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
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):

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>abelo</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>lot</td>
<td>share</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>abhalala</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>striped Sotho blanket</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>abhobho</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>chasm</td>
<td>yawning cavity</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>abi</td>
<td>noun</td>
<td>um</td>
<td>1 ab</td>
<td>2</td>
<td>divider</td>
<td>distributor</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>abonkolo</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>tadpole</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>aci</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>motto</td>
<td>pet saying</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>adlunge</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>wagon tree</td>
<td>Protea grandiflora</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>atobe</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>idioms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>akhi</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>formative</td>
<td>morpheme</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>akhiwo</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>building</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>akhiwo</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>structure</td>
<td>building</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>akhiwo sesikhumbuzo</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>monument</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>akhono</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>dexterity</td>
<td>ability to manipulate</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>alam</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>poor person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>alathandawo</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>agentative</td>
<td>demonstrative pronoun</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>alathisi</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>demonstrative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>alelo</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>prohibition</td>
<td>opposition</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>ambalo</td>
<td>noun</td>
<td>is</td>
<td>7 iz</td>
<td>8</td>
<td>necklace</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>ambantlanya</td>
<td>noun</td>
<td>is</td>
<td>7</td>
<td></td>
<td>outcry</td>
<td>gist</td>
<td></td>
</tr>
</tbody>
</table>
Bantu Language Model (1/2)

- Structured using Bantu Language Model (BLM)
  - RDF/OWL-based ontology
  - Based on the *Multilingual Morpheme Ontology* (MMoOn)
Bantu Language Model (2/2)

- Access via dedicated Web page, SPARQL endpoint and downloads

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?

<table>
<thead>
<tr>
<th>Virtual Language Observatory</th>
<th>Search</th>
<th>Help</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLD</td>
<td>Faceted search</td>
<td>Search results</td>
</tr>
</tbody>
</table>

Record 1 of 11

**Lexical and Morphological isiXhosa Dataset**

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

<table>
<thead>
<tr>
<th>Record details</th>
<th>Resources (1)</th>
<th>Availability</th>
<th>All metadata</th>
<th>Technical details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Lexical and Morphological isiXhosa Dataset</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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. Louise, 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 Eckert and Bettina Klimke (Leupolz 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 (SADiLeR). More information about the resource can be found in the following publications: Sonja Bosch, Thomas Eckert, Bettina Klimke, Dirk Goldmann and Uwe Quasthoff (2013): '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).
Dissemination
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?