

Enriching Lexicographical Data for Lesser Resourced Languages: A Use Case

Dirk Goldhahn¹, Thomas Eckart¹, Sonja Bosch²

¹Natural Language Processing Group, Institute of Computer Science, University of Leipzig, Germany

²Department of African Languages, University of South Africa, South Africa



- Motivation
- Lexical Data
- Crawling Under-Resourced Languages-Portal
- Collecting Usage Samples
- Conclusion

Motivation

General problem:

- Availability of contemporary text material is a prerequisite for a variety of applications and research scenarios
- Only for a small number of languages the situation is satisfactory
- Even for most of the languages with more than one million speakers no reasonably sized textual resources or tools like POS taggers adapted to these languages are available

→ Widespread need for digital language resources for many languages of the world

Goal: Create a lexicon app for lesser resourced languages of South Africa (among others Bantu languages)

Starting point: Lexical data for Xhosa and Kalanga

- Available as Linguistic Linked Open Data
- Integrated into CLARIN (see Annual Conference 2018)

Problem:

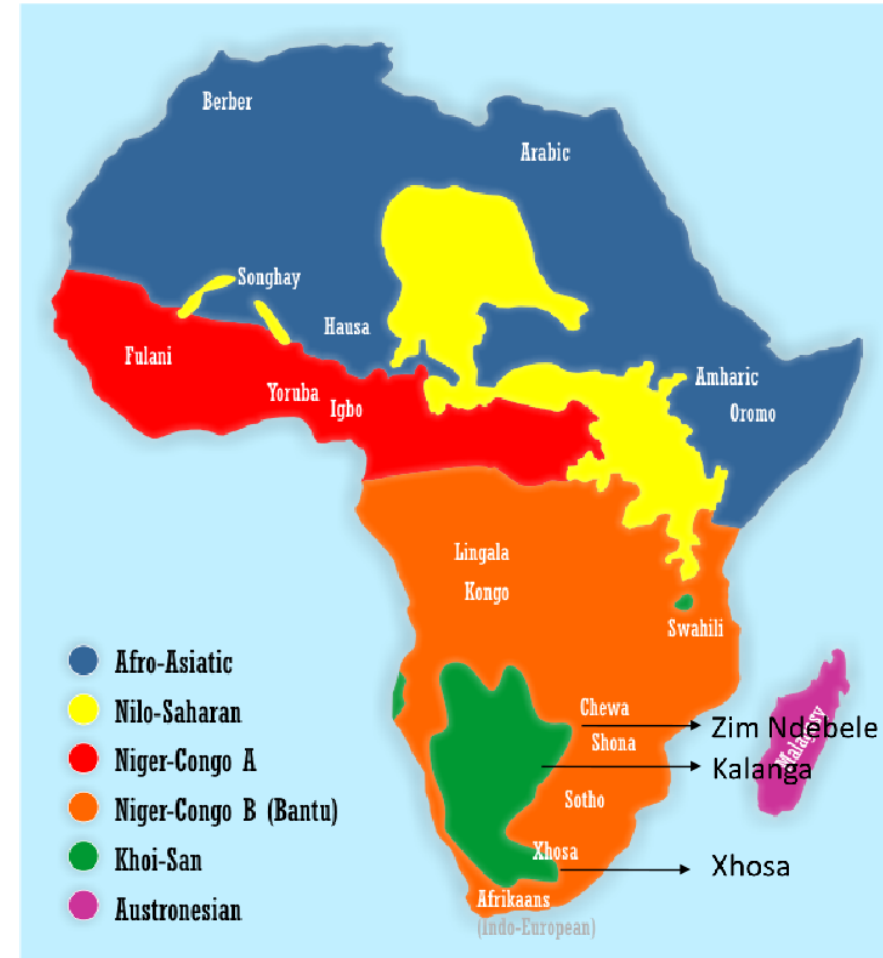
- Bantu languages are typically under-resourced
- Missing usage samples

Approach: CLARIN-services!

- CURL portal to automatically generate valuable textual data from online resources

Lexical Data

- Language family spoken in Sub-Saharan Africa
- 440–690 distinct languages with 240 million speakers
- Share similarities in phonetics, morphology, syntax, and vocabulary
- Examples: Swahili, Shona, Zulu
- Most Bantu languages considered as resource scarce



Lexicographical resources used:

- Comparative Bantu OnLine Dictionary (CBOLD)
- Lexicographical Xhosa dataset (J.A. Louw)

Resources vary in:

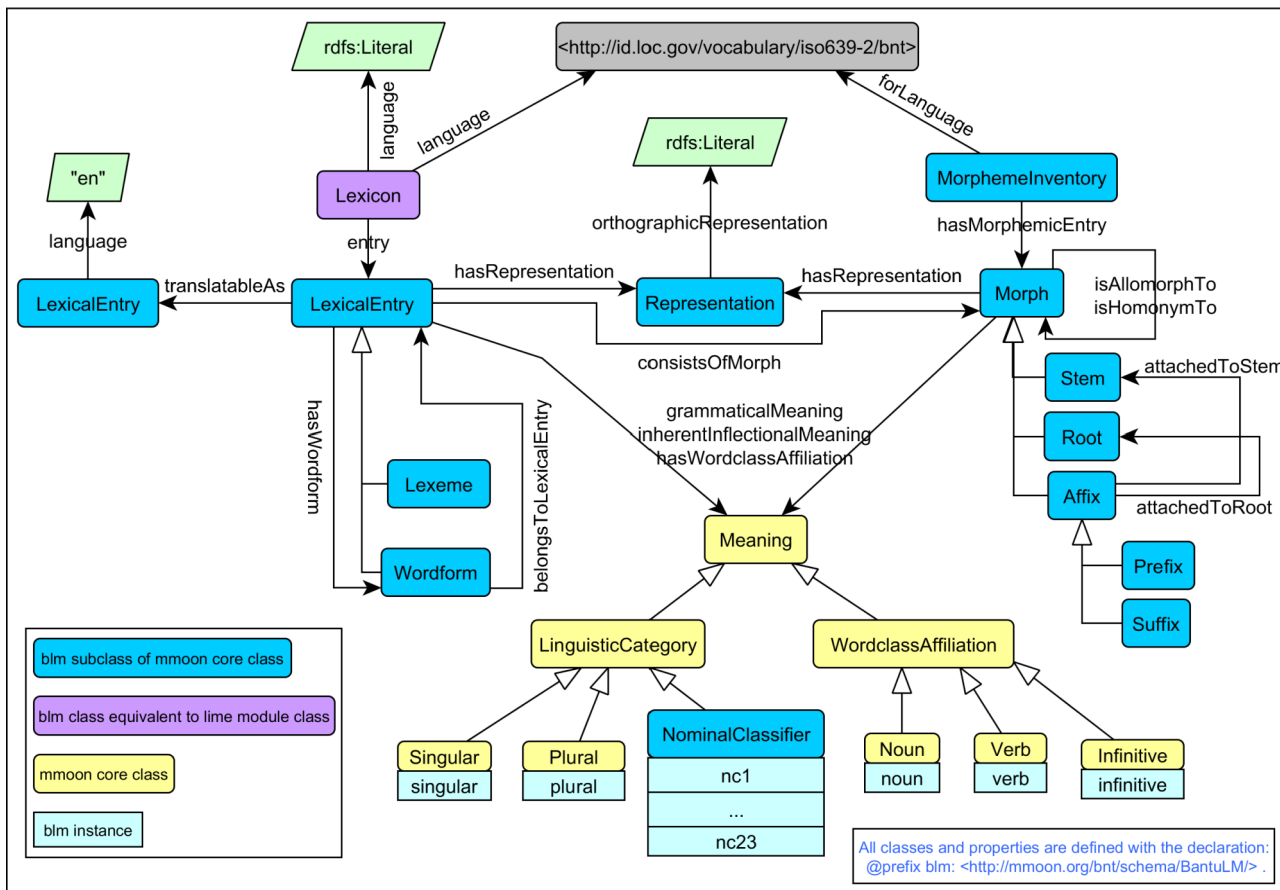
- Quality
- Formats used (CBOLD: Plain text, MS Word, FileMaker, HyperCard)

→ Demand for harmonized data model

Based on *MmoOn* (Multilingual Morpheme Ontology)

Reflects Bantu language characteristics while preserving compatibility with non-Bantu languages

Model for lexical and morphological data of Bantu languages.



Access via:

- Web portal
- SPARQL endpoint
- Download

CURL

Crawling Under-Resourced Languages Portal



Crawling

Underresourced Languages

[Home](#)[Submissions / Process Status](#)[Corpora Downloads](#)[Help](#)

Collecting Web Pages for



Under-Resourced Languages

On this website you can contribute to corpus collection for under-resourced languages by simply entering a URL. The languages are chosen so that they have more than one million of speakers, but up to now there are less than one million of sentences in the Leipzig Corpora Collection. The URLs or domains you provide will be crawled and reviewed for text data in the respective language. After processing you will be presented with statistics for the URLs you provided. The created corpora will be freely available.

[Flyer with more information about this project](#)

Thank you for your support!

Step 1: Select the language

Xhosa [xho] - South Africa

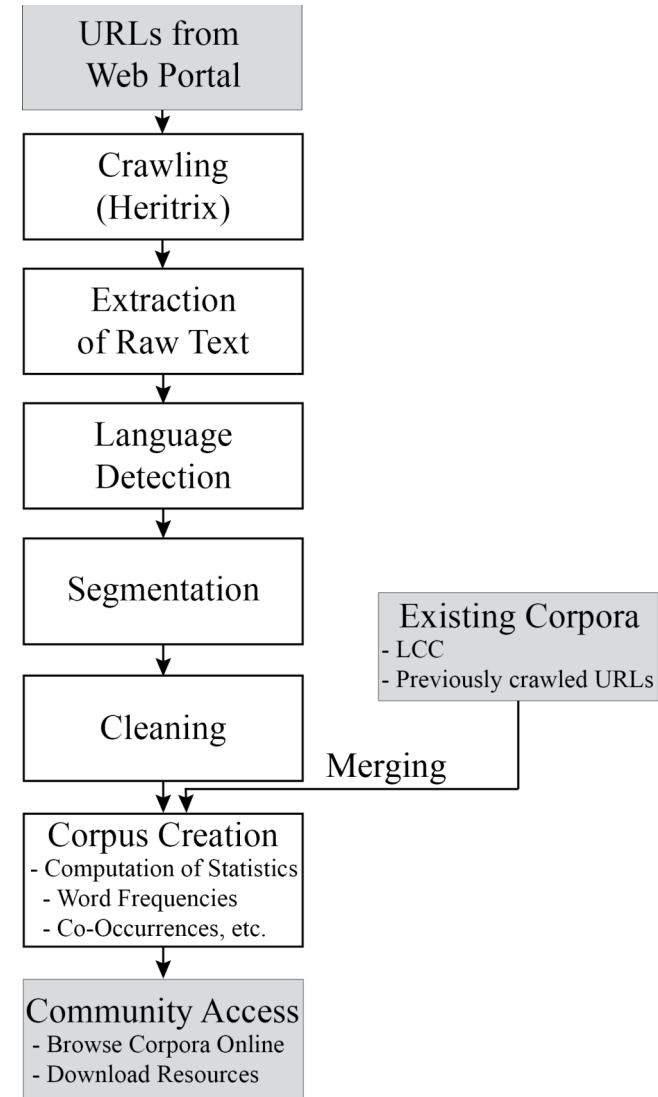
Corpus: **xho_community_2019**
Types: **172520**, Tokens: **972301**, Sentences: **63387**

Step 2: Insert the URLs / Upload a URL list

Text File

<https://www.isolezwelesixhosa.co.za/>

Process URLs



Home	Submissions / Process Status	Corpora Downloads	Help
------	------------------------------	--------------------------	------

<p>Northern Sotho [nso] South Africa</p> <p>Corpus: nso_community_2017 Sentences: 4746, Types: 10166, Tokens: 108522 URLs: 430</p> <p style="text-align: right;">more details ...</p>
<p>Southern Ndebele [nbl] South Africa</p> <p>Corpus: nbl_community_2017 Sentences: 318, Types: 2643, Tokens: 5424 URLs: 29</p> <p style="text-align: right;">more details ...</p>
<p>Southern Sotho [sot] South Africa, Lesotho</p> <p>Corpus: sot_community_2017 Sentences: 9773, Types: 17421, Tokens: 238709 URLs: 542</p> <p style="text-align: right;">more details ...</p>
<p>Swati [ssw] South Africa, Swaziland</p>

Collecting Usage Samples

Xhosa:

- 18,000 unique sentences

Kalanga:

- 621 unique sentences
(only Watchtower / Jehova's Witnesses)


Joint effort with Pretoria (Sonja Bosch and colleagues)


- 180 seed URLs for Xhosa
- 1 URL for Kalanga

[Home](#)[Submissions / Process Status](#)[Corpora Downloads](#)[Help](#)

Corpus Information for Xhosa [xho] South Africa

Language Xhosa
ISO Code xho [Wikipedia](#) , [Ethnologue](#) 
Country South Africa

Corpus Name xho_community_2017 [Corpus Information](#) 
Tokens 401067
Types 89008
Sentences 23993
Sources (URLs) 2503
Build date 2017-10-16

Corpus Name xho_community_2019 [Corpus Information](#) 
Tokens 972301
Types 172520
Sentences 63387
Sources (URLs) 4227
Build date 2019-03-29

URLs [List of URLs](#) download
[List of Domains](#) download

Download [xho_community_2017](#) 2017-10-16
[xho_community_2019](#) 2019-03-29

Xhosa:

45,000 unique sentences added

Kalanga:

1,000 unique sentences added

Xhosa: Sample sentences for about 25% of the lexical entries

Results: CLARIN-Integration

VLO / Faceted search / Search results / Record: Xhosa community corpus from 2019 (xho_community_2019)



Record 4 of 911

< previous next >



Xhosa community corpus from 2019 (xho_community_2019)

- Search page for this record
- Plain text search via Federated Content Search

Record details Links (1) Availability All metadata Technical details Hierarchy

Name	Xhosa community corpus from 2019 (xho_community_2019)
Description	Xhosa community corpus based on material from 2019
Collection	Leipzig Corpora Collection
Language	Xhosa
Genre	community text
Organisation	CLARIN-D center, Natural Language Processing Group, University of Leipzig
National project	CLARIN-D
Resource type	text corpus
Data provider	ASV Leipzig



Conclusion

We presented a use case:

- Enriching lexical resources for lesser resourced languages with sample sentences
- Based on the CLARIN-infrastructure (tools and data)

Open work / next steps:

- Integration of usage samples into BLM
- Meeting with South African lexical units, aiming for making more lexical data freely available in cooperation with SADiLaR
- Development of lexicon app based on these datasets

Thank you!

Dirk Goldhahn¹, Thomas Eckart¹, Sonja Bosch²

¹Natural Language Processing Group, Institute of Computer Science, University of Leipzig, Germany

²Department of African Languages, University of South Africa, South Africa

