Thinking globally, acting locally – progress in the African Wordnet Project

Marissa Griesel
University of South Africa
(UNISA)
Pretoria, South Africa
griesm@unisa.ac.za

Sonja Bosch
University of South Africa
(UNISA)
Pretoria, South Africa
boschse@unisa.ac.za

Mampaka L. Mojapelo
University of South Africa
(UNISA)
Pretoria, South Africa
mojapml@unisa.ac.za

Abstract

The African Wordnet Project (AWN) includes all nine indigenous South African languages, namely isiZulu, isiXhosa, Setswana, Sesotho sa Leboa, Tshivenda, Siswati, Sesotho, isiNdebele and Xitsonga. The AWN currently includes 61 000 synsets as well as definitions and usage examples for a large part of the synsets. The project recently received extended funding from the South African Centre for Digital Language Resources (SADiLaR) and aims to update all aspects of the current resource, including the seed list used for new development, software tools used and mapping the AWN to the latest version of PWN 3.1. As with any resource development project, it is essential to also include phases of focused quality assurance and updating of the basis on which the resource is built. The African languages remain under-resourced. This paper describes progress made in the development of the AWN as well as recent technical improvements.

1 Introduction

The African Wordnet Project (AWN) has seen various phases of development with different funding cycles and collaborators (see Bosch & Griesel, 2017 for a comprehensive breakdown of previous phases). The most recent cycle is funded by the South African Centre for Digital Language Resources (SADiLaR)[1] and will run from 2018 to the end of February 2020, with an extension to 2022 currently under consideration. The most notable change to the project in the past two years is the addition of four further languages to include the full range of nine indigenous South African languages, namely isiZulu (ZUL), isiXhosa (XHO), Setswana (TSN), Sesotho sa Leboa (NSO), Tshivenḓa (VEN), Siswati (SSW), Sesotho (SOT), isiNdebele (NDE) and Xitsonga (TSO). The number of synsets, usage examples and definitions for all languages included in the AWN have also been substantially increased. As with any resource development project, it is essential to include phases of focused quality assurance and updating of the basis on which the resource is built. For the AWN, this meant reassessing several core aspects, including the seed terms used for further development, software to assist linguists to develop and structure the wordnets, as well as the process by which development is managed.

The African languages remain under-resourced despite progress being made with a resource catalogue hosted by the Resource Management Agency of SADiLaR. Currently there are still no freely available dictionaries for any of the languages and as Oliver (2014:7) notes: “The most commonly used strategy within the expand model is the use of bilingual dictionaries”. In this paper, key aspects pertaining to the development of a multilingual wordnet for such under-resourced languages will be highlighted and our solutions to challenges that emerged as a result of the growth in the scope of the project, will be discussed. The last section of the paper will mention smaller challenges and project specific matters that might be of interest to other projects with similar restrictions.

2 Recent progress

The AWN team first began the development of wordnets for South African languages in 2010 and has grown slowly but consistently. Currently, the AWN includes 61 000 synsets across the nine identified languages. The number of synsets per language varies from nearly 17 000 for Setswana, to only 600 each for isiNdebele and Xitsonga. This variation is due to the amount of time linguists have available to work on the project as well as the incremental addition of languages to the project (see below). In addition to the basic synsets, the AWN also includes 26 500 definitions

[1] https://www.sadilar.org/

and 37 000 usage examples across the nine languages.

One of the most significant expansions to the AWN over the past two years has been the addition of four new languages. This means that all nine indigenous South African languages are now represented[2], although not in equal numbers yet. The current funding phase will see isiNdebele and Xitsonga also grow to 1 000 synsets each, with definitions and usage examples.

In addition to this, the AWN has also added definitions to the synsets already captured in previous phases. Where the developers initially focussed only on synsets with their usage examples, feedback from the South African Digital Humanities and Human Language Technology communities indicated that definitions would make the AWN even more useful in language learning applications – an ever-growing research and development area given the multilingual nature of the country. An initial experiment into this application is described in Bosch and Griesel (2018). In the second application, data from the AWN has also been used experimentally in the Kamusi GOLD project[3] to populate an online dictionary for which definitions and usage examples are important.

Section 3 describes another significant decision regarding the content of the AWN – moving away from relying on the Eurocentric core base concepts[4] (CBC) to a more localised wordlist to be used as seed terms for new synsets.

3 The SIL list as seed terms

3.1 Contextualisation

The SIL Comparative African Wordlist (SIL-CAWL) was compiled in 2006 by Keith Snider (SIL International and Canada Institute of Linguistics) and James Roberts (SIL Chad and Université de N'Djaména). It is a list of lexical data consisting of 1 700 words with both English and French glosses which resulted from linguistic research in Africa. The items are organised semantically under 12 main headings which generally move on a continuum from items relating to human domains on the one extreme, via animate domains, to items relating to non-human domains on the other extreme, and then from concrete items to more abstract items. The following are the 12 main headings:

1. Man's physical being
2. Man's non-physical being
3. Persons
4. Personal interaction
5. Human civilisation
6. Animals
7. Plants
8. Environment
9. Events and actions
10. Quality
11. Quantity
12. Grammatical items

Table 1. Headings in the SIL CAWL list

Each of the above headings is then subdivided into second and third level headings. For instance, in the case of Persons, the following first level headings are distinguished: STAGES OF LIFE, BLOOD RELATIONS, MARRIAGE RELATIONS, RELATIONS, EXTENDED AND SOCIAL, and PROFESSIONS. A third level, for example, in the case of PROFESSIONS includes divisions such as: farmer, fisherman, hunter, blacksmith, potter, weaver, medicine man etc. The parts of speech covered in the SIL list are nouns, verbs, adjectives, adverbs, pronouns, interrogatives and conjunctions. Although Snider and Roberts (2006:4) concede that they still notice "imperfections and room for improvement (e.g. words that could be deleted, words that could be added, words that could be moved to different semantic domains etc"), the SIL list has proven to be a welcome improvement on the CBC list used in the past in the development of the AWN that follows the expand model (Vossen, 1998) and is based on the English Princeton WordNet (PWN) (Fellbaum, 1998). The most significant improvement is observed against the background of localisation where the content (of the entries) would be lexicalised within an African environment.

3.2 Comparison of the SIL list to the core base concepts

The CBC list is a combination of seed lists extracted from European language corpora for the EuroWordNet and BalkaNet projects (see a description of the core base concepts list at http://globalwordnet.org/gwa-base-concepts/). The CBC aims at covering terms that display many relations with other terms (synsets) and are also placed high in the semantic structure of a wordnet. It includes very basic terminology such as "light" (noun), "Earth" (noun), "catch" (verb) and "shake" (verb), but also less frequently used terms such as "actinic radiation" (noun) and "protozoan" (noun). As discussed in Bosch and

[2] An Afrikaans wordnet already exists independently from this project but is not currently under active development. See https://hdl.handle.net/20.500.12185/158.

[3] https://kamusi.org

[4] As found on http://globalwordnet.org/?page_id=68

Griesel (2017), these unfamiliar terms caused some problems for the African language team, resulting in wasted time and lost momentum in the early phases of development and the team decided to investigate an alternative seed list such as the SIL list described in Section 3.1, drawn up from local African sources.

All terms in the CBC can be found in the PWN and therefore have a direct mapping to the larger wordnet structure with a unique identifying number. The SIL list, however, includes 41 terms that have no equivalents in the PWN. These terms are not necessarily foreign to an English native speaker but might be more frequently used in the African context. They include terms such as "cooking stone" (noun) and "thorn tree" (noun).

Another noteworthy category of terms that is present in the SIL list but not in the CBC includes various terms where African languages make a distinction based on usage that other languages might not make but are well known (lexicalised) to native speakers. The South African languages, for instance, distinguish between harvesting by digging up versus harvesting by cutting or plucking, etc. The subtle differences between these terms in isiZulu and Sesotho sa Leboa are illustrated in Table 2.

SILCAWL	**ZUL**	**NSO**
0757 harvest (maize) (v)	*ukuvuna* *ukukwica* *ukucasa* (while still green, harvest green corn before it has hardened) *Comment:* synonyms, or near synonyms in the case of *ukuvuna* and *ukucasa*	*buna* *Comment:* general concept related to the time of harvest
0758 harvest, dig up (yams)	*ukuvuna* *Comment:* same as harvesting crops that grow above the ground	*bupula* *Comment:* harvest groundnuts
0760 harvest, collect (honey from hive)	*ukuthapha* *Comment:* extract, take out honey from a hive.	*rafa* *Comment:* extract honey from a hive.

Table 2. Harvesting in isiZulu and Sesotho sa Leboa

Kinship terms are another instance of a very intricate system in the African languages as illustrated in a few examples in Table 3.

SILCAWL	**ZUL**	**NSO**
BLOOD RELATIONS		
0348 father's brother (uncle)	*ubabamkhulu* (big father) 'father's elder brother' *ubabomncane* (small father) - 'father's younger brother'	*ramogolo* 'father's elder brother' *rangwane* 'father's younger brother'
0351 father's sister (aunt)	*ubabekazi* (female father) 'father's sister'	*rakgadi* 'father's sister'
0349 mother's brother (uncle)	*umalume* (male mother) 'mother's brother'	*malome* 'mother's brother'
0350 mother's sister (aunt)	*umamekazi* (female mother) or *umame* 'mother's sister'	*mmamogolo* 'mother's elder sister' *mmane* 'mother's younger sister'
MARRIAGE RELATIONS		
0365 father-in-law	*ubabezala* 'father-in-law' used by Zulu-speaking woman *umukhwe* 'father-in-law' used by Zulu-speaking man	*ratswale* 'father-in-law'
0366 mother-in-law	*umkhwekazi* 'mother-in-law' used by Zulu-speaking man *umamezala* 'mother-in-law' used by Zulu-speaking woman	*mmatswale* / *mogwegadi* 'mother-in-law' (man speaking – dialectal) *mmatswale* 'mother-in-law' (woman speaking)
0367 brother-in-law	*umfowethu* 'husband's brother' *umkhwenyawethu* 'sister's husband' (man speaking) *umlamu* 'wife's brother' *umkhwenyana* 'sister's husband' (woman speaking)	*molamo, sebara* 'sister's husband' (man and woman speaking) *molamo, sebara* 'wife's brother' (man speaking)
0368 sister-in-law	*udadewethu* 'husband's sister'	*mogadibo*

	umakoti, umlobokazi, umkami ‘brother’s wife’ (man speaking) *umlamu* ‘wife’s sister’ *umakoti womfowethu, umakoti womnewethu* ‘brother’s wife’ (woman speaking)	‘husband’s sister’/ ‘brother’s wife’

Table 3. Kinship terms in isiZulu and Sesotho sa Leboa

3.3 Translation of the expanded SIL list

One of the advantages of a common seed list such as the SIL list across all the languages in the AWN, is that it enables the creation of a parallel corpus within the larger wordnet structure. Parallel synsets are not only useful for language learners, but also in applications such as multilingual information retrieval, semantic analysis and machine translation. The AWN team therefore decided to incorporate the terms in the SIL list using the following steps: first, the English term in the SIL list was compared to the PWN and an ID to the corresponding synset was added to each term in the SIL list. If available, the definition and usage example from the PWN was also extracted to a simple spreadsheet. This document was next presented to an expert South African English lexicographer to a) fill in any gaps there might still be so that each term has a part of speech tag, definition and usage example; and b) edit the existing PWN data to fit the South African context better.

The African language translators were briefed on the nature of the project and specifically on the unique characteristics of a wordnet with a strict protocol to follow. Translation of the first 1 000 synsets from the expanded SIL list dataset took roughly five months, including internal quality assurance. The output of this process was a multilingual parallel corpus of common terms, each with a clear definition, usage example and part of speech tag. This is already a valuable resource, but for inclusion into the AWN, we will now need to incorporate this data into the hierarchical structure of a wordnet, identify the relations within this structure and perform formal quality assurance. This process is currently ongoing.

4 Visualisation of the AWN in WordnetLoom

Developing data to populate a wordnet offline in spreadsheets has its advantages, most notably fast tracking of development because it is a familiar process for inexperienced linguists, ease of applying spell checking or other quality assurance, no delays due to interruptions in internet connectivity or access to a central server, etc. However, it is very difficult to see the true nature of a wordnet with connecting relations and multilingual similarities. The AWN previously used the DEBVisDic editor (see Rambousek & Horak, 2016) to facilitate development and align work across the different languages. At the onset of the current phase, however, it became clear that a focus on quality assurance of, especially the semantic relations, was needed and it was decided to port the AWN to WordnetLoom (WNL) (cf. Naskret *et al.*, 2018) – an editor with advanced visualisation of wordnets. While preparing the data for use in this tool, the AWN was also mapped to the PWN 3.0 to ensure the latest format and most up to date English equivalents. To move from DEBVisDic to WNL involved extracting the AWN database in LMF format, whereafter a programmer could map the AWN to PWN 3.0 using an in-house script. Where there was no PWN 3.0 equivalent or ID, the PWN 2.0 ID was retained.

Advantages of this tool are that it speaks to the organic growth development style that the AWN teams have always favoured (see Bosch & Griesel, 2017) and also adds the ability to perform more productive searches when working in a specific domain or looking for a specific semantic relation. The addition of a multilingual relation also means that specific senses in different languages can be connected to each other without having to connect the entire synset. The subtle differences between isiZulu and Sesotho sa Leboa verbs and kinship terms mentioned in Section 3.2 can, for instance, be represented more accurately. Discussions with the development team behind this state-of-the-art software tool led to an intense two-day workshop in South Africa, facilitated by members of the WNL and Polish Wordnet development team where linguists were introduced to WNL and its many advanced features. The workshop, which was hosted by SADiLaR, was attended by at least two linguists from each of the nine languages included in the AWN and took on a very hands-on approach. Figure 1 shows the “harvest” example from Table 2 as it was added to the Sesotho sa Leboa wordnet using WNL.

5 Conclusion and future work

Many challenges, including the low resource nature of the languages in the AWN, restraints on funding, a part-time development team etc. were reported on extensively in Bosch and Griesel (2014). The AWN team have managed to mitigate these risks to a large extent and porting development of the AWN to WNL played a large part in this. The porting process described in Section 4 did not come without some initial challenges and adaptations needed. Most notable is that the visualisation of the AWN now draws our attention to the lack of proper definition and application of the semantic relations between terms. Relations were previously automatically carried over from the PWN as is common when following the expand model. The AWN team was always aware that this method was not fool proof and that relations would need revision. WNL enables linguists to see immediately all synsets connected with any of the predefined relations as well as the lacking relations within the South African context. Some terms also need to be moved from an independent synset to a more accurate embedded synonym position and vice versa. Lexical gaps between the (American) English PWN and the African languages can now also be addressed more effectively by eliminating the need to link a synset in an African language to a synset in the PWN as synsets can either stand independently in WNL or be linked to another African language. Again, the visualisation of the synsets within the larger structure is key in this process of identifying the lexical gaps, as can be seen in Figure 2, a representation of the isiZulu marriage relations discussed in Table 3 above. These aspects will receive priority attention during the quality assurance phase that is underway.

With continued research, collaboration with other developers and an invested interest in growing the African languages as digital language resources, we believe that this project will soon be of significant academic and industrial interest to members of the global wordnet community.

Acknowledgements

The African Wordnet project (AWN) 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).

References

Bosch, Sonja and Griesel, Marissa. 2018. African Wordnet: facilitating language learning in African Languages. *Proceedings of Ninth Global WordNet Conference 2018 (GWC2018*), 12 January 2018, Nanyang Technological University (NTU), Singapore. Available at http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/GWC2018_paper_22.pdf

Bosch, Sonja and Griesel, Marissa. 2017. Strategies for building wordnets for under-resourced languages: the case of African languages. *Literator* 38(1), a1351. https://doi.org/10.4102/lit.v38i1.1351

Fellbaum, Christiane, (ed), 1998. *Wordnet: An electronic lexical database*. The MIT Press, Cambridge, Mass.

Griesel, Marissa and Bosch, Sonja. 2014. Taking stock of the African Wordnet project: 5 years of development. *Proceedings of the Seventh Global WordNet Conference 2014* (GWC2014), pp. 148-153. Tartu, Estonia. Available at http://gwc2014.ut.ee/proceedings_of_GWC_2014.pdf

Naskręt, Tomasz, Dziob, Agnieszka, Maciej Piasecki, Maciej, Saedi, Chakaveh and Branco, António. 2018. WordnetLoom - a Multilingual Wordnet Editing System Focused on Graph-based Presentation. *Proceedings of Ninth Global WordNet Conference 2018 (GWC2018*), 12 January 2018, Nanyang Technological University (NTU), Singapore. Available at http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/gwc-2018-proceedings.pdf

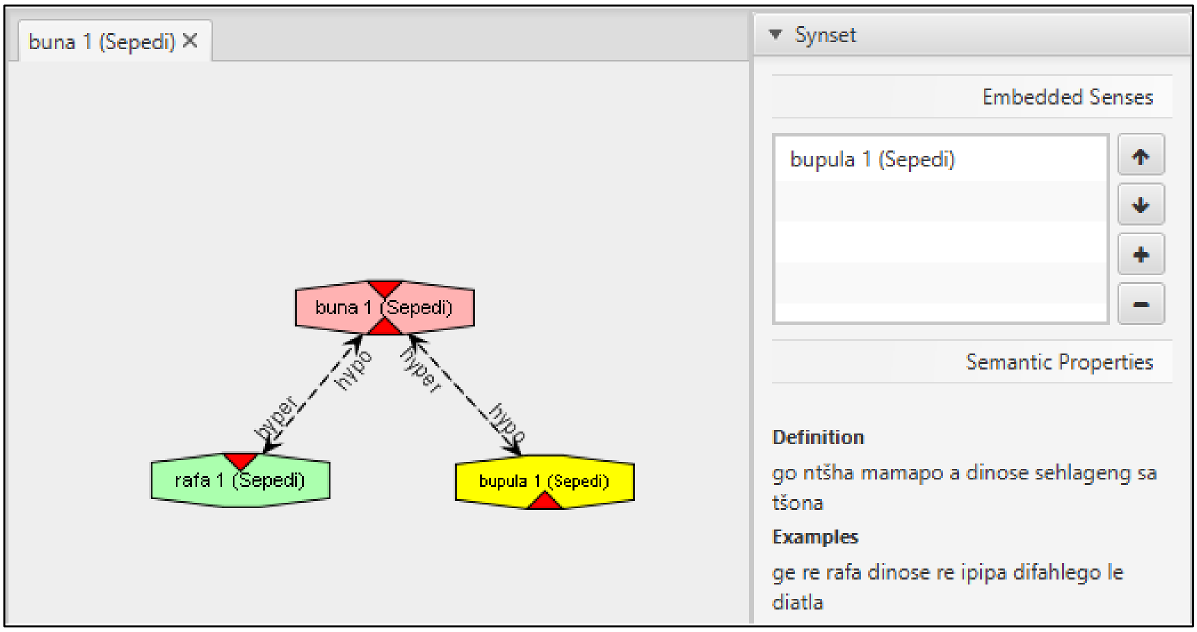
Oliver, Antoni. 2014. WN-Toolkit: Automatic generation of WordNets following the expand model. *Proceedings of the Seventh Global WordNet Conference 2014* (GWC2014), Tartu, Estonia. Available at http://gwc2014.ut.ee/proceedings_of_GWC_2014.pdf

Princeton University. 2017. WordNet – A lexical database for English. https://wordnet.princeton.edu/ Accessed on 12 March 2019.

Rambousek, Adam and Horák, Aleš. 2016. DEBVisDic: Instant Wordnet building. In V. Mititelu, C. Forăscu, C. Fellbaum and P. Vossen (eds.), *Proceedings of the Eighth Global WordNet Conference 2016* (GWC2016), Bucharest, Romania, January 25–29, pp. 317–321.

Snider, Keith and Roberts, James. 2006. *SIL Comparative African Wordlist (SILCAWL).* Available at https://www.eva.mpg.de/lingua/tools-at-lingboard/pdf/Snider_silewp2006-005.pdf)

Vossen, Piek. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic, Dordrecht.



bupula 1 (Sepedi)

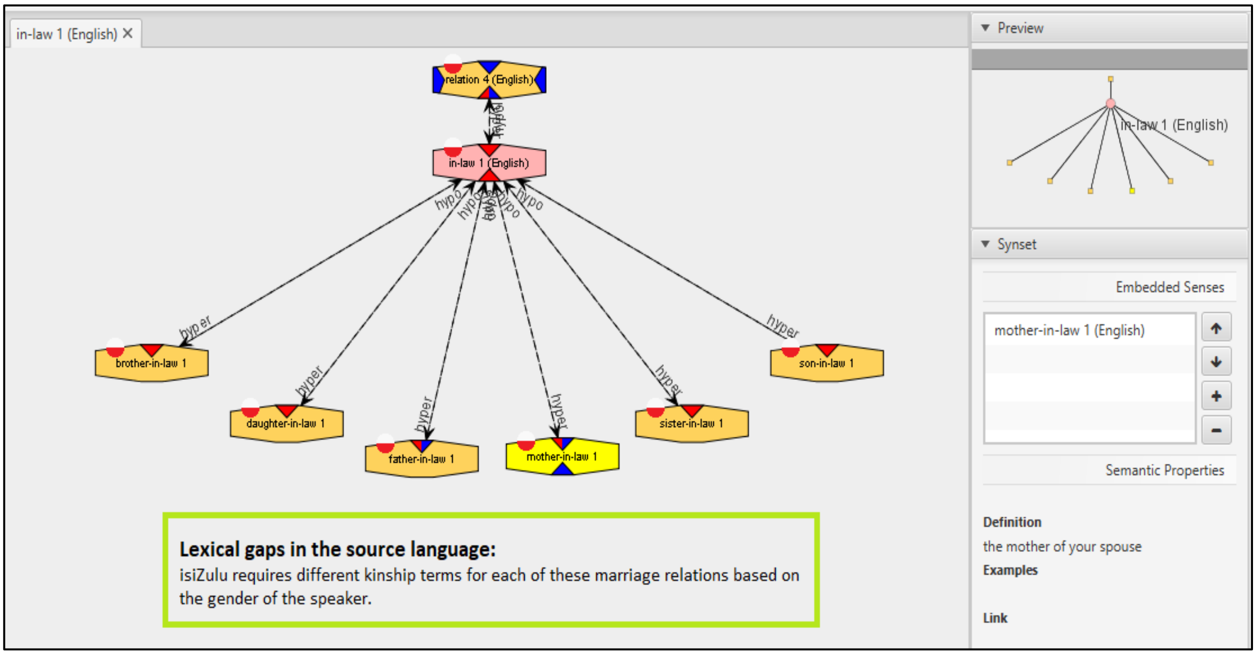
Definition

go ntšha mamapo a dinose sehlageng sa tšona

Examples

ge re rafa dinose re ipipa difahlego le diatla

Fig. 1. *"Harvest" (buna) as included in the Sesotho sa Leboa* wordnet



Lexical gaps in the source language:
isiZulu requires different kinship terms for each of these marriage relations based on the gender of the speaker.

mother-in-law 1 (English)

Definition

the mother of your spouse

Examples

Link

Fig. 2. *Lexical gaps between English and isiZulu for kinship terms*