Using Linked Data Techniques for Creating an isiXhosa Lexical Resource - a Collaborative Approach

Thomas Eckart

Dirk Goldhahn

Natural Language Processing Group

University of Leipzig, Germany

Bettina Klimek

AKSW Group

University of Leipzig, Germany

Sonja Bosch

Department of African Languages

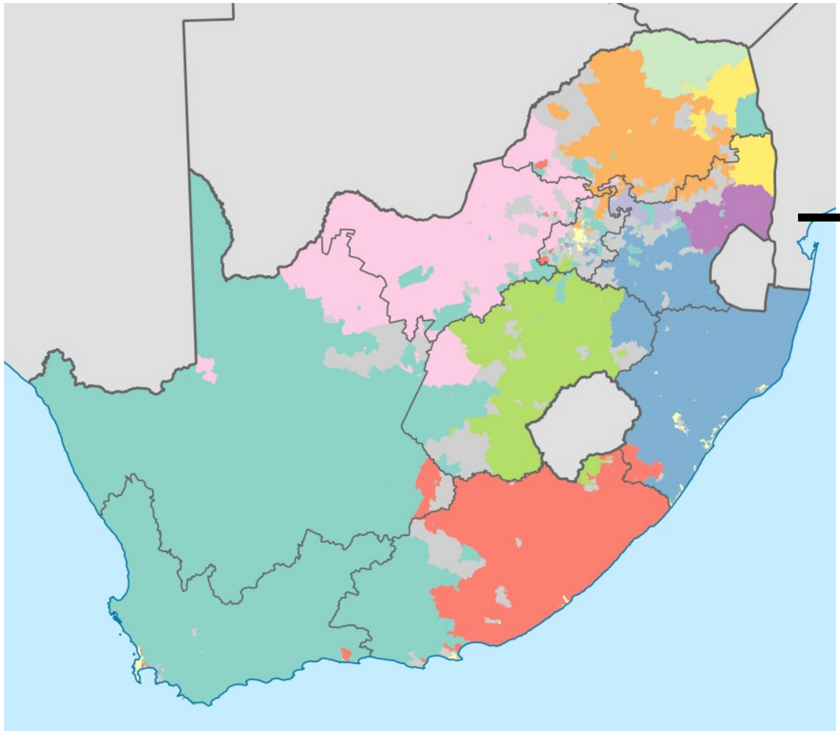
University of South Africa

Agenda

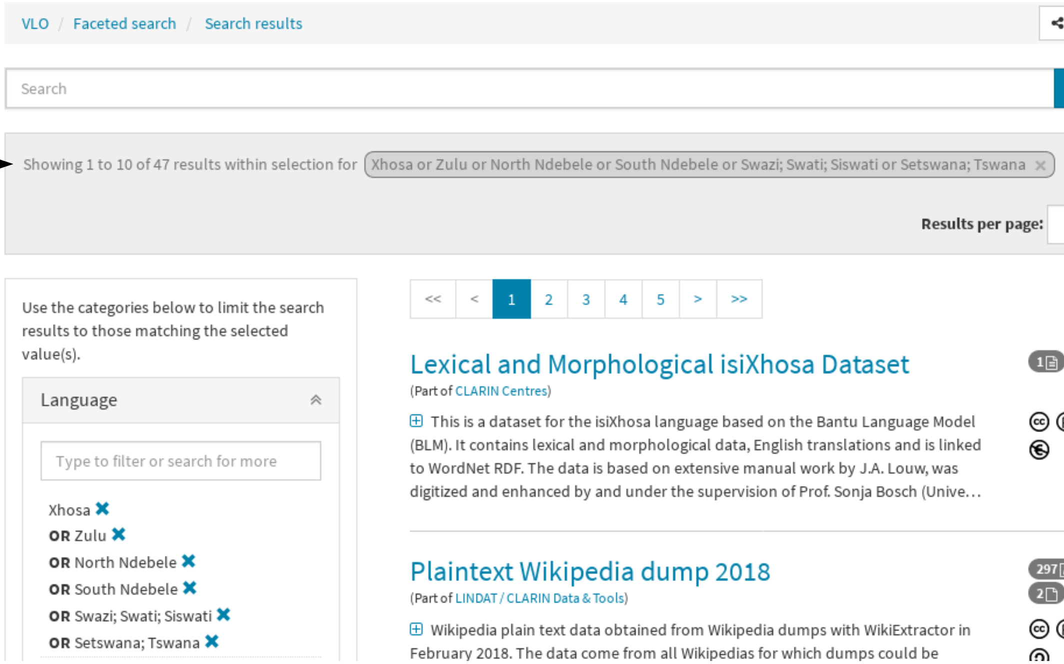
- Language Diversity in South Africa
- isiXhosa Data Set
- Bantu Language Model (BLM)
- CLARIN integration and Dissemination
- Further work

Language Diversity in South Africa

- 11 official languages, including 9 Bantu languages



Source: Stats SA



VLO / Faceted search / Search results
Search

Showing 1 to 10 of 47 results within selection for Xhosa or Zulu or North Ndebele or South Ndebele or Swazi; Swati; Siswati or Setswana; Tswana ×

Results per page:

Use the categories below to limit the search results to those matching the selected value(s).

Language
Type to filter or search for more

Xhosa ✖
OR Zulu ✖
OR North Ndebele ✖
OR South Ndebele ✖
OR Swazi; Swati; Siswati ✖
OR Setswana; Tswana ✖

Lexical and Morphological isiXhosa Dataset
(Part of CLARIN Centres)

⊞ This is a dataset for the isiXhosa language based on the Bantu Language Model (BLM). It contains lexical and morphological data, English translations and is linked to WordNet RDF. The data is based on extensive manual work by J.A. Louw, was digitized and enhanced by and under the supervision of Prof. Sonja Bosch (Unive…

Plaintext Wikipedia dump 2018
(Part of LINDAT / CLARIN Data & Tools)

⊞ Wikipedia plain text data obtained from Wikipedia dumps with WikiExtractor in February 2018. The data come from all Wikipedias for which dumps could be

isiXhosa in South Africa

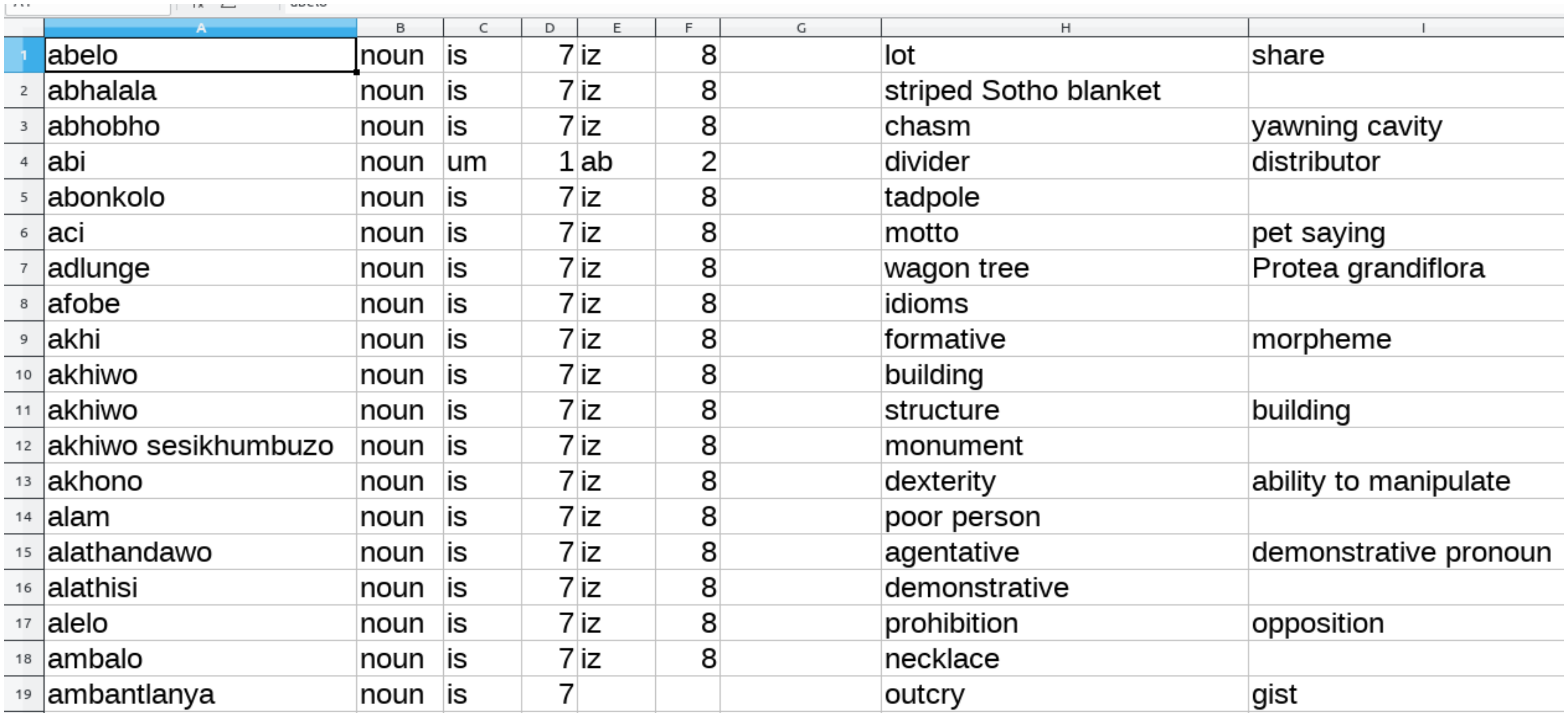
- Spoken in the Eastern Cape and Western Cape regions
- Approximately 8.1 million first language speakers (16% of South Africa's population)
- Nguni language: conjunctive orthography where bound morphemes are attached to words
- Resource scarce language, especially regarding lexical resources; no machine-readable lexicons freely available

isiXhosa Data Set (1/2)

- isiXhosa dictionary
  - Based on data by J.A. Louw and S. Bosch
  - Digitization starting with index cards
  - Still work in progress, finished until Mid 2019
  - Available data:
    - Part of speech, noun classes and prefixes, singular/plural forms, English translations
  - Current status:
    - ~4,000 nouns, ~2,700 verbs, ~8,000 translations

isiXhosa Data Set (2/2)

- isiXhosa nouns (excerpt):

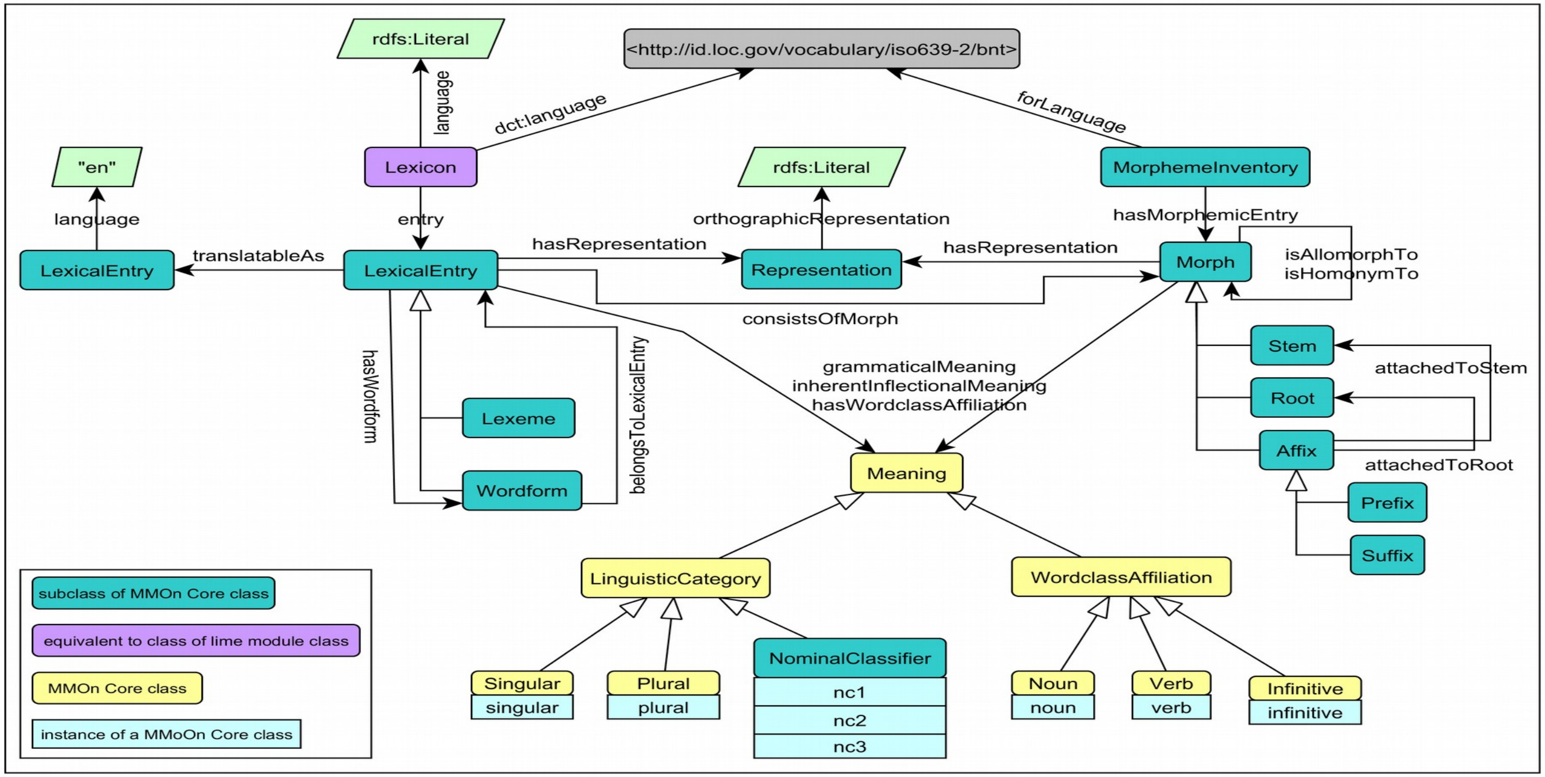


	A	B	C	D	E	F	G	H	I
1	abelo	noun	is	7	iz	8		lot	share
2	abhalala	noun	is	7	iz	8		striped Sotho blanket	
3	abhobho	noun	is	7	iz	8		chasm	yawning cavity
4	abi	noun	um	1	ab	2		divider	distributor
5	abonkolo	noun	is	7	iz	8		tadpole	
6	aci	noun	is	7	iz	8		motto	pet saying
7	adlunge	noun	is	7	iz	8		wagon tree	Protea grandiflora
8	afobe	noun	is	7	iz	8		idioms	
9	akhi	noun	is	7	iz	8		formative	morpheme
10	akhiwo	noun	is	7	iz	8		building	
11	akhiwo	noun	is	7	iz	8		structure	building
12	akhiwo sesikhumbuzo	noun	is	7	iz	8		monument	
13	akhono	noun	is	7	iz	8		dexterity	ability to manipulate
14	alam	noun	is	7	iz	8		poor person	
15	alathandawo	noun	is	7	iz	8		agentative	demonstrative pronoun
16	alathisi	noun	is	7	iz	8		demonstrative	
17	alelo	noun	is	7	iz	8		prohibition	opposition
18	ambalo	noun	is	7	iz	8		necklace	
19	ambantlanya	noun	is	7				outcry	gist

Bantu Language Model (1/2)

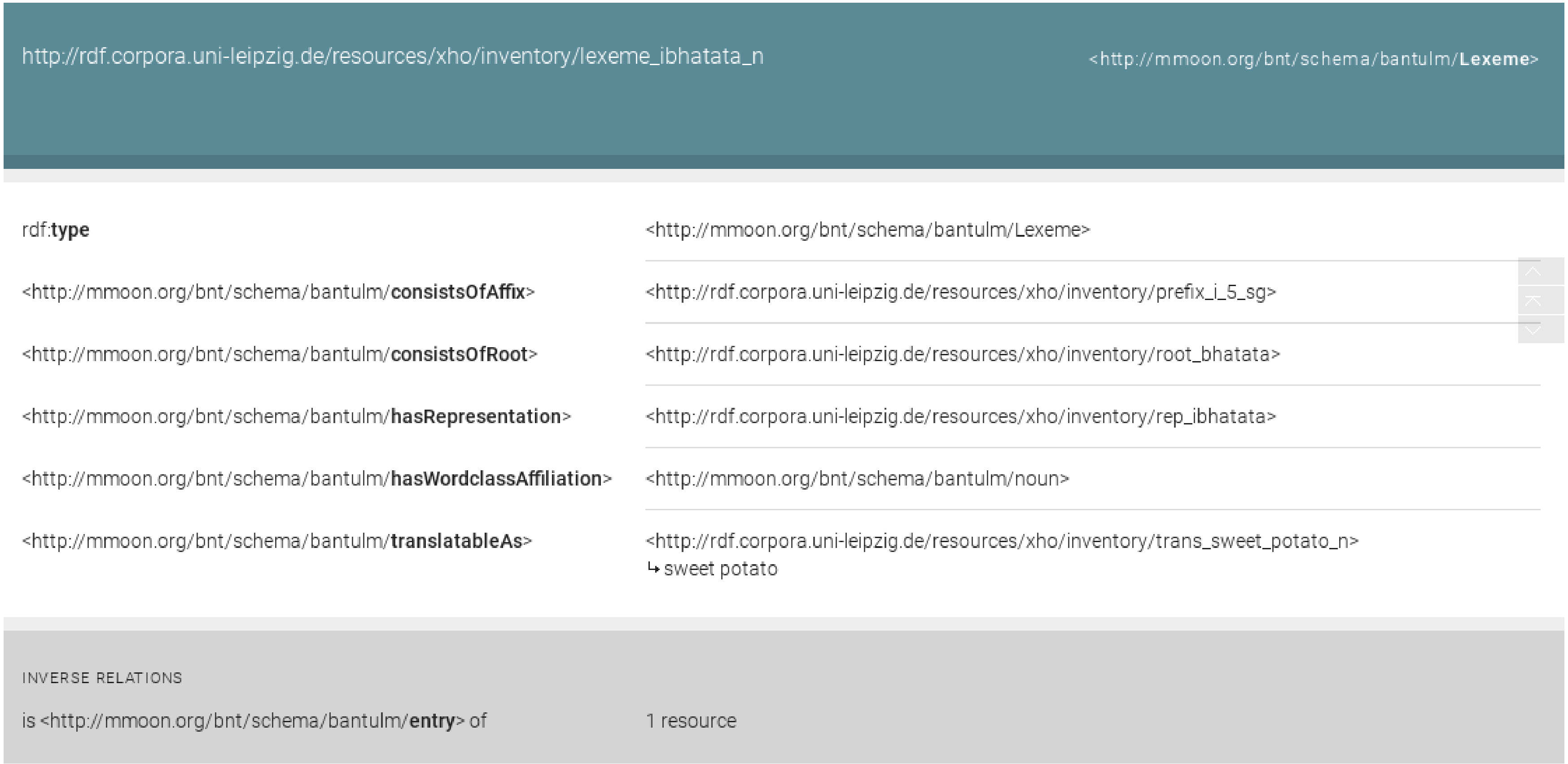
- Structured using Bantu Language Model (BLM)
  - RDF/OWL-based ontology
  - Based on the *Multilingual Morpheme Ontology* (MMoOn)

Model for lexical data of Bantu languages.



Bantu Language Model (2/2)

- Access via dedicated Web page, SPARQL endpoint and downloads



http://rdf.corpora.uni-leipzig.de/resources/xho/inventory/lexeme_ibhatata_n

<http://mmoon.org/bnt/schema/bantulm/**Lexeme**>

rdf:**type**	<http://mmoon.org/bnt/schema/bantulm/Lexeme>
<http://mmoon.org/bnt/schema/bantulm/**consistsOfAffix**>	<http://rdf.corpora.uni-leipzig.de/resources/xho/inventory/prefix_i_5_sg>
<http://mmoon.org/bnt/schema/bantulm/**consistsOfRoot**>	<http://rdf.corpora.uni-leipzig.de/resources/xho/inventory/root_bhatata>
<http://mmoon.org/bnt/schema/bantulm/**hasRepresentation**>	<http://rdf.corpora.uni-leipzig.de/resources/xho/inventory/rep_ibhatata>
<http://mmoon.org/bnt/schema/bantulm/**hasWordclassAffiliation**>	<http://mmoon.org/bnt/schema/bantulm/noun>
<http://mmoon.org/bnt/schema/bantulm/**translatableAs**>	<http://rdf.corpora.uni-leipzig.de/resources/xho/inventory/trans_sweet_potato_n> ↳ sweet potato

INVERSE RELATIONS

is <http://mmoon.org/bnt/schema/bantulm/**entry**> of	1 resource

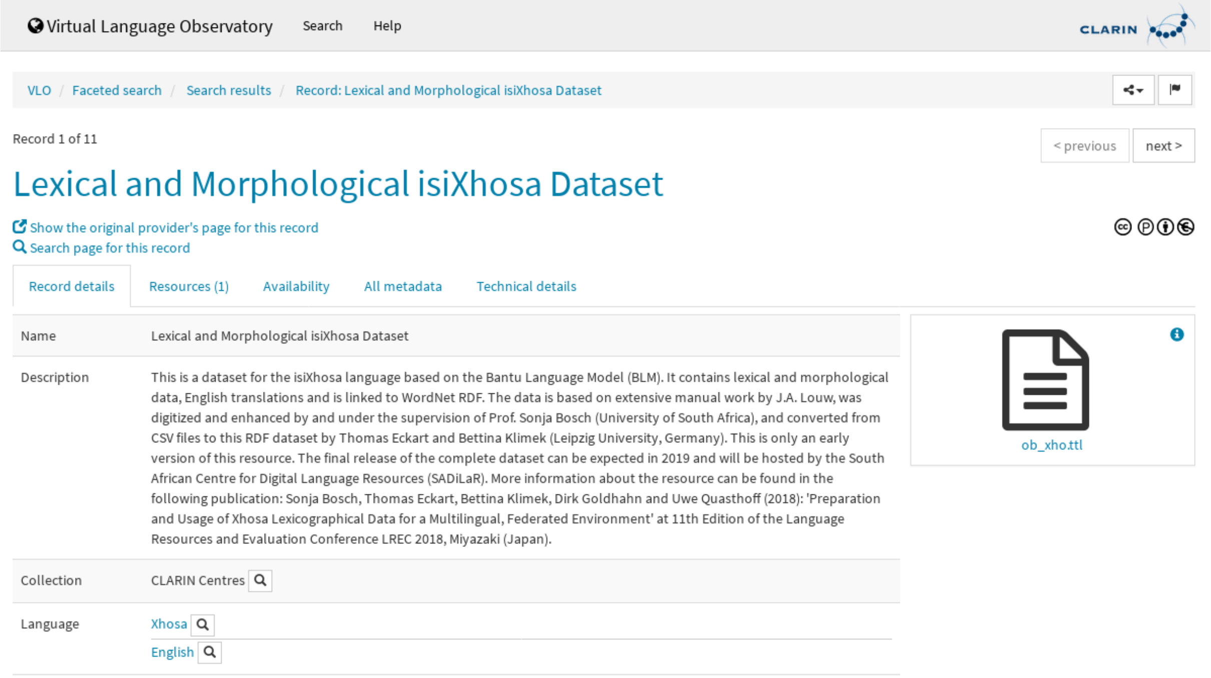
Source: https://rdf.corpora.uni-leipzig.de/lexeme_ibhatata_n.html

CLARIN Integration and Dissemination

- Joint work of three partners, providing their linguistic and technical expertise
- Resource currently hosted at CLARIN-D centre Leipzig and GitHub
  - https://rdf.corpora.uni-leipzig.de
  - https://github.com/MMoOn-Project/OpenBantu
- "Final" resource hosted at *SADiLaR Resource Catalogue*

CLARIN Integration and Dissemination

- Usage of Handles & CMDI metadata
- Strategy for integrating LLOD in CLARIN?



VLO / Faceted search / Search results / Record: Lexical and Morphological isiXhosa Dataset

Record 1 of 11

< previous next >

Lexical and Morphological isiXhosa Dataset

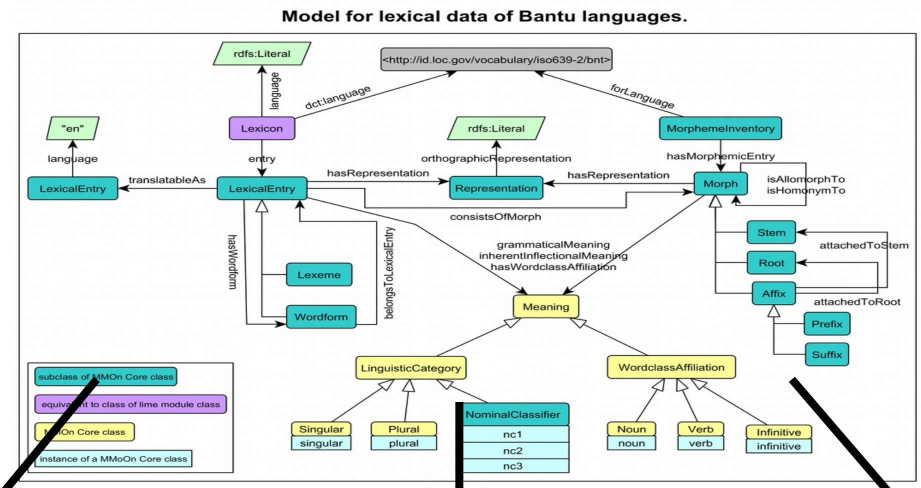
Show the original provider's page for this record
Search page for this record

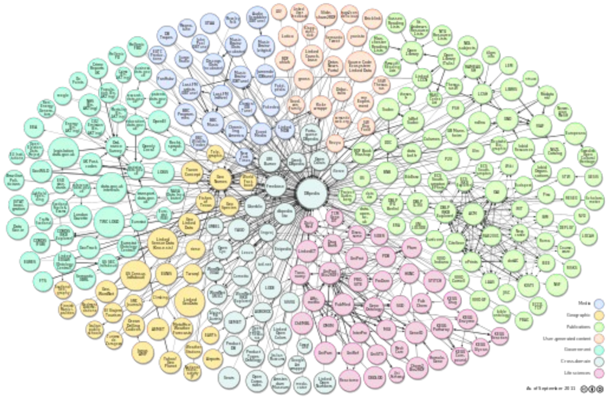
Record details | Resources (1) | Availability | All metadata | Technical details

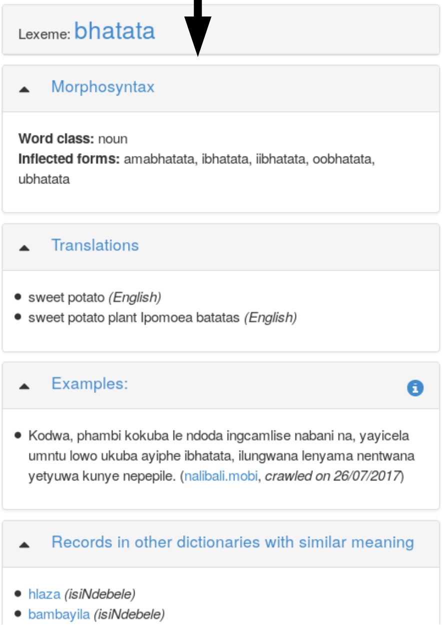
Name	Lexical and Morphological isiXhosa Dataset
Description	This is a dataset for the isiXhosa language based on the Bantu Language Model (BLM). It contains lexical and morphological data, English translations and is linked to WordNet RDF. The data is based on extensive manual work by J.A. Louw, was digitized and enhanced by and under the supervision of Prof. Sonja Bosch (University of South Africa), and converted from CSV files to this RDF dataset by Thomas Eckart and Bettina Klimek (Leipzig University, Germany). This is only an early version of this resource. The final release of the complete dataset can be expected in 2019 and will be hosted by the South African Centre for Digital Language Resources (SADiLaR). More information about the resource can be found in the following publication: Sonja Bosch, Thomas Eckart, Bettina Klimek, Dirk Goldhahn and Uwe Quasthoff (2018): 'Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment' at 11th Edition of the Language Resources and Evaluation Conference LREC 2018, Miyazaki (Japan).
Collection	CLARIN Centres
Language	Xhosa English

ob_xho.ttl

Dissemination







Lexeme: bhatata

▲ Morphosyntax

Word class: noun
Inflected forms: amabhatata, ibhatata, iibhatata, oobhatata, ubhatata

▲ Translations

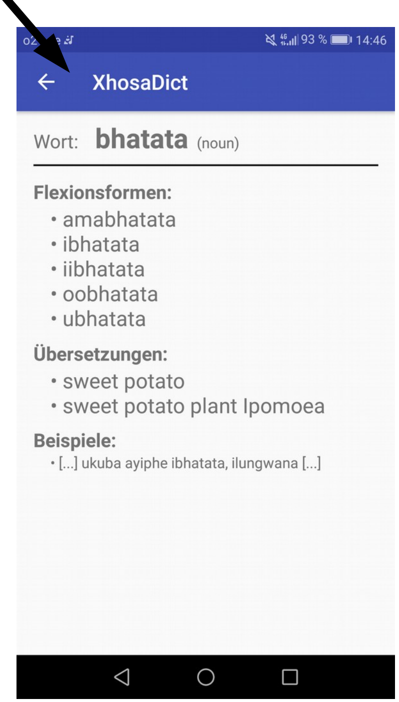
- sweet potato *(English)*
- sweet potato plant Ipomoea batatas *(English)*

▲ Examples:

- Kodwa, phambi kokuba le ndoda ingcamlise nabani na, yayicela umntu lowo ukuba ayiphe ibhatata, ilungwana lenyama nentwana yetyuwa kunye nepepile. (nalibali.mobi, *crawled on 26/07/2017*)

▲ Records in other dictionaries with similar meaning

- hlaza *(isiNdebele)*
- bambayila *(isiNdebele)*



XhosaDict

Wort: **bhatata** (noun)

Flexionsformen:

- amabhatata
- ibhatata
- iibhatata
- oobhatata
- ubhatata

Übersetzungen:

- sweet potato
- sweet potato plant Ipomoea

Beispiele:

- [...] ukuba ayiphe ibhatata, ilungwana [...]

Further Work

- Continuing work on the data and QA
- Combination with open text corpora
- Open data services for a federated lexicographical infrastructure
  - Prototype: CBOLD data for lexical alignment based on English translations
- Annotation services: WebLicht?
- ?

Thank you!

Questions? Remarks?