User Manual for Corpus of Spoken isiXhosa

Contents

1. Corpus data

The Corpus of Spoken isiXhosa (referred to in this manual by its exonym Xhosa) consists of transcribed and annotated recordings which have been made in the Eastern Cape in South Africa from 2015 onwards. Currently, recordings from 11 different geographical sites are included in the corpus, these are all listed in Table 1. The recordings have been made as part of three different research projects led by Eva-Marie Bloom Ström at the University of Gothenburg. All projects, 'Morphosyntactic variation in the dialects of Xhosa', 'The role of the verb phrase and word order in the expression of definiteness in Bantu languages', and 'How do words get in order? The role of speaker-hearer interaction in languages of southern Africa', were funded by the Swedish Research Council.

Place Names	
Baleni	(BLN)
Bulungula	(BU)
Cata	(CA)
Gusi	(GU)
Gxulu	(GX)
Mnaymeni	(MN)
Mount Frere	(MTF)
Ncera	(NC)
Ndibela	(ND)
Port St Johns	(PSJ)
Sterkspruit	(STP)

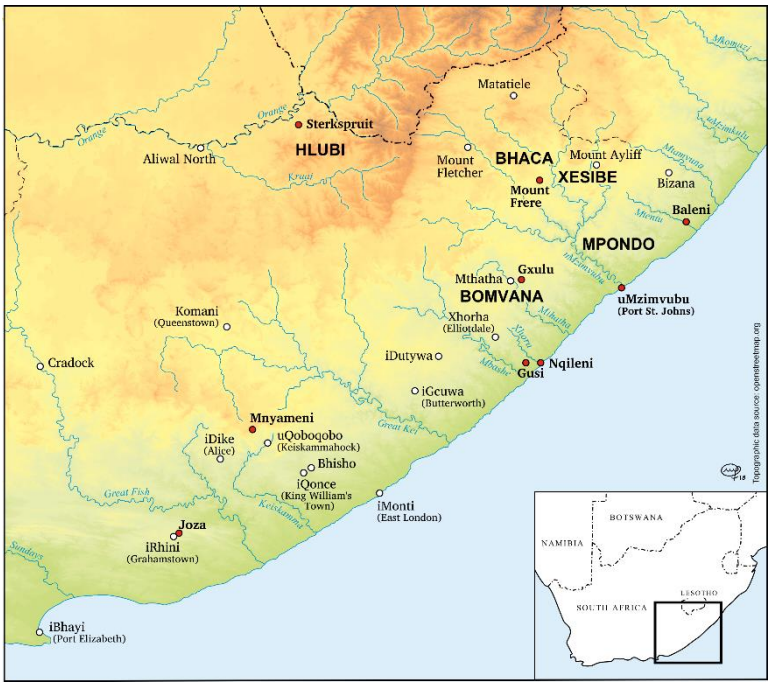


Table 1: Locations of recordings

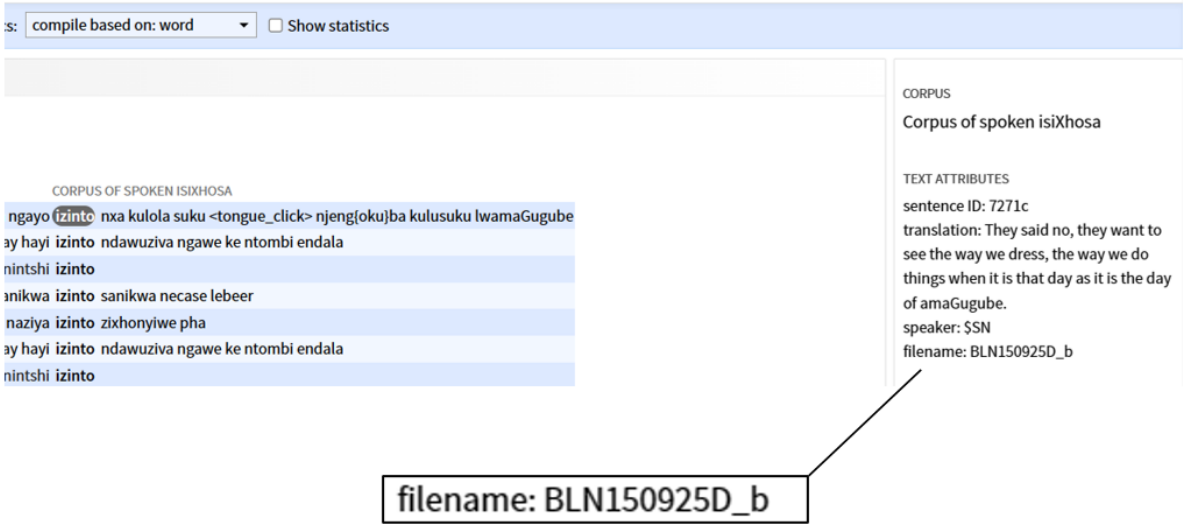
Broadly, the recordings may be either 'staged' or 'observed' events. Staged events means that the researcher has prompted the speech in the recording either through stimuli, questions, or by giving the speakers some sort of task. Observed events are events in which the researcher has simply recorded a speech event without actively influencing its nature. This two-way distinction describes the kinds of recordings in the most general way. In the corpus, a more fine-grained distinction is made between four different types of recordings. These types are listed below in Table 2:

Types of Recordings	
Dialogue (D)	Several participants engaging in dialogue. Sometimes staged, mostly observed.
Monologue (M)	Mainly a single speaker talking about any type of topic. Sometimes the researcher or a second speaker will ask follow-up questions or make small comments. These include staged events such as procedural accounts (e.g., how to cook a certain dish), and open questions like “What did you do this weekend?”
Oral tradition (O)	Oral tradition are mainly stories and folktales told by one person with or without the presence of an audience.
Stimuli (S)	Recordings of speech events in which the speaker talks about some sort of provided material such as cartoons or videos. Generally much shorter than the other types of recordings.

Table 2: Types of Recordings in the Corpus of Spoken isiXhosa

1.1. File names and metadata

All the data in the Xhosa corpus are linked to an original recording. The name of the recording is shown under ‘filename’ in the right column when clicking on an example (cf. picture 1).



Picture 1: Finding file names in the Corpus of Spoken isiXhosa

The file names contain metadata about the geographical place where the recording was made, the date, and the type of recording. The first two or three letters is an abbreviation indicating the place name:

BLN150925D_b

The following digits indicate the date with the format YYMMDD:

BLN150925D_b

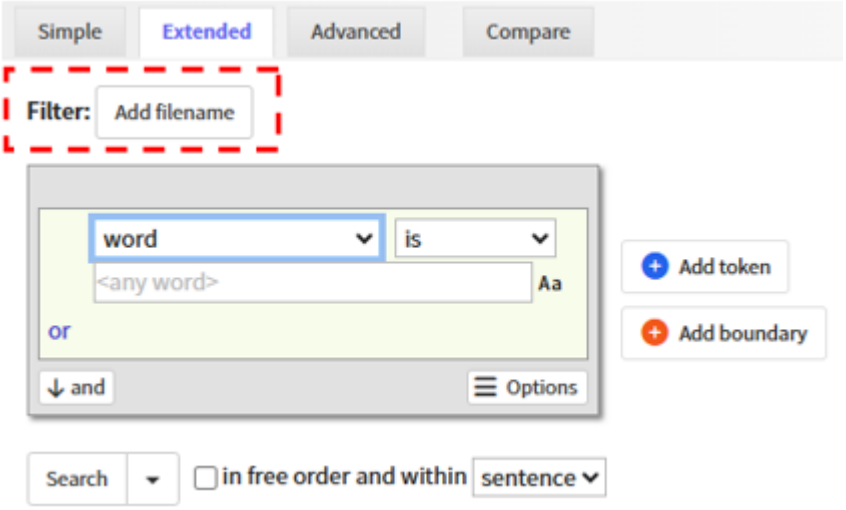
The last capital letter indicates the type of recording:

BLN150925D_b

If there is more than one recording of the same type from the same place and same date, the different files are distinguished by a lowercase letter from a–z:

BLN150925D_b

In the Korp search interface, there is a filter function that allows the user to filter the search by file name (see Picture 2). Using this filter, the user may choose to only look at results from a specific type of recording, e.g., monologues, or only look at results from a certain geographical location.



Picture 2: Filter search results by recording

2. Annotation

The data in the Corpus of Spoken isiXhosa is annotated on 5 different levels:

1. Transcription (surface form)
2. Segmentation (normalized word form, underlying form, segmented into morphemes)
3. Glossings (morphemic annotation and lexical sense)
4. Part of speech (token level)
5. Free translation in English (phrase level)

Some parts of the corpus have been annotated manually, while other parts are pre-annotated automatically before being corrected manually. Levels 1, 2, and 5 have been done manually, while levels 3 and 4 have been partially automatized. More details can be found in the relevant subsections, and in Bloom Ström et al., (2023). With standardized protocols for annotation and several control stages, the annotation can be considered reliable, with the inevitable reservation for human error.

2.1. Transcription

All recordings have been transcribed manually by a team of transcribers, all of whom are based in the Eastern Cape in South Africa. The transcriptions having been made manually by several different people, they are bound to contain a degree of subjectivity, variation, and human error.

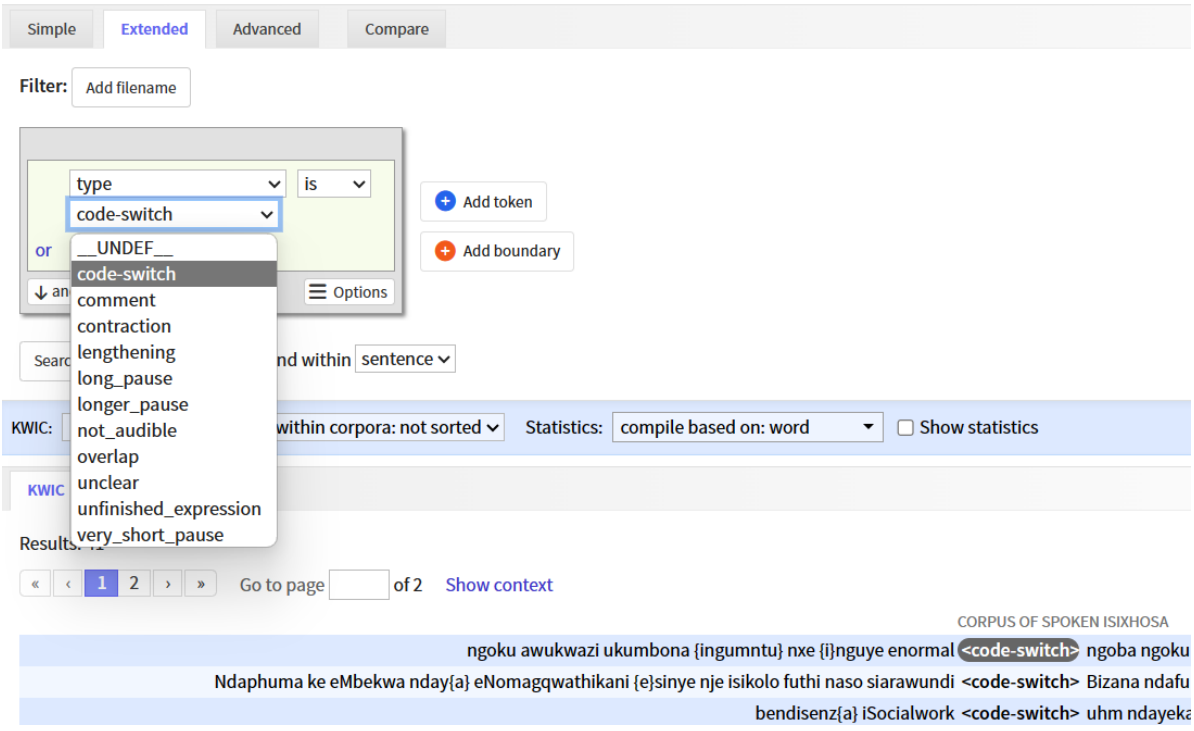
To capture variation in the data, the recordings have been transcribed in a way that is aimed to be true to what is actually said, based on “Guidelines for developing spoken language corpora” (Allwood, 2005). This means that for example non-standard realizations and slips of the tongue are

included in the corpus without any judgement of “correctness” or whether it adheres to standard Xhosa. When certain sounds are elided in fast speech, the transcription will show the omitted sounds within curly brackets ‘{ }’, so both the realized and supposed full forms can be seen. The transcriptions include several other symbols, all transcription symbols and their respective functions are listed in Table 3: Transcription symbols (adapted from Allwood, 2005)Table 3:

Symbol	**Type / Use**	**Example**
<code-switch>	Code-switch	“It's either **<code-switch>** ziziqhamo okanye yimifino”
< >	Comment – Any comment about the speech or noise in the recording	“Phela phela ngantsomi **<laughs>**”
{}	Contraction	”Hayi qhuba **ndifun{a}** ukuva nje”
:	Lengthening – Lengthening of a vowel or consonant for some pragmatic or semantic effect (not used for regular long vowels).	“Kwenzakele abantu abahlanu **kulo:** ngozi”
/	Short pause. Length of pauses based on transcribers subjective judgement.	”Okay **/** ngubani obegwaba?”
//	Medium pause	”Ke ngoku **//** iphelile intsomi”
///	Long pause	”Kwa[kusi+] **///** kwakusilwelwa phi?”
(...)	Not audible	
[ ]	Overlap – Used for overlapping speech	Speaker 1: “**[hayi]**” Speaker 2: “**[ukuba]**”
( )	Unclear – Words that are not clearly heard	“Phosan{i} intambo **(nephu)** nephu”

Table 3: Transcription symbols (adapted from Allwood, 2005)

It is possible to conduct a search for Xhosa examples including any of these specialized transcription symbols. For the researcher interested in these examples, they are listed under ‘type’ in the extended search function (see Picture 3).



Picture 3: Searching for 'types' in the Corpus of Spoken isiXhosa

2.2. Underlying Forms and Glossing

2.2.1. Basic principles

The data in the Corpus of Spoken isiXhosa is morphologically annotated with morpheme-by-morpheme glosses, largely following the Leipzig Glossing Rules (Comrie et al., 2015). Many areas of Xhosa grammar remain under-described. Morphological annotation of the corpus has required extensive searches in existing publications combined with grammatical analysis of all relevant aspects of the language (cf. Bloom Ström et al., 2023).

When deciding on morpheme abbreviations, several features have been considered. Besides the most obvious requirement of attempting to provide accurate descriptions of the Xhosa morphemes, other important considerations include:

i. Uniformity with the Leipzig glossing rules
ii. Uniformity with common labels in the Bantuist literature
iii. Searchability in the corpus

To ensure searchability in the corpus, both the surface form and the underlying form are represented, with the morphological annotation being based on the underlying form. This allows for transparency of contracted surface forms and provides the user with the possibility of searching for morphemes present in the underlying forms. A case in point is the occurrence of adjacent vowels across morpheme boundaries. When two different vowels occur next to each other they coalesce in predictable ways, e.g., *a + i = e* (example (1)); when the two vowels are identical, they merge into one short vowel, e.g., *a + a = a* (example (2)). With the use of underlying forms, all morphemes and their glosses can be represented:

1) badibana **ne**ndoda
   ba-dib-an-a **na-i**-ndoda
   SM.PST.2-meet-RECP-FV **COM-AUG**-9.man
   V N
   'they met with a man' [BU160401O]

2) **na**makhwenkwe
   na-a-ma-khwenkwe
   COM-AUG-NCP.6-6.boy
   N
   'with the boys' [BU160401D_e]

2.2.2. Microgloss and macrogloss

Certain morphemes have been analysed by us as being fusional, i.e., a single morpheme expressing several different meanings or grammatical functions. Although the composition of such morphemes is sometimes rather transparent, we have judged them to be grammaticalized and labeled them macrogloss, to enhance searchability in the corpus.

A microgloss is the smallest possible unit of glossing, e.g., PST 'past tense'. A macrogloss is any gloss of a non-segmentable morph. Macroglosses are separated by hyphens and may contain either a single microgloss (e.g., COM 'comitative' in example (2)), or multiple microglosses if a single morph expresses several grammatical meanings. For example, the first morpheme *ba-* in example (1) consist of the micro glosses SM, PST, and 2, denoting the subject marker of noun class 2 in the past tense.

Similarly, the morpheme *bendi* in example (3), is an even more complex macrogloss consisting of 4 different microglosses (SM.IPFV.REC.1SG[1]). In other words, a macrogloss may consist of any combination of microglosses expressed by a single fusional morpheme.

3) ngulo bendithetha
ngu-lo **bendi**-theth-a
COP.3-DEM.PROX.3 **SM.IPFV.REC.1SG**-speak-FV
COP V
'The one I was talking about' [GX150515M_c]

2.2.3. Lexical sense

Lexical items are translated to English based on context. This means that a single Xhosa root, may have various lexical glosses, e.g., either 'walk' or 'go' for *-hamba*.

Some lexical items that occur in our data, especially from the far north of the Eastern Cape, are not recognised as Xhosa words in published works such as dictionaries. Since Xhosa is part of a dialect continuum within the Nguni group of Bantu languages, we have in such cases tried to find that specific lexical entry in the dictionary of another relevant Nguni variety, e.g., Zulu. These words are not annotated in any special way in our corpus.

Lexical items that are not recognizable by the Xhosa speaking transcribers, and that cannot be found in any dictionaries, are glossed with whatever discernible morpheme they might have. The lexical root is kept as is, e.g., *e-kuphahl-eni* / LOC-kuphahl-LOC.

2.2.4. Automatic glossing

The segmentation of Xhosa tokens with underlying forms has been done manually. Initially, the glossing was also done manually. At a second stage, researchers from Språkbanken SBX have produced automatic glossings of the segmented Xhosa files, which have then been manually corrected. The automatic glossing has been done with the Marmot tagger (Mueller et al., 2013), which was trained on previously annotated data (for more information on this process (cf. Bloom Ström et al., 2023, pp. 64–65)).

2.3. Part of Speech Tags

Each token in the corpus is annotated with a Part of Speech (POS) tag. The three most important factors when deciding which POS categories to use in the corpus, have been: 1) ensuring that the tag labels would provide accurate descriptions of the Xhosa tokens; 2) consider how these tags can be combined with glossings to provide maximum searchability; and 3) providing relative uniformity with typologically recognized POS categories (cf. de Marneffe et al., 2021). The tags are listed in Section 4.2.

The part of speech tagging is also done automatically by Språkbanken SBX. The automated tagger which is used is a modified version of the annotation tool developed by SADiLaR (du Toit & Puttkammer, 2021), for their corpus of written Xhosa (Gaustad & Puttkammer, 2022). The automatic annotation is than manually revised (for more information on the automated annotation of POS tags (cf. Bloom Ström et al., 2023, pp. 64–65)).

[1] 'subject marker, imperfective, recent past tense, 1st person singular'

2.4. Free Translation

The translations are free English translations of a whole utterance. In general, these translations aim to be fairly close to the literal Xhosa construction, while still being acceptable in English. When this balance is hard to achieve, for example with metaphorical expressions, the translation includes both an idiomatic expression and a literal translation in brackets, e.g.:

4) Eh mnye unyana ngulo unguNtakozuko uzothatha iintambo

eh	m-nye	u-nyana	ngu-lo	u-ngu-Ntakozuko	u-zo-thath-a	ii-ntambo
eh	1a-one	AUG-1a.son	COP.1-PRO.5	SM.1-COP.1-1.Ntakozuko	SM.1-FUT-take-FV	AUG-10.rope
INTJ	NUM	N	COP	PROPN	V	N

'Eh, there is one son, the one who is Ntakozuko, who is going to take over the reign' (lit. 'take the ropes'). [GU151208D_d]

3. Searchability

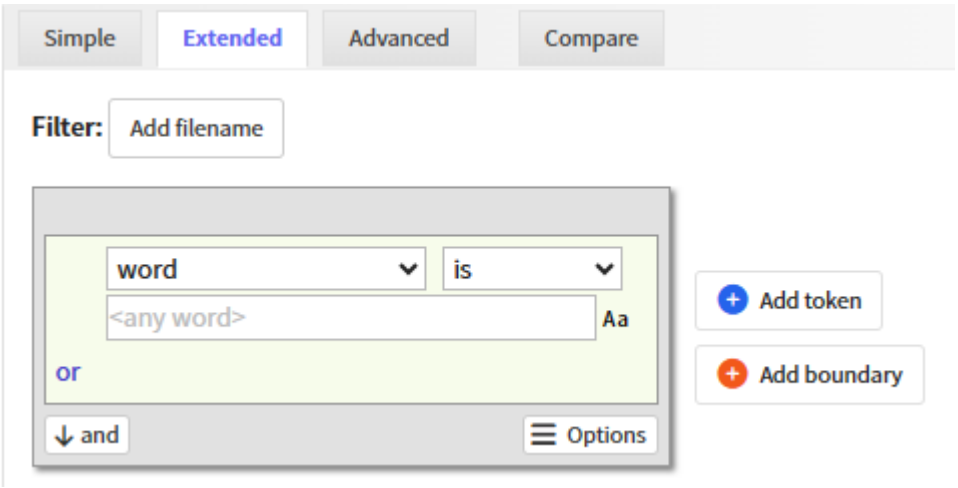
The search interface in Korp can be used to perform a wide range of search queries in the corpus. Queries can be conducted with each of the individual annotation levels as a search parameter, i.e., transcription, segmentation, glossing, POS, and translations. These parameters may also be combined in various ways in order to refine the search.

Below are some illustrations of the 'extended' search function, and some of the different search queries one can perform in the Xhosa Corpus combining different annotation levels.

For general information about the Korp interface and more information on it's various search functions, please refer to the Korp User Manual, accessible via Språkbanken Text's website: [Korp user manual | Språkbanken Text (gu.se)]

3.1. Extended searches – parameters

In the extended search function, each grey box (Picture 4) represents a token (a word or punctuation). The search criteria for the token can be specified in a variety of ways and multiple tokens may be added for a single search query.



Picture 4: Token - Extended search

The extended search function has three different macro levels, 'word', 'word attributes', and 'text attributes'.

3.1.1. 'Word'

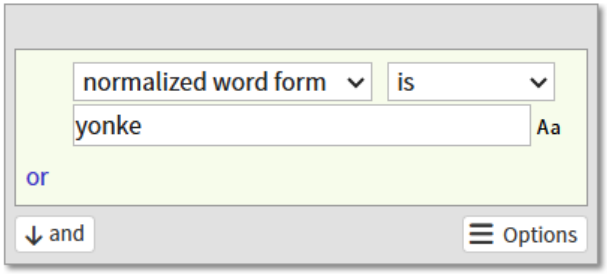
The 'word' search refers to any token as it is transcribed in the surface form, including transcription symbols. In other words, a search for ["word is *yonke*"], will provide all results of *yonke* 'all' but not e.g., *yonk{e}*. In this sense, the 'word' search refers to the first annotation level (repeated below from section 2):

Annotation levels

1. Transcription (surface form)
2. Segmentation (normalized word form, underlying form, segmented into morphemes)
3. Glossings (morphemic annotation and lexical sense)
4. Part of speech (token level)
5. Free translation in English (phrase level)

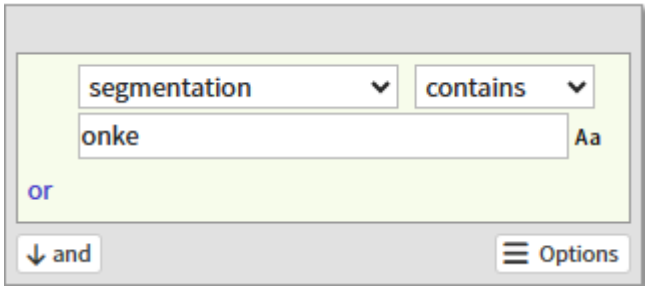
3.1.2. 'Word attributes'

The macro level 'word attributes' is divided into 8 different categories that relate to the annotations of a token. The function 'normalized word form' can be used to search for any word, regardless of the transcription symbols used in the surface form. A search for ["normalized word form is *yonke"*] (cf. Picture 5) will thus include all occurences of *yonke,* regardless of surface forms with elided vowels such as *yonk{e}*.



Picture 5: Extended search – Normalized word form

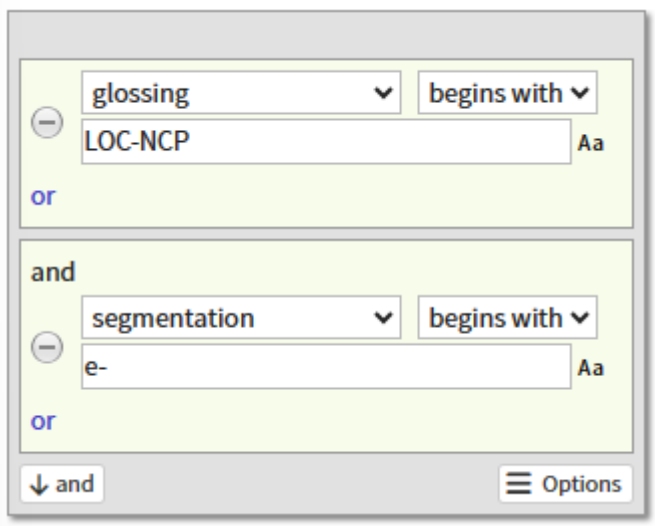
'Segmentation' refers to the second annotation level, i.e., the Xhosa segmented morphemes with their underlying forms. This search function can be used for example to find all examples of a specific underlying sound, root, or morpheme. By selecting [segmentation contains *onke*] (cf. Picture 6), the user can find all the different inflections of the quantifier root *-onke* 'all'.



Picture 6: Extended search – Segmentation

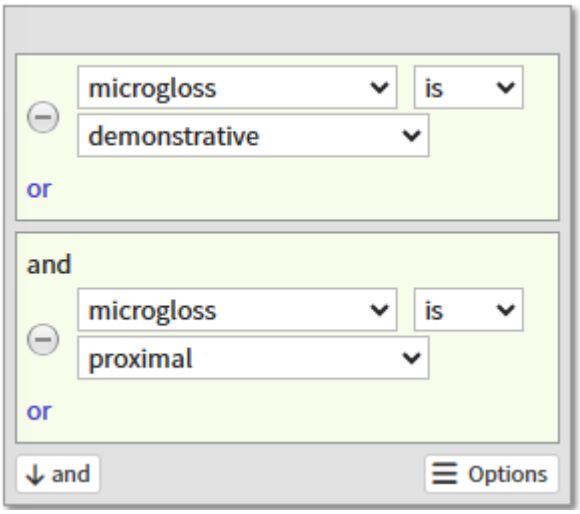
Another example use of the 'segmentation' search query is to search for strings of morphemes or sounds. If one is interested in for example what happens with different vowels over morpheme boundaries, one could search for [segmentation contains *a-i*], which will provide all the instances (and surface forms) which underlyingly has *a* followed by *i* across a morpheme boundary.

The ‘glossing’ attribute combines all the information on the third annotation level, i.e., morphemic annotation and lexical sense. This function allows the user to easily search for different strings of morpheme abbreviations and/or (English) lexical items. Different annotation levels can always be combined. For example, this third annotation level can easily be combined with the previous, second, annotation level of ‘segmentation’. If the user searches for something like [glossing begins with *LOC-NCP*], they will find all instances in which the first two morphemes of a token are a locative marker followed by a noun class prefix. Looking at the results, the user will then discover that there are different locative markers (e.g., *e-, ku-*). The user can then further refine their search by clicking the “and-button” (in the bottom left corner) and specifying the search query as [segmentation begins with *e-*]:



Picture 7: Extended search - Combining glossing and segmentation

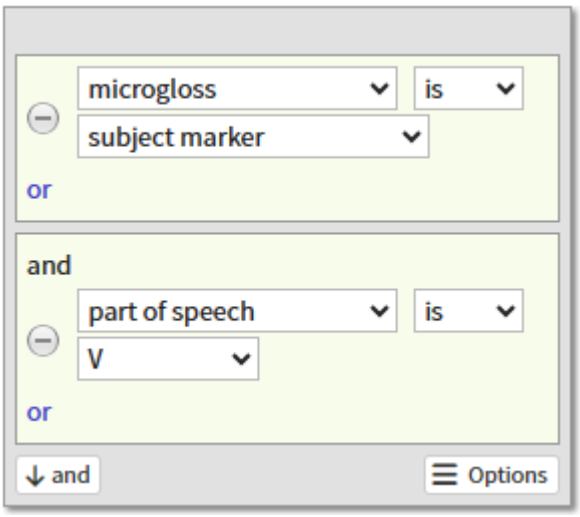
The ‘microgloss’ and ‘macrogloss’ attributes also correspond to the third annotation level, but specifically to the morphemic annotation. All microglosses used in the corpus are displayed in the scroll down menu in their un-abbreviated form (also listed here in 4.1). All the macroglosses, i.e., the existing combinations of microglosses, are displayed in their abbreviated form in the scroll down menu when ‘macrogloss’ has been selected. Macrogloss is the easiest way to find a specific grammatical morpheme, whereas Microgloss is very useful to find a lot of different results within a broader category. This idea can be illustrated with demonstratives; Xhosa has a three-way distinction of demonstratives (proximal, medial, and distal), which are also inflected according to noun class. With macrogloss, you can easily find a specific one, e.g., DEM.PROX.1 (demonstrative, proximal, noun class 1). With microgloss, you can get a very broad overview by searching for [microgloss is demonstrative], or a more narrowed down overview by searching for [microgloss is demonstrative “and” microgloss is proximal], see Picture 8:



Picture 8: Extended search - combining microglosses

The attribute 'lexical sense' also refers to the third annotation level, but specifically to the lexical sense. When this attribute is selected, all lexical senses in the corpus are listed in the scroll down menu.

The 'part of speech' attribute refers to the fourth annotation level. This attribute is useful for large overview results such as all nouns or all verbs etc. It can of course also be conveniently combined with other search parameters for refined searches. Certain morphemes may occur with multiple part of speech categories, one such example is subject markers which may occur both on nouns (with nominal predication) and on verbs. If the user is only interested in verbal subject markers, they may for example combine the searches [microgloss is subject marker "and" part of speech is V], see Picture 9:



Picture 9: Extended search - combining part of speech and microgloss

3.1.3. 'Text attributes'

'Text attributes' has four sub-categories. Three of them, 'sentence id', 'speaker', and 'filename', refer to different meta-data, whereas the fourth, 'translation' refers to the free English translation, i.e., the fifth annotation level. As mentioned in 2.4, some translations (especially metaphorical expressions),

include both an idiomatic expression and a literal translation in brackets. If one searches for [translation contains *lit.*], one will find all such translations.

4. List of Annotations

4.1. List of Microglosses (Morpheme Abbreviations)

Abbreviation	**Gloss**
1SG	1st person singular
2SG	2nd person singular
1PL	1st person plural
2PL	2nd person plural
1	noun class 1
1a	noun class 1a
2	noun class 2
2a	noun class 2a
3	noun class 3
4	noun class 4
5	noun class 5
6	noun class 6
7	noun class 7
8	noun class 8
9	noun class 9
10	noun class 10
11	noun class 11
14	noun class 14
15	noun class 15
17	noun class 17
AC	adjectival concord
ALR	already
APPL	applicative
AS	associative
AUG	augment
CAUS	causative
CJ	conjoint

COM	comitative
COP	copulative
CT	continuous
DEM	demonstrative
DIM	diminutive
DIST	distal
DJ	disjoint
EMPH	emphatic
FE	formulaic expression (accompanied with footnote on meaning)
FUT	future
FV	final vowel
HRT	hortative
IDEO	ideophone[2]
IMP	imperative
IND	indicative
INF	infinitive
INTJ	interjection
INSTR	instrumental
IPFV	imperfective
IR	ironic negative
IT	itive
LOC	locative
MED	medial (demonstrative)
NEG	negative
NCP	noun class prefix
OM	object marker
ONOM	onomastic
PASS	passive
PERS	persistive (*-sa*)
POSS	possessive
POSSIB	possibility (modal marker)

[2] When applicable, ideophones have an English translation/gloss in addition to the gloss IDEO, e.g., 'walk.slowly.IDEO'. Otherwise, we have glossed with the ideophone itself e.g., nephu.IDEO

POT	potential
PRO	pronoun
PROX	proximal
PRSV	presentative
PRT	participial
PST	past
Q	question particle/marker
RV	relative vowel
REC	recent past
RECP	reciprocal
REF	referential
REFL	reflexive
REL	relative -yo
RV	relative vowel prefix
STAT	stative
SBJV	subjunctive
SM	subject marker
VEN	ventive
VOC	vocative

4.2. List of Part of Speech Tags (POS)

TAG	Part of Speech	Comment
ADJ	Adjective	ADJ includes adjectives and nominal relatives, e.g., *mnyama* 'black'. Verbs that are used in a qualifying way are tagged as V. This distinction is sometime hard to make.
ADV	Adverb	Adverbs like *namhlanje* 'today'. Demonstratives from the locative series (*apha/pha/phaya* 'here, there'), and instrumentals with *nga-*.
DEM	Demonstrative	Where a demonstrative is the head.
CONJ	Conjunction	E.g. *ukuba* 'then', *okanye* 'but', *ke* 'then'.
COP	Copula	E.g. *-kho/-khona*. Other POS that take copulative prefixes are tagged as their head, e.g. Noun.
FOR	Foreign	Mainly used for code-switches. If there is Xhosa morphology we have used the tag of the relevant POS. No attempt at a more informed distinction code-switch vs. loan has been made.
IDEO	Ideophone	Ideophones are treated like a word class of its own.
INTJ	Interjection	E.g. *heke* 'exactly', *hayi* 'no'.
INTER	Interrogative	E.g. *-njani* 'how', *-ntoni* 'what'.
N	Noun	Where a noun is the head. Regardless of copulative, associative and other prefixes.
NLOC	Noun with locative marking	Restricted to nouns that are turned into adverbs, through locative marking.
NUM	Numeral	Also lower numerals which are formally adjectives, e.g. *-bini* 'two'.
POSS	Possessive	Possessives include possessive pronouns, but not nouns preceded by an associative (since those are tagged as nouns).
PRO	Pronoun	
PROPN	Proper noun	Name for a person, a place.
PROQUANT	Quantitative pronoun	Used for *-onke* 'all' and *-odwa/-edwa* 'only', and also for *-enye* when it means 'another' (when it precedes the noun) and not as the numeral one.
V	Verb	Also relative verbs.
VAUX	Auxiliary verb	We use this tag for all auxiliaries and do not segment the last vowel in the auxiliary verb. This is because the AUX can be full verbs like *phinda* (and here it would make sense to segment), but often they are 'deficient' and occur in a fixed form, sometimes without inflection (e.g. *zange*).

5. References

Allwood, J. (2005). Guidelines for Developing Spoken Language Corpora. *Spoken African Language Corpora Series, UNISA, Dept. of Linguistics, Pretoria, South Africa*. https://www.academia.edu/85538133/Guidelines_for_Developing_Spoken_Language_Corpora

Bloom Ström, E.-M., Slater, O., Zahran, A., Berdicevskis, A., & Schumacher, A. (2023). Preparing a corpus of spoken Xhosa. *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, 62–67. https://aclanthology.org/2023.clasp-1.7

Comrie, B., Haspelmath, M., & Bickel, B. (2015). *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses*. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics, University of Leipzig.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

du Toit, J. S., & Puttkammer, M. J. (2021). Developing Core Technologies for Resource-Scarce Nguni Languages. *Information*, *12*(12), Article 12. https://doi.org/10.3390/info12120520

Gaustad, T., & Puttkammer, M. J. (2022). Linguistically annotated dataset for four official South African languages with a conjunctive orthography: IsiNdebele, isiXhosa, isiZulu, and Siswati. *Data in Brief*, *41*, 107994. https://doi.org/10.1016/j.dib.2022.107994

Mueller, T., Schmid, H., & Schütze, H. (2013). Efficient Higher-Order CRFs for Morphological Tagging. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 322–332). Association for Computational Linguistics. https://aclanthology.org/D13-1032