Creating linguistic resources for use in digital humanities: notes from one proudly South African adventure

Marissa Griesel

15 September 2021
DH Colloquium
In this presentation...

• Quick introduction to the AfWN and the MLTB respectively
• Our approach and experiments
• What we learnt from it
• Suggestions and conclusion
Wordnets in a nutshell

- Princeton WordNet for English
  (https://wordnet.princeton.edu/)
- Open Multilingual Wordnet
  (http://compling.hss.ntu.edu.sg/omw/)
AfWN development: Background (1)

- The African Wordnet has steadily added more synsets, definitions and usage examples to wordnets for 9 South African languages
  - 2007: Initial workshop with all South African languages represented
  - 2010: Development starts for isiZulu, isiXhosa, Setswana and Sesotho sa Leboa
  - 2013: Addition of Tshivenđa
  - 2018: Addition of Sesotho & Siswati
  - 2019: Addition of Xitsonga & isiNdebele
  - 2020-2022: Expansion of these languages and new release
- South African Digital Language Resource Centre (SADiLaR) became the long-term host for the AfWN in 2018
AfWN development: Background (2)

- The *expand approach* is suggested for less-resourced languages
- Uses the tried and tested ontological structure and content of an existing wordnet (often the English Princeton WordNet (PWN)) as a template
- Translation of synsets from (American) English to the target language
- Advantages: no advanced language resources needed & a less complicated approach for inexperienced teams
- Disadvantages: labour intensive task; requires advanced linguistic knowledge from the team & time spent on unlexicalised concepts

AfWN followed this advice and is still using the PWN as guide, although the need for more intricate localization has become obvious

- The AfWN currently includes roughly 70 000 synsets across the 9 languages
First experiment: Linking wordlists to the PWN (1)

- The AfWN team acquired the following resources:
  - English-Venda Bilingual Dictionary (Murphy; 1997)
  - The Comprehensive Northern Sotho Dictionary (Ziervogel and Mokgokong; 1975)
    - trilingual monodirectional dictionary for Northern Sotho, Afrikaans and English;
  - Setswana-English Dictionary (Brown; 1925);
  - For isiXhosa and isiZulu, a selection of private bilingual wordlists were made available to the project team
- Since language resources are a scarce commodity, we have to make the most of what we have
Current strategy: SILCAWL list and translations (1)

• An African alternative to Eurocentric seed lists can be found in the SIL Comparative African Wordlist (SILCAWL), compiled by Keith Snider and James Roberts (2006)
  • It is a list of lexical data (1700 words with both English and French glosses) resulting from linguistic research in Africa
  • The items are organized semantically under twelve main headings from man’s physical being to animals and grammatical items
• The hypothesis: Creating a more localised (South African) English source text and then translating this into all 9 South African languages not only fast-tracks development of new synsets, but also forms a multilingual parallel sub-corpus of commonly used concepts
Current strategy: SILCAWL list and translations (2)
<table>
<thead>
<tr>
<th></th>
<th>English Synonym</th>
<th>PWN ID</th>
<th>POS</th>
<th>NSO Synonym</th>
<th>English Definition</th>
<th>NSO Definition</th>
<th>English Usage</th>
<th>NSO Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>body</td>
<td>ENG20-05232131-n</td>
<td>n</td>
<td>kseoto</td>
<td>the body excluding the head and neck and limbs</td>
<td>mmelo go se akorešewe hlogo le molela le maeso</td>
<td>the dancers moved their bodies to the music</td>
<td>babini ba sepedisitsi mebele ya bona ba binela rminono</td>
</tr>
<tr>
<td>3</td>
<td>skin (of man)</td>
<td>ENG20-04996703-n</td>
<td>n</td>
<td>letialo</td>
<td>a natural protective covering of the body, site of the sense of touch</td>
<td>selo sa tihago seo se pipilego mmelo, lehelakele le setwi sa go kgongwa</td>
<td>your skin is sensitive to the sun</td>
<td>letialo la gago le tšhaba lefatsi</td>
</tr>
<tr>
<td>4</td>
<td>head</td>
<td>ENG20-05221598-n</td>
<td>n</td>
<td>hiogo</td>
<td>the upper part of the human body or the front part of the body in animals, contains the face and brains</td>
<td>karolo ya godingwona ya mmelo wa motho goba</td>
<td>he banged his head on the low door</td>
<td>o le a bethša hlogo ya gagwe lebating la fasana</td>
</tr>
<tr>
<td>5</td>
<td>forehead</td>
<td>ENG20-05282535-n</td>
<td>n</td>
<td>phetlo</td>
<td>the part of the face between the hair and the eyes.</td>
<td>karolo ya sefahlelo gareng ga moriri le maelo. Tšhalošo e tšwa go OERTB (<a href="http://oertb.titerm.com/">http://oertb.titerm.com/</a>))</td>
<td>I have wrinkles on my forehead</td>
<td>ke na le mašelo phatšeng ya ka</td>
</tr>
<tr>
<td>6</td>
<td>face</td>
<td>ENG20-05280860-n</td>
<td>n</td>
<td>sefahlelo</td>
<td>the front of the human head from the forehead to the chin and ear to ear</td>
<td>motho go tšopa phatšeng go ya setšedišong le go tšopa tsebeng go ya go ye ngwe</td>
<td>I wish I had seen the look on his face when he got the news</td>
<td>ke dumga ge nhà le ke bone sefahlelo sa gagwe go a skwa dītša</td>
</tr>
</tbody>
</table>
kitchen utensil

ibhodwe [pot]

isiphakuluzi [ladle]

imvogoqo 2 (isiZulu)
imvongoqo 1 (isiZulu)

ikhanzi 1 (isiZulu)

isico 1 (isiZulu)
isiyoco 1 (isiZulu)

isikhetho 1

imvogoqo 2

isibambo 3

Definition
isipini sokuphaka esikhulu esibaziwe esinesitsha esinomgodi oshonayo noma esinomgodi ovulekile

Examples
ubaba ubaza imvogoqo

Link
Lessons learnt: technical errors

- The existing AfWN data contains errors that can be identified and corrected (semi-) automatically.
- Although the translators did an excellent job, many errors still occurred.
- Language independent SQL scripts were developed to identify and correct these errors.
  - Capital letters and punctuation;
  - Empty mandatory fields;
  - Invalid/incorrect information in fields with set values (POS and semantic relations for instance); etc.
Lessons learnt: language specific decisions (1)

- The South African English source text certainly made a difference but we soon realised that each language would also present unique challenges.
- Translation for wordnets was definitely a specialised activity due to the hierarchical nature and technical constraints.
- African languages often use euphemisms to refer to taboo terms and are the preferred form in some languages.
- Xitsonga: the concept of “breaking wind/farting”
  - Biological translation = *tamba*
  - Preferred euphemism = *humesa moya* – literally translated as “to kill an insect”
- AfWN team decided to include both as synonyms with a tag to indicate the taboo term.
Lessons learnt: language specific decisions (2)
Lessons learnt: language specific decisions (3)

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Lexical gap: no distinction between *gourd* and *calabash* in (American) English

Word to search for: gourd

Noun

- **S**: (n) *gourd*, *calabash* (bottle made from the dried shell of a bottle gourd)
  - *direct hypernym* / *inherited hypernym* / *sister term*
- **S**: (n) *bottle* (a glass or plastic vessel used for storing drinks or other liquids; typically cylindrical without handles and with a narrow neck that can be plugged or capped)
- **S**: (n) *gourd* (any of numerous inedible fruits with hard rinds)
- **S**: (n) *gourd*, *gourd vine* (any vine of the family Cucurbitaceae that bears fruits with hard rinds)
Term banks in a nutshell

Open Educational Resource Term Bank - UP

(http://oertb.tlterm.com/)
Background

• Need for standardisation of terms in African languages
• Different source language for students meant confusion
• Freely available, editable and standardised term bank with linguistic terminology needed
  • To be used in classrooms from a very early age
Process

- Initially the plan was to perform (semi-) automatic term extraction after digitisation of old study guides
- Strange fonts and other technical problems meant this was impossible
- Setting up a manual term list by team of experts was the only way
- Loosely based on terms used in guides
- Multilingual is NB
- Difference – not focussed on English base meaning
<table>
<thead>
<tr>
<th>number</th>
<th>term</th>
<th>definition of the term</th>
<th>source of the definition of the term</th>
<th>ENG</th>
<th>example of use</th>
<th>source of example</th>
<th>question date</th>
<th>notes</th>
<th>see also</th>
<th>internal</th>
</tr>
</thead>
<tbody>
<tr>
<td>66</td>
<td>borrowings</td>
<td>mantsu a setshwene paling e teleng a le tshwane</td>
<td>borrowings</td>
<td>Manetsu a Sesotho a la bokago a la tshwane</td>
<td>borrowing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>category shifting</td>
<td>go theothe ya gana, gana ya theothe a ba</td>
<td>category shifting</td>
<td>Gana theothe ya bokago a la ba</td>
<td>gana theothe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>causative</td>
<td>lela theothe ya la bokago</td>
<td>causative</td>
<td>Lela theothe ya la bokago</td>
<td>causative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>chronological succession</td>
<td>lela theothe ya la bokago</td>
<td>chronological succession</td>
<td>Lela theothe ya la bokago</td>
<td>chronological succession</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>classification</td>
<td>lela theothe ya la bokago</td>
<td>classification</td>
<td>Lela theothe ya la bokago</td>
<td>classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>clique</td>
<td>lela theothe ya la bokago</td>
<td>clique</td>
<td>Lela theothe ya la bokago</td>
<td>clique</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>click sound</td>
<td>lela theothe ya la bokago</td>
<td>click sound</td>
<td>Lela theothe ya la bokago</td>
<td>click sound</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>click</td>
<td>lela theothe ya la bokago</td>
<td>click</td>
<td>Lela theothe ya la bokago</td>
<td>click</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>connectedness</td>
<td>lela theothe ya la bokago</td>
<td>connectedness</td>
<td>Lela theothe ya la bokago</td>
<td>connectedness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>cognitive meaning</td>
<td>lela theothe ya la bokago</td>
<td>cognitive meaning</td>
<td>Lela theothe ya la bokago</td>
<td>cognitive meaning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>collective noun</td>
<td>lela theothe ya la bokago</td>
<td>collective noun</td>
<td>Lela theothe ya la bokago</td>
<td>collective noun</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Welcome to Lexonomy, a cloud-based, open-source platform for writing and publishing dictionaries.

**Gentle introduction to Lexonomy**
A short and sweet guided tour of Lexonomy for beginners.

**Introducing Lexonomy: an open-source dictionary writing and publishing system** [PDF]
A conference paper offering a detailed review of Lexonomy's features.

**Reference**

If you are referring to Lexonomy from an academic publication, it is recommended that you cite this paper.
```xml
<entry>
  <english_term/>
  <english_definition/>
  <zulu_translation>
    <zul_term/>
    <zul_definition/>
    <zul_definition_source/>
    <zul_usage/>
    <zul_sealso/>
  </zulu_translation>
  <nso_translation>
    <nso_term/>
    <nso_definition/>
    <nso_definition_source/>
    <nso_usage/>
    <nso_sealso/>
  </nso_translation>
  <xhosa_translation>
    <xhosa_term/>
    <xhosa_definition/>
    <xhosa_definition_source/>
    <xhosa_usage/>
    <xhosa_sealso/>
  </xhosa_translation>
  <tsonga_translation>
    <tsonga_term/>
    <tsonga_definition/>
    <tsonga_definition_source/>
    <tsonga_usage/>
    <tsonga_sealso/>
  </tsonga_translation>
  <venda_translation>
    <venda_term/>
    <venda_definition/>
    <venda_definition_source/>
    <venda_usage/>
    <venda_sealso/>
  </venda_translation>
</entry>
```

**Child elements**
- Text
- Text with markup
- Value from list
- Empty
- Media

<table>
<thead>
<tr>
<th>Child elements</th>
<th>Text</th>
<th>Text with markup</th>
<th>Value from list</th>
<th>Empty</th>
<th>Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;english_term&gt;</td>
<td>min 1</td>
<td>max 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;english_definition&gt;</td>
<td>min 1</td>
<td>max 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;zulu_translation&gt;</td>
<td>min 1</td>
<td>max 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;nso_translation&gt;</td>
<td>min</td>
<td>max</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;xhosa_translation&gt;</td>
<td>min</td>
<td>max</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;tsonga_translation&gt;</td>
<td>min</td>
<td>max</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;venda_translation&gt;</td>
<td>min</td>
<td>max</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
absolute pronoun
abstract noun
acoustic phonetics
acronym
active articulators
active sentence
actuality
adaptation
addressee; second person
adjectival concord
adjectival phrase
adjunct
adoptive
adverbial phrase

---

**<english_term>adverbial phrase</english_term>**

**<zulu_translation>**

---

**<zul_term>ibinzana lesandiso</zul_term>**

**<zul_definition>ingxenye yomusho esebenza njengesandiso</zul_definition>**

**<zul_definition_source>**

---

**<zul_usage>Emshweni othi Ubaba uhaba ngebhasi elisuka kusempendozankomo uma eya emsebenzini ibinzana elithi ngebhasi elisuka kusempendozankomo lisebenza njengesandiso.</zul_usage>**

---

**<zul_notes>**
adverbial phrase

An adverbial phrase is a group of words that functions as an adverb.

**isiZulu** ibinzana lesandiso

- ingxenye yomusho esebenza njengesandiso
  - *Emshweni othi Ubaba uhamba ngebhasi elisuka kusempondozankomo uma eya emsebenzini ibinzana elithi ngebhasi elisuka kusempondozankomo lisebenza njengesandiso.*

**Sesotho sa Leboa** sekafo sa lehlati

- sekafo se se tlaeleletša go tiro ya lediri
  - *sekafo sa lehlati ke se bjalo ka godimo ga magala mo lefokong le: Bana ba dutše godimo ga magala*
Suggestions (1)

• Some improvements could be implemented immediately:
  • Better understanding of the use of the translations by both the English lexicographer and the African language translators is essential;
  • The style guide was updated to include more detailed instructions with examples in various languages;
  • Where possible, closely related languages such as isiZulu and isiXhosa could benefit from shared translations.
Suggestions (2)

- Other measures will take some more work to implement:
  - AfWN: Adding morphological analysis or as a minimum lemmatisation to the quality assurance pipeline to check that the lemma is used in the usage example, for instance, will reduce the amount of confusing sentences;
  - Both projects: Including quality assurance measures such as spell checking, checking for mandatory fields left open or invalid values for some fields directly into the development interface is a long-term goal.
Conclusion

• All data developed in the project will be made available under a Creative Commons license via the SADiLaR language resource repository
  • Hope to increase NLP development particularly for the African languages;
  • All users have to be aware of certain linguistic and technical decisions made during development so that they can also make provision for certain aspects in their systems;

• Nothing beats a good old round table discussion to sort out some of the standardisation problems or find more creative solutions.
This presentation is based on previously published work. For a full list of research outputs in the project, please see https://africanwordnet.wordpress.com/research/

The African Wordnet (AfWN) and Multilingual Linguistic Terminology project was made possible with support from the South African Centre for Digital Language Resources (SADiLaR). SADiLaR (www.sadilar.org) is a research infrastructure established by the Department of Science and Technology of the South African government as part of the South African Research Infrastructure Roadmap (SARIR).