Building resources for African languages

Karel Pala[1], Sonja Bosch[2], Christiane Fellbaum[3]

[1]Faculty of Informatics, Masaryk University, Botanicka 68a, 602 00 Brno, Czech Republic
pala@fi.muni.cz
[2]Department of African Languages, University of South Africa, PO Box 392, 0003 Pretoria, South Africa
boschse@unisa.ac.za
[3]Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08544, USA
fellbaum@princeton.edu

Abstract

We report on work towards the creation of African Languages WordNet, comprised of interlinked semantic networks in several African languages that are known to have limited language resources. Adding these languages to the WordNet family will enable NLP applications for each of the languages in isolation. Moreover, linking the African Wordnets to one another and to the many global WordNets will make crosslinguistic information retrieval and question answering possible, and significantly aid machine translation. In this paper it is demonstrated how collaborative work between people, using existing tools, can contribute to the building of large text corpora and subsequently address the challenge of limited availability of language resources. The long term aim is the development of aligned WordNets for Bantu languages spoken in South Africa as multilingual knowledge resources which could be extended to include a wide variety of related languages from other parts of Africa.

1. Introduction

Many Natural Language Processing (NLP) applications that require word sense disambiguation rely on WordNet as an essential lexical resource. WordNets have been created for dozens of languages, primarily those spoken by large populations and in technologically advanced countries where funding for resource development is relatively easily available (Miller, 1995; Fellbaum, 1998; Vossen, 1998).

We report on work towards the creation of the African Languages WordNet, comprised of interlinked semantic networks in several African languages. Adding these languages to the WordNet family will enable NLP applications for each of the languages in isolation. Moreover, linking the African Languages Wordnets to one another and to the many global WordNets will make crosslinguistic information retrieval and question answering possible, and significantly aid machine translation. WordNets have also been shown to be very useful for language learning.

Besides these practical considerations, there are many purely linguistic motivations for building African languages WordNets. WordNets currently exist in some 50 languages, many of them typologically and historically unrelated. But no African language, and no language with the particular linguistic features of the languages of South Africa, has developed a WordNet. Doing so will force a new and broader perspective of the lexicon and will enrich our understanding of this component of human language.

1.1 WordNet

All present WordNets are modelled on the Princeton WordNet developed in the mid-1980s (Miller, 1995; Fellbaum, 1998). A WordNet is a large semantic network where words and groups of words are interlinked by means of lexical and conceptual relations represented by labelled arcs. Like a dictionary, WordNet's units are words, and its aim is to provide semantic information about words. This information is given in a form resembling a thesaurus, though the network of words is more rigorously structured than in a thesaurus.

WordNet's building blocks are unordered sets of synonymous words and phrases, dubbed "synsets". Synset members are denotationally equivalent and substitution of a synset member by another does not change the truth value of the context, though stylistic infelicity may result from such substitution. WordNet provides some information on how synset members are used; register tags are given ("colloquial," "slang" etc.), and example sentences accompany most synsets illustrating the synonyms' usage.

A synset is said to lexically express a concept. Examples of synsets are {mail, post}, {hit, strike} and {small, little}. All synsets further contain a brief definition. A domain label (sports, medicine, biology) marks many synsets.

Concepts expressed by nouns are densely interconnected by the hyponymy relation (or hyperononymy, or subsumption, or the ISA relation), which links specific concepts to more general ones. For example, the synset {gym shoe, sneaker, tennis shoe} is a hyponym, or subordinate of {shoe}, which in turn is a hyponym of {footwear, footgear}, etc. Hyponymy builds hierarchical "trees" up to fifteeen layers deep with increasingly specific "leaf" concepts growing from an abstract "root".

Crosslinguistic WordNets share the same structure and can be interlinked, allowing for the identification of

equivalent words and synsets and enabling translation. The technical instrument for interlinking and thus capturing multilinguality of WordNets is Interligual Index (ILI) developed in the EuroWordNet project (Vossen, 1998). Languages differ in their lexical make-up, and words (or entire areas of the lexicon) that are expressed in one language may be „missing“ in another. WordNet's systematic structure identifies both crosslinguistic matches as well as mismatches. WordNet construction is therefore a way to compare not only the lexicons of African languages with one another but also with those of dozens of other languages.

2. WordNet training workshop

In the light of the crucial contribution of global Wordnets to NLP, an infrastructure for WordNet development for African languages was created by means of a week long training workshop. The aim of the workshop was to develop a platform for WordNet development for African languages.

Seed research funding for the project was obtained from the Meraka Institute (2007) to enable facilitation by international experts namely Christiane Fellbaum (Wordnet, 2006) as one of the pioneers of WordNets, Piek Vossen (1998) as project coordinator of the EuroWordNet project, and Karel Pala as participant in the Czech WordNet (cf. Pala & Smrž, 2004) and developer of the lexicographer's editing tools DEBVisDic in particular (cf. DEBVisDic Manual, 2008). The project afforded linguists, translators and lexicographers representing the 9 official African languages in South Africa, as well as computer scientists, the opportunity of high level multi-disciplinary training.

The nine official African languages of South Africa are Zulu (isiZulu), Xhosa (isiXhosa), Swati (siSwati), Ndebelele (isiNdebele), Venda (Tšhivenda), Tsonga (Xitsonga), Southern Sotho (Sesotho), Northern Sotho (Sesotho sa Leboa) and Tswana (Setswana). These languages all belong to the Bantu language family and are grammatically closely related. The Nguni languages, i.e. Zulu, Xhosa, Swati and Ndebele form one group. The Sotho languages, viz. Southern Sotho, Northern Sotho and Tswana form another group with Venda and Tsonga being more or less on their own

2.1 Accomplishments

The facilitators each gave lectures on the area of their speciality that related to the methods, theory, and practical steps for WordNet construction. Christiane Fellbaum (Princeton) lectured on the design of WordNet and invited the participants to reflect on specific questions from the viewpoint of their native languages. A number of hands-on exercises were carried out, where the participants built "toy" WordNets for their languages. Piek Vossen (Amsterdam) lectured on his experiences with EuroWordNet, where he introduced some fundamental changes to the original Princeton WordNet.

Karel Pala (Brno) introduced his editing tool and the participants trained on it under his guidance.

The user manual for the editing tool DEBVisDic was updated after contributions made by workshop participants regarding the user friendliness of the software tool (cf. DEBVisDic Manual, 2008). Extensive reading matter was distributed among participants before, during and after the workshop. A CD ROM containing the reading matter as well the presentations of the facilitators was handed to participants after the workshop.

The African languages WordNets are still in a conceptualisation phase although experimental work on noun and verb synsets has begun (cf. le Roux et al., 2008).

2.2 Challenges specific to African languages

During the workshop various challenges for WordNets specific to the African languages were identified, the first and foremost being the morphological complexities of agglutinative languages centred around a noun class system and roots. WordNets for such languages pose novel challenges, especially with respect to the concept of "word," which must be defined to determine synset membership. The conjunctive and disjunctive orthographies of the various language groups contribute to this challenge. For example, the orthographic word *ngiyabathanda* (“I like them”) in Zulu corresponds to four orthographic words or separate orthographic entities in Northern Sotho, viz. *ke a ba rata* (“I like them”).

A feature particular to the Bantu languages is the POS known as “ideophone”, a term proposed by Doke (1935:118) for a word category which describes a predicate, qualificative or adverb in respect to manner, colour, sound, smell, action, state or intensity. In contrast to the linguistic word in the Bantu languages, which is characterised by a number of morphemes such as prefixes and suffixes, as well as a root or stem, the ideophone consists only of a root which simultaneously functions as a stem and a fully-fledged word. The following are some Zulu examples:

> *Bathula bathi **du*** (They kept **completely quiet**)
> *Ingilazi iwe yathi **phahla** phansi* (The glass fell **smashing** on the floor)
> ***Kubomvu klubhu!*** **(It is blood red)**

These can be accommodated in the WordNets in the following way. Often, they can map to the "canonical" parts of speech (nouns, verbs, adjectives, adverbs) in the existing WordNet. For example, the workshop participants cited over 200 verbs denoting manners of motion, many encoding ideophones. These can be entered as manner-specific subordinates ("troponyms") in the WordNets, and, wherever possible, mapped to the corresponding manner-of-walking verbs in other languages. Similarly, colour

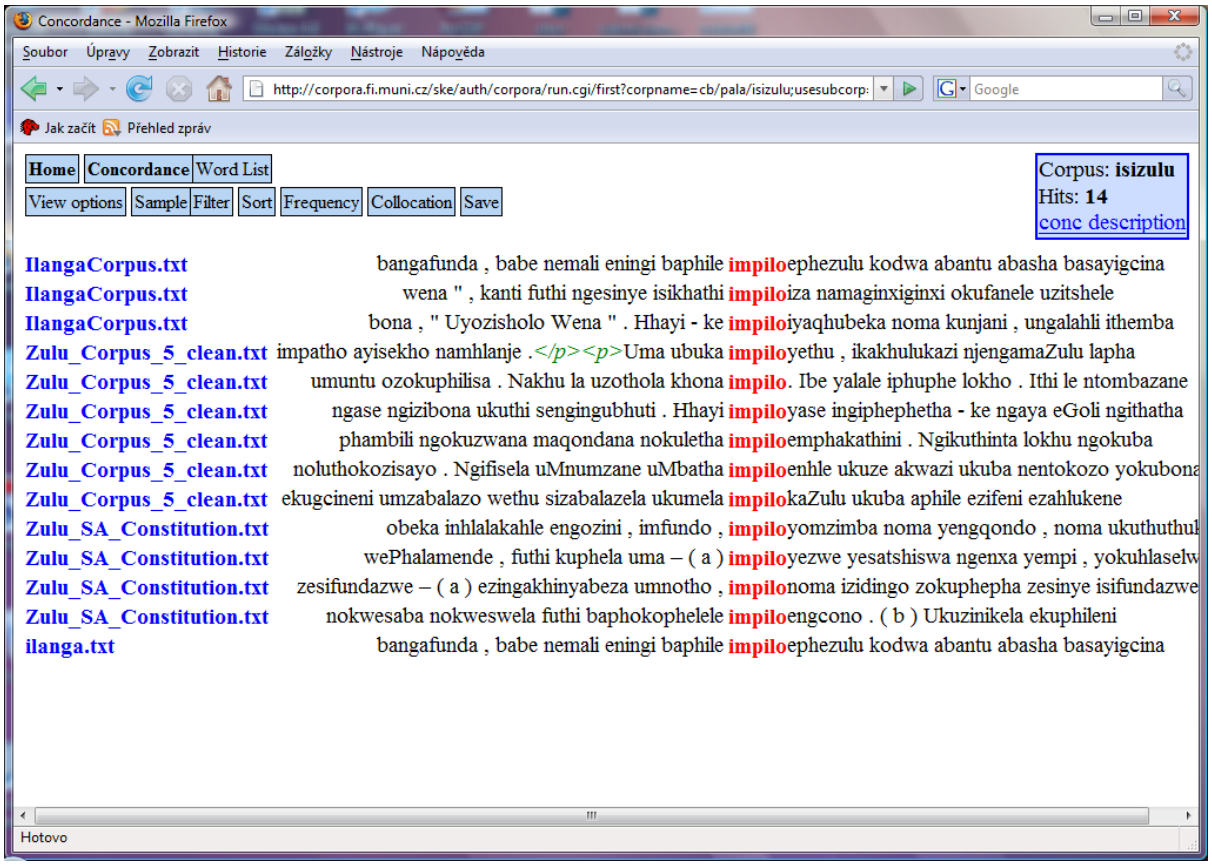

Figure 1: A concordance list from an experimental Zulu corpus

words involving ideophones can be linked to colour words (adjectives and nouns) in existing the WordNet. Wherever necessary, basic ideophones or words including an ideophone will be accommodated as a new lexical category.

In the course of the workshop, some noun and verb synsets were created in the various African languages. Secondly, limited availability of electronic language resources, such as large corpora, parallel corpora, electronic dictionaries and machine-readable lexicons was identified as a stumbling block, particularly in comparison to the generous availability of language recources for other WordNets in the world. Workshop participants relied on both monolingual and bilingual dictionaries available in their particular languages for semantic information. A need for corpus compilation was expressed, as corpora enable researchers to find and examine a particular word in context. Corpus data captures language use by many people in different contexts over time, and thus corpus data is more reliable than introspection by a few linguists or lexicographers. Corpus data is vital for determining the sense inventory of a language. General corpora are available for all nine mentioned African languages at the University of Pretoria (University of Pretoria, 2003), but with access restrictions which involve on site computer processing of the corpus and down-loading only the results of the analyses. The sizes of the various corpora currently range from 1 million tokens for Ndebele to 5.8 million tokens for Northern Sotho.

3. Working collaboratively to build text corpora with few existing language resources

In order to address the challenge of limited availability of electronic language resources, this section demonstrates how collaborative work between people, using existing tools, can contribute to the building of large text corpora.

3.1 Building corpora

Tools exist that allow us to almost automatically build text corpora for any language, and for African languages in particular. The only condition is the availability of a collection of texts in plain format. To demonstrate this we used the Corpus Builder tool developed in the NLP Centre at the Faculty of Informatics Masaryk University (Baroni et al., 2006) and created a small Zulu text corpus containing approximately 80 000 tokens in a very short time (approx. 30 minutes).

To visualise concordance lists we used a corpus manager tool, Manatee/Bonito2 (Rychlý, 2000). This tool is also integrated with the Corpus Builder, thus the newly built corpus can be immediately inspected.

If appropriate collections of texts are available, for instance from Web pages that are freely accessible the corpus can be enlarged in next to no time.

When larger plain text corpora are built, the need arises to tag them. Thus, the next step is to build taggers and tagsets for African languages. Work on taggers for Northern Sotho is reported on in Prinsloo and Heid (2006) and de Schryver and de Pauw (2007). This is a relatively independent enterprise but it will contribute to enriching electronic African language resources considerably. For this purpose automatic morphological analysis and analysers are being developed (cf. Bosch et al., 2006). Without morphological analysers, building high quality mono- and multilingual lexical resources including WordNets and other lexical databases will not be possible. This applies especially to the Bantu languages, the rich morphology of which calls for these tools to be developed as soon as possible.

One further tool that needs to be mentioned is the BootCat (Baroni et al., 2006) which allows one to build rather small domain corpora directly from Web pages, if they are at one's disposal.

Finally, it should be remarked that the described way of building corpora can be applied to all African languages mentioned above since the techniques are language independent.

3.1.1. Corpora and WordNets

Experience with building WordNets in the Balkanet project (Pala & Smrž, 2004) has shown that the evidence obtained from corpora can be profitably exploited for making them empirically more reliable and descriptively adequate. This means that corpora are very helpful for compiling the representative list of synsets on the ground of frequency considerations obtained from corpora. It also means that the evidence obtained from corpora is useful for making decisions about the senses that have to be associated with the respective synsets. This applies fully to all the considered African languages – we have shown that corpora for these languages can be built cost effectively by using the tools that are easily accessible.

It should be noted that corpora exploited for developing WordNets have to be of a general nature. In other words, texts from which such corpora are created should come either from newspaper resources or they can be appropriately selected novel texts (in the Balkanet project it was the novel 1984 by G. Orwell which existed as a parallel corpus for all Balkanet languages). Specialized corpora containing specialized technical or terminologically oriented texts are not appropriate.

Obviously, the next step would be to try to build and then exploit parallel corpora for Bantu languages. This will be useful not only for developing African WordNets as such but also for promoting the cultural relations between them. More information about the mentioned corpus tools can be obtained from: http://www.textforge.cz/products, or http://www.fi.muni.cz/~thomas/corpora/cb_text_upload.htm

3.2 A tool for building WordNets – DEBVisDic

For editing and browsing WordNets one needs a tool that can serve this purpose. DEBVisDic (Horák et al, 2006) is a tool for building WordNets and it works with lexical data in XML format. It is a browser and editor exploiting client/server architecture so that more developers and/or lexicographers can work on their WordNets simultaneously. This is an important requirement for the people working on African languages WordNets, since the lack of resources calls for constant sharing of information. The tool is built on the DEB (Dictionary Editing and Browsing) platform and is equipped with all features for supporting the linguistic work on WordNets.

DEBVisDic uses a versatile interface that allows the user to arrange the work without any limitations. It displays the following main functions:

- multiple views of multiple WordNets (multilingual view)
- freely defined text views
- synset editing and introducing semantic relations between them
- building and visualising hypero-hyponymic trees
- query result lists
- plain XML view of a synset
- synchronization of the synsets
- inter-dictionary linking
- consistency checks and journaling
- user configuration ensuring exchange of data
- entry locking for concurrent editing
- links display preview caching (speeds up the processing)

Presently, DEBVisDic is being supplemented with several other features that are currently accessible only as separate tools or resources. This functionality includes:

- link to a morphological analyser (for languages, where it is available)
- connection to language corpora, including Word Sketches statistics (for languages with accessible Word Sketch Engine, (cf. Kilgarriff et al., 2004).
- access to any electronic dictionaries stored in XML format within the DEB server
- searching for literals within encyclopaedic web sites.

Existing WordNets as well as those under development are stored in the DB XML database which is well suited for processing complicated XML structures. Simple processing of the data (like export or import of the whole dictionary) is not a problem as the whole English WordNet export (over 100.000 entries) takes less than 1 minute. However, searching for values of specific subtags can take several seconds in such a large dictionary even when indexes are used. We are currently working on several solutions for this,

which include link caching, specific DB XML indexing and also experimenting with a completely different database backend.

The DEBVisDic client is continuously being developed in the NLP Centre, Faculty of Informatics Masaryk University according to specific needs of running projects. These needs are determined to a great extent by the people doing the lexicographic work in the projects. We can mention the work on the PolishWordNet (Pala, Vetulani et al., 2007) and on the Dutch Cornetto project (Vossen, 2007). The African languages are accommodated here as well, as mentioned previously. The DEBVisDic tool will be made available freely and is to be installed at the University of South Africa for the use of the WordNet builders in a next stage of the project.

More information about the DEBVisDic tools can be found in DEBVisDic Manual (2008) (see the References).

3. Future work

The long term aim of this project is the development of aligned WordNets for African languages spoken in South Africa (i.e. languages belonging to the Bantu language family) as multilingual knowledge resources which could be extended to include a wide variety of related languages from other parts of Africa.

The construction of WordNets in a number of African languages presents exciting prospects since these languages unfortunately have not yet been as widely studied as European and Asian languages. At the same time, their sophisticated properties are of great linguistic interest. Building WordNets will necessitate the careful investigation of linguistic phenomena like classifiers that have not yet been explored in the context of non-Asian languages and may force a re-examination and broadening of the WordNet structure. Undoubtedly, the crosslinguistic study of lexicalization patterns, an interest of the GWA (Global WordNet Association), will benefit greatly from the addition of the new perspectives afforded by African languages. Like the original Princeton WordNet as well as WordNets in many other languages, African WordNet will be made freely and publicly available. Information is available on the Global WordNet website (http://www.globalwordnet.org).

Needs that have been identified for the successful continuation of the project, are the following:

Language resources: corpora and (electronic) dictionaries

Hardware: laptops for lexicographers and coders

Human resources: database manager with appropriate technical training

Modules for future integration into comprehensive NLP systems: POS-taggers, morphological analysers and syntactic parsers.

Finally, such research and development would depend on the commitment of linguists, translators and lexicographers to continue the work begun with great enthusiasm, the co-operation of numerous language institutions, the availability of a variety of language resources as well as further financial support following the seed research funding.

4. Conclusion

Follow-up research funding by the National Human Language Technology Network of the Meraka Institute for the development of African languages WordNets has just been announced. WordNets consisting of either 10 000 synsets each for four of the above mentioned languages (two Nguni and two Sotho languages), or two Wordnets (for one Nguni and one Sotho language), containing 20 000 synsets each, will be developed. These decisions will be taken during the planning phase of the project and will depend on the availability of resources for the involved languages. Where deemed necessary and possible, international collaborators will also be involved. This will extend current networks of collaboration, and will also extend the knowledge-base of HLT in South Africa.

5. Acknowledgement

The African Languages Wordnet Project (2006-2007) was funded by the National Human Language Technology Network of the Meraka Institute (South Africa).

This work has also been partly supported by the Academy of Sciences of Czech Republic under the project 1ET200610406 and by the Ministry of Education of Czech Republic within the Center of Computational Linguistics Project LC536.

6. References

Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P. (2006). WebBootCat: a Web Tool for Instant Corpora. In Proceedings of the Twelfth EURALEX International Congress, Turin, 2006, pp. 123--132.

Bosch, S., Jones, J., Pretorius, L., Anderson, W. (2006). Resource Development for South African Bantu Languages: Computational Morphological Analysers and Machine-Readable Lexicons. In Proceedings on the Workshop on Networking the Development of Language Resources for African Languages 5th International Conference on Language Resources and Evaluation, 22 May 2006, Genoa, Italy, pp. 38--43.

de Schryver, G-M., De Pauw, G.. (2007). Dictionary Writing System (DWS) + Corpus Query Package (CQP): The case of TshwaneLex. *Lexikos* 17 (AFRILEX-reeks/series 17: 2007), pp.226--246.

DEBVisDic Manual. (2008). Available at: http://nlp.fi.muni.cz/trac/deb2/wiki/DebVisDicManual. Accessed on: 22 February 2008.

Doke, C.M. (1935). Textbook of Zulu Grammar. Johannesburg: University of the Witwatersrand Press.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Horák A., Pala, K., Rambousek, A., Povolný, M. 2006. First version of new client-server wordnet browsing and editing tool. In Proceedings of the Third International WordNet Conference – GWC 2006, pp. 325-328, Jeju, South Korea, Masaryk University, Brno.

Kilgarriff A., Rychlý P., Smrž P., Tugwell D. (2004). The Sketch Engine. In Proceedings of the Eleventh EURALEX International Congress, Lorient, 2004, pp. 105--116

le Roux, J., Moropa, K., Bosch, S., Fellbaum, C. (2008). Introducing the African Languages Wordnet. Proceedings for the Fourth Global WordNet Conference, Szeged, Hungary, January, 2008.

Meraka Institute. (2007). African Advanced Institute for Information and Communication. Available at: http://www.meraka.org.za/index.htm
Accessed on: 28 February 2008.

Miller, G..A. (1995). WordNet: a lexical database for English. In Communications of the ACM. 38(11), pp. 39--41.

Pala, K., Smrž, P. (2004). Building Czech WordNet. *Romanian Journal of Information Science and Technology*, Romanian Academy Bucharest, 7, 1-2, pp. 79--88.

Pala, K., Vetulani, Z., Horák, A., Rambousek, A., Konieczka, P., Marciniak, J., Obrębski, T., Rzepecki, P., Walkowska, J. (2007). DEB Platform tools for effective development of WordNets in application to PolNet. In Vetulani, Zygmunt (ed.) Human Language Technologies as a Challenge for Computer Science and Linguistics, Proceedings of 3rd Language and Technology Conference, Poznan, 2007, pp. 514--518.

Prinsloo, D.J., Heid, U. (2006). Creating Word Class tagged Corpora for Northern Sotho by Linguistically Informed Bootstrapping. In: Isabella Ties (Ed.)/ Proceedings of the Conference for Lesser Used Languages and Computer Linguistics, EURAC research, European Academy. Bolzano, Italy. 27th October - 28th October 2005, (Bolzano: EURAC), 2006, pp. 97-- 115.

Rychlý, P. (2000). Corpus managers and their effective implementation, Ph. D. Thesis, Masaryk University, Brno, Czech Republic.

University of Pretoria Department of African Languages. (2003). Available: http://www.up.ac.za/academic/humanities/eng/eng/afrlan/eng/initiative.htm
Accessed on 13 April 2006.

Vossen, P. (1998). (Ed.). EuroWordNet. Dordrecht: Holland: Kluwer.

Vossen, P. (2007). The Cornetto project. Available: http://www.let.vu.nl/onderzoek/projectsites/cornetto/start.htm
Accessed on 28 February 2008.

WordNet a lexical database for the English language. (2006). Available: http://wordnet.princeton.edu/
Accessed on 28 February 2008.