpus, the paper copy of the transcription in a file and the copy of the transcription on CD or stiffer in the media archive.

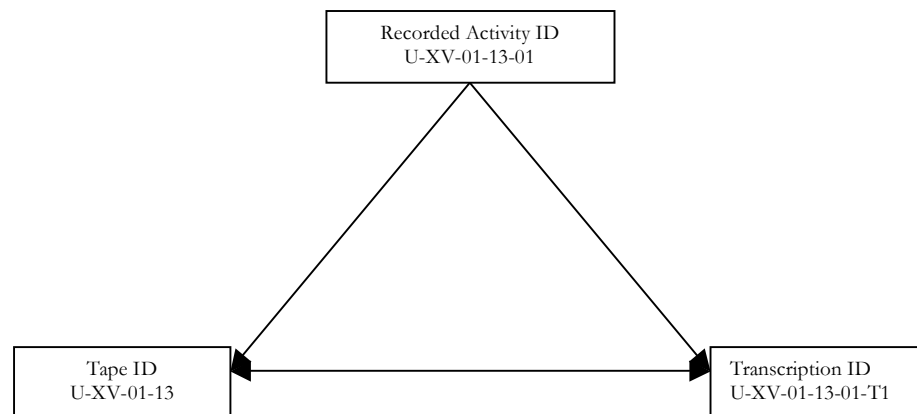
And yet, we would like to keep this interlocking system intact for various reasons albeit in a different format. On the one hand, we may want to bring these various

items or at least some of them again together at same point. For instance, we may want to look at specific gestures on a tape to compare male - female gesturing patterns in certain activity types. In order to look at these gestures systematically a hard copy of the transcription must be at hand. We therefore need a consolidated system where all the various items of all recorded activities are represented in an easily accessible and manageable format.

On the other hand, we need to be able to keep track of the progress of the project both as regards the recordings of activity types as well as the transcription of these recordings. In addition, we need to keep track of the scope (e.g. the range and number of activity types) and volume (e.g. number of tokens) of the corpus as a whole.

In the light of the reasons mentioned above a proper recordkeeping system as well as an archiving system are essential. In the following sections the design of these systems and the procedures for maintaining them will be outlined.

The key in linking all the various items of a recorded activity as well as keeping track of the progress of the project as a whole is the **recorded activity ID**. The recorded activity ID forms the basis for tape and backup CD / DVD names as well as the transcription ID as is illustrated in the diagram below.



Recording-related recordkeeping

In order to keep track of the progress of the recordings (i.e. both the audio and audio-video recording tapes) of speech activities in a particular language at any one of the participating institutions it is important that a form of bookkeeping should be instituted and maintained at each institution. The transcribers are responsible for this bookkeeping.



How to set up and maintain a recordkeeping system for recordings

Step 1. Set up a table in Microsoft Excel according to the template below on the computer where the corpus will be kept. If you don't have access to Microsoft Excel an ordinary table will do. (Excel is recommended simply because it enables automatic calculations on numerical types of information in the relevant columns.)

Step 2. Prepare a record book according to the same template mentioned above.

Step 3. When a transcriber receives a Digital Video (DV) tape containing a recording he/she transfers the recording to the computer as a video file as described in *Step 4*.

Step 4. Set the camera to VCR mode. Link the camera to the Firewire card on the computer and transfer the recording to the computer.

Step 5. Assign an appropriate tape ID code to the tape, i.e. the same code as the one in the header of the transcription of the recorded activity on the tape.

Step 6. Make a back-up of the recording tape on a CD or DVD and assign the relevant identification code, i.e. the same code as the one assigned to the recording tape.

Step 7. Enter the codes of the tape and its back-up in the table on the computer as well as in the record book.

Step 8. Put the tape and its back-up copy in the correct sequence according to the tape number in the respective archiving cupboards for original tapes and backups.