Literator - Journal of Literary Criticism, Comparative Linguistics and Literary Studies
ISSN: (Online) 2219-8237, (Print) 0258-2279

Strategies for building wordnets for under-resourced languages: The case of African languages

Authors:
Sonja E. Bosch[1]
Marissa Griesel[1]

Affiliations:
[1]Department of African Languages, University of South Africa; African Wordnet Project, South Africa

Corresponding author:
Sonja Bosch,
boschse@unisa.ac.za

Dates:
Received: 26 Sept. 2016
Accepted: 09 Jan. 2017
Published: 31 Mar. 2017

How to cite this article:
Bosch, S.E. & Griesel, M., 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

The African Wordnet Project (AWN) aims at building wordnets for five African languages: Setswana, isiXhosa, isiZulu, Sesotho sa Leboa (also referred to as Sepedi or Northern Sotho) and Tshivenda. Currently, the so-called expand model, based on the structure of the English Princeton WordNet (PWN), is used to continually develop the African Wordnets manually. This is a labour-intensive work that needs to be performed by linguistic experts, guided by several considerations such as the level of lexicalisation of a term in the African language. Up to now, linguists were responsible for identifying and translating appropriate synsets without much help from electronic resources because in the case of African languages even basic resources such as computer readable and electronic bilingual wordlists are usually not freely available. Methods to speed up the manual development of synsets and ease the workload of the human language experts were recently investigated. These centred around utilising the minimal amount of information available in bilingual dictionaries to identify synsets in the PWN that should be included in the AWN, transferring information from dictionaries to the wordnet and presenting the potential synsets to linguists for final approval and inclusion in the wordnets. In this article, we describe the methodology developed for building the African Wordnets, a potentially significant resource for natural language processing applications. Available resources that could be taken advantage of and resources that had to be developed are investigated, and initial results and future plans are explained.

Strategieë om woordnette vir hulpbronskaars tale te ontwikkel: 'n gevallestudie vir Afrikatale. Die African Wordnet Projek (AWN) het ten doel om woordnette vir vyf Afrikatale te ontwikkel. Die tale sluit Setswana, isiXhosa, isiZulu, Sesotho sa Leboa (ook Sepedi of Noord-Sotho genoem) en Tshivenda in. Die sogenaamde uitbreidingsmodel, wat op die struktuur van die Engelse Princeton WordNet (PWN) gebaseer is, word tans gebruik om die AWN deurlopend handmatig uit te brei. Hierdie metode is baie arbeidsintensief en moet deur linguiste uitgevoer word. Die linguiste word deur verskeie kriteria, soos die vlak van leksikalisering van 'n woord en die geskiktheid van die sinstel vir die taal, gelei. Linguiste moes tot nou toe hierdie besluite sonder veel ondersteuning in die vorm van elektroniese hulpmiddels maak, aangesien daar vir baie Afrikatale nog nie eers basiese hulpbronne soos vrylik beskikbare, rekenaarleesbare en elektroniese tweetalige woordelyste bestaan nie. Metodes om die handmatige ontwikkeling van sinstelle te bespoedig en die werkslading op die taalspesialiste te verlig, het onlangs baie aandag geniet. Die eksperimente het daaroor gegaan dat die minimale hoeveelheid bronne wat wel beskikbaar is, ingespan word om sinstelle in die PWN te identifiseer wat na die AWN oorgedra behoort te word. Inligting uit die tweetalige woordelyste word op sinvolle wyse onttrek en aan die linguiste voorgehou om die finale seleksie te maak. In hierdie artikel word die metodologie wat gebruik is om die AWN te ontwikkel, voorgelê. Beskikbare hulpbronne wat in die verskillende eksperimente gebruik of ontwikkel is, word beskryf, voorlopige resultate word gegee en toekomstige planne word beskryf.

Introduction and aims

A wordnet is an electronic lexical database consisting of words that are grouped into sets of synonyms called synsets and linked by conceptual-semantic and lexical relations (Miller 1995). Examples of synsets are {car, automobile} and {shut, close}. Synsets are interrelated by means of semantic relations, such as the superordinate versus subordinate or hyperonymy versus hyponymy relation (car-convertible), the part-whole or meronymy relation (tyre-car), and antonymy (open-close). The interlinked synsets form an extensive semantic network, the digital format of which allows both manual and automatic searches for words that are meaningfully related to one another. In this regard, Fellbaum (1998:7) explains:

> WordNet is a semantic dictionary that was designed as a network, partly because representing words and concepts as an interrelated system seems to be consistent with evidence for the way speakers organise their mental lexicons.

An example of a very comprehensive synset from the Princeton WordNet (PWN) (Princeton University, 2016) for English is shown in Figure 1. This example demonstrates the semantic nature of a wordnet and the hierarchical relations captured therein (more detail on this is provided in the Semantic Domains section). The PWN contains extensive synsets such as the one shown in Figure 1 not only for nouns but also for verbs, adjectives and adverbs, each articulating a distinct concept.

Since the 1990s, wordnets have been built for more than 150 languages worldwide, including many that are genetically and typologically unrelated to the original English wordnet. The first step towards creating cross-lingual wordnets was EuroWordNet (Vossen 1998), which encompasses eight languages, followed by the Multilingual Open Wordnet Project (Bond & Paik 2012) where 34 open wordnets have been merged and are housed in a central repository for easy use. The latest venture is the Global WordNet Grid (Vossen, Bond & McCrae 2016) that aims at providing a platform for centralising all existing wordnets.

Development of a wordnet typically follows one of two distinct methods, as discussed by Ordan and Wintner (2007) and Vossen (1998). New wordnets are usually constructed from the ground up as a stand-alone resource and subsequently aligned with the PWN (Fellbaum 1998) in the so-called *merge* approach as in the case of PolNet, a Polish wordnet (Vetulani, Kubis & Obrębski 2010) that is based on a high-quality monolingual Polish lexicon. Alternatively, the

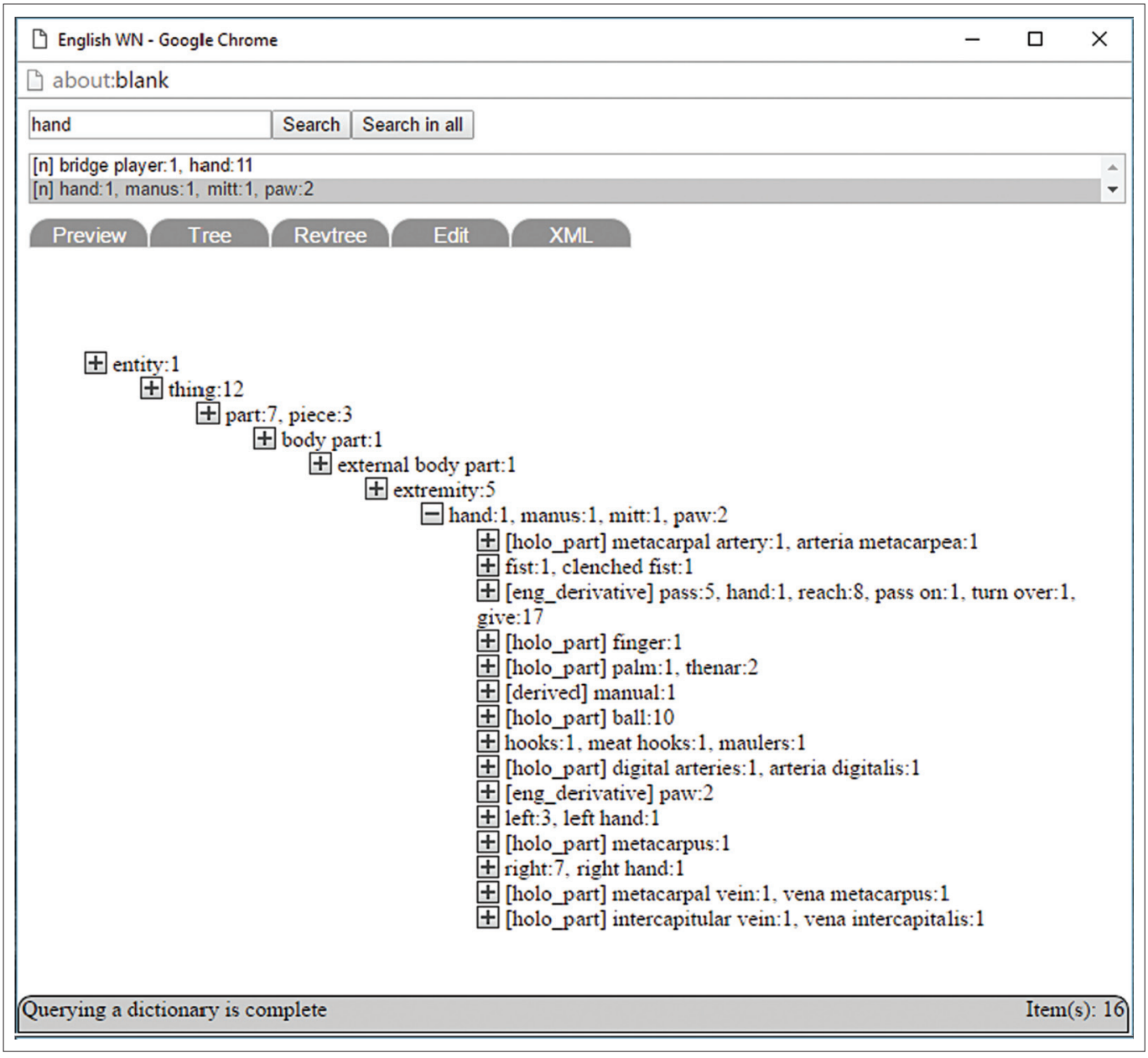

FIGURE 1: The first sense for the noun synset 'hand' in PWN.

alignment to PWN can be used as the basic structure on which to build a new wordnet from the onset. This latter method is referred to as the expand model which translates the English wordnet into the target language, and assumes that the new language shares an underlying structure with PWN. An example is the Hungarian wordnet (HuWN) (Vincze & Almási 2014) for which PWN 2.0 served as the basis.

Although a wordnet is accessible to human users via a web browser for the study of lexical structure and lexicalisation patterns, wordnets are also an essential resource for natural language processing applications that, for instance, require lexical disambiguation. Semantic relations in a wordnet can be exploited for word sense discrimination (cf. the 2013 shared task for SemEval, reported on by Navigli, Jurgens & Vannella 2013), which is one of the core technologies for many other natural language processing applications. The usefulness of wordnets is further described by Abdullah and Ibrahim (2015), who applied wordnets to improve the accuracy of information retrieval in a semantically driven search engine. Regarding language learning applications, Susanti, Iida and Tokunaga (2015) reported on the use of wordnets to automatically generate vocabulary tests for second language acquisition.

In light of the foregoing description of wordnets as significant resources for natural language processing applications, the aim of this article is to present various strategies used for building the African Wordnets, as well as explaining the available resources that were taken advantage of and the resources that had to be developed in the case of these under-resourced languages. In the next section, we briefly present the African Wordnets and discuss their status quo. Then, various development strategies are discussed critically and illustrated with concrete examples. We conclude and point to future work in the final section.

Background and status quo of the African Wordnet Project

The languages in this project are considered resource scarce compared to most other languages listed by the Global WordNet Association (2016), in the sense that lexical resources are very limited and that there are no machine-readable lexicons freely available. The monolingual wordlists without lexical or grammatical information and relatively small, domain-specific corpora available at, for example, the Language Resource Management Agency (RMA) (2013) are insufficient for semi-automatic wordnet construction in the African languages. The agglutinating nature of the African languages belonging to the Bantu language family, particularly for those with a conjunctive orthography, such as isiZulu and isiXhosa, calls for morphological annotation for accurate corpus searches. Spiegler, van der Spuy and Flach (2010:1022) point out that the complex morphology of isiZulu is a challenge in particular for computational analysis, because:

> Words usually incorporate both prefixes and suffixes, and there can be several of each. This makes it hard to identify the root by mechanical means, as the root could be the first, second, third, or even a later morpheme in a word. The complexities involved are exacerbated by the fact that a considerable number of affixes, especially prefixes, have allomorphic forms.

Although prototypes of rule-based morphological analysers have been developed for the mentioned two languages, these are not freely available yet (cf. Bosch & Pretorius 2011).

The purpose of the African Wordnet Project (AWN) 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 also from other parts of Africa. Linking such wordnets to one another and to the many global wordnets makes cross-linguistic research and development possible. The first step towards developing such a rich resource for African languages was a training workshop for linguists, lexicographers and computer scientists which took place in 2007. As a direct result, development of wordnet prototypes for five official South African languages commenced as the AWN. Currently, the project includes isiXhosa, isiZulu, Setswana, Sesotho sa Leboa and Tshivenda[1] and has consisted of phases as described in Griesel and Bosch (2014). Throughout the development, the AWN used the DEBVisDic editor tools (DEBVisDic: WordNet editor and browser n.d.) which are distributed as freeware and aim to be user-friendly and intuitive for linguists building semantic networks. DEBVisDic has been used in more than 20 projects and was recently re-launched as a web application (Rambousek & Horak 2016). An example of the DEBVisDic interface for the Setswana wordnet is given in Figure 2. The definition, usage example, domain and other linguistic data fields for *seatla* (hand) can all be seen in one view.

Because of the resource scarceness of African languages, it was decided to follow the expand model for the development of the African Wordnet. As indicated by Ordan and Wintner (2007), the expand model provides a tested structure on which to build a new resource and is therefore typically the choice for less resourced languages. Furthermore, the focus would be on the noun part included in PWN. The focus is mainly on the nouns because a starting point was motivated by the assumption that this would be a gentle introduction to wordnet development for our relatively inexperienced team of linguists. Also, nouns make up the bulk of the lexicon and would therefore, see the wordnets grow at a steady pace. The deliverables were divided into three categories: a basic synset, a definition and a usage example. The team started out by creating a basic synset in a first pass, with definitions and usage examples being added in subsequent iterations. It soon became clear after discussions with experienced wordnet developers at the Global Wordnet Conference in 2014 that the strength of a wordnet for further use lies more in the usage of examples than in definitions, and the focus

1.The ISO 639-2 codes as found on http://www.loc.gov/standards/iso639-2/php/code_list.php are used for ease of reference. These codes are XHO (isiXhosa), ZUL (isiZulu), TSN (Setswana), NSO (Sesotho sa Leboa) and VEN (Tshivenda).

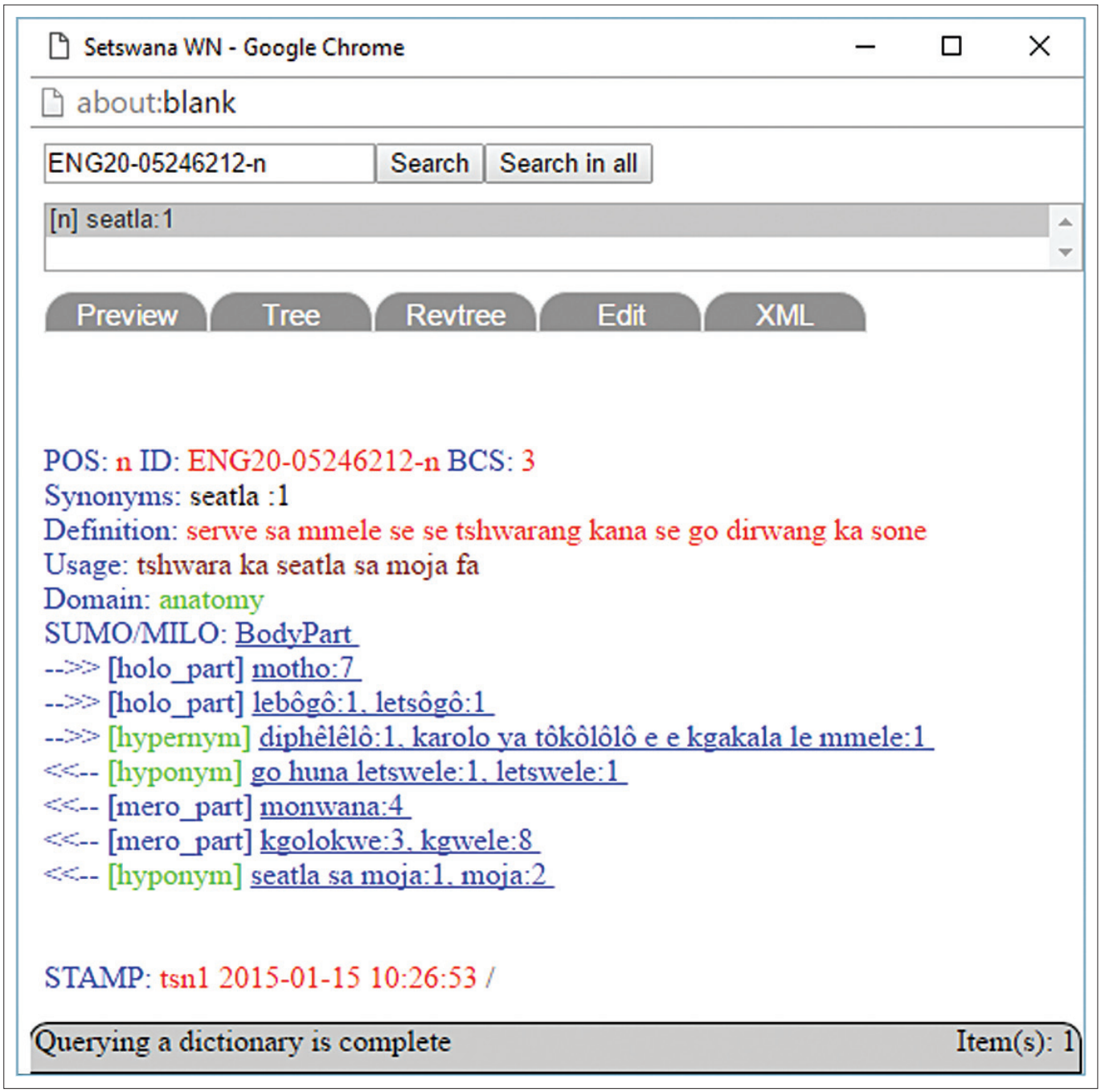

FIGURE 2: An example in DEBVisDic (seatla ~ hand).

then shifted in this direction. Currently, the development team is engaged in formal quality assurance and further experiments on providing each synset with at least one usage example (see the Conclusion and Future Work section for more details in this regard).

Table 1 reflects the status quo of the data contained in the AWN. The figures reported here were gradually built up over an 8-year period (2008–2016) and involved a large group of linguists working part-time on the project.

The only other wordnet covering a South African language is for Afrikaans (Kotzé 2008) and includes 10 068 synsets developed using a combination of manual and automatic methods. The AWN still has some way to go to reach the number of synsets encapsulated in larger projects such as:

- Princeton WordNet – 117 659 synsets for English (http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html)
- FinnWordNet – 120 449 synsets for Finnish (http://www.ling.helsinki.fi/en/lt/research/finnwordnet/news.shtml)
- plWn – 178 000 synsets for Polish (http://plwordnet.pwr.wroc.pl/wordnet/about)
- Chinese WordNet – 150 400 synsets (http://universal.elra.info/product_info.php?cPath=0_42_45&products_id=1637)

TABLE 1: Status quo of the African Wordnet Project.

Language	Synsets	Definitions	Usage examples
Northern Sotho (NSO)	8412	1178	5253
Tshivenda (VEN)	4270	209	4270
IsiZulu (ZUL)	10 782	2179	5112
IsiXhosa (XHO)	14 715	2198	7015
Setswana (TSN)	15 803	3515	7203
Total	**53 982**	**9279**	**28 853**

Source: African Wordnet Project information (unpublished)

- MultiWordnet (http://multiwordnet.fbk.eu/english/whatin.php), including new data for Italian and with mappings to other European languages such as Portuguese and Spanish – 32 673 synsets linked to PWN and 2 825 synsets which could not be linked.

Given the data scarceness of the African languages, manual development has been the only feasible option for continued development up to now; however, experiments to exploit the relatively little data available for the five African languages in our project have already begun. The next section describes each of our development strategies in more detail.

Development strategies

Base concepts

When following the expand model in wordnet development, the most important consideration is to decide how to identify the concepts that should be included in the wordnet first. Because manual development is a slow and labour-intensive task, one would aim at including concepts that are used frequently and in a broad spectrum of domains. Such a resource can be found in the Princeton Core Concepts list (Available at http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt). In addition to this set of seed terms, the AWN looked towards two international projects, the BalkaNet Project (See http://www.dblab.upatras.gr/balkanet/) and the EuroWordNet Project (See http://www.illc.uva.nl/EuroWordNet/), for guidance in this regard. In the EuroWordNet Project, a list of Common Base Concepts containing roughly 1024 concepts was identified and mapped to synsets in the PWN 2.0. This list was later expanded to about 5000 synsets and applied in the BalkaNet Project. Both the Princeton Core Concepts and BalkaNet Common Base Concepts aim at providing a starting point for new development. The Common 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' (Weisscher 2013). Some examples of the application of these lists are discussed below.

Lindén and Niemi (2014:191) base the creation of an extensive Finnish wordnet, directly aligned with the PWN, 'on the assumption that most synsets in PWN represent language-independent real-world concepts'. In creating a wordnet for Catalan (Benítez et al. 1998), the development team also used the Common Base Concepts as starting point because most of these concepts could be found in monolingual dictionaries for the language and thus transferred to wordnet format with relative ease. This automatic transfer only needed the manual addition of the hierarchical structure which is unique to ontologies, such as a wordnet to be done by linguists. In the IndoWordNet project (Prabhu et al. 2012), which also followed the expand model, the team creatively turned to poetry (some of which translated from English) to find equivalents for all the terms in the Common Base Concepts and Princeton Core Concepts lists.

During the first development phases, the AWN followed suit and used the extended Common Base Concepts list as well as the Princeton Core Concepts list to extract English synsets for linguists to include and translate into the African languages concerned. However, in the case of the AWN, it soon became clear that a more localised approach was needed. The seed lists described above contain many concepts that are not lexicalised in the African context. Linguists were forced to create new terminology or do time-consuming searches in their own text collections and online to match the foreign concepts. This not only resulted in many lengthy descriptions for unfamiliar terms being included as synonyms but also hampered progress as our inexperienced team were discouraged by the time-consuming work.

Some slight meaning differences between concepts in the African languages from those captured in PWN also came to light. The Setswana word *utlwa*, for instance, points to four of the five qualities which could be found under the English concept 'sensation; sense experience'. In English, this concept would include perception of taste, sight, hearing, smell and feeling, but in Setswana, the concept excludes the perception of sight and includes the meaning of understanding and listening. Such a small difference in meaning would be very confusing for a user of an English-Setswana wordnet if we were to simply link the synsets for 'sensation' with *utlwa*.

Developers of wordnets in other languages seemed to experience the same difficulties with non-lexicalised terminology (cf. Vincze & Almási 2014) and, like the IndoWordNet, needed to find creative methods for solving this problem. The most popular solutions seemed to be (1) coining of new terms to cover the entire PWN scope and (2) creating language-specific synsets which are not linked to PWN as a subset in the larger database. This presents challenges for the storage and multilingual nature of the new wordnets but ensures a thorough coverage in the target language.

In Table 2, we give some examples of nouns included in the Princeton Core Concepts and Common Base Concepts lists that are not lexicalised in the African languages. From the table, it can be seen that linguists sometimes provided very long descriptions for non-lexicalised terminology, rather than a precise concept. For instance, the translation provided for 'apparatus' in isiZulu literally means 'equipment to make something specific', while the translation of 'buffet' in isiXhosa denotes 'food that is

TABLE 2: Examples of Princeton Core Concepts and Common Base Concepts not lexicalised in African languages.

Concepts list with ID in PWN	African language translations
Princeton core	
apparatus (ENG20-02633705-n) POS: n ID: ENG20-02633705-n BCS: 1 Synonyms: apparatus: 1, setup: 1 Definition: equipment designed to serve a specific function Domain: factotum SUMO/MILO†: + Device	NSO: diaparata VEN: tshishumuswa ZUL: impahla yokwenza okuthile XHO: izixhobo TSN: aparata
buffet (ENG20-07108586-n) POS: n ID: ENG20-07108586-n Synonyms: buffet: 2 Definition: a meal set out on a buffet at which guests help themselves Domain: gastronomy SUMO/MILO: + Food -->> [hypernym] meal:1, repast:1 <<-- [hyponym] smorgasbord:2	ZUL: isiphihli Definition: ukudla okwenzelwe ukuthi abantu baziphakele Usage: abantu babusiswe ngesiphihli sokudla kwakusihlwa XHO: ukudla okubekwe ukuba abantu baziphakele TSN: boitsholêlô Usage: Baeng ba nnake ba jele dijô tse di monate tsa boitsholêlô.
sable (ENG20-02362706-n) POS: n ID: ENG20-02362706-n BCS: 3 Synonyms: sable: 5, Martes zibellina :1 Definition: marten of northern Asian forests having luxuriant dark brown fur Domain: zoology SUMO/MILO: + Carnivore -->> [hypernym] marten:1, marten cat:1	NSO: mohuta wa kgano: 4
Common base	
earldom (ENG20-08035206-n) POS: n ID: ENG20-08035206-n Synonyms: earldom: 2 Definition: the domain controlled by an earl or count or countess Domain: administration SUMO/MILO: + GeopoliticalArea -->> [hypernym] domain:2, demesne:2, land:4	NSO: tikologo ya mohlomphegi: 1
earldom (ENG20-13615298-n) POS: n ID: ENG20-13615298-n Synonyms: earldom: 1 Definition: the dignity or rank or position of an earl or countess Domain: heraldry SUMO/MILO: + SocialRole -->> [hypernym] rank:2	ZUL: isikhundla seyeli: 1 TSN: botlhokamolêmô: 1 Usage: O tlogile seriti ka ntlha ya botlhôkamôlemô jwa gagwe.

†, SUMO (Suggested Upper Merged Ontology) and MILO (Mid-level Ontology) are used to organise synsets into common categories in a hierarchical manner. The Semantic Domains section discusses these categories in more detail.

Source: African Wordnet Project information (unpublished)

displayed so that people can help themselves'. Lexicalisation challenges are also encountered in the synsets 'earldom' in both the administrative and heraldry domains because these concepts are foreign to the African context.

Following the example of the HuWN (Vincze & Almási 2014:119), we subsequently made provision for linguists to indicate that an English noun is not lexicalised in the African languages by introducing a special field in the DEBVisDic editor where linguists can clearly indicate the non-lexicalisation of concepts in their particular language. In this way, superfluous non-lexicalised synsets are eliminated so that the wordnet of the language concerned eventually takes on a natural way of lexicalising concepts.

Organic growth

Some linguists working on the AWN quickly gained confidence and a thorough understanding of the goal of a wordnet. These linguists realised that adhering to the lists discussed in the Base Concepts section would not result in a truly African database but instead be nothing more than a translation of foreign (European) terminology. Especially, the Setswana language team ventured off this list quickly and began including synsets that the researchers found interesting and applicable to other areas of their work. Linguists would start by including the most prototypical sense of a frequent word and allow this sense to guide them to the next. Because of this organic style, the Setswana wordnet includes many figurative meanings and unexpected relations in the ontology structure. The Setswana team also simultaneously included adjectives and verbs semantically related to the nouns that they were developing. One such example can be seen in the Setswana word *sebete* (n: liver, adj: brave) which is semantically linked to *bogale* (adj: sharpness, brave, angry, strong). Both words are synonyms for being brave, but have other meanings that place them in a variety of domains, including BodyParts (see the Semantic Domains section for more information on how the team used this characteristic). The result is a semantically narrower wordnet when considering the different domains or levels included but a much richer resource in the semantic domains that are covered. The only restriction which was placed on the team was that each synset they included should still be linked to a synset and sense in the PWN, thus still following the expand model and negating the need for either creating a new ontology to fit the data or manually aligning with the PWN at a later stage.

Because this style of expansion did not suit all linguists, we continued to supply lists of seed terms, but tried to extract those synsets in the PWN that had a solid localised base in the African languages. The most rudimentary way of doing this was to provide a continually updated list of terms that other languages in the AWN had already included in their wordnets. Languages such as isiXhosa and isiZulu that belong to the same language group show similar morphological and orthographic patterns as demonstrated by Pretorius and Bosch (2009). It therefore makes sense to use the data developed for the related language productively and

expand on it, rather than starting new development from the ground up. This is not an unusual approach, particularly in under-resourced languages, as recommended by Alberts and Mollema (2013:46) for the harmonisation of terminology in the South African Bantu languages. Linguists working on languages in the same language group were encouraged to work together in the development of new synsets for the AWN, and information was shared among the groups at regular project meetings. Thus, the isiZulu and isiXhosa wordnets now contain a shared batch of 3746 synsets, whereas Setswana and Sesotho sa Leboa share a batch of 1527 synsets. The first four languages that were added to the AWN, isiXhosa, isiZulu, Setswana and Sesotho sa Leboa, share 2532 synsets among all four languages, which is indeed a constructive step towards multilingual applications of the wordnets. A language learner can, for instance, see translations for a word in different languages when consulting a semantically tagged corpus. The organic growth style encouraged not only shared synsets but also leads to valuable comparative insights between Tshivenda, isiZulu and Sesotho sa Leboa, for example, in the case of the naming of body parts (see Madonsela et al. 2016).

As mentioned previously, Tshivenda was later added as a fifth language when the development for the other four languages had already been under way for some time. To ease and speed up the development, the linguists were not only provided with the localised base concepts list as described in the Corpus Frequencies section, but also with the completed synsets from Sesotho sa Leboa as examples. The choice of language for this fast tracking was purely a practical consideration – both linguists working on Tshivenda also had a very good knowledge of Sesotho sa Leboa and felt confident to use this data when creating new synsets for their language. Tshivenda linguists, therefore, did not only have a list of lexicalised terms in English to start incorporating into their wordnet, but the seed list was further enhanced with the lemmas, usage examples and definitions in Sesotho sa Leboa. This approach worked exceptionally well and witnessed the Tshivenda wordnet grow to nearly 5000 synsets within 3 years.

Corpus frequencies

As the project progressed, more resources became available, for instance, via the RMA (2013). These resources could be used to extract our own lists based on real-world parallel corpora for the languages included in AWN, and therefore allowed us to follow new methods. To test this approach, a multilingual parallel corpus, including all 11 official South African languages, 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 with the Princeton Core Concepts. From the multilingual corpus, we extracted a frequency list for Tshivenda and compared the 5000 most frequent terms in the multilingual African wordlist with the list of (English) base and core concepts mentioned above.

This frequency list extracted from the above mentioned multilingual corpus includes concepts that reflect unique African language usage but are also skew in terms of domain representation. Most of the data in the parallel corpus were sourced from government domain web pages and freely available online newspapers. As the data were mostly sourced from the web, a platform that is quite new for the African languages, it also does not reflect older but still acceptable word forms. The domains included also do not provide many figurative interpretations (see Eiselen & Puttkammer 2014 for a complete description of the corpora developed in the NCHLT project). This can be seen in the fact that some of the more frequent words found in the corpora included 'benefit' (2042 occurrences, translated as *uncedo* in XHO and *kholo* in NSO) and 'money' (1592 occurrences, translated as *imali* in ZUL and XHO and as *tšhêlêtê* or *madi* in TSN). It is therefore clear that the lists are by no means a well-rounded representation of the language usage, but at least the linguists in our team were acquainted with the concepts, which allowed for further exploiting and organic growth (see the Base Concepts section for more discussion on this point). The approach utilising frequency lists from language-specific corpora proposed here was also followed in the development of the Romanian WordNet (Tufiş et al. 2006:337). In the case of Tshivenda, the result was a list of concepts that are internationally regarded and commonly occur in modern African corpora. The list of roughly 1000 concepts was shared with the linguists as a starting point for Tshivenda wordnet development.

It should, however, be noted that the disjunctive orthography of Tshivenda lends itself to straightforward extraction of frequency lists from corpora, particularly in the case of nouns. Extraction of frequency lists in isiZulu and isiXhosa, with their conjunctive orthography, is more complex and requires preprocessing, such as morphological analysis of text corpora, to identify word roots (cf. Bosch & Pretorius 2011).

Semantic domains

Development speed was significantly higher in the teams that focussed on lexicalised terminology and were following the direction each term led them in to exploit more meanings and senses. The organic growth style with seed lists that were extracted from local corpora suited the AWN team. Concurrently, some teams needed guidance in the form of seed lists more than others as the sheer quantity of work still to be done for AWN was overwhelming.

While individual methods and workflows were respected and linguists were encouraged to follow whichever method suited their situation (taking time constraints, research involvement and experience into consideration), the problem remained that we aimed to create an African wordnet where a significant set of synsets would at least be shared across all languages in the project. According to Anderson, Pretorius and Kotzé (2010:3757), 'the establishment of inter-lingual indices and ontologies would make cross-linguistic information retrieval and question answering possible, and significantly aid machine translation'. It was therefore decided to exploit the existing structure of the wordnet as a

semantic ontology and the hierarchical structure of the Suggested Upper Merged Ontology (SUMO 2002) and Mid-Level Ontology (MILO; Niles & Pease 2001) that is already represented in the wordnet design.

Niles and Pease (2001:3) describe the ontologies as follows: 'The SUMO provides definitions for general-purpose terms and acts as a foundation for more specific domain ontologies'. The MILO is described as 'an ontology that is being developed as a bridge between the abstract content of the SUMO and the rich detail of the various domain ontologies'. Combining these two frameworks aims at putting forth a hierarchical categorisation that is both machine-readable and easily understood by human interpreters. After discussing the different development methods used by the language teams in the AWN with the entire group, a compromise between more free-flowing development as used by the Setswana team and the stricter coherence to the PWN was suggested – all linguists would work in a specific SUMO and MILO domain and exploit all interesting terminology within that domain for a specified time frame. After this period, the larger project team would re-evaluate the direction this domain took them in and consider moving to the next. As a first suggestion, the team agreed to work on 'Body Parts'. This domain is much less abstract, concepts are shared across the languages to a large degree, and it was therefore hoped that the synsets could be added faster than domains lower down (and thus more abstract) in the ontologies.

This approach seemed to work especially well for encouraging collaboration among the language groups. Although the number of synsets developed did not increase significantly, constructive discussions dealing with the comparison of linguistic phenomena took place and resulted in various research outputs (cf. Mabusela 2013; Madonsela & Mahonga 2013; Madonsela et al. 2016; Mojapelo 2013, 2016).

Linking

For the first 5 years of development, linguists were responsible for identifying and translating appropriate synsets without much help from electronic resources. Over that period, the African Wordnets only grew with an average of 1000 synsets per language per year (see Griesel & Bosch 2014 for a detailed introduction). Regarding the Catalan Wordnet, Benítez et al. (1998:1) confirm that although manual construction of lexicons is the most reliable technique, it is costly and highly time-consuming. They continue by giving this as the reason why researchers rather focus 'on the massive acquisition of lexical knowledge and semantic information from pre-existing structured lexical resources as automatically as possible'.

Recently, research was done to speed up the manual development of synsets in the AWN in order to ease the workload of the human language experts. The investigations centred around utilising the minimal amount of information available in limited bilingual dictionaries to identify synsets in the PWN that could be included in the AWN semi-automatically. After identifying appropriate and still missing synsets, key pieces of information from the dictionary can be transferred to the wordnet presented to linguists for final approval and inclusion in the wordnets.

For the experiments described here, a few basic bilingual dictionaries that were made available for research purposes were used. These resources ranged in scope from a few hundred terms in a bilingual wordlist with little more than a translated lemma to a more comprehensive bilingual dictionary with at least a part of speech tag and some indication of the meaning. Many of the dictionaries were not in machine-readable format and required extensive proofreading to ensure a usable data source. The dictionaries were also often older manuscripts and therefore did not include newer terminology or word forms (cf. the Setswana-English dictionary by Brown [1925] that is freely available from https://archive.org/details/secwanadictiona00browgoog).

Similar studies using bilingual dictionaries have been conducted for a variety of languages. Oliver (2014) describes various methods for automatic expansion of wordnets using Wikipedia, bilingual dictionaries, BabelNet, machine translation and other resources to identify and validate possible synsets. His methods depend heavily on the availability of online data sources in the languages for which he proposes to build wordnets and deliver promising results. Montazery and Faili (2010) used bilingual corpora and a large dictionary for Persian-English to map PWN synsets with Persian words, based on a score calculated for each synset and the possible Persian translations linked to it. With a precision score of 82% on unambiguous synsets, this method shows great promise but requires quite a large amount of data for the target language.

The biggest difference between these languages and the African languages, however, is the wealth and quantity of data contained in dictionaries which are freely accessible, often in machine-readable formats. Most of the entries in the dictionaries used in the above studies had at least searchable definitions and examples for each lemma. In the case of African languages, even basic resources like computer readable and electronic dictionaries are not always freely available. Given this resource scarceness, we had to develop a semi-automatic method of extracting possible synsets from the data listed above. It was decided to still include manual verification in the methodology as the data available were either very small or outdated and would, therefore, be more difficult to map to the PWN.

As described in the Introduction section, a basic synset is made up of a literal, a part of speech tag and the different semantic relations deemed necessary by the SUMO and MILO categorisation. It is also linked to the PWN by a unique identification code (ENG ID). By virtue of this ENG ID, the five African language wordnets are then connected to form a multilingual resource. Utilising the minimal amount of information available in the electronic resources listed above

TABLE 3: Examples of the spreadsheets used in the manual verification of the linking technique.

Language	English	English definition and identification	Match? (yes or no)	If Yes	
				NSO usage example	VEN usage example
Northern Sotho					
moamandêlê	almond tree	ENG20-11896052-n: any of several small bushy trees having pink or white blossoms and usually bearing nuts	yes	Kenyo ya moamandele e na le koko ye e jewago.	-
alefabete	alphabet	ENG20-06096415-n: a character set that includes letters and is used to write a language	yes (*alternative spelling = alfabete*)	Ge motho a nyaka go ngwala le go peleta ka nepagalo o swanetše go tseba alefabete.	-
moagôtlhatlaganô	storey	ENG20-03243815-n: structure consisting of a room or set of rooms comprising a single level of a multilevel building	no (*= 'multilevel building', 'NOT storey' or 'single level'*)		-
Tshivenda					
agere	acre	ENG20-12847449-n: a unit of area (4840 square yards) used in English-speaking countries	yes	-	Tsimu ya Vho-Vele ndi khulu, ndi agere mbili
volenga	arum lily	ENG20-11047703-n: South African plant widely cultivated for its showy pure white spathe and yellow spadix	yes	-	Maluvha a volenga a na muvhala mutshena
babalasi	hangover	ENG20-13628315-n: disagreeable aftereffects from the use of drugs (especially alcohol)	yes	-	Denga o farwa nga babalasi nge a nwa halwa vhunzhi mulovha
belekedzo	animal	ENG20-00012748-n: a living organism characterised by voluntary movement	no (*=that accompanies wife when she returns to the husband she wrongfully deserted*)	-	-

(sometimes as little as a lemma and its translation), we identified synsets in the PWN as potential links. One such example from the Sesotho sa Leboa dictionary is 'almond tree' with its translation *moamandêlê*. Using the English lemma, a possible match could easily be found in the PWN, from which an ENG ID and definition were extracted. A simple spreadsheet was drawn up with all of the possible matches like that for 'almond tree' (see Table 3). This sheet has a column each for the Sesotho sa Leboa and the English lemmas, as found in the bilingual dictionary, the ENG ID and definition from the PWN, and open columns for the linguist to indicate a true match and provide a usage example if the definition and lemma are indeed a match. Linguists simply had to indicate whether the definition matched the dictionary entry with a 'yes' or 'no' classification. Because of various copyright issues, no usage examples from external sources could be used, but linguists rather had to provide a novel example for each matched synset. The information provided by the linguists was then automatically carried over to the appropriate XML database format, with additional information on the SUMO and MILO classification, domain and hierarchical structure which could be garnered from the English PWN.

The methodology was kept simple while utilising as much of the dictionaries as possible. Mappings where linguists marked 'no' will be evaluated at a later stage and might still be included in the wordnet, but linked to a different PWN counterpart. Although the addition of a manual verification step seems unnecessary given the promising results (see Table 4), Kotzé (2008:180) warns that a single lexicographical resource does not suffice for this type of experiment. It was therefore deemed necessary to ensure the quality of the AWN that linguists verified and in some cases corrected the entries before we included these in the larger database. Efforts to acquire more (bilingual) wordlists and corpora are ongoing.

TABLE 4: New synsets added semi-automatically.

Language	Nouns in resource	Linked nouns	Successful links (%)
Setswana	905	786	86.8
isiZulu	382	345	90.3
isiXhosa	1294	1108	85.6
Tshivenda	5117	3218	62.8
Sesotho sa Leboa	316	301	95.2
Total added	8014	5758	71.8

Source: AWN Project information (unpublished)

Extracting usage examples

In addition to the basic synset, each sense should be further enriched by a usage example in the target language showing the use of the sense in context. In most wordnets of languages which are highly resourced, usage examples are semi-automatically extracted from available (tagged) corpora as demonstrated, for instance, by Broda, Maziarz and Piasecki (2012:3648) with regard to the Polish wordnet plWordNet. It is interesting to note that Broda et al. op cit. advise that linguists should not depend exclusively on intuition, but should also consult available corpora for usage examples. However, they caution that 'finding examples of rare senses of words in a large corpus is difficult and time-consuming'.

In the case of the AWN, linguists did not use corpora to find usage examples, but either created their own examples or translated the English usage examples if these were available. For example, in the case of:

- buffet (ENG20-07108586-n), there is no usage example in the PWN, but in the Setswana synset, a usage example has been provided, viz. *Baeng ba nnake ba jele dijô tse di monate tsa boitsholêlô.*
- newspaper (ENG20-07573103-n), the English usage example is not translated into isiZulu, but a new usage example is created: *ngifunda iphephandaba*.

- standing (ENG20-13156245-n) the English usage example has simply been translated into Sesotho sa Leboa viz. *leloko la boemo bja godimo setšhabeng* 'a member in good standing (in society)'.

A first attempt was then made to fast-track the process for the isiZulu wordnet by using the recently compiled isiZulu Wortschatz corpus (Universität Leipzig http://corpora.informatik.uni-leipzig.de/) to semi-automate the process of extracting usage examples. The corpus that contains 100K sentences with an average length of 12 words per sentence implements NoSketchEngine as concordance user interface or corpus browser and has a basic lemmatiser for isiZulu built-in (see http://cql.corpora.uni-leipzig.de/?corpusId=zul_mixed_2014). This semi-automatic process is illustrated in Figure 3.

The search on word form (*igatsha*) results in nine hits; however, because of the complex morphology and conjunctive orthography of isiZulu, it might be necessary in some cases to widen the search to lemma (*gatsha*) because it results in 59 hits, presenting a wider range of usage examples to choose from. Future research will include thorough evaluation of the corpus and its suitability to automatic extraction of usage examples.

Conclusion and future work

Much has been written about the resource scarceness of the African languages (cf. the resource audit performed by Grover, Van Huyssteen & Pretorius 2010). Wordnets not only aim to serve as direct sources of data for further human language technology and linguistic research, but also to create more intricate resources such as semantically tagged corpora, information retrieval systems and the like. The AWN is a first step in creating valuable databases for five South African languages and will soon be made available to the larger linguistic community for further research.

The development strategies used for building a first version of the AWN for isiXhosa, isiZulu, Setswana, Sesotho sa Leboa and Tshivenda were described. Despite the varying strategies implemented to build wordnets for the five African

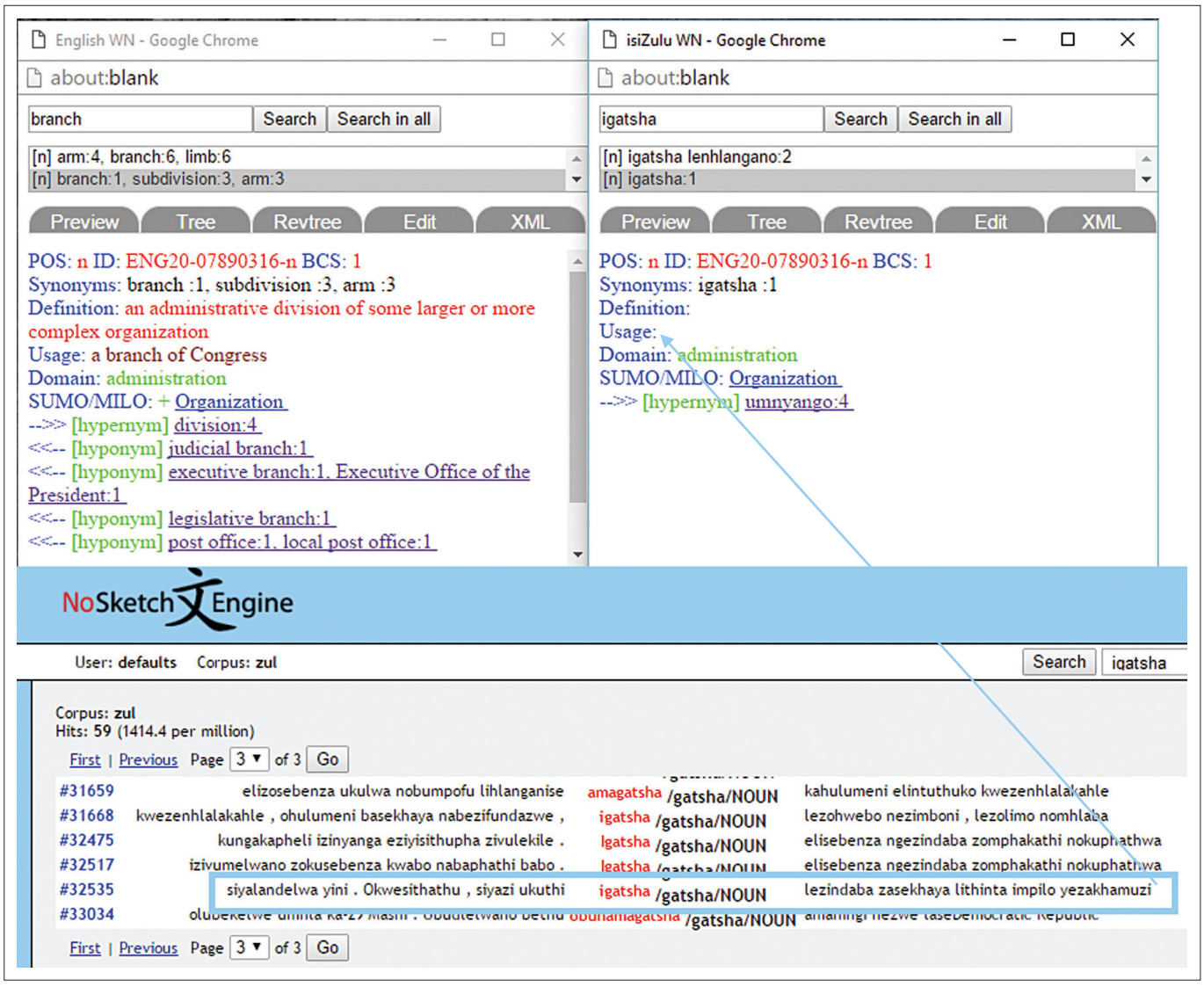

FIGURE 3: Extraction of usage example from online corpus.

languages concerned, the similarities shared on levels such as morphology or grammar and semantics allow the language teams to learn from one another, to share and thus to fast-track the development of the individual wordnets in this way.

The dilemma of under-resourced African languages further called for the implementation of various methods during the early stages of development. For example, much of the initial work was done manually following the expand model from the PWN. As the team gathered more experience and suitable lexical resources became available, more localised guidance could be given in the form of frequency-based seed terms and semi-automatic linking of lemmas from bilingual wordlists and the PWN. Experiments to speed up the collection of usage examples from online corpora also show promising results.

Each of the various development strategies plays a part in creating the unique AWN (Figure 4). These methods were employed simultaneously by the different language groups, depending on their level of experience, the available language resources and the individual preferences of the linguists. This qualitative approach to development worked well in a diverse team, building wordnets for five African languages in parallel.

During the development, a few questions also arose as to the best practise for handling conceptual and lexical gaps which exist between English and the African languages. Bentivogli and Pianta (2000:665) point out that even among culturally related languages, such as Italian and English, it has been shown that a medium-sized dictionary of English to Italian contains around 7.8% lexical gaps, where there is no equivalence and a free translation is needed. One possible solution for the AWN might lie in participating in the Global WordNet Grid initiative (Fellbaum & Vossen 2007; Vossen et al. 2016) which aims at creating a centralised platform for all wordnets with focussed efforts in including new concepts in multiple languages.

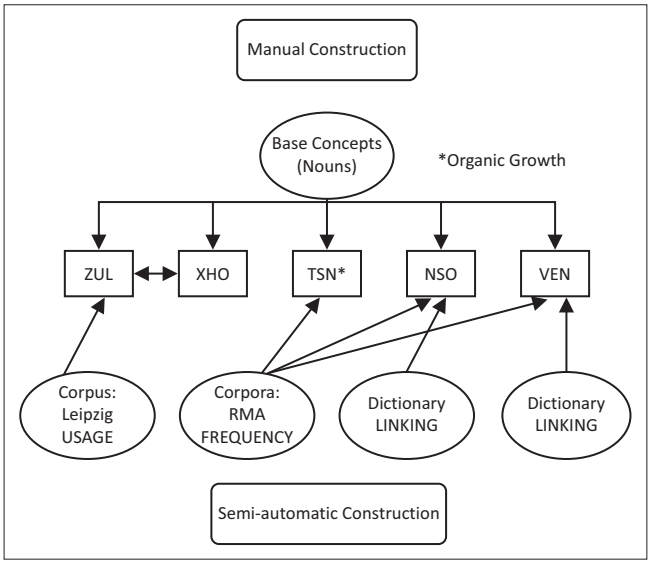

FIGURE 4: Summary of different components used in creating AWN.

The AWN is currently undergoing extensive quality assurance. Aspects that need attention are to verify that the African language concept is linked to the correct English PWN synset and sense, spellchecking and normalisation of the content to ensure uniformity. Smrz (2004) also lists several categories that can be tracked and corrected automatically, including omitted POS tags, SUMO and MILO categories, etc. Future work will include creating a user interface to not only browse and develop new synsets but also to check the quality of the completed work in time as a linguist verifies the data. When freely available text manipulation tools such as lemmatisers and spellcheckers become available for the African languages, these tools should also be incorporated into such an interface.

Because of the limited availability of lexicographic and basic language resources for the African languages, wordnet construction presents a challenging and time-consuming manual task for linguists. As it stands, notwithstanding the different strategies implemented in the development of wordnets for the five African languages concerned, the AWN provides a solid base for future development of new synsets, expansion of the synsets with usage examples and definitions and inclusion of further African languages. It is foreseen that this first version will soon become a useful tool in the creation of more complex applications.

Acknowledgements

The authors acknowledge the South African National HLT Network, Department of Arts and Culture, and Women in Research Fund (University of South Africa) for providing funding in the various phases of the AWN project; Christiane Fellbaum (Princeton University) for constructive feedback on an earlier draft of this article; and the AWN Development Team for linguistic expertise.

Competing interests

The authors declare that they have no financial or personal relationship(s) that may have inappropriately influenced them in writing this article.

Author's contributions

S.E.B. was the project leader, linguistic coordinator and senior researcher in the African Wordnet Project. She conceptualised the idea for the research, extracted language-specific examples, analysed the relevant linguistic data and wrote large parts of the manuscript. M.G. was the project manager, technical coordinator and research assistant in the African Wordnet Project. She co-developed large parts of the manuscript and performed experiments of a technical nature.

References

Abdullah, N. & Ibrahim, R., 2015, 'Managing information by utilizing WordNet as the database for semantic search engine', *International Journal of Software Engineering and Its Applications* 9(5), 193–204. https://doi.org/10.14257/ijseia.2015.9.5.19

Alberts, M. & Mollema, N., 2013, 'Developing legal terminology in African languages as aid to the court interpreter: A South African perspective', *Lexikos* 23, 29–58. https://doi.org/10.5788/23-1-1203

Anderson, W., Pretorius, L. & Kotzé, A., 2010, 'Base concepts in the African languages compared to upper ontologies and the WordNet Top Ontology', in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, et al. (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 17–23, pp. 3757–3764.

Benítez, L., Cervell, S., Escudero, G., López, M., Rigau, G. & Taulé, M., 1998, 'Methods and tools for building the Catalan Wordnet', in *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain, May 28–30, viewed 22 August 2016, from http://www.cs.upc.edu/~escudero/wsd/98-lrec.pdf

Bentivogli, L. & Pianta, E., 2000, 'Looking for lexical gaps', in *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*, Stuttgart, Germany, August 8–12, pp. 663–669.

Bond, F. & Paik, K., 2012, 'A survey of wordnets and their licenses', in V. Mititelu, C. Forăscu, C. Fellbaum & P. Vossen (eds.), *Proceedings of the Sixth Global WordNet Conference 2012 (GWC2012)*, Matsue, January 25–29, pp. 64–71.

Bosch, S.E. & Pretorius, L., 2011, 'Towards Zulu corpus clean-up, lexicon development and corpus annotation by means of computational morphological analysis', *South African Journal of African Languages* 31(1), 138–158.

Broda, B., Maziarz, M. & Piasecki, M., 2012, 'Tools for plWordNet development: Presentation and perspectives', in N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, et al. (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 21–27, pp. 3647–3652.

DEBVisDic: WordNet editor and browser, n.d., viewed 19 September 2016, from http://deb.fi.muni.cz/clients-debvisdic.php

Eiselen, E. & Puttkammer, M., 2014, 'Developing text resources for ten South African languages', in N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, et al. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 26–31, pp. 3698–3703.

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

Fellbaum, C. & Vossen, P., 2007, 'Connecting the universal to the specific: Towards the global grid', in *Proceedings of The First International Workshop on Intercultural Collaboration (IWIC 2007)*, Kyoto, Japan, January 25–26.

Global Wordnet Association, 2016, viewed 19 September 2016, from http://globalwordnet.org/

Griesel, M. & Bosch, S., 2014, 'Taking stock of the African Wordnet project: 5 years of development', in H. Orav, C. Fellbaum & P. Vossen (eds.), *Proceedings of the 7th Global WordNet Conference 2014 (GWC2014), Demonstration Session*, Tartu, Estonia, January 25–29, pp. 148–153, viewed 19 September 2016, from http://gwc2014.ut.ee/proceedings_of_GWC_2014.pdf

Grover, A.S., Van Huyssteen, G.B. & Pretorius, M.W., 2010, 'South African human language technologies audit', in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, et al. (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 17–23, pp. 2847–2850.

Kotzé, G., 2008, 'Ontwikkeling van 'n Afrikaanse woordnet: Metodologie en integrasie' [Development of an Afrikaans wordnet: Methodology and integration], *Literator* 29(1), 163–184. https://doi.org/10.4102/lit.v29i1.105

Language Resource Management Agency, 2013, viewed 12 February 2016, from http://rma.nwu.ac.za/index.php/

Lindén, K. & Niemi, J., 2014, 'Is it possible to create a very large wordnet in 100 days? An evaluation', *Language Resources & Evaluation* 48(2), 191–201. https://doi.org/10.1007/s10579-013-9245-0

Mabusela, M., 2013, 'Meaning equivalence of adjectives within the African Wordnet context', presented at the 17th Biennial International Conference of the African Language Association of Southern Africa (ALASA), Pretoria, South Africa, 17–19 July, viewed 20 September 2016, from http://www.unisa.ac.za/chs/news/wp-content/uploads/2012/11/ALASA-Final-Programme-2013.pdf

Madonsela, S. & Mahonga, L., 2013, 'The problems of translating words with extended meanings: Lessons from isiZulu synsets', presented at the 17th Biennial International Conference of the African Language Association of Southern Africa (ALASA), Pretoria, South Africa, 17–19 July, viewed 20 September 2016, from http://www.unisa.ac.za/chs/news/wp-content/uploads/2012/11/ALASA-Final-Programme-2013.pdf

Madonsela, S., Mojapelo, M.L., Mafela, M.J. & Masubelele, R., 2016, 'African WordNet: A viable tool for sense discrimination in the indigenous African languages of South Africa', in V. Mititelu, C. Forăscu, C. Fellbaum & P. Vossen (eds.), *Proceedings of the Eighth Global WordNet Conference 2016 (GWC2016)*, Bucharest, Romania, January 25–29, pp. 192–198.

Miller, G.A., 1995, 'WordNet: A lexical database for English', *Communications of the ACM* 38(11), 38–41. https://doi.org/10.1145/219717.219748

Mojapelo, M., 2013, 'Morphological considerations for encoding the qualificative in African WordNet with reference to Northern Sotho', presented at the 17th Biennial International Conference of the African Language Association of Southern Africa (ALASA), Pretoria, South Africa, July 17–19, viewed 20 September 2016, from http://www.unisa.ac.za/chs/news/wp-content/uploads/2012/11/ALASA-Final-Programme-2013.pdf

Mojapelo, M.L., 2016, 'Semantics of body parts in African WordNet: A case of Northern Sotho', in V. Mititelu, C. Forăscu, C. Fellbaum & P. Vossen (eds.), *Proceedings of the Eighth Global WordNet Conference 2016 (GWC2016)*, Bucharest, Romania, January 25–29, pp. 233–241.

Montazery, M. & Faili, H., 2010, 'Automatic Persian WordNet construction', in C.R. Huang (ed.), *Proceedings of the 23rd International Conference on Computational Linguistics: Posters Volume (COLING'10)*, Beijing, China, August 23–27, pp. 846–850.

Navigli, R., Jurgens, D. & Vannella, D., 2013, 'Semeval-2013 task 12: Multilingual word sense disambiguation', in *Second Joint Conference on Lexical and Computational Semantics*, Atlanta, GA, June 13–14, pp. 222–231.

Niles, I. & Pease, A., 2001, 'Toward a standard upper ontology', in B. Smith & C. Welty (eds.), *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, ME, October 17–19, pp. 3–9.

Oliver, A., 2014, 'WN-Toolkit: Automatic generation of WordNets following the expand model', in H. Orav, C. Fellbaum & P. Vossen (eds.), *Proceedings of the 7th Global WordNet Conference 2014 (GWC2014), Demonstration Session*, Tartu, Estonia, January 25–29, pp. 7–15.

Ordan, N. & Wintner, S., 2007, 'Hebrew WordNet: A test case of aligning lexical databases across languages', *International Journal of Translation, special issue on Lexical Resources for Machine Translation* 19(1), 39–58.

Prabhu, V., Desai, S., Redkar, H., Prabhugaonkar, N., Nagvenkar, A. & Karmali, R., 2012, 'An efficient database design for IndoWordNet development using hybrid approach', in *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP) at COLING 2012*, Mumbai, December, pp. 229–236.

Pretorius, L. & Bosch, S., 2009. 'Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology', in *Proceedings of the Workshop on Language Technologies for African Languages, 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Technografia Digital Press, March 31, 2009, pp. 96–102.

Princeton University, 2016, *WordNet – A lexical database for English*, viewed 12 August 2016, from https://wordnet.princeton.edu/

Rambousek, A. & Horak, A., 2016, 'DEBVisDic: Instant Wordnet building', in V. Mititelu, C. Forăscu, C. Fellbaum & P. Vossen (eds.), *Proceedings of the Eighth Global WordNet Conference 2016* (GWC2016), Bucharest, Romania, January 25–29, pp. 317–321.

Smrz, P., 2004, 'Quality control for Wordnet development', in P. Vossen (ed.), *Proceedings of the 2nd International Wordnet Conference*, Brno, Czech Republic, January 20–23, 2004.

Spiegler, S., van der Spuy, A. & Flach, P.A., 2010., 'Ukwabelana – An open-source morphological Zulu corpus', in *Proceedings of the 23rd International Conference on Computational Linguistics* (Coling 2010), Beijing, August, pp. 1020–1028.

SUMO, 2002, *Suggested upper merged ontology*, viewed 22 August 2016, from http://ontology.teknowledge.com

Susanti, Y., Iida, R. & Tokunaga, T., 2015, 'Automatic generation of English vocabulary tests', in M. Helfert (ed.), *Proceedings of the 7th International Conference on Computer Supported Education (CSEDU 2015)*, Lisbon, Portugal, May 23–25, pp. 77–78.

Tufiş, D., Mititelu, V.B., Bozianu, L. & Mihăilă, C., 2006, 'Romanian WordNet: New developments and applications', in P. Sojka, K. Choi, C. Fellbaum & P. Vossen (eds.), *Proceedings of the 3rd Global Wordnet Conference 2006 (GWC2006)*, Masaryk, Czech Republic, January, pp. 337–344.

Vetulani, Z., Kubis, M. & Obrębski, T., 2010, 'PolNet – Polish WordNet: Data and tools', in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, et al. (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 17–23, pp. 3739–3797.

Vincze, V. & Almási, A., 2014, 'Non-lexicalized concepts in Wordnets: A case study of English and Hungarian', in H. Orav, C. Fellbaum & P. Vossen (eds.), *Proceedings of the 7th Global WordNet Conference 2014 (GWC2014), Demonstration Session*, Tartu, Estonia, January 25–29, pp. 118–126.

Vossen, P. (ed.), 1998, *EuroWordNet: A multilingual database with lexical semantic networks for European Languages*, Kluwer, Dordrecht.

Vossen, P., Bond, F. & McCrae, J., 2016, 'Toward a truly multilingual GlobalWordnet Grid', in V. Mititelu, C. Forăscu, C. Fellbaum & P. Vossen (eds.), *Proceedings of the Eighth Global WordNet Conference 2016 (GWC2016)*, Bucharest, Romania, January 25–29.

Weisscher, A., 2013, *Global Wordnet Association base concepts*, viewed 22 August 2016, from http://globalwordnet.org/gwa-base-concepts/