Corpus of Spoken Xhosa

Eva-Marie Bloom Ström, University of Gothenburg & Rhodes University, eva-marie.strom@sprak.gu.se
Nozibele Nomdebevana, University of South Africa, nomden@unisa.ac.za

A multi-modal corpus of transcribed spoken Xhosa

Design

The corpus is multi-modal, i.e. it incorporates audio as well as video, and the transcribed text. We aim at recording speech of different types and social activities, like informal conversations, inter¬views, story-telling and descriptions of for example how to cook a certain dish or build a house. Speakers have to give their consent for inclusion of the recorded conversation in the corpus. In some cases the aim of recording language in a natural setting conflicts with the aims of getting high-quality recordings, asking for the consent, explaining what the research is about and so on. We try to avoid an environment created by the researcher, however this is a challenge!

Variation

The corpus explicitly incorporates variation of any kind, be it geographical, socio-economic, gender- or age-based. The corpus can therefore be used for the study of variation, but as variation is intrinsic to language and can depend on many different factors, we do not tag a text as being of a certain variety. This means that we do not classify the variation.

Metadata

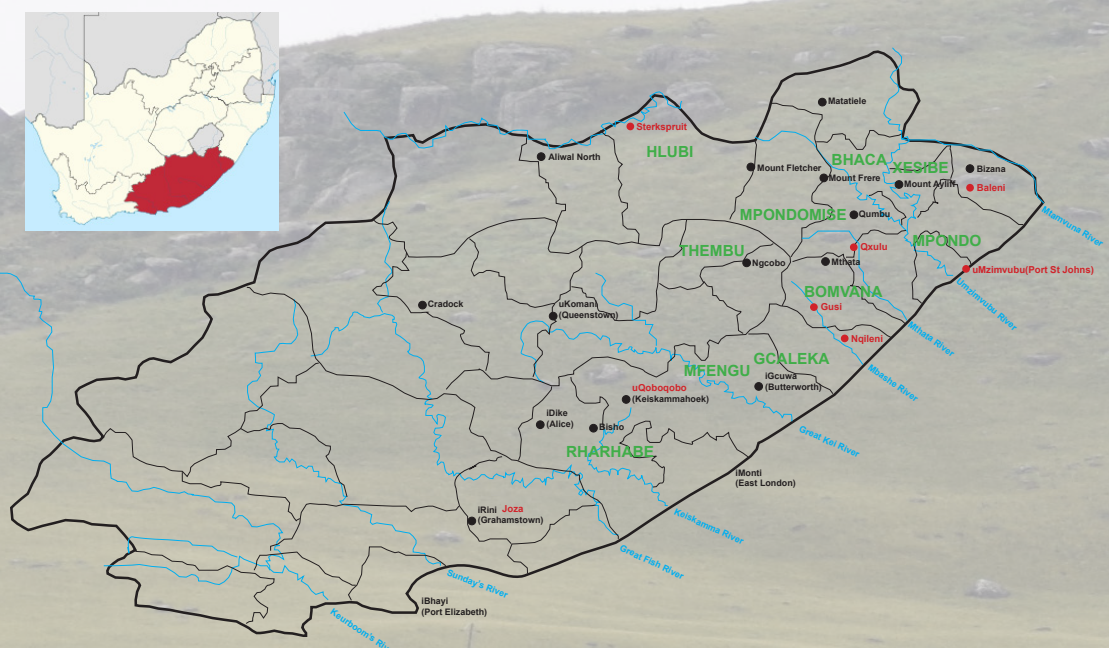
Importantly, we include detailed meta-information about the recording event as well as on the speaker(s). Therefore, the interested researcher can - amongst other details provided - find out that for the recording in question, the speaker is e.g. male, from Mthatha, and educated until grade 9. Moreover, s/he will know that the recording was made in Gxulu, in a kitchen rondavel in the morning, and with a H4nZoom recorder.

Morpho-syntactic variation in the Eastern Cape

The development of the corpus takes place within the project Morpho-syntactic variation in the dialects of Xhosa, funded by the Swedish Research Council. The project aims at analysing gram-matical differences that occur in the Eastern Cape, in the dialect cluster of Xhosa . This means that it will not focus on phonology or lexicon. The reason for this is that the existing literature on variation gives us some information on just that, but not on grammar. Moreover, initial fieldwork for this project has not revealed the same phonological differences as those reported. Maybe such differences have lost in significance?

Points of departure

- Explicitly includes variation and makes no judgement about whether this is the standard or any specific dialect
- Multimodal: video, audio, transcribed text
- A wide variety of text-genres and social activities
- Aims at recording language in a setting that is as natural as possible
- High quality recordings



Fieldwork

Transcribed Text

@Recorded Activity ID: GU151208D_e
@Name of the recording person: Eva-Marie Bloom Ström
@ Duration: 00:59
@Recorded activity date: 2015-12-08
@ Recorded activity type: dialogue (interview)
@ Subject: about King Gambushe
@ Activity mode: face to face
@ Recorded activity location: Gusi, Elliotdale
@ Participant/interviewer: KT= m1 (Kjetil Torp)
@ Participant: NG =f1 (Nobangile Gwebindlala, Queen of Bomvana)
@ Transcriber: Babalwa Resha
@ Transcription date: 2016-02-14
@ Transcription segment: All
@ Transcription system: Standard isiXhosa orthography
@ Recorder: audio H4nZoom, video Nikon D5300
@ Comment: Clear

$NG: Kwabe kukho umntu oza{ku}lilanda eli lizwe, ok{u}ba eli lizwe liphantsi kuka/ kuka nantsika kaGambushe yikumkani leyo nathi ke simel{u}ba sibezikumkani kuba uGambushe lo walilwelay' ilizwe walilelwa waze waba nobuKhosi/ obukhulu obuyikumkani. Yiyo le nto kuth{i}we pha phandle King Gambushe

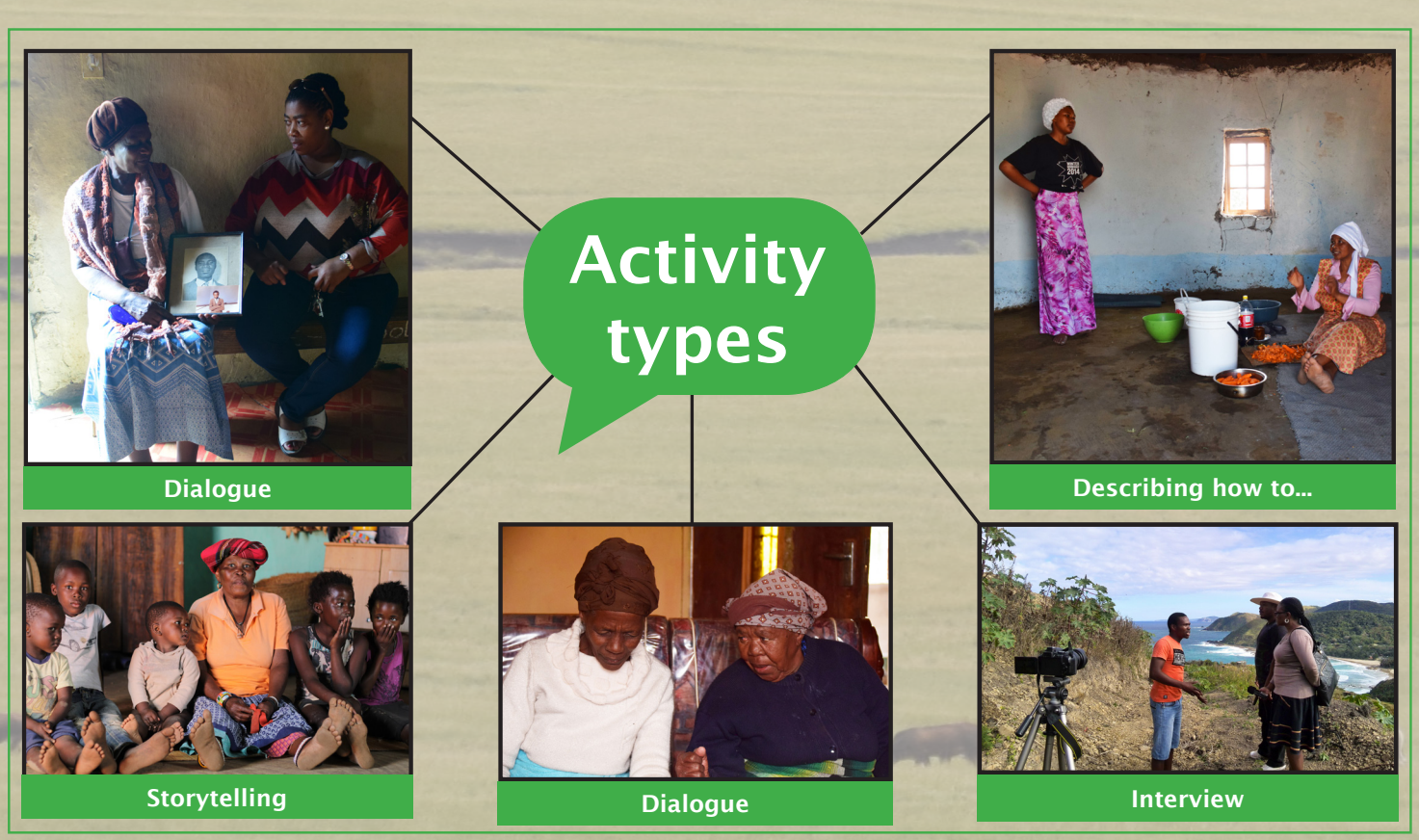
$KT: Mh mh oh/ Nguwe lo waqala ukufika aph'{a} (eNxusini)// uGambushe?

$NG: (…) Kweli lizwe lo:nke, phakathi koMbhashe noMthatha/ uMbhash'{e} ungapha uMthath'{a} ungapha///

$KT: Mh mh lilizwe elihle

$NG: Elihle <they both laugh>

$NG: Li{li}zwe elihle/ eh ndiyabulela nam



The transcription is based on:
Allwood, Jens and Rusandre Hendrikse. 2005. Guidelines for developing spoken language corpora. In: J. Allwood et al. (eds.) Spoken African Language Corpora Series Pretoria, South Africa: UNISA, Dept of Linguistics.