Preparing a corpus of spoken Xhosa

Eva-Marie Bloom Ström[1], Onelisa Slater[2], Aron Zahran[1], Aleksandrs Berdicevskis[3] and Anne Schumacher[3]

[1]Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
[2] Department of Linguistics and Applied Language Studies, Rhodes University
[3] Språkbanken Text, Department of Swedish, Multilingualism, Language Technology, Gothenburg University

eva-marie.strom@gu.se, onelisaslater@outlook.com, aron.zahran@hotmail.com, aleksandrs.berdicevskis@gu.se, anne.schumacher@svenska.gu.se

Abstract

The aim of this paper is to describe ongoing work on an annotated corpus of spoken Xhosa. The data consists of natural spoken language and includes regional and social variation. We discuss encountered challenges with preparing such data from a lower-resourced language for corpus use. We describe the annotation, the search interface and the pilot experiments on automatic glossing of this highly agglutinative language.

1 Introduction

Xhosa, or isiXhosa, is a Bantu language of the Nguni sub-group, spoken in South Africa. Approximately 16 percent of South Africa's population speak the language as their first language, and it is one of 11 official languages of the country (Statistics South Africa 2012). Xhosa is to a large extent mutually intelligible with the other Nguni languages Ndebele, Swati and especially Zulu. Although a relatively large language, it can be considered a lower-resourced language in several respects including in terms of its digital resources. There exist unannotated text collections made available through the South African Centre for Digital Language Resources (SADiLaR), and since recently also an annotated parallel corpus for the four Nguni languages (Gaustad and Puttkammer 2022). The corpus consists of ca. 50 000 tokens of government texts for each language (translated from English) (Gaustad and Puttkammer 2022). Annotated spoken language corpora are lacking altogether. There exists a small collection of audio resources (available through SADiLaR) such as orthographically transcribed audio recordings (6 hours) for the development of text-to-speech (Louw and Schlünz 2018). None of these resources contain natural conversation data. The aim of the current project is to fill this gap by creating an annotated corpus of spoken Xhosa. One important reason for this is that many speakers who were recorded, especially those belonging to minority communities in the area, requested that their contributions of data be well preserved and disseminated.

Consequently, a collaboration was initiated with the aim to make the data available and searchable. Besides providing the digital infrastructure, we aim to explore the possibilities of reducing the manual workload by using automated annotation tools.

2 Fieldwork and content of the data

The recordings included in the corpus all stem from fieldwork by the first author. These recordings have been made in different parts of the Eastern Cape, the province in South Africa where a majority of the population speak Xhosa. Not all speakers identify as Xhosa, however, since the identification as Xhosa implies a certain ancestral line. They identify as belonging to other communities with their own languages. In present-day South Africa, however, differences between these varieties are small as evidenced from the collected data (Bloom Ström 2018). Our material is therefore not necessarily in accordance with standard Xhosa norms. This gives a unique opportunity to study the language in all its facets, as the language is actually used in the communities. The recordings vary in spontaneity. The collection of texts includes dialogues with several speakers. Some of these dialogues are about a certain topic and others are

Proceedings of the 2023 CLASP Conference on Learning with Small Data, pages 62–67
September 11–12, 2023.

completely free. There are monologues in which one speaker explains a certain procedure (e.g., cooking), or tells a traditional story, mostly including an audience. A minority of recordings are more controlled and based on stimuli, i.e., the speaker explains the content of a series of pictures or a film.

This is still a small corpus, with the aim of expanding when the infrastructure is in place. At present, there are approximately 10 hours of transcribed recordings. This is estimated to sum up to ca. 40 000 tokens. Metadata for each recording is noted, including the date, location, speaker information, topic of discussion, length of recording, and audio quality.

3 Premises for data preparation

The overall guideline in the process of making the data available has been maximal searchability for linguistic researchers.

3.1 Transcription Process

The time-consuming transcription process by language students at Rhodes University ensures that recorded audio is represented as accurately as possible in written form. Although standard orthography has been used for transcriptions, we take a descriptive approach to language. This means that we do not adjust the transcribed speech to prescriptive norms.

The idea is that this approach will provide potential corpora users with a rich set of data in which one can investigate things like phonological and/or morphosyntactic variation, but also potential developments and grammaticalization processes based on systematic distribution of different forms encoding the same function. A good illustration of this concerns future tense marking, see Example (1) (glossing follows the Leipzig glossing rules (Comrie et al. 2008/2015); abbreviations are listed at the end of this paper). A construction that originally involved the auxiliary verb *-za* 'come' followed by a verb in the infinitive, has then evolved into a verb form that to different degrees retains the infinitival marker *uku-/ku-*.

1. a) ba-**za** **uku**-fika
   SM.2-come INF-arrive
   'They will arrive'

   b) *si-**zaku**-ya* *kwa-malume!*
   SM.1PL-come.INF-go LOC-1a.uncle
   'We are going to (our) uncle!'

While in the utterance (1a) we can assume, based on phonological criteria, that the verb *-za* and the infinitival marker *-uku* follow each other as segmentable morphemes, in (1b) they are fused into a non-segmentable future marker. Further evidence for this fusion or grammaticalization is that this future marker, originating in a verb meaning 'to come', can in (1b) unproblematically be used with a verb 'to go' due to semantic bleaching of the original meaning of *-za*.

Hence, due to the variation in our data in the realization of this marker, e.g., *zu-, zaku-, zoku-, za-, zo-, zau-,* the grammaticalization process can be investigated. This variation is likely to be higher in our spoken data, than if the corpus was based on standardized written Xhosa.

3.2 Annotation

The morphemic annotation, or glossing, has proven to be a challenge since many areas of Xhosa grammar remain un(der)described. Deciding on a suitable translation for a certain morpheme has more often than not implied thorough investigation of available publications on the language, in combination with our own analysis together with mother tongue speaker and team member Onelisa Slater. There is no modern and comprehensive reference grammar of the language in which one can search for the right abbreviation. All decisions have been made with consideration to the Leipzig glossing rules (Comrie et al. 2008/2015), while also adhering to conventions used by researchers in Bantu linguistics. Ensuring searchability includes, for example, finding a balance between making the glossing general enough to include comparable forms, but also specific enough for the user to be able to unambiguously find what they are looking for. One example is the so-called augment, a vowel that in certain environments occurs before the noun class prefix or the nominal root. While it can certainly be interesting to consider all occurrences of the augment, the researcher might also be interested in only looking at the occurrences of the augment in more restricted settings, say in one specific noun class at a time. Since the augment itself is not noun class specific, search features can be combined to include only those augments that are followed by a nominal prefix or root of a certain noun class.

Another challenge with the glossing stems from the fact that surface forms of (especially spoken) Xhosa do not always show all the information contained in the underlying form, for example because of vowel elision. For this reason, we make use of underlying forms in our glossing, while also showing the surface form in transcription. Again, the augment provides an interesting case in point. In example (2a), the vowel of the comitative marker *na-* coalesces with the augment vowel *i*, forming *e*. In example (2b), the augment vowel of noun class 6 is *a*, i.e. the same as the vowel of the comitative. In (2b), it is therefore not transparent in the surface form that the augment occurs, although it would definitely be in the interest of the researcher to find these constructions as well, when looking for environments with the augment.

2. a) badibana *nendoda*
ba-dib-an-a na-i-ndoda
SM.PST.2-meet-RECP-FV COM-AUG-9.man
'they met with a man'

b) *namakhwenkwe*
na-a-ma-khwenkwe
COM-AUG-NCP.6-6.boy
'with the boys'

While this is a very effective way of making forms more transparent to the user, and making underlying morphemes searchable in the corpus, it also requires further analysis and decision making on the extent to which these underlying forms can be safely assumed.

Moreover, considerations are made on how the glossing conventions can be combined with part-of-speech (POS) tags when searching in the corpus, as these combinations can serve to make searches more inclusive or exclusive depending on the aim of the user. POS tags add information that is not encoded in the glossing, which could help the potential corpus user to identify the functions of different constructions in Xhosa. In cases where tokens are homonymous, POS-tagging can help disambiguate. Example (3), for instance, demonstrates that the token *ukuhamba* from the lexical root *hamb-* 'walk', can be labelled either as verb or a noun based on its syntactic properties. In (3a) *ukuhamba* is a verbal noun/gerund, tagged as a noun in the corpus, while in (3b) it is a verbal infinitive following the inflected first verb 'want' and tagged as a verb:

3. a) *u-ku-hamba* *kw-am*
AUG-NCP.15-walk 15-POSS.1
'my walking'

b) *ndi-fun-e* *uku-hamba*
SM.1SG-want-REC.CJ INF-walk
'I wanted to walk'

One of the main challenges in this regard has been the universality of established part of speech categories, and to what extent tags like the ones used by Universal Dependencies (de Marneffe et al. 2021) are applicable to Xhosa. A relevant example concerns non-verbal predication in Xhosa, in which a copula is prefixed to a noun as in example (4). The copula *ngu-* is verb-like in that it takes some inflectional morphology, although there is not enough diachronic nor synchronic evidence for it to be considered a verb. For example, it does not possess other verbal properties like taking derivational morphology or having an infinitival form. Tagging the whole construction as either a copula or a noun would however not make it justice, but rather, we identify the need of a specialized part-of-speech category called "nominal copula"; NCOP in this case (while the morpheme abbreviation remains COP):

4. Ndandi-ngu-m-ntu
SM.PST.IPFV.1SG-COP.1-NCP.1-1.person
'I am a person'

4 Pilot experiments on automatic annotation

Automatic annotation of spoken Xhosa texts faces several challenges: first, the small amount of data available, second, frequent variation and usage of non-standard forms. Third, the annotation guidelines are being finalized as the manual annotation progresses, which means that the tag sets have not been finalized yet. Despite that, we make a preliminary attempt to estimate whether parts of the pipeline can be automated.

As mentioned above, a corpus of written Xhosa (Gaustad and Puttkammer 2022) has recently been released, and an annotation tool used to create it have also been made available by SADiLaR (du Toit and Puttkammer 2021). SADiLaR, however, uses different glossing

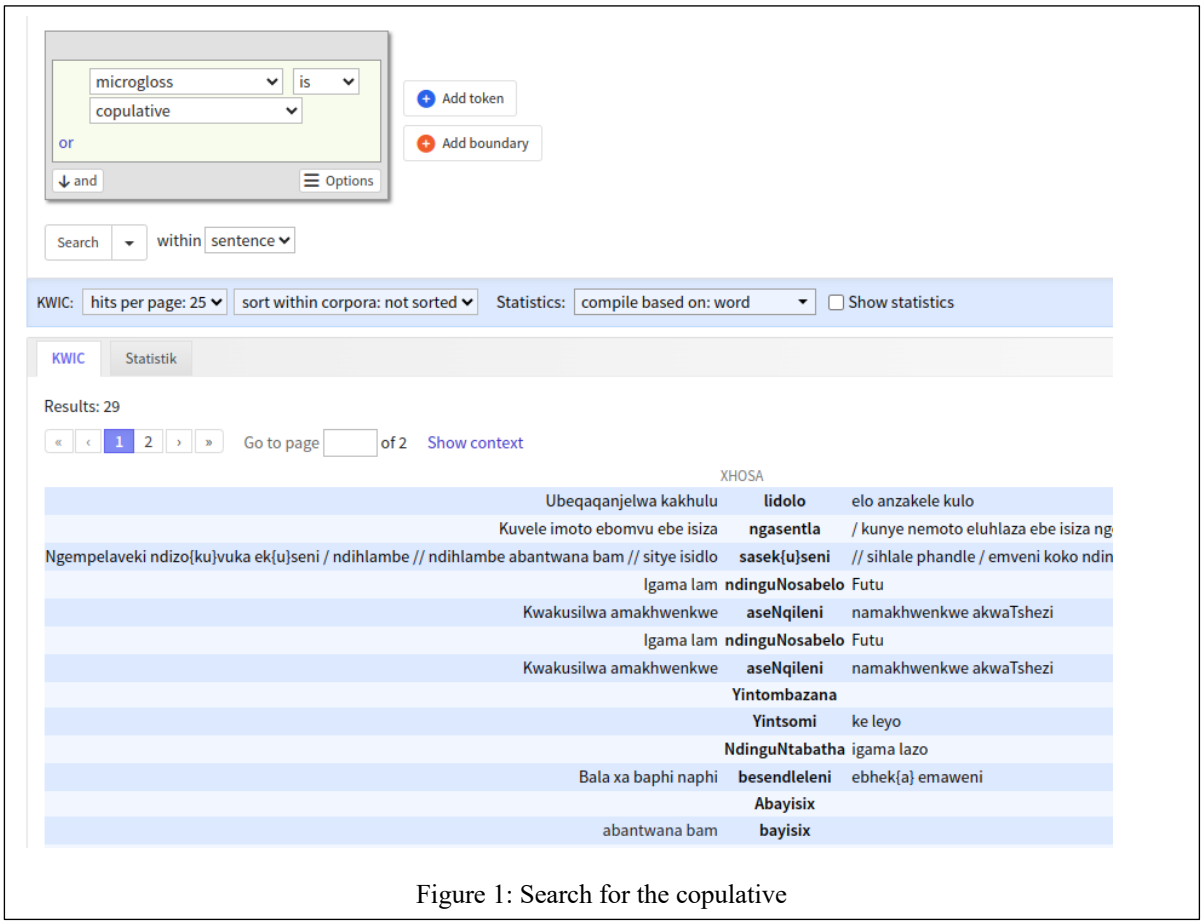

Figure 1: Search for the copulative

principles. The POS tag set, on the other hand, was judged to be compatible with the purposes of the current project. Du Toit and Puttkammer (2021) report the accuracy of their POS tagger, based on the Marmot tagger (Mueller et al. 2013) and trained on the parliamentary texts, to reach 96% in the same domain. On our data, the accuracy is 74%. The drop in accuracy is unsurprising, given the high number of out-of-vocabulary items and the systematic differences in the usage of grammatical forms.

Since the SADiLaR corpus cannot be used to train a morphemic (glossing) tagger, we ran a pilot experiment, training Marmot on our own data. Despite a very small training set of 1122 morphemes, Marmot achieves 67% on the test set (267 morphemes). As is common in such tasks (Barriga Martínez et al. 2021), we did not attempt glossing stems, using the LEX tag for all stems instead.

On the grammatical morphemes only, the accuracy is 51%, with some of the ambiguous morphemes being correctly tagged.

Pre-annotating the texts automatically and manually post-correcting them is likely to be more efficient than manually annotating them from scratch. As the amount of manually annotated data increases, the performance of the tagger will hopefully improve. It remains to be seen whether, given the small training set, "fast learners" like Marmot can be beaten by large languages models (e.g. Eiselen 2023), fine-tuned on the same data.

We have not yet attempted automatically segmenting words into morphemes.

5 The search interface

The corpus is hosted by Språkbanken Text (SBX) and available [1] through the corpus search tool Korp (Borin et al. 2012). Korp can be used to perform advanced corpus search queries where

[1] https://spraakbanken.gu.se/korp/?mode=xhosa

transcriptions along with their annotations (segmentation, glosses, POS, lexical meanings etc.) can be used as search parameters. The parameters can be combined in various ways in order to refine the search.

Note that the parameters apply to different levels of analysis; some are on sentence level (e.g., idiomatic translation of the whole sentence), some are on token level (e.g., POS, lexical meaning), some are on sub-word (morpheme) level (gloss).

For querying purposes, we distinguish between "microglosses" and "macroglosses". A microgloss is any single gloss, the smallest possible unit of glossing, e.g., PST: 'past tense'. Macrogloss is any gloss of a non-segmentable morph. It may contain one microgloss (e.g., RECP for *an* in example 2a) or several microglosses if the morph expresses several grammatical meanings at once, e.g., SM.PST.2 (the gloss for *ba* in example 2a) or SM.PST.IPFV.1SG (the gloss for *ndandi* in example 4). Depending on the users' needs, they may either search for a micro- or macrogloss. The search for microgloss PST, for instance, would return both example (2a) and (4), but it is also possible to search specifically for the macrogloss SM.PST.2.

As an example, Figure 1 shows a search for all copulatives in the corpus (COP).

For this particular corpus a special button was added to the interface which allows the user to copy a traditional four-row representation of glossed examples in linguistics (surface form, underlying form, glossing and translation, cf. example 2b). This was done to facilitate using examples from search results in publications or for teaching purposes.

The corpus will be publicly available, both in Korp and as a downloadable data set.

In the future we will also incorporate the original audio recordings into Korp, and, ideally, synchronize them with the transcriptions (cf. the implementation in the IVIP corpus [2]).

Limitations

The limitations of this project first and foremost concern the amount of data. As automatic annotation starts to improve, the idea is to keep adding transcribed texts to the corpus and this is expected to improve accuracy. Further tests of different kinds of automatic annotation are required.

Ethics Statement

The project includes data provided by speakers of the relevant language and varieties. The conversations concern everyday issues of the people, but also traditional stories and more controlled speech based on stimuli. Recordings which nevertheless ended up containing possibly sensitive information have been removed. All speakers have given informed consent for the use of the recorded data for research and publication purposes.

Acknowledgments

The authors are immensely thankful to all speakers who have contributed data; transcribers; as well as to those who have assisted with contacts and have helped during recording sessions.

In the meticulous process of researching the grammar of Xhosa, we are thankful to Stefan Saviç, Thera Crane, Jochen Zeller and Fahima Ayub Khan for generously sharing their expertise.

We thank the South African Centre for Digital Language Resources (SADiLaR) for their time and assistance at different stages of this project.

The Swedish Research Council is gratefully acknowledged in making this data collection and the spoken language corpus possible in funding the following projects: Morphosyntactic variation in the dialects of Xhosa (VR2014-00244); The role of the verb phrase and word order in the expression of definiteness in Bantu languages (VR2017-01811); How do words get in order? The role of speaker-hearer interaction in languages of southern Africa (VR2021-03125). For the development of the corpus, this work has been supported by Nationella språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (VR 2017-00626).

Abbreviations

AUG augment, a nominal prefix combined with the noun class prefix
CJ conjoint; one of two morphological forms in certain tenses
COM comitative

[2] https://spraakbanken.gu.se/korp/#?corpus=ivip-demo

COP	copulative
FV	final vowel, indicative mood
INF	infinitive prefix
IPFV	imperfective
LOC	locative
NCP	noun class prefix
POSS	possessive
PST	past
REC	recent past
RECP	reciprocal
SM	subject marker

Numbers not followed by SG or PL identify noun class.

References

Barriga Martínez, Diego, Victor Mijangos and Ximena Gutierrez-Vasques. 2021. Automatic Interlinear Glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43, Online. Association for Computational Linguistics.

Bloom Ström, Eva-Marie. 2018. Linguistic and sociolinguistic aspects of variation in the Eastern Cape: complexities of Xhosa language use. *Studia Orientalia Electronica* 6:90-120.

Borin, Lars, Markus Forsberg, Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*. Istanbul: ELRA, volume Accepted, pages 474–478.

Comrie, Bernard, Martin Haspelmath and Balthasar Bickel. 2008, updated 2015. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses, edited by M. P. I. f. E. Anthropology. Leipzig.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 2021; 47 (2): 255–308.

du Toit, Jakobus S. and Martin J. Puttkammer. 2021. Developing core technologies for resource-scarce Nguni languages. *Information* 2021 *12*, 520.

Eiselen, Roald. 2023. NCHLT isiXhosa RoBERTa language model. North-West University; Centre for Text Technology (CTexT).

Gaustad, Tanja and Martin J. Puttkammer. 2022. Linguistically annotated dataset for four official South African languages with a conjunctive orthography: IsiNdebele, isiXhosa, isiZulu, and Siswati. *Data in brief* 41.

Louw, Aby and Georg Schlünz. 2018. Lwazi III isiXhosa TTS Corpus. Meraka Institute, CSIR.

Mueller, Thomas, Helmut Schmid and Hinrich Schütze. 2013. Efficient Higher-Order CRFs for Morphological Tagging. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Statistics South Africa. 2012. *Census 2011, census in brief.* Retrieved from www.statssa.gov.za.