Base Concepts in the African Languages Compared to Upper Ontologies and the WordNet Top Ontology

Winston Anderson, Laurette Pretorius, Albert Kotzé

University of South Africa
School of Computing and Department of African Languages, PO Box 392, UNISA, 0003, South Africa,
8999104@mylife.unisa.ac.za, pretol@unisa.ac.za, kotzeae@unisa.ac.za

Abstract

Ontologies, and in particular upper ontologies, are foundational to the establishment of the Semantic Web. Upper ontologies are used as equivalence formalisms between domain specific ontologies. Multilingualism brings one of the key challenges to the development of these ontologies. Fundamental to the challenges of defining upper ontologies is the assumption that concepts are universally shared. The approach to developing linguistic ontologies aligned to upper ontologies, particularly in the non-Indo-European language families, has highlighted these challenges. Previously two approaches to developing new linguistic ontologies and the influence of these approaches on the upper ontologies have been well documented. These approaches are examined in a unique new context: the African, and in particular, the Bantu languages. In particular, we address the following two questions: Which approach is better for the alignment of the African languages to upper ontologies? Can the concepts that are linguistically shared amongst the African languages be aligned easily with upper ontology concepts claimed to be universally shared?

1. Introduction

Berners-Lee (2006) envisioned a new generation of the web, called the Semantic Web that would enable the automated access of information and the use of this information based on machine-processable data. Such proposed use requires the definition of robust ontologies for reliable inference.

Benjamins et al (2004) have shown that ontology development and multilingualism are two of the six challenges confronting the Semantic Web. With regards to multilingualism and the Semantic Web, various more detailed challenges have been highlighted by others. These include the use of ontologies to integrate the Semantic Web with language technologies (Gatius et al., 2006), the use of semi-formal natural language descriptions to navigate and interpret services on the Semantic Web (Ding et al., 2003), and the challenges of trying to align linguistic base concepts and ontologies with the upper ontologies required for inference on the Semantic Web (Gangemi, 2004). The the progress of the implementation of HLT and the Bantu languages (Bosch et al., 2006) compounds these challenges for resource development for the African languages.

The South African Bantu languages have a solid documented grammatical and lexical foundation. These serve as traditional language resources supporting humans in creating and processing text in human language technologies today (Bosch, 2007). Halfway through the nineteenth century interest in the field of Bantu grammars was sparked off by the work of missionaries whose primary task was to reach the people in their own languages. One of the treasures that emerged from these studies in the middle of the last century was the establishment of a broad taxonomy of all the Bantu languages by the linguistics department of Oxford University (Guthrie, 1948) and by European researchers (Meinhof, 1932; Meeussen and Rodegem, 1969). This research for a common lexical base for the all the Bantu languages mirrored the similar previous studies into Indo-European lexical relations.

1.1. WordNets, top ontologies and upper ontologies

WordNet describes itself as a large lexical database (Miller et al., 1990; Miller, 1995). Lexemes are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. WordNet is a combination of cognitive and linguistic ontology (Fellbaum, 1998) and is based on a taxonomical structure with the concepts of hyponyms, synonyms, meronyms and antonyms at its core. For the purposes of this paper we refer to WordNet as all WordNets developed for multiple languages under the Global WordNet (GWN) project.

The "top ontology" is the 64 concepts based on existing linguistic classifications and adapted to represent the diversity of the Base Concepts (BCs) by the EuroWordNet and GWN projects (Vossen et al., 1997; Vossen, 1998). The 64 concepts are based on the fundamental semantic distinctions used in various semantic theories and paradigms forming a hierarchy of language-independent concepts reflecting the distinctions between, for example, object and substance or dynamic and static.They have explicitly been defined in terms of hyponymy and opposition relations.

An "upper ontology" is an upper level ontology that provides definitions for general-purpose terms. It acts as a foundation for linking more specific domain ontologies for computational usage such as cross-domain inference. Our specific focus for upper ontology in this research is SUMO (Suggested Upper Merged Ontology).

Much of the international work around WordNet and SUMO has been connected to inter-lingual indices (ILIs) and top ontologies (Niles and Pease, 2003b) or WordNet and OWL (van Assem et al., 2006). There are already 40 existing global WordNet databases, and the establishment of inter-lingual indices and ontologies would make cross-linguistic information retrieval and question answering possible, and significantly aid machine translation. This would also be the case for the African language WordNets.

The appropriate approach to this upper/top ontology linkage is the key research question of this paper. In other words, should the nature of the alignment of the linguistically common Bantu language concepts with previous efforts at selecting and defining base concepts and upper ontologies influence the approach to development of the African language WordNets?

The structure of the remainder of the paper is as follows: Section 2 addresses the construction of WordNets and the role that key role that base concepts play. In section 3 introduces so-called main Bantu Lexical Reconstructions or main entry roots. Our approach in addressing the key research question and the alignment methodology used, are explicated in section 4. A detailed discussion of the results follow in section 5. Section 6 contains the conclusion and suggested future research directions.

2. WordNet construction

GWN has defined synsets that are most important in 4 WordNets for different languages (English, Spanish, Dutch and Italian), the so-called base concepts (BCs). EuroWordNet was developed with a shared set of 1024 so-called Common Base Concepts (CBCs),which were classified using a common shared semantic framework. These BCs were chosen as the most significant meanings in the local European WordNets (Vossen et al., 1997). The BalkaNet project extended the list by considering Greek, Romanian, Serbian, Turkish and Bulgarian to 4689 synsets and upgraded the mapping of the CBCs to Princeton WordNet 2.0. The latest published list contains 5000 CBCs.

The BCs are the major building blocks on which the other word meanings in the WordNets depend. They were introduced to reach maximum overlap and compatibility across WordNets in different languages, allowing for the distributive development of WordNets in the world, with each WordNet being a language specific structure and lexicalization pattern. The BCs are supposed to be the concepts that play the most important role in the various WordNets of different languages. This role was measured in terms of two main criteria: a high position in the semantic hierarchy and having many relations to other concepts.

This approach is a similar approach often used in the construction of upper ontologies - concepts that have high agreement between ontologies and have high positions in their hierarchy.

The 1024 CBCs have a reduced set of 164 core base concepts that occur in 3 or more WordNets as important meanings. The Global WordNet project further defined an ontology of 71 base types (a reduction of the 164 core base concepts). The reduction involved removing unbalanced hyponyms (when both the hypernym and hyponym are present but not other co-hyponyms) and by replacing closely related synsets (e.g. act and action) by a single type. The base types are a minimalized list of fundamental concepts. The base types (the semantic primitives or taxonomy top nodes) play a key role in large-scale semantic networks like the Semantic Web. By providing clear definitions or features for these base types GWN has stated that it is possible to augment a large-scale lexicon with rich feature structures, via (multiple) hyponymy relations that connect each word meaning to the relevant base types.

Subsequent to the Euro WordNet project, which started the drive towards Global WordNet, there has been significant development of WordNets for other languages - BalkaNet (Balkanet, 2001) and Slovene WordNet (Fišer, 2007) also developed a mapping to a top ontology.

We analyzed approaches to top ontologies and BCs in those languages that fall outside the Indo-European family, specifically the Arabic WordNet (Black et al., 2006), Hebrew WordNet (Ordan and Wintner, 2007) and Chinese WordNet (Wong and Pala, 2001; Huang et al., 2004). Of particular interest to the Semantic Web, is that all of these three latter developments were done in conjunction with ontology development.

In the construction of the Hebrew WordNet, Ordan (2007) discusses two paradigms for constructing WordNets – either construction from scratch followed by alignment (see the merge approach below) or alternatively, strict alignment with Princeton WordNet as the base under the assumption that those concepts are universally shared (see the the expand approach below). The latter approach involves the potential risk that the resulting hierarchy will be influenced by Princeton WordNet. Oran proposes that the expand approach is still a better approach for languages poor in resources.

A similar argument for the two different WordNet construction paradigms is proposed by Vossen (2007). In the expand approach WordNet synsets are translated to another language and the structure is then inherited and managed. An advantage of this approach is that it is an "easier and more efficient method" and compatible with Princeton WordNet, which allows the exploitation of many resources already linked to Princeton WordNet, for instance SUMO, WordNet domains and selection restriction from the British National Corpus. The disadvantage is that it will be biased.

In the merge approach, there is the creation of an independent WordNet in another language which is then aligned with the Princeton WordNet by generating the appropriate translations. This approach has the disadvantage of being more complex and labour intensive and will create a structure different from that of the Princeton WordNet, but the advantage is that the language specific patterns can be maintained.

The African WordNet project was initiated by Bosch (2007). The aim of the project is to create a platform for WordNet development for African languages, based on existing global networks such as the English WordNet (Princeton), the EuroWordNet and BalkaNet. Linking the African language WordNets to one another is strategic.

3. Main entry roots in the Bantu languages

In the linguistics of the Bantu languages, there have been projects over the last 50 years to align the core concepts of all the languages. The Comparative On-line Bantu Dictionary (CBOLD) project has taken the initial linguistic unification work and extended it (Schadeberg, 2002). The CBOLD project was started in 1994 by Larry Hyman and John Lowe to produce a lexicographic database in Berkeley to support and enhance the theoretical, descriptive, and historical linguistic study of the languages in the Bantu language family. CBOLD includes a list of reconstructed Proto-Bantu roots, thousands of additional reconstructed regional roots called Bantu Lexical Reconstructions 2 (BLR2), and reflexes of these roots for a substantial subset of the more than 500 daughter languages.

Of these roots the CBOLD project has selected 10000 BLR3 reconstructions (Bastin et al., 2003) that represent so called main entries of which there are 1400. These main entries are referred to as basic reconstructed etymons.

The main entries have been further categorized by Maho (2001) to isolate all main entries that have modern reflexes in Zone A and Zone S, as shown in figure 1 (Zone S is the region containing all the Southern African Bantu languages).

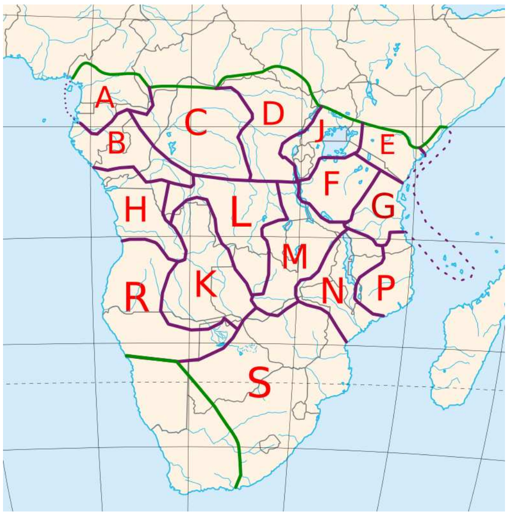

Figure 1: Bantu Language zones in Sub-Saharan Africa

4. Basic approach taken

In order to answer our key research question we examine the mentioned two approaches to developing ontologies in a Bantu language context and reformulate our question as two subquestions: Can the linguistically common or "upper" ontological concepts be aligned readily with those claimed to be universally shared? Moreover, are there core concepts in the Bantu languages that are not core concepts in the Princeton WordNet and vice versa?

Our modus operandi was as follows: the 1400 main entries of the CBOLD BLR3 list of 10 000 suggested Proto-Bantu reconstructions were utilized as the theoretical base, then further reduced to the subset proposed by Maho for Zone A and S languages, resulting in 375 entries. Of these Maho determined which main entries have modern reflexes with a claimed total zone-spread covering at least 14 zones of a total of 16 zones, yielding 231 entries. These were then further reduced by the authors to words that have zone spread across all 16 zones and are therefore also in Zone S, where equivalent modern reflexes can be found in Northern Sotho and Zulu (by reference to the predominant local dictionary for each). Northern Sotho and Zulu are representative of two significant different large groups within Zone S[1].

These (80 concepts in number) were mapped to their Princeton WordNet equivalents if they existed or marked if no mappings were found. The mappings were verified and the phonetic mapping to BLR3 quality assured, which reduced the final list to 67 words. The principles used for the mapping were the ILI and EuroWordNet base concept methodology, and existing SUMO mapping verification.

4.1. Alignment methodology

Ontological tree comparison measurements have been proposed for measuring the similarity of concept trees (Xue et al., 2009). We have reused their definitions for calculations of alignment with Princeton WordNet concepts and thus the Core concept alignment.

They describe a mechanism for comparing ontologies. Whereas the classical methods used structural and geometric characteristics of trees, focusing on the nodes affected, they propose more attention to the concepts represented by internal nodes. Specifically they consider the position and conceptual similarities of the affected nodes to be considered in the comparison. They achieve this by defining four distinct tree transformation operations, each which has a different transformation cost.

Of interest to us are the insert, move and relabelling operations. The reason for using these costs is that at the completion of all our research we can determine a final transformation cost for our Bantu language core concept tree, in comparison to the Global WordNet Base Concept tree as represented in Princeton WordNet. For the purposes of this paper, we will not look at the final cost, but at the individual operations per concept to compare the alignment of individual concepts.

5. Results

The final quality assured concept list was analyzed. A subset of this Bantu concept list is shown in Table 1. The proto-Bantu refers to the original root concept that has been attested in all 16 Bantu languages zones, including Zone A and S, and verified that it has a local Northern Sotho lexicalization. The BLR3 reference refers to the reference number for the proto-Bantu root on the CBOLD project. The attested meaning is the meaning provided by Maho. The POS indicates the part of speech for the proto-Bantu root. The WordNet sense is the English Princeton WordNet closest equivalent mapped via ILI. The tree operation indicates the base operation required to calculate the ontological similarity measurement. The word is the noun stem or verb root or adjective in Northern Sotho. The noun stem is shown independent of nominal class. The core set indicates whether the English Princeton concept is in the Balkanet Common Synset (BCS) list (Smrž, 2004), and in which list. Being in the BCS incorporates being in the Glabal WordNet Core Concept list. The SUMO domain is the mapping of the concept to SUMO as provided via the ILI link to Princeton WordNet. The SUMO operation indicates the WordNet

[1] The examples given here and results are shown for Northern Sotho only, since its lexicalization has been verified and quality assured. The Zulu lexicalizations are still being quality assured at the time of publication and have therefore not been referenced.

mapping operation to SUMO and the SUMO node indicates the mapped node.

5.1. Sense alignment with Princeton WordNet

The majority of the 67 concepts (62 or 93%) do align well with an English Princeton WordNet concept already defined. Alignment means that the major sense of the word (the first listed sense of the word in at least the 2 most authoritative dictionaries (Kriel, 2003; Ziervogel and Mokgokong, 1985) in a lexicalized form (Northern Sotho) has one-to-one synonymy with a Princeton WordNet sense.

For, example, the noun '-bàdí', which BLR3 represents as 'pool; pond; deep water; well' and which is lexicalized in Northern Sotho as *sediba:1*, maps to the Princeton WordNet noun sense *pool:2*.

The verb '-łánɪk-', which BLR3 represents as 'to spread to dry in the sun; to spread out' and which is lexicalized in Northern Sotho as *anega:1*, maps to the Princeton WordNet verb sense *air:1*.

The adjective '-łíŋí', which BLR3 represents as 'many, much' and which is lexicalized in Northern Sotho as *ntšhi:1, ntši:1*, maps to the Princeton WordNet adjective sense *many:1*.

In terms of our alignment methodology, this one-to-one alignment is referred to as "relabelling" in the context of ontological comparison metrics.

If we consider more complicated sense alignments (the remaining 7%), then there are 3 other potential scenarios - insert, move, and combinations of insert and move. This is as a result of the concept either not fully lexicalized in English (insert) or the sense in Northern Sotho of the English equivalent sense does not align with the current position of that English sense in the WordNet concept tree (move or insert and move).

There are three insert operations of new concepts - one verb and two nouns.

Consider the verb example of '-pép', which BLR3 represents as 'to blow as wind; to winnow; to smoke tobacco; to breathe', lexicalized in Northern Sotho as *fefera:1* and described by the comprehensive dictionary as primary sense: 'winnow (stamped corn is shaken in a lesêlô until the chaff lies on top)'. This is a hyponym of the Princeton WordNet sense *winnow:1, fan:4*, as its meaning is more specific than the Princeton WordNet closest equivalent.

A complex transformation (move and insert) is required for the Northern Sotho word "kgaka:1, Numida meleagris coronata:1" which has sense *Numida coronata, crowned guinea-fowl* in the comprehensive dictionary. The complexity is that this would be inserted as hyponym under a tree structure of *bird, fowl, landfowl, poultry, Numididae, Numida, Numida maleagris*. The Princeton WordNet is quite specific on European and New World birds, but could represent African birds better. The current guinea fowl in WordNet is defined as a West African bird under the structure "bird, gallinaceous bird, domestic fowl ". The guinea fowl is regarded by mother tongue speakers as both a wild fowl and a domestic fowl. Inserting it under landfowl in a WordNet tree would make more sense. In fact, this confirms a former conclusion made about the heterogeneity in the intuitive level of generality in WordNet (Oltramari et al., 2002). Specifically they have shown that for animals there is ontological confusion in WordNet between types (landfowl versus waterfowl) and rôles (domestic fowl versus gamefowl).

Beside the 67 quality assured concepts, there were other concepts that were inserted into the African language WordNet with the same or a similar problem. Interestingly enough, the broad pattern is that the complex transformation is often required for animals that are African specific, e.g. the Northern Sotho words *lehoho:1, lekhukhu:1*, which is *Francolinus swainsonii* and *kwale:1* which is *Francolinus lavaillantoides*. They are both types of francolin, which is a small type of partridge indigenous to Africa. The concept "francolin", which does exist in most English dictionaries, is not a Princeton WordNet lexicalized concept.

These complex transformations appear to be rare and specific, so we do not use these examples to detract from the broader fit to the BCs, but merely to highlight that there will be obvious divergence for African specific concepts.

There are no complex transformations for verbs or adjectives.

There are two nouns and one verb that require move operations in the Northern Sotho WordNet tree from the corresponding position of the concept in the Princeton WordNet tree.

The BLR3 entry (BLR3 ref 2071) "-kʊ́pá", which represents 'tick; insect' and is lexicalized in Northern Sotho as *kgofa:1, Ixodida:1* has a sense of "parasite" more than Arachnid, so it has been mapped to *tick:2* in Princeton WordNet using ILI, but has the hypernym structure *Arachnida:1, Acari:1, Parasitiformes:1/kgofa:2, kgofa:1* rather than the current Princeton WordNet hypernym structure *arachnid:1, acarine:1, tick:2*

The sense alignments of the BLR3 concepts, when locally lexicalized into Northern Sotho, therefore, largely map well via the ILI to Princeton WordNet, with a few notable exceptions.

5.2. Base Concept alignment of BLR3 with Balkanet Common Synsets

The alignment of the Bantu language concepts in our study (which, to repeat for emphasis, are words that occur in over 500 languages across 16 Bantu language zones in Africa and are lexicalized in Zone A and S at the furthest geographical extremes) to the Global WordNet BCs is not as good as the individual word sense alignment to Princeton WordNet.

The Bantu language concepts cover 35 of the BalkaNet Common Synsets (BCS) in Global WordNet. The Bantu language concepts cover 15 level 1 BCS in Global WordNet,12 level 2 BCS and 8 level 3 BCS.

The rest of the 67 Bantu language concepts (32 or 49%) do not match the BCS.

Of the matching level 1 BCS nine are verbs and six are nouns. In level 2, seven are nouns and five are verbs and in level 3 there are six nouns, no verbs and two adjectives mappings.

So there is only a half set correspondence of Bantu language core concepts to Balkanet Common Synsets. The other half is unique to the Bantu languages.

Proto-Bantu	BLR3 Ref	Attested and/or reconstructed meaning	POS	WordNet sense	Tree Operation	Word or Stem	Core Set	SUMO Domain	SUMO Operation	SUMO Node
łáná	3203	‘child’	n	**Child:2**	relabelling	ngwana	1	person	+	Human
łókà	3536	‘snake; intestinal worm’	n	**Snake:1**	relabelling	noga	3	zoology	=	Snake
łíkí	1622	‘bee’	n	**Bee:1**	relabelling	nose	2	entomology	=	Bee
ntʊ̀	4807	‘some entity; any’	n	**Person:1**	relabelling	motho	1	biology	=	Human
łíŋí	3485	‘many, much’	adj	**Many:1**	relabelling	-ntši	None	factotum	=	Subjective Assessment Attribute
ɲó-	7047	‘to drink’	v	**Drink:1**	relabelling	-nwa	1	alimentation	=	Beverage
łót-	3579	‘to warm oneself’	v	**Bask:2**	relabelling	-ora	None	factotum	+	Process

Table 1: Sample BLR Roots and Meanings

5.3. Base Concept alignment of BLR3 with Global WordNet Base Concepts

The goal of the BCs in Global WordNet is to represent core concepts that have a high position in the semantic hierarchy or many relations to other concepts. The universality of Global WordNet focusses on specific BCs of differing types:

- Common Base Concepts (CBC): concepts that act as BCs in at least two languages;
- Local Base Concepts (LBC): concepts that act as BCs in only a single language;
- Global Base Concepts (GBC): concepts that act as BCs in all languages of the world.

The 5000 Balkanet Common Synsets include all the original EuroWordNet and Global WordNet BCs.

The mismatch of the 49% of the concepts here means they do not occur in the full 5000 CBCs determined by EuroWordNet and BalkaNet for Global WordNet.

These Global WordNet BCs were used to construct the WordNet Top Ontology, so the significance of this mismatch is important specifically for the development of African language WordNets.

5.4. Top ontology comparison

Of the 64 top ontology concepts, the Bantu BCs concepts map to 25 1st Order Entities and 42 2nd Order Entities. There are no mappings to 3rd Order Entities (Figure 2).

The lack of mappings to 3rd order entities corresponds to findings in mapping the Top Ontology to Chinese where similarly no linkage was found between the Chinese BCs (radicals in Chinese) and the 3rd order entities (Pala and Wong, 1999).

In terms of qualia rôles within the Top Ontology, the majority rôles mapped are Physical, Dynamic, BoundedEvent, Object and Agentitive. For the comprehensive list refer to Table 2.

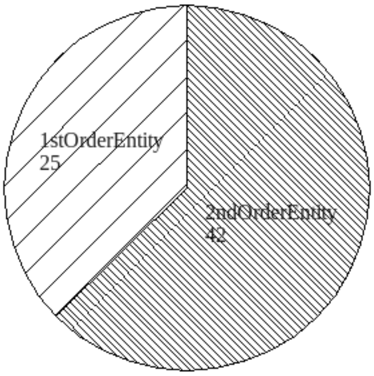

Figure 2: Bantu Base concepts by Top Ontology Entity Orders

5.5. Upper ontology comparison

Of the 33 Bantu Language concepts not aligned to BCs in Global WordNet, the majority have a hypernym relationship to SUMO (not synonymy to a SUMO node, but subsumption).

In terms of comparison of the full 67 items in the list of Bantu language concepts aligning with SUMO, the coverage is reflected over 30 factota as the majority ontological mapping. Following factotum, the significance of coverage by the number of concepts covering that domain are:

1. anatomy and gastronomy
2. entomology, number and zoology
3. quality, biology and number

The rest of the domains are covered by one concept in the list only. These domains are: alimentation, botany, dance, geography, industry, medicine, meteorology, person, physiology and play.

Thirty six of the Bantu language concepts are linked to SUMO via synonymy. To accomplish this linkage we used

Qualia rôle	Bantu concepts mapped
Physical	25
Dynamic	12
BoundedEvent	11
Object	11
Agentive	10
Animal	9
Condition	8
Location	8
Quantity	8
UnboundedEvent	8
Cause	7
Experience	7
Part	7
Purpose	6
Living	5
Property	5
Static	5
Human	4
Phenomenal	4
Solid	4
Comestible	3
Relation	3
Social	3
Usage	3
Existence	2
Manner	2
Natural	2
Artifact	1
Covering	1
LanguageRepresentation	1
Liquid	1
Mental	1
Place	1
Plant	1
Possession	1
Substance	1
Time	1

Table 2: Bantu concept mapping to top ontology qualia rôles

the linkage methodology of Niles (Niles and Pease, 2003a). Twenty eight of the concepts are linked via an hypernym to a SUMO node. Three of the concepts have neither equivalence in meaning in SUMO nor subsumption in meaning (all of these are numbers - adjectival concepts mapped to the SUMO 'Integer' node).

The only concept that mapped to Princeton WordNet, for which Princeton WordNet does not have an existing SUMO mapping, is *sangoma:1*. The mapping was made to the WordNet hypernym's SUMO mapping (TherapeuticProcess in domain medicine), but changing the operator from synonymy with SUMO to hyperonomy with SUMO.

In terms of the classes in SUMO, the attribute and process class are the most well represented in their sub-classes. Between physical and abstract concept classes, the physical class is well represented. Within the physical class, of the four types of object sub-classes, three are represented. All of the process sub-classes are represented by concepts. The abstract class is not as well represented. Figure 3 illustrates the subsumption of the Bantu core concepts in the SUMO top level classes. The dotted nodes reflect classes not covering any core concepts. For deeper sub-class levels, these are just summarised by number of nodes.

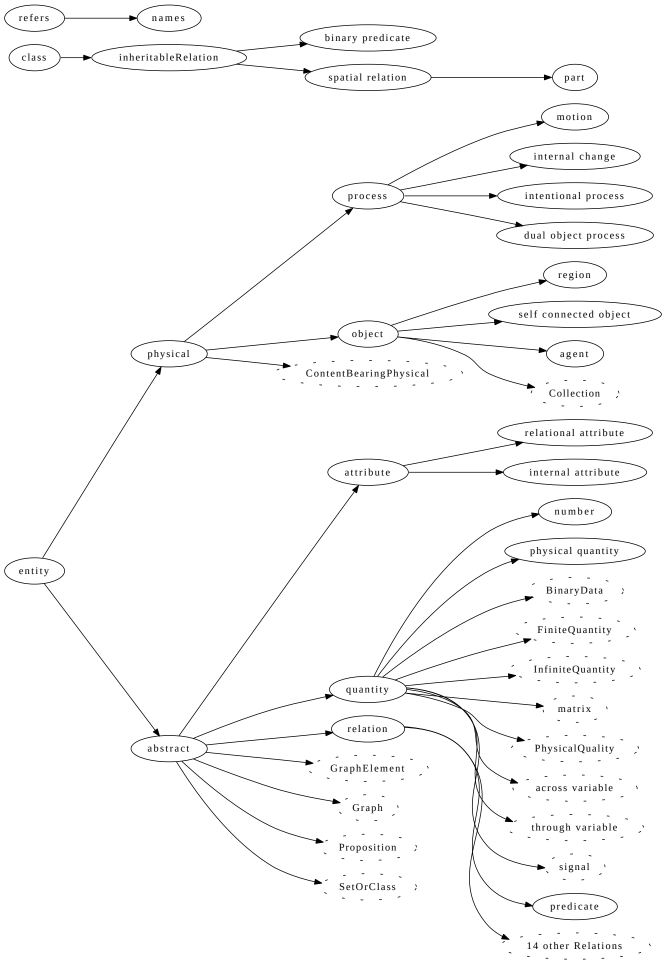

Figure 3: Bantu base concept subsumption in SUMO

6. Conclusions and future work

Our examination has highlighted a number of key issues. The construction from scratch followed by alignment was used on a subset of concepts, strategically chosen as already recognised core concepts in the Bantu languages by many linguists over years of research.

We aimed to produce an informed approach to alignment to WordNet and upper ontologies. It is clear from our results that alignment at the word sense level is good (93% fit), but alignment between the BC set proposed by Global WordNet and our BC set is not good (only half fit).

Use of the Global WordNet BCs as a starting point will not necessarily be a good idea for the African languages. This approach uses strict alignment with Princeton WordNet as the base. Within Africa, this strict alignment was used for the construction of the Afrikaans WordNet (Kotzé, 2008). This made sense as the core concepts in Afrikaans would probably closely align with the core concepts in Dutch which was used as one of the inputs to the Global

WordNet BCs. The advantages of the strict alignment approach used with Afrikaans is that bootstrapping is made easier, and automation can be utilised to advantage with a less resourced language - the advantages proposed by Ordan (2007). This is only beneficial if the core concepts of the language, particularly those words that are used as the base for most morphological derivation, are not decidedly different from the Global WordNet BCs.

The disadvantage of that approach is that the fundamental WordNet base will be biased to those concepts that are not necessarily core in the new target language.

Since the focus in WordNet has always been on concept hyponymy based on mother tongue speaker understanding, we propose a *hybrid* approach to building future African language WordNets.

The first step would be to build the core concepts from scratch, or use the current BLR3 lists as a base, and the second step to build out the WordNet structure using automation and alignment with Princeton WordNet (first expand and then merge approach (Vossen, 2007)). Both fundamental steps here should use the ILI as a bridging mechanism. This should provide the advantage that the core base concepts will be more appropriate, but that the rest of the concepts will be mapped well in an automated approach.

This approach could also be used for other language families initiating WordNets that are not related to the Indo-European family.

An interesting observation is that the alignment to Global WordNet of the BCs was "better" at the top levels for verbs, and "better" at the lower levels for nouns. This could indicate that for the Global WordNet BC requirement that the concept occupies a "high position in the semantic hierarchy", the importance of verbs will need to be considered. It might be appropriate to focus on the verb structure first in terms of BCs.

The result in terms of alignment with upper ontology concepts claimed to be universally shared is not as conclusive. 53% of the Bantu language concepts had synonymy with SUMO. The obvious nodes, such as "entity" match well, but it is not immediately clear why "Bee" (a Global WordNet Base Concept and a Bantu language core concept) has synonymy with SUMO but "Tick" (only a Bantu language core concept) does not. Should they be part of SUMO or rather part of a domain specific ontology?

Consider the verb examples of "heat" and "cool". Both words exist as BCs in Global WordNet BCs and in our Bantu language core concepts. The one is regarded in Princeton WordNet as the antonym of the other Process, but the WordNet mapping to SUMO regards *heat:1* as subsumption of SUMO node Heating, but *cool:1* as synonymy with SUMO node Cooling. This is either a misalignment between Princeton WordNet and SUMO, or if aligned correctly would produce different logical interpretation of OWL and RDF results for these concepts. Logical discrepancies can result from this – alignment of one concept via synonymy and the opposite concept by a sub-class relationship.

Further research would need to be done on the SUMO alignment to produce more conclusive results.

This research has produced peripheral resource artefacts that are useful for further research. An open available base for the Northern Sotho WordNet is now available as part of the DEB Visdic project. To link the BCs to Princeton WordNet, not only were the 67 concepts mentioned here created, but many other related concepts to complete the tree in terms of hyponomy, meronomy and morphological derivation. The ILI linkage allows for these concepts to be easily added into the related African language WordNets.

The list chosen as a subset, discussed in this the paper, is the quality assured list. It can still be the case that, after expert opinion, we add to this core list. Part of the ongoing project is to continually add to this list. Significant further comparison work to SUMO can be done once the African language WordNets are more substantial in terms of concepts. Once a number of different languages are completed, it will be worthwhile to revisit this core concept list.

Even though the mapping via WordNet to SUMO raises interesting questions, the actual mapping of Northern Sotho words to SUMO appears successful and confirms what the original mapping of SUMO to Princeton Wordnet ascertained - that most nouns map to classes, most verbs map to &%subclasses of Process and most adjectives map to a &%SubjectiveAssessmentAttribute. The mapping directly from each concept to SUMO was clear, and therefore we can conclude that though there are linguistic mapping challenges to the WordNet Top Ontology, the Bantu languages can be aligned easily with upper ontology concepts claimed to be universally shared.

7. Acknowledgements

This material is based upon work supported by the South African National Research Foundation under grant number 2053403. Any opinion, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Research Foundation.

8. References

Balkanet. 2001. Annex i part b: Description of scientific/technological objectives and work plan. Technical report, Information Society Technologies, 02. BalkaNet Ref.No: 29388.

Y. Bastin, A. Coupez, E. Mumba, and T. C. Schadeberg. 2003. Reconstructions lexicales bantoues 3/Bantu lexical reconstructions 3. *Available on-line at http://linguistics. africamuseum. be/BLR3. html, accessed*, 3(4):4.

V. R. Benjamins, J. Contreras, O. Corcho, and A. Gomez-Perez. 2004. Six challenges for the semantic web. *AIS SIGSEMIS Bulletin*, 1(1):24–25.

Tim Berners-Lee. 2006. Semantic web roadmap. http://www.w3.org/DesignIssues/Sematnic.html.

W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum. 2006. Introducing the Arabic WordNet project. In *Proceedings of the third International WordNet Conference (GWC-06)*.

Sonja Bosch, Jackie Jones, Laurette Pretorius, and Winston Anderson. 2006. Resource development for the South African Bantu languages: Computational morphological analysers and machine-readable lexicons. In Justus

Roux and Sonja Bosch, editors, *Fifth International Conference on Language Resources and Evaluation : Workshop 3 - Networking the Development of Language Resources for African Languages*, pages 38–43. ELRA.

S. Bosch. 2007. African languages—is the writing on the screen? *Southern African Linguistics and Applied Language Studies*, 25(2):169.

Ying Ding, Cornelis J. van Rijsbergen, Iadh Ounis, and Joemon Jose. 2003. Report on ACM SIGIR workshop on "semantic web" SWIR 2003. *SIGIR Forum*, 37(2):45–49.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.

D. Fišer. 2007. Leveraging parallel corpora and existing WordNets for automatic construction of the Slovene wordnet. *Proceedings of the 3rd Language and Technology Conference*, 7:3–5.

Aldo Gangemi. 2004. Porting WordNets to the semantic web. Technical report, W3C, July.

Marta Gatius, Meritxell González, Sheyla Militello, and Pablo Hernández. 2006. Integrating semantic web and language technologies to improve the online public administrations services. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 847–848, New York, NY, USA. ACM.

M. Guthrie. 1948. *The classification of the Bantu languages*. Published for the International African Institute by the Oxford University Press.

C. R. Huang, R. Y. Chang, and S. B. Lee. 2004. Sinica BOW (bilingual ontological WordNet): Integration of bilingual WordNet and SUMO. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 26–28.

G 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.

T. J. Kriel. 2003. *Popular Northern Sotho Dictionary*. J. L. Van Schaik, Pretoria.

J. Maho. 2001. The Bantu area: towards clearing up a mess. *Africa and Asia: Goteborg working papers on Asian and African languages and literatures*, pages 1–40.

A. E. Meeussen and F. Rodegem. 1969. *Bantu lexical reconstructions*. Musée royal de l'Afrique centrale.

C Meinhof. 1932. *Introduction to the phonology of the Bantu languages*. Dietrich Reiner/Ernst Vohsen, Johannesburg. Translated, revised and enlarged in collaboration with the author and Dr. Alice Werner by N.J. Van Warmelo.

George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Five papers on WordNet. Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton University.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

I. Niles and A. Pease. 2003a. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, pages 412–416.

I. Niles and A. Pease. 2003b. Mapping WordNet to the SUMO ontology. In *Proceedings of the IEEE International Knowledge Engineering conference*, pages 23–26.

A. Oltramari, A. Gangemi, N. Guarino, and C. Masolo. 2002. Restructuring WordNet's top-level: The OntoClean approach. *Co-operating Organisations*, page 17.

N. Ordan and S. Wintner. 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.

K. Pala and S. H. S. Wong. 1999. Chinese characters and top ontology in eurowordnet. In *Proceedings of the Global WordNet Conference'2002*, pages 224–233. Mysore University Further information.

T. C. Schadeberg. 2002. Progress in Bantu lexical reconstruction. *Journal of African Languages and Linguistics*, 23(2):183–195.

P. Smrž. 2004. Quality control and checking for WordNet development: A case study of BalkaNet. *Romanian Journal of Information Science and Technology*, 7(1-2):173–182.

M. van Assem, A. Gangemi, and G. Schreiber. 2006. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy*.

P. Vossen, L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge, and W. Peters. 1997. The EuroWordNet base concepts and top ontology. Technical report, EuroWordNet. Deliverable D017,D034,D036 EuroWordNet LE2-4003.

P. Vossen. 1998. EuroWordNet: a multilingual database with lexical semantic networks. *Computational Linguistics*, 25(4).

Piek Vossen. 2007. African WordNet: EuroWordNet project. presented at CSIR at launch of African WordNet Project, March.

Shun Ha Sylvia Wong and Karel Pala. 2001. Chinese radicals and top ontology in EuroWordNet. In *TSD '01: Proceedings of the 4th International Conference on Text, Speech and Dialogue*, pages 313–322, London, UK. Springer-Verlag.

Y. Xue, C. Wang, H. H. Ghenniwa, and W. Shen. 2009. A tree similarity measuring method and its application to ontology comparison. *Journal of Universal Computer Science*, 15(9):1766–1781.

D. Ziervogel and P. C. Mokgokong. 1985. *Pukuntšu Ye Kgolo Ya Sesotho Sa Leboa/Groot Noord-Sotho Woordeboek/Comprehensive Northern Sotho Dictionary*. J. L. Van Schaik, Pretoria, second corrected edition.