Taking stock of the African Wordnets project: 5 years of development

Marissa Griesel
University of South Africa (UNISA)
Pretoria, South Africa
griesel.marissa@gmail.com

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

Abstract

This paper reports on the development of the prototype African Wordnet (AWN) which currently includes four languages. The resource has been developed by translating Common Base Concepts from English, and currently holds roughly 42 000 synsets. We describe here how some language specific and technical challenges have been overcome and discuss efforts to localise the content of the wordnet and quality assurance methods. A comparison of the number of synsets per language is given before concluding with plans to fast-track the development and for dissemination of the resource.

1 Introduction

Wordnets for African languages were introduced with a training workshop for linguists, lexicographers and computer scientists facilitated by international experts in 2007. The development of wordnet prototypes for four official South African languages started in 2008 as the African Wordnet Project. This project was based on collaboration between the Department of African Languages at the University of South Africa (UNISA) and the Centre for Text Technology (CTexT) at the North-West University (NWU), as well as support from the developers of the DEBVisDic tools at the Masaryk University[1]. The initiative resulted in first versions of wordnets for isiZulu [zul], isiXhosa [xho], Setswana [tsn] and Sesotho sa Leboa (Sepedi) [nso][2], all members of the Bantu language family. An expansion of the African Wordnet followed in 2011, and currently the development has entered a third phase that aims at solidifying the African Wordnets as a valued resource with formal quality assurance, as well as further expansion of the synsets, definitions and usage examples. Figure 1 gives an overview of the development, as well as the deliverables in each phase.

In this paper, we reflect critically on the previous phases in development including challenges faced and solutions to some common problems. Section 3 gives a brief report on the current standing of the African wordnets and sections 4 and 5 give details regarding future work.

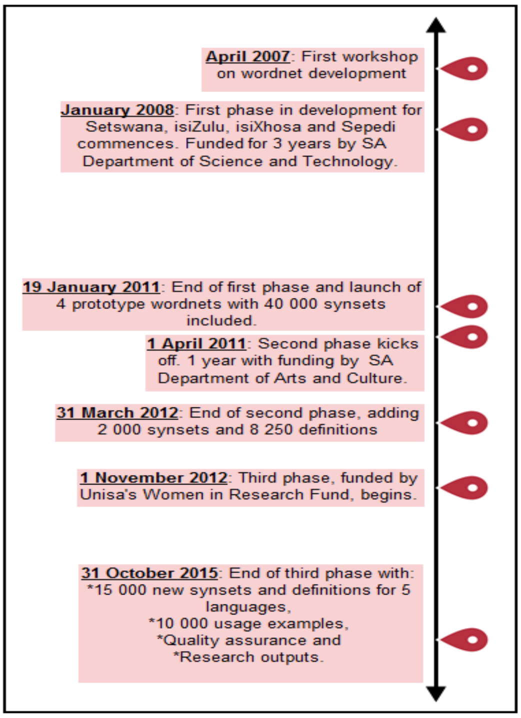
April 2007: First workshop on wordnet development
January 2008: First phase in development for Setswana, isiZulu, isiXhosa and Sepedi commences. Funded for 3 years by SA Department of Science and Technology.
19 January 2011: End of first phase and launch of 4 prototype wordnets with 40 000 synsets included.
1 April 2011: Second phase kicks off. 1 year with funding by SA Department of Arts and Culture.
31 March 2012: End of second phase, adding 2 000 synsets and 8 250 definitions
1 November 2012: Third phase, funded by Unisa's Women in Research Fund, begins.
31 October 2015: End of third phase with: *15 000 new synsets and definitions for 5 languages, *10 000 usage examples, *Quality assurance and *Research outputs.

Figure. 1. Timeline of development in the African Wordnet Project.

2 Status quo after the first 2 phases

During the first phase (2008-2010), linguists who had participated in the introductory workshop were invited to partake in the project. Linguists representing the four languages mentioned above, volunteered and since then, the development has been constant with two phases completed. Table 1 gives a

[1] See http://deb.fi.muni.cz/clients-debvisdic.php

[2] Each language is followed by its ISO 639-3 code (ISO 2012) in order to distinguish one language from other languages with the same or similar names and to identify the names of cross-border languages.

summary of the total number of synsets and definitions that have been developed thus far.

Language	Synsets	Definitions
isiZulu	10 000	2563
isiXhosa	10 000	2370
Setswana	15 000	1755
Sesotho sa Leboa (Sepedi)	7005	2062
Total	**42005**	**8250**

Table 1. Total number of synsets and definitions developed for four African languages.

As will be mentioned in section 3, the team faced many challenges and had to apply some creative problem solving at times. During the first two phases, important fundamental training and development had to be done, for instance a second workshop, again facilitated by international wordnet experts was held at the beginning of 2011, followed by training on more technical aspects of wordnet development such as automated quality control, in 2012. The core project team has stayed largely unchanged and renewed funding for a third phase of development contributed to the continued growth of the African Wordnets.

3 Challenges to the development of African Wordnets

3.1 Availability of resources

The languages in this project are considered resource scarce compared to most other languages listed in The Global WordNet Organization[3] in the sense that lexical resources are relatively limited. The four languages included in the project so far, however, each have at least one or two paper dictionaries available, ranging from monolingual to bilingual general purpose or learners' dictionaries. Apart from a basic on-line dictionary for Sesotho sa Leboa[4] and isiZulu.net[5], which is an online isiZulu-English dictionary that anyone can contribute to, containing bidirectional lookups as well as basic morphological decomposition, there are no online or machine-readable lexicons available for any of the languages.

Currently only relatively restricted unannotated and not freely accessible corpora are available. For example, the University of Pretoria Corpora (Prinsloo & de Schryver, 2005:101) range from approximately two to nine million tokens for the various South African languages. Three types of corpora have been collected, viz. general purpose (LGP) corpora, special-purpose (LSP) corpora and true parallel corpora. The main characteristics of the eleven South African LGP corpora, which are the biggest of the three types built, are shown in Table 2.

Corpus Name	Acronym	Tokens	Types
Pretoria isiNdebele Corpus	PNC	1,959,482	250,990
Pretoria siSwati Corpus	PSwC	4,442,666	293,156
Pretoria isiXhosa Corpus	PXhC	8,065,349	846,162
Pretoria isiZulu Corpus	PZC	5,783,634	674,380
Pretoria Xitsonga Corpus	PXiC	4,556,959	115,848
Pretoria Tshivenda Corpus	PTC	4,117,176	118,771
Pretoria Setswana Corpus	PSTC	6,130,557	157,274
Pretoria Sesotho sa Leboa Corpus	PSC	8,749,597	165,209
Pretoria Sesotho Corpus	PSSC	4,513,287	107,102

Table 2. Pretoria LPG corpora.

Smaller, unannotated parallel corpora are freely available from the newly established Resource Management Agency (RMA). Recently the NLP Group of the University of Leipzig has also made corpora for most of the languages in the African Wordnet project, freely available (Wortschatz Universität Leipzig, 2013). Although these corpora are unannotated and still relatively small, the development work seems promising.

The agglutinating nature of the African languages belonging to the Bantu language family, particularly for those with a conjunctive orthography e.g. isiZulu and isiXhosa, call for morphological annotation for the purposes of accurate corpus searches. Although prototypes of rule-based morphological analysers have been developed for the mentioned two languages, these are not freely available yet (cf. Bosch et al., 2008).

[3] See http://globalwordnet.org/?page_id=38

[4] See http://africanlanguages.com/sdp/

[5] See http://isizulu.net/

Due to the limited availability of lexicographic and basic language resources for the African languages, wordnet construction thus presents a challenging and time-consuming task for the linguists.

3.2 Language specific challenges

A number of language specific challenges anticipated at the beginning of the project are discussed in Le Roux et al. (2007) and will not be repeated here. However, a number of additional challenges were encountered, some of which are dealt with in more detail in a parallel paper (cf. Mojapelo, 2014). For example, consider the following synset for "breaststroke"[6]:

> {00572097} <noun.act>[04] S: (n) breaststroke#1 (a swimming stroke; the arms are extended together in front of the head and swept back on either side accompanied by a frog kick)

A whole discussion arose around the isiZulu version of the above synset since a dictionary entry of the verb *-gwedla* (swim by breaststroke OR paddle/row) was found in a bilingual dictionary (Doke & Vilakazi, 1964:285). The debate among linguists was whether *-gwedla* in the infinitive, i.e. *ukugwedla* (lit. to swim by breaststroke) would be a suitable representation in isiZulu. Some felt that *-gwedla* is more commonly used in the context of 'rowing an actual boat'. To complicate matters, no equivalents for other swimming strokes such as butterfly, backstroke, freestyle etc. are lexicalised in isiZulu, or for that matter, any of the languages in the project.

3.3 Technical challenges

One of the major worries for the African Wordnets team, was securing continual funding for the very important base work. Not only was funding needed to provide technical assistance and project management, but also to reimburse linguists for the linguistic development of the wordnets. All of the linguists involved with this project are employed full time at academic institutions and are not able to devote much of their workday to development of the wordnets, slowing progress almost to a standstill. The BalkaNet project, for instance, also incentivised or contracted the initial development of wordnets for Bulgarian, Greek, Romanian, Serbian and Turkish. The core wordnets delivered at the end of the 3 year project contained roughly 8000 synsets, developed in 3 years – comparative to our 10 000 synsets in each of our African wordnets. The Serbian team then continued development on a voluntary basis and in the next 2 years (2006 – 2008) could only add another 2240 synsets (Krstev et al., 2008). This supports our decision to apply for further funding and continue incentivising the development in order to speed up the process to a point that the wordnets are a truly useful tool for the creation of other NLP applications (where an excess of 200 000 synsets have proven to make a considerable difference in the quality those applications can deliver).

A number of problems with the connection to the server were reported by the linguists. These problems related mainly to the high level of network security and restricted access at the universities involved. The project team was dependent on collaboration of IT-departments from three universities, as we had no direct control over security policies, firewalls, etc. The distance between the linguists (mostly at UNISA in Pretoria, South Africa) and the support team (NWU in Potchefstroom, some 160 km away) also posed a threat to project progress. This risk was managed through an intent focus on regular communication between the sites, and the implementation of a backup plan, namely reverting to working on Microsoft® Excel spreadsheets during 'down-time', and then importing them to the database and online DEBVisDic environment afterwards. Some linguists also experienced regular interruptions in internet connectivity due to a weaker infrastructure in the whole of South Africa. Being able to revert to this offline method, meant that they could continue working from home without needing a constant internet connection.

Human capital development also took time and since this is the first project of its kind for African languages, new technical skills, like working with the DEBVisDic tools, had to be learnt. Because of the slow progress in the first project, the project team had to include more

[6] See http://wordnetweb.princeton.edu/perl/webwn?c=6&sub=Change&o2=1&o0=1&o8=1&o1=1&o7=1&o5=1&o9=&o6=1&o3=1&o4=1&i=0&h=10000&s=breaststroke

linguists in the development of synsets and definitions than initially planned. The advantages of this were twofold. Not only did the progress speed up significantly and were we able to deliver the contracted number of new synsets and definitions on time, but more South African linguists were trained in development of wordnets.

4 Current development

4.1 Introducing the third phase

The aim of the current third project is an extended scope of the African Wordnet Project which gained considerable momentum over the past 4 years. Our primary aim is to develop at least 15 000 new synsets and definitions, to add usage examples to existing synsets and to do continual quality assurance on the wordnets. Most importantly, a 5th African language, Tshivenda [ven], is being added to the project. From the previous phases, it became clear that a stronger emphasis needs to be placed on localisation of the wordnet. It was found that many synsets in the English wordnet are not concepts that belong in the African environment (lexicalised items). During this phase, greater care will be taken to ensure that truly African synsets are included.

4.2 Quality assessment and semi-automatic assistance

As mentioned earlier, very few core technologies exist for the resource scarce African languages. For this reason, many of the internationally proven methods to do quality assessment on wordnets could not be applied (cf. Smrz, 2004 and Kotzé, 2008). The team did have access to proprietary spelling checkers developed for Microsoft® Office. These spelling checkers can be seen as so called first generation technologies, since very little language analysis like with grammar or morphological analysers is available and they rely strongly on lexicon lookup.

The Excel sheets and online versions of the wordnets were consolidated in a single XML file per language before three categories of possible errors were identified automatically. Cells with potential problems were indicated with coloured formatting and linguists were asked to pay special attention while doing quality assessment to these cells. The error categories are:

- Possible spelling errors,
- Empty (critical) fields, and
- Formatting errors (i.e. missing or invalid sense numbers, English IDs and SUMO/MILO relations, recognised with a simple Perl script).

4.3 Localisation of the base concepts

Most of the initial decisions made regarding the design of the African wordnets, were based on the experiences of 2 international projects, namely the BalkaNet project and the EuroNet Project. In both these successful endeavours, the project teams drew up an initial list of the most important concepts to use as seed terms to start building wordnets. These so-called Base Concepts are regarded as “the fundamental building blocks for establishing the relations in a wordnet and give information about the dominant lexicalization patterns in languages” (GWA, 2013). The list of Common Base Concepts created in the EuroNet project contains roughly 1024 synsets. These Common Base Concepts were extended to 5000 synsets and mapped to the Princeton WordNet 2.0 in the BalkaNet project, using a similar approach, but applied to other (mostly European) languages.

During the first 2 phases, we followed the guidance given in the extended Common Base Concepts lists. It soon became clear that a more localised approach was needed, as this and the Princeton Core Concepts list[7] contain concepts that do not accurately describe the African context. Linguists were spending too much time on foreign concepts and especially the less experienced linguists did not have the confidence to venture off this list too far. Table 3 gives some examples of nouns that are not lexicalised in the African languages.

Princeton core set	EuroNet base concepts
abbey	abnegator
apparatus	bellyacher
aquarium	calligrapher
baseball	gasbag
bishop	mesomorph
buffet	scaremonger
kit	slowcoach
mars	tiger
mosaic	twerp
soprano	urchin

Table 3. Unfamiliar words in international standards.

[7] See `http://wordnet.princeton.edu/`

When adding the new Tshivenda wordnet to this project, we decided to take a careful look at the concepts we use as the seed terms. Our premise was that more localised terms might be extracted from real-world parallel corpora. To examine the difference, a multilingual parallel corpus, including English, Setswana, isiZulu, isiXhosa, Sesotho sa Leboa and Tshivenda equivalents was acquired from the RMA. The English version of the parallel corpus contained 50 000 tokens and was used to compare the African languages data to the Princeton core concepts.

From the multilingual corpus, we extracted a frequency list for Tshivenda and Setswana. The next step was to compare the 5000 most frequent Tshivenda and Setswana words in the multilingual African wordlist to the list of (English) base and core concepts mentioned above. Table 4 below shows some of the concepts unique to the African language list. The frequency of the word is given in brackets.

Noun	Frequency
benefit	2042
basket	71
conflict	419
lodge	177
malaria	355
mandate	838
mine	321
money	1592
soil	104
water	2964

Table 4. Frequent nouns from a large multilingual African language corpus.

It is clear that our frequency list includes concepts that reflect unique African language usage. The Princeton and EuroNets lists both include concepts that might not be completely unknown in an African context, but that certainly are less commonly used.

The new approach proposed in this third phase of the African Wordnets project entails extracting a subset of concepts that were present in this list. We now have a list of concepts that are both internationally regarded and frequent in African corpora. This new list of roughly 1000 concepts was shared with the linguists as a starting point for Tshivenda. For the other four languages, we extracted the list of concepts that were not added in the previous projects to use as a starting point for new development in this phase.

5 Conclusion and future work

5.1 Comparing development in the 4 languages

Figures 2 and 3 represent the total number of synsets and definitions for each language combination. This comparative review gives a clear indication of fast-tracking possibilities for each language by using the synsets/definitions of its closely related counterpart language. For example, synsets or definitions developed for isiZulu and not for isiXhosa, can be fast-tracked for the latter since both languages belong to the Nguni language group, and vice versa. On the other hand, synsets or definitions developed for Setswana and not for Sesotho sa Leboa (Sepedi), can be fast-tracked for the latter since both languages belong to the Sotho language group, and vice versa.

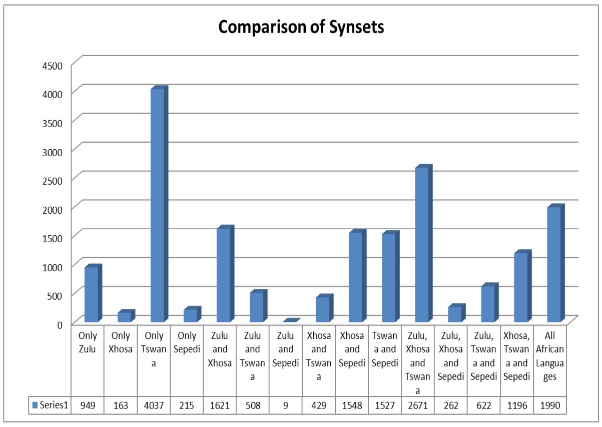

Figure 2. Comparison of synsets completed for each language.

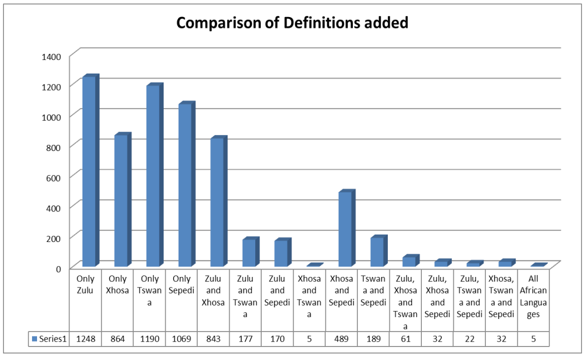

Figure 3. Comparison of definitions completed for each language.

5.2 Dissemination of the information

Since the resource that will be further developed in this project is vital to so many linguistic and language technology endeavours, it is essential that it be accessible to all researchers in the field. After quality assurance (see section 4.2) the wordnets will be included in the repository of the RMA, who will advertise and make available the wordnets for others to use. The appropriate licensing options and usage rights (most probably under one of the Creative Commons licenses[8]), will also be determined in conjunction with the RMA.

5.3 Conclusion

The African Wordnet project is unique in its approach to create wordnets for several languages in parallel, resulting in a very important language resource. This approach allows team members to share experiences during the process and thus build the lexicon more effectively. It also allows for a multilingual resource that can be applied in various other technologies, such as for machine translation, extracting content for learner's dictionaries and other teaching material, but also as a reference for linguists. There is still much work to be done, but by learning from previous projects and keeping the ultimate goal of a rich linguistc resource in mind, we trust that this work will fill many gaps in NLP in South Africa and Africa as a whole.

Acknowledgements

The authors would like to thank the UNISA Women in Research Fund for financial support in this third project, as well as Christiane Fellbaum, Karel Pala, Piek Vossen and Adam Rambousek for their invaluable advice. Without the vast linguistic knowledge and dedication of the linguists working on the development of the African Wordnet, this project would also not be as successful. Comments of two anonymous reviewers are also appreciated.

References

Sonja Bosch, Laurette Pretorius and Axel Fleisch. 2008. Experimental Bootstrapping of Morphological Analysers for Nguni Languages. *Nordic Journal of African Studies,* 17(2):66-88. http://www.njas.helsinki.fi/

Clement Doke and Benedict Vilakazi. 1964. *Zulu–English Dictionary*. Witwatersrand University Press, Johannesburg.

Global Wordnet Association. 2013. GWA Base Concepts.

ISO 639-3.http://www.sil.org/iso639-3

Cvetana Krstev, Bojana Đorđević, Sanja Antonić, Nevena Ivković-Berček, Zorica Zorica, Vesna Crnogorac and Ljiljana Macura. 2008. Cooperative work in further development of Serbian Wordnet. *INFOTHECA – Journal of Informatics and Librarianship.* 1(2):59-78. http://infoteka.bg.ac.rs/PDF/Eng/2008/INFOTHECA_IX_1-2_May2008_59a-78a.pdf

Gideon Kotzé. 2008. Ontwikkeling van 'n Afrikaanse woordnet : metodologie en integrasie. *Literator : Journal of Literary Criticism, Comparative Linguistics and Literary Studies : Human language technology for South African languages,* 29(1):168 – 184.

Jurie le Roux, Koliswa Moropa, Sonja Bosch & Christiane Fellbaum. 2007. Introducing the African Languages Wordnet, in Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, Piek Vossen (eds.) *Proceedings of The Fourth Global WordNet Conference*, Szeged, Hungary, January 22-25, 2008, pp 269-280. Szeged: University of Szeged, Department of Informatics (ISBN 978-963-482-854-9).

Mampaka Mojapelo. 2013. Morphological considerations for encoding the qualificative in African Wordnet with reference to Northern Sotho. *To be presented* at the Global Wordnet Conference 2014. Tartu, Estonia. 25 – 29 January 2014.

Princeton Wordnet. 2013. http://wordnetweb.princeton.edu

Danie Prinsloo & Gilles-Maurice de Schryver. 2005. Managing eleven parallel corpora and the extraction of data in all official South African languages. In W. Daelemans, T. du Plessis, C. Snyman & L. Teck (eds.) *Multilingualism and Electronic Language Management. Proceedings of the 4th International MIDP Colloquium*, Bloemfontein, South Africa, 22-23 September 2003. pp 100–122. Pretoria: Van Schaik Publishers.

Resource Management Agency (RMA). 2013. http://rma.nwu.ac.za/

Pavel Smrz. 2008. Quality Control and Checking for Wordnet Development: A Case Study of BalkaNet. *Romanian Journal of Information Science and Technology*, 7(1):173-181.

Wortschatz Universität Leipzig. 2013. http://corpora.informatik.uni-leipzig.de/

[8] See http://creativecommons.org/licenses/