Investigating the Universality of a Semantic Web
Upper Ontology in the Context of the African Languages

by

WINSTON NOËL ANDERSON

submitted in accordance with the requirements
for the degree of

Master of Science

in the subject

Computer Science

at the

UNIVERSITY OF SOUTH AFRICA

Supervisor: Prof. Laurette Pretorius
Co-supervisor: Prof. Albert Kotzé

August 2016
I declare that

**Investigating the Universality of a Semantic Web Upper Ontology in the Context of the African Languages**

is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

(Mr) W.N. Anderson

2016-08-17
And although learned men have long since thought of some kind of language or universal characteristic by which all concepts and things can be put into beautiful order, and with whose help different nations might communicate their thoughts and each read in his own language what another has written in his, yet no one has attempted a language or characteristic which includes at once both the arts of discovery and judgment, that is, one whose signs and characters serve the same purpose that arithmetical signs serve for numbers, and algebraic signs for quantities taken abstractly.

Philosophical Papers and Letters, Gottfried Wilhelm Leibniz (Loemker, 1976)
Acknowledgments

This material is partially based upon work supported by the South African NRF (grant number 2053403). Any opinion, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NRF.

A sincere thanks for the wisdom, insight and advice provided by my supervisor, Laurette, and co-supervisor, Albert. You patiently taught me the skill of good research writing. Your ideas helped refine my research interests and challenged my thinking. As a computer scientist, and not a linguist, I needed assistance in reviewing the data used in my research. Thank you to Mampaka Mojapelo, who sacrificed much of her valuable time to assist me and provided valuable insights, and to Jouni Maho who reviewed my key data.

A deep thanks to my family for their encouragement. Particularly to Hyreath who advised me and gave up personal time to help edit the drafts. Our Egypt trip, where you joined me for a brief holiday on the way to publically present this research in Malta, will always be remembered. Thank you also to Hyreath and our children for sacrificing so much of their time with me during this research. My mother, Florence, though no longer with us, was an encouragement via calls and visits to start and continue with this. She remains an inspiration.
Abstract

Ontologies are foundational to, and upper ontologies provide semantic integration across, the Semantic Web. Multilingualism has been shown to be a key challenge to the development of the Semantic Web, and is a particular challenge to the universality requirement of upper ontologies. Universality implies a qualitative mapping from lexical ontologies, like WordNet, to an upper ontology, such as SUMO. Are a given natural language family’s core concepts currently included in an existing, accepted upper ontology? Does SUMO preserve an ontological non-bias with respect to the multilingual challenge, particularly in the context of the African languages? The approach to developing WordNets mapped to shared core concepts in the non-Indo-European language families has highlighted these challenges and this is examined in a unique new context: the Southern African languages. This is achieved through a new mapping from African language core concepts to SUMO. It is shown that SUMO has no significant natural language ontology bias.
Diontholotši ke motheo wa Semantic Web, e lego mararankodi a tlhalošo, gomme diontholotši tša ka godimo di tlabakela ka togaganyo ya tlhalošo go kgabaganya mararankodi ao. Tšhomišo ya dipolelo ka bontši e laeditšwe e le tl-hohlo ye kgo mo tšweletšong ya Semantic Web, diontholotši, gomme kudukudu e tloga e le tlhohlo mo go beng le dinyakwakakaretšo ga diontholotši tša ka godimo. Kakaretšo mo e šupa boleng bja nyalantšho go tloga diontholotšing tša tlotlontšu bjalo ka WordNet, go ya ontholotšing ya ka godimo bjalo ka SUMO. Na dikgopolotheo tša leloko le le itšego la maleme ga bjale di akaretšwa mo ontholotšing ya ka godimo, ye e lego gona gomme e amogelwago? Na SUMO e kgonthiša go se sekamele ka lehlakoreng le le itšego ga diontholotši mabapi le ditlhohlo tša bolementši, kudu ge re lebetše seemotikologo sa maleme a Afrika? Tebanyo ya go tšweletša di-WordNet ka go di nyalantšha le dikgopolotheo tša mohlakanelwa melokong ya maleme ao e sego a Indo-European e tšweleditše ditlhohlo tše nyanyeng, gomme taba ye e tsinkelwa seemotikologong se sefsa sa moswananoši: Maleme a Afrika-Borwa. Se se phethagatšwa ka nyalantšho ye mpsha go tšwa dikgopolotheong tša maleme a Afrika go ya go SUMO. Go laetšwa gore SUMO ga e sekamele ka lehlakoreng le le itšego ka mokgwa wo o kwagalago mo diontholotšing tša maleme a tlhago.
Key Terms

Upper Ontology; Suggested Upper Merged Ontology (SUMO); Tree comparison; Ontology; Resource Description Framework (RDF); Lexical semantics; Semantic networks; Language resources; Open environment; WordNet; Extensible Markup Language (XML); African languages of Sub-Saharan Origin; Proto-Bantu language.
ACM CCS

The following are the key terms used in this dissertation as organized according to the Association of Computing Machinery Computing Classification System 2012:

1. *Mathematics of computing* → *Discrete Mathematics* → *Graph theory* → *Trees*;

2. (a) *Information systems* → *Information retrieval* → *Document representation* → *Ontologies*;
   
   (b) *Information systems* → *World Wide Web* → *Web data description languages* → *Semantic Web description languages* → *Resource Description Framework (RDF)*;

3. (a) *Computing methodologies* → *Artificial intelligence* → *Natural language processing* → *Lexical semantics*;
   
   (b) *Computing methodologies* → *Artificial intelligence* → *Knowledge representation and reasoning* → *Semantic networks*;
   
   (c) *Computing methodologies* → *Artificial intelligence* → *Natural language processing* → *Language resources*;

4. (a) *Proper nouns: People, technologies and companies* → *Technologies* → *WordNet*;

5. *Applied computing* → *Document management and text processing* → *Document preparation* → *Mark-up languages* → *Extensible Mark-up Language (XML)*;
6. *Social and professional topics* → *User characteristics* → *Cultural characteristics.*
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of figures</td>
<td>xv</td>
</tr>
<tr>
<td>List of tables</td>
<td>xvi</td>
</tr>
<tr>
<td>List of listings</td>
<td>xviii</td>
</tr>
<tr>
<td>List of algorithms</td>
<td>xix</td>
</tr>
<tr>
<td><strong>I Contextualisation</strong></td>
<td>1</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>2</td>
</tr>
<tr>
<td>1.1 Background</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Problem statement and research question</td>
<td>14</td>
</tr>
<tr>
<td>1.3 Research objectives and methods</td>
<td>18</td>
</tr>
<tr>
<td>1.4 Delineation of the research</td>
<td>20</td>
</tr>
<tr>
<td>1.5 Style conventions</td>
<td>21</td>
</tr>
<tr>
<td>1.6 Significance of the contribution</td>
<td>22</td>
</tr>
<tr>
<td>1.7 Structure of the dissertation</td>
<td>23</td>
</tr>
</tbody>
</table>
2 Semantic Web architecture

2.1 Semantic Web layered architecture .......................... 27
2.2 The foundational layers ........................................ 30
   2.2.1 Layer 1 – The Web platform .......................... 31
   2.2.2 Layer 2 – The syntax .................................. 37
2.3 The core layers .................................................. 40
   2.3.1 Layer 3 – Knowledge representation structure .......... 40
   2.3.2 Layer 4 – Semantics and rules ......................... 42
2.4 The top layers of the Semantic Web architecture .......... 56
   2.4.1 Layers 5, 6 and 7 – Logic, proof and trust .......... 57
2.5 Goals of the Semantic Web architecture ...................... 58

3 Lexical core concepts and lexical ontologies ................. 61

3.1 Introduction .................................................... 61
3.2 Semantic concepts in linguistics ............................... 62
3.3 Lexical ontologies ............................................. 63
3.4 WordNet base concepts ....................................... 67
3.5 Qualia rôles .................................................. 69
3.6 African language concepts .................................... 70
3.7 African WordNet construction ................................. 74
3.8 WordNet concepts and top lexical ontologies ............ 77

II Research design and implementation .......................... 84

4 Ontology comparison ............................................ 85

4.1 Introduction .................................................... 85
4.2 Ontology comparison ........................................ 86
  4.2.1 Concept tree ........................................... 93
  4.2.2 Conceptual similarity measures ...................... 94
  4.2.3 Tree operations: deletion .............................. 95
  4.2.4 Tree operations: insertion ............................. 95
  4.2.5 Tree operations: re-labelling ......................... 97
  4.2.6 Tree operations: movement ........................... 97
4.3 Limitations of calculations ............................... 102
4.4 Comparison principles .................................... 102

5 Ontology mapping approach ................................. 104
  5.1 Introduction ............................................. 104
  5.2 Methodological approach ............................... 105
  5.3 Quality assurance ....................................... 109
  5.4 Meta-data documentation ............................... 110
  5.5 SUMO mapping confirmation ............................ 111
  5.6 Applying ontology comparison ......................... 115
  5.7 Methodological questions .............................. 118

III Contribution and conclusion ............................ 120

6 Results ...................................................... 121
  6.1 Introduction ............................................. 121
  6.2 Final word list .......................................... 122
  6.3 Qualitative comparison results ......................... 126
    6.3.1 Sense mapping with WordNet .................... 139
6.3.2 Mapping of BLR3 with Balkanet common synsets . . . . 141
6.3.3 Mapping of BLR3 with Global Base Concepts . . . . . . 142
6.3.4 Top Ontology comparison . . . . . . . . . . . . . . . . . 143
6.3.5 Upper Ontology comparison . . . . . . . . . . . . . . . . 150
6.4 Quantitative ontology comparison . . . . . . . . . . . . . 152

7 Conclusion and future work 155

7.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 155
7.2 Answering the research questions . . . . . . . . . . . . . . . 157
   7.2.1 Research sub-questions . . . . . . . . . . . . . . . . . 157
   7.2.2 Main research question . . . . . . . . . . . . . . . . . 158
7.3 Reflection . . . . . . . . . . . . . . . . . . . . . . . . . . . . 159
7.4 Recommendations . . . . . . . . . . . . . . . . . . . . . . . 162
   7.4.1 Policy and practice . . . . . . . . . . . . . . . . . . . . 162
   7.4.2 Evaluation . . . . . . . . . . . . . . . . . . . . . . . . 163
   7.4.3 Future research . . . . . . . . . . . . . . . . . . . . . 164
   7.4.4 Further development work . . . . . . . . . . . . . . . 164

References 166

IV Additional information 213

A Word and concept lists 214
   A.1 Original word list . . . . . . . . . . . . . . . . . . . . . . . 215
   A.2 Attested word list . . . . . . . . . . . . . . . . . . . . . . . 220
   A.3 Quality assured word list . . . . . . . . . . . . . . . . . . . 224
A.4 Variant BLR3 list ................................................. 228

B  Web Ontology Language results ......................... 231

B.1 Sample WordNet RDF results ............................. 232
   B.1.1 Nouns .................................................. 232
   B.1.2 Verbs .................................................. 234
   B.1.3 Adjectives .............................................. 237

B.2 Sample SUMO results ....................................... 240
   B.2.1 Nouns .................................................. 240
   B.2.2 Verbs .................................................. 241
   B.2.3 Adjectives .............................................. 247

C  Ontology comparison calculations ...................... 250

C.1 Final word list comparison values ........................ 250

C.2 Calculation details .......................................... 251

D  Abstract of publication .................................... 255

E  Bantu Base Concept subsumption in SUMO ............ 257
LIST OF FIGURES

1.1 Kinds of ontologies ............................................. 5
1.2 Structure of dissertation ................................. 25

2.1 The common, layered Semantic Web technology stack .... 29
2.2 Graph example of RDF ................................. 43
2.3 The relationship between SUMO and mid-level ontologies ... 60

3.1 Ontology learning layer cake ................................. 64
3.2 Bantu language zones in Sub-Saharan Africa ... 83

4.1 Hyponymy tree for the noun bee ........................ 88
4.2 Hyponymy tree for the noun sangoma .................... 89
4.3 Hyponymy tree for the verb roast ...................... 90
4.4 Hyponymy tree for the verb bite ..................... 91
4.5 Deleting a node ........................................... 94
4.6 Inserting a node .......................................... 96
4.7 Example of node insertion and movement .......... 97
4.8 Re-labelling a node ...................................... 98
4.9 Example of node re-labelling ............................ 98
LIST OF TABLES

1.1 Research questions .................................................. 18

6.1 BLR roots and meanings ........................................... 122
6.2 Sample BLR roots and meanings ................................. 128
6.3 BLR verb roots and meanings ................................... 129
6.4 BLR adjective roots and meanings .............................. 134
6.5 BLR noun stems and meanings .................................. 135
6.6 BLR3 to Kāngxī radical mapping ............................... 145
6.7 Bantu Concept mapping to Top Ontology qualia rôles .... 148

7.1 Research questions .................................................... 157

A.1 Original word list ..................................................... 215
A.2 Attested word list .................................................... 220
A.3 Quality assured word list ......................................... 225
A.4 BLR variants ............................................................ 229

C.1 Transformation cost and similarity index ....................... 251
C.2 Tree cost calculations ................................................ 252
LISTINGS

2.1 SUMO Bee class ................................................. 34
2.2 Bee Synset ................................................... 38
2.3 XML header encoding example ............................... 40
2.4 RDF TURTLE gloss example ................................. 41
2.5 WordNet Bee synset .......................................... 44
2.6 WordNet synset Class example ............................... 49
2.7 SUMO Bee Class ............................................... 53
3.1 WordNet synset relations ..................................... 79
B.1 The synset for Sangoma ....................................... 232
B.2 The synset for Entity .......................................... 232
B.3 The synset for Numida ........................................ 233
B.4 The synset for Bee ............................................ 234
B.5 The synset for Dance .......................................... 235
B.6 The synset for Carry .......................................... 236
B.7 The synset for Winnow ....................................... 237
B.8 The synset for Bad ............................................ 237
B.9 The synset for Many .......................................... 239
B.10 The Bee Class ........................................... 240
B.11 The Tongue Class ......................................... 241
B.12 The Weeping Class ......................................... 241
B.13 The Giving Class ........................................... 242
B.14 The Positive Integer Class ................................. 247
LIST OF ALGORITHMS

1. The transformation pre-processing phase ........................................ 117
2. The transformation cost computing phase ..................................... 119
Part I

Contextualisation
CHAPTER 1

Introduction

The principle of universality allows the Web to work no matter what hardware, software, network connection or language you use and to handle information of all types and qualities.

Long live the Web, Berners-Lee (2010, p. 82)

1.1 Background

The Semantic Web as touted by Tim Berners-Lee in various sources (Berners-Lee, 2000, 2005; Berners-Lee et al., 2001) was a vision to extend the World Wide Web to a new generation of technology to enhance the current architecture to make it more machine-readable, and not just human-readable, as the original Web based on the html standard (World Wide Web Consortium, 2001a,b, 2013). Human-readable information means traditional electronic documents on the Web which are intended for human use, whereas machine-readable documents means
data which has explicitly been prepared for machine access and use: part of a Semantic Web (Zou et al., 2005).

In order to achieve this machine-readability of data required in the Semantic Web there are stringent meta-data requirements. Meta-data is the additional information about the data which allows a computer to interpret that data. Computationally, the meaning of data, or meta-data, is represented using meta-data mark-up. Mark-up is a sequence of characters, called tags, hidden to humans, but visible to computers, which explicitly shows the logical structure and the meaning of the document data. There is a hierarchy of meaning or computational semantics established by the different levels of meta-data mark-up (Geroimenko, 2013).

Furthermore, in order to enable machine-readability, various explicit data standards have been defined for the Semantic Web. Whereas, on the original Web the notion of resources is almost always a reference to documents, images or other content, on the Semantic Web the notion of resources is broader. For example, resources could be concepts or the actual relationships between the concepts. The need to define and designate resources – concepts and their relationships – and their descriptions is fundamental to the Semantic Web. The Resource Description Framework (RDF) is the standard on the Semantic Web that allows this vision to be achieved (Cyganiak et al., 2014; Hayes and Patel-Schneider, 2014).

Informally, a grouping of resources – precise concepts and their explicit relationships – in a particular knowledge domain, utilizing a resource description framework, based on a vocabulary, is termed an ontology. As such, ontologies become key to the architecture of the Semantic Web (Berners-Lee, 1999), since
they are the mechanism that allows the interpretation of resources on a machine-readable level. Originally ontology was used in philosophy to refer to the study of the kinds of entities in the world and how they are related (Geroimenko, 2013). Ontologies allow the Semantic Web to not only cater for just machine-readability but also semantics and rules, thus enabling semantic computing or machine-understandability. For the purpose of this work the definition of an ontology is that of Studer et al. (1998) as quoted in (Guarino et al., 2009, p. 2)\(^1\):

> An ontology is a formal, explicit specification of a shared conceptualization (Guarino et al., 2009, p. 2).

An ontology is specified using an ontology definition language, such as the W3C Web Ontology Language (OWL) (Grau et al., 2008; Krötzsch et al., 2012; Patel-Schneider and Motik, 2012; Patel-Schneider et al., 2012a,b; Schneider, 2012).

In the Semantic Web there is a further distinction between different types of ontologies, specifically whether they are domain specific or not. The majority of ontologies are domain specific. Four ontology types are often distinguished: top-level (or now upper ontologies), domain ontologies, task ontologies and application ontologies as in Figure 1.1 (Guarino, 1997a, 1998)\(^2\). Upper ontologies, as viewed from a top-down perspective, are meant to provide semantic integration across the Semantic Web architecture by removing the knowledge domain focus and by being universally applicable. Upper ontologies provide definitions for general-purpose terms, and aim to be the foundation, as viewed from a bottom-
Upper ontologies are important to the Semantic Web for the following reasons. Firstly, new ontologies can be constructed by starting off using a previous base of common terminology, thus enabling the possibility of a bootstrapping, or borrowing approach to ontology construction. Secondly, the re-use of data is possible by doing a mapping from existing data to a common ontology which provides the data with an accurate context. Lastly, upper ontologies allow the semantic interoperability of existing ontologies.

*Semantic interoperability* implies the ability to practically integrate the usage
of different existing ontologies. This is often achieved by defining the semantic equivalence or subsumption of concepts across two different ontology definitions. This process of determining the semantic interoperability could be seen as a decision process. The first choice is concept alignment where there is equivalence between source and target concepts. The second choice is concept linkage where there is a subsumption relation between the source and target concepts. In this dissertation, the term map will be used for the process of determining semantic interoperability. Mapping will encompass both the processes of alignment and linkage of ontologies. Semantic interoperability can also be achieved through mid-level ontologies. A mid-level ontology is not domain specific but has far more detailed concepts than the general entities of an upper ontology (Fellbaum and Vossen, 2007; Soroa et al., 2010).

Upper ontologies, since they follow the ideals of the Semantic Web, should be open and universal (Berners-Lee, 2010). While open and universal are broad concepts, they have a particular interpretation in terms of ontologies used in this dissertation.

Firstly, open, in this context, is best described in Section 7, Clause 3 of the Internet Engineering Task Force’s RFC 2026 Standard as a standard that is internationally recognized through standards bodies and freely available in order to be practically implementable (Bradner, 1996). This openness is in

---

4. Note that what WordNet regards as synonymy relates to equivalence in ontologies, and hyponymy and instantiation relate to subsumption.

5. Various national and international standards bodies, such as ANSI, ISO, IEEE, and ITU-T, develop a variety of protocol and service specifications that are similar to Technical Specifications defined here. National and international groups also publish “implementors’ agreements” that are analogous to Applicability Statements, capturing a body of implementation-specific
contrast to proprietary or closed standards. Additionally, open standards also means “standards that can have any committed expert involved in the design, that have been widely reviewed as acceptable, which are available for free on the Web, and that are royalty-free (no need to pay) for developers and users” (Berners-Lee, 2010). So, in principle, for an upper ontology to be open it should be internationally recognized and freely available.

Secondly, universal inherits the definition from formal or symbolic logic, in turn inherited from philosophy, of the idea that something is true for every entity or every relevant entity (Ackrill, 1963). Concepts and entities used in an upper ontology should therefore be universal. Moreover, universality is also a goal of the Semantic Web:

The principle of universality allows the Web to work no matter what hardware, software, network connection or language you use and to handle information of all types and qualities. (Berners-Lee, 2010, p. 82)

Thus, the concepts in the upper ontology need to be universal and the upper ontology as a whole should follow the Semantic Web principal of universality. So, in principle, for an upper ontology to be universal it should conform to three criteria. It should allow for the consistent implementation of logically universal statements. Secondly, it should also function consistently across all hardware, software, network and language contexts. Lastly, it should consistently handle all data types and qualities.

detail concerned with the practical application of their standards. All of these are considered to be “open external standards” [my emphasis] for the purposes of the Internet Standards Process (Bradner, 1996).
Upper ontologies should, besides being open and universal, also be designed to adequately deal with natural language (Pease and Niles, 2002; Pease et al., 2002) and there should be a comprehensive mapping between lexical ontologies and upper ontologies (Niles and Pease, 2003). Moreover, multilingualism has been shown to be one of the key challenges to the development of ontologies (Benjamins et al., 2002; Fellbaum and Vossen, 2012).

For ontologies to adequately represent concepts that are realized in different languages, these concepts need to be shown to be shared and ideally to be realized (via a specific word or words) in many natural languages. In this context, the sharing of concepts implies an open concept definition, and the multilingual realization implies a universal concept definition. This multilingual challenge therefore re-enforces the requirement of an upper ontology being open and, in particular, universal.

While upper ontologies are, therefore, meant to be global open standards to define concepts on the Semantic Web, the following question now arises: are upper ontologies biased or restricted due to the natural language or languages that are used for either their specification or construction?

An ontology is a formal representation that has two key components: its ontology specification and the concepts and relations it defines as concept and relation specifications. Whereas the concepts and relations it defines, or the conceptualization it represents, should be natural language independent (Cimiano et al., 2011), the concept and relation specification has to be language dependent in order to be realized into text. The realization depends on whether and how the lexicon of the language used for the specification makes provision for the realization of the concepts and relations. In other words, the realization de-
pends on the lexicalization. Therefore, all ontologies, including upper ontologies, are inherently language dependent (Guarino, 1997b) since they depend on the realization of a concept in natural language. So, in order to achieve universality, upper ontologies should not be ontologically biased due to the choice of language used for their specification, lexicalization and realization. Investigating this, in the context of the African languages, is the main focus of this dissertation.

A lexical ontology is not a formal ontology, but is regarded as an informal ontology:

Whereas most ontologies are constructed for a given domain and contain relations between concepts, a lexical ontology is intended to provide structured information on words of a given language and their semantic relatedness; meaning is encoded by relating a given lexical item to others. Also, the main goal of a lexical ontology is not to store general encyclopædic or ontological knowledge, but to serve as common database, assembling lexical and semantic information (Wandmacher et al., 2007, p. 61).

It is therefore clear that the notion of a formal ontology cannot be applied to a lexical ontology since the relationships are linguistic and not conceptual relationships and there are often inherent inconsistencies in a lexical ontology (Oltramari et al., 2002), whereas consistency (by virtue of its formal nature) is foundational to formal ontologies.

Previous approaches in the literature to developing lexical ontologies mapped to upper ontologies, particularly in the non Indo-European language family, has further highlighted the Semantic Web multilingual challenge (Benjamins et al.,
The place to start examining this mapping from a lexical ontology to an upper ontology is to examine WordNet. Continuing on from the quotation above:

In the past years a number of projects have been presented that try to achieve this goal (of a lexical ontology), of which the most prominent one is the Princeton WordNet (Fellbaum, 1998). It represents domain independent, lexical-semantic knowledge in a network-like structure which makes taxonomic relationships explicit. However, it cannot be considered as an ontology in the formal sense, since the relations are based on linguistic evidence rather than on formal ontological principles, and it does not guarantee any kind of consistency ...

The main problem, however, remains data coverage. Even though WordNet and its cousins are considered as broad coverage resources, many NLP applications run into problems of data sparsity when relying on such resources only, which are all developed manually at great cost (Wandmacher et al., 2007, p. 61).

Therefore, although WordNet is termed a lexical ontology, it is not an ontology in the formal sense.

At present, mappings from lexical ontologies to upper ontologies assume the realization and lexicalization of the concepts. WordNet, a semantic network originally developed in US English at Princeton, which has its roots in cognitive psychology, has been used as a base hierarchy of concepts based on a lexical framework (Miller et al., 1993). Subsequent to similar WordNet implementations
in many other languages through projects such as Euro WordNet and BalkaNet, an interlingual and core concept alignment process has been developed. The interlingual and language core concept process has produced an alignment of WordNets with formal upper ontologies that supports the Semantic Web: in particular, the alignment with the Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2003; Reed and Pease, 2015). In other words, the interlingual and multilingual approach to aligning WordNets has become integrally linked with the alignment, or mapping, of Princeton WordNet to SUMO, an upper ontology (Bond et al., 2014).

Therefore, although WordNet is often called an ontology, ... its creators did not have in mind a philosophical construct. WordNet merely represents an attempt to map the English lexicon into a network by means of a few semantic relations. Many of these relations are implicit in standard lexicographic definitions (Prévot et al., 2010, p. 27).

In this dissertation I will refer to WordNet as a *lexical ontology* (Fellbaum and Vossen, 2012, p. 313–316).

Whereas EuroWordNet and BalkaNet connected WordNets regionally, Global WordNet is the project that connects various WordNets internationally. This linkage is achieved through the “standardization of an Inter-Lingual-Index (my emphasis) for inter-linking the WordNets of different languages, as a universal

---

6. The word *alignment* is common in the linguistic literature for WordNet and the word *mapping* is common in the Semantic Web computer science literature. From this point forwards, I will use the two words interchangeably depending on the context, either linguistic or computational.
index of meaning” (Vossen and Fellbaum, 2014b). Therefore, similarly in this dissertation, the WordNets in other languages and the interlingual index will also be referred to as lexical ontologies (Fellbaum and Vossen, 2012; Prévot et al., 2010). This term (lexical ontology) will conform with the literature but will not make any formal ontology claims about WordNet/s or Global WordNet.

In the context of the Global WordNet Project, concepts that are shared across WordNets and are foundational are termed Base Concepts and in EuroWordNet were termed Common Base Concepts. These 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 (Vossen and Fellbaum, 2014a).

Similarly Princeton WordNet has a list of Princeton Core Concepts, besides the Global WordNet Base Concepts and the EuroWordNet Common Base Concepts (Bosch et al., 2008; Griesel and Bosch, 2013, 2014; Lindén and Niemi, 2014; Vasiljevsa et al., 2012). Furthermore, in this dissertation I will be comparing these to Proto-Bantu reconstructed roots called Bantu Lexical Reconstructions 3 (Bastin et al., 2005; Bostoen and Bastin, 2016; Fleisch, 2008; Lesage, 2016), that are categorized, where common, across languages as main entries or reconstructed etymons. The distinction between language dependence and language independence is critical for this research. To ensure consistency in referring to these common/core/base/main concepts consistently, I will adopt a term: natural language core concepts. This term will feature in the research questions about these lexical ontology, language dependent concepts. My use of the term natural language core concepts will be used to distinguish them from the upper
ontology concepts, which claim to be language independent.

In many ways, as an upper ontology is used for the integration of concepts and relations within different ontologies, so the natural language core concepts in Global WordNet and its interlingual index are used as a semantic integration between different language WordNets. This semantic interoperability, also referred to as the integration or mapping, from WordNet, a language-dependent informal ontology, to SUMO, a language-independent formal upper ontology, is the basis for the research question. In particular, the question arises whether the ontological non-bias, with respect to the African languages, has been preserved by SUMO. The universality claim made by SUMO for the Princeton WordNet mapping that:

... we believe that we have refined the SUMO into an ontology that can be used to express anything that anyone would ever want to say in a formal context (Niles and Pease, 2003, p. 415).

becomes the claim from which the research question emerges.

In the linguistics of the African languages, there have been projects over the last 50 years to align the natural language core concepts of the Bantu languages. The Comparative Bantu On-line Dictionary (CBOLD) has strategically

7. In fact, the Global WordNet Project has three terms: Common Base Concepts (CBC): concepts that act as Base Concepts in at least two languages; Local Base Concepts (LBC): concepts that act as Base Concepts in only a single language; and Global Base Concepts (GBC): concepts that act as Base Concepts in all languages of the world (Vossen and Fellbaum, 2014b).

8. Due to historic sensitivity about the term Bantu in South Africa, I have used the term African languages which is used locally instead, except in contexts where Bantu has to be used to avoid possible confusion with other African language families.
unified most of the natural language core concepts that are lexicalized\(^9\) in the majority of Bantu languages. Further research in the last few years has now isolated core concepts that have current lexical alignment in the African languages (Maho, 2001). These common lexicalizations are referred to as Bantu Lexical Reconstructions, and the project is termed the BLR project.

Lexical reconstruction has been an important enterprise in Bantu historical linguistics since the earliest days of the discipline ... the Comparative Method has been and can be applied to reconstruct ancestral Bantu vocabulary via the intermediate step of phonological reconstruction and ... the study of sound change needs to be completed with diachronic semantics in order to correctly reconstruct both the form and the meaning of etymons (Bostoen and Bastin, 2016).

BLR went through different phases, each of which served as a database containing reconstructions. BLR3 is the current database. Although the African experience of unifying core concepts has had different driving factors, its outcomes have similarities to the core concept construction for EuroWordNet and BalkaNet.

1.2 Problem statement and research question

The context of this research are the premises that upper ontologies are largely universal and that lexical ontologies such as WordNet could be comprehensively

\(^9\) Style: For detail on spelling style used for -ise or -ize in this document please refer to Section 1.5
mapped to upper ontologies (Cimiano et al., 2011). Since the original mapping of WordNet to SUMO was done from one linguistic base only\textsuperscript{10}, the general research challenge, or the problem, is as follows: is this assumption that the universality of the upper ontology is preserved for the concepts realized in other languages, particularly in other language families, true (Pease et al., 2002)? Furthermore, does the language chosen to do the upper ontology specification and construction affect the concepts that are chosen for inclusion in the upper ontology?

The main research question emanating from this problem is: are core concepts\textsuperscript{11}, from a proposed natural language family, currently included in an existing, accepted upper ontology? Specifically, is every one of these core concepts equivalent to or subsumed by a concept in a defined upper ontology? These mappings from a computational perspective or alignments from a linguistic perspective are from fundamental, acknowledged core concepts in a natural language to concepts existing in upper ontologies.

The focus in this dissertation is on non-Indo-European language families. In order to answer this core research question, two further aspects are investigated:

- The state of the art of mappings from other, specifically non-Indo-European, language family concepts, to upper ontology concepts and

- mappings from the core concepts of an African language family, specifically the Bantu languages, to an upper ontology.

\textsuperscript{10} Note that the original Wordnet to SUMO mapping was done from Princeton WordNet (Reed and Pease, 2015).

\textsuperscript{11} see 1.1 on page 13 for definition
As already alluded to in Section 1.1, the inclusion of a natural language core concept in an upper ontology will be affirmed in this study in one of two ways:

1. The core concept is equivalent to an upper ontology concept, or

2. The core concept is subsumed by an upper ontology concept.

So, if a core concept is found not to occur in an upper ontology, then no equivalence could be established between the core concept and any concept in the upper ontology. Furthermore, there is a possibility that although no equivalence can be established, a subsumption relation can be established between the core concept and a broader concept in the upper ontology.

This means that in order to answer the research question, there is an obligation to identify, inspect and count those natural language core concepts that either are equivalent to concepts in the upper ontology; or are subsumed by broader concepts in the upper ontology; or have no mapping possibility at all. A research outcome where this count results in either a few mapping possibilities or many subsumption relations would mean that there would be little equivalence overall, and would provide a qualitatively, negative answer to the research question. Alternatively, a large proportion of equivalence relations would provide a qualitatively positive answer to the research question.

For the language family under investigation, that is, the Bantu languages, this is, as far as we know, the first study of its kind. This research is therefore novel and exploratory, and the results largely qualitative.

Research sub-questions that follow from this main research question are:

- What is the state of the art of the natural language core concept definition
in WordNets? This provides the linguistic background to the mapping process proposed.

• What is the state of the art of the upper ontology usage in the context of these natural language core concepts? This provides the computational background to the mapping process proposed.

• How do existing mappings of non-Indo-European language family core concepts to upper ontologies compare to that of Princeton WordNet? This provides the background to related work.

• What will a new structure of core concepts, from an African linguistic base, look like and how can it be compared to existing structures? Addressing this research sub-question is key in providing the practical results of a mapping process, which, once completed, contributes to answering the main research question. This sub-question is therefore intrinsically linked to the significance (or contribution) of this dissertation.

The accepted upper ontology used in this dissertation is the Suggested Upper Merged Ontology (SUMO), since this is the most common upper ontology to which WordNets are mapped. SUMO is also broadly representative (Mascardi et al., 2007) of other upper ontologies. Therefore similar results should apply to other upper ontologies:

SUMO and its domain ontologies ... form one of the largest formal public ontology(sic) in existence today. They are being used for research and applications in search, linguistics and reasoning (Mascardi et al., 2007, p. 5).
Number | Main Research Question
---|---
1 | Are the core concepts from a proposed natural language family currently included in an existing, accepted upper ontology?
1a | Is every one of these core concepts equivalent to or subsumed by a concept in a defined upper ontology?

<table>
<thead>
<tr>
<th>Number</th>
<th>Research Sub-Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>What is the state of the art of the natural language core concept definition in WordNets?</td>
</tr>
<tr>
<td>3</td>
<td>What is the state of the art of the upper ontology usage in the context of these natural language core concepts?</td>
</tr>
<tr>
<td>4</td>
<td>How do existing mappings of non-Indo-European language family core concepts to upper ontologies compare to that of Princeton WordNet?</td>
</tr>
<tr>
<td>5</td>
<td>What will a new structure of core concepts, from a novel African linguistic base, look like, and how can it be compared to existing structures?</td>
</tr>
</tbody>
</table>

Table 1.1: Research questions

1.3 Research objectives and methods

The first objective is to provide a contextualization for the research contribution through answering three related research sub-questions (research sub-questions numbered 2, 3 and 4 in table 1.1). It accomplishes this objective by examining the state of the art and the existing core concept mappings done prior to this study. The three related sub-questions are therefore answered by a literature investigation, the purpose of which is to provide an overview of scholarship in
a certain domain (Mouton, 2001). The method applied is non-empirical using secondary data. The objective of describing the state of the art is presented as a broader context in covering the topic of ontologies and particularly upper ontologies in more detail, within the Semantic Web architectural context (in Chapter 2). A more detailed domain context of the relation between ontologies and linguistic core concepts follows (in Chapter 3).

The final research sub-question (research question 5 in table 1.1) requires a new investigation. The second objective is to answer this question through the creation of a new concept mapping and to validate the mapping, by means of an accepted methodology (in Chapter 4). The approach used for this objective is a design and creation research strategy (Oates, 2005). The design allows for an initial design and an iterative refinement thereof. This conforms to the approach of research in design science (Hevner and Chatterjee, 2010). The method applied is empirical (Mouton, 2001) and uses secondary data and primary data. The secondary data are previous artifacts such as Princeton WordNet, BalkaNet and EuroWordNet and literature describing Chinese WordNet. The primary data are new mappings and verifications verified by a linguist. The methodology is to firstly use existing research to choose the core concepts, to secondly use existing mappings to qualify that they hold for Bantu languages and then lastly, to validate these mappings by language experts (discussed in Chapter 5).

There are two accepted approaches to developing new lexical ontologies, and the influence of these approaches on the upper ontologies have been well documented (Ordan and Wintner, 2007; Pala and Wong, 2001; Vossen, 2007b). A focus in answering research sub-question 5 will be to compare the two accepted approaches of creating lexical core concepts. This is done through the
construction of African Language core concepts. The results of the mapping and comparison will be highlighted and conclusions drawn in the context of the main research question (in Chapter 6).

The results of the research sub-questions and how they work together to answer the main research question is discussed in Chapter 7. Finally, a conclusion to this research and an exploration of future work is provided.

1.4 Delineation of the research

The research reported on in this dissertation broadly concerns the use of a specific upper ontology in the Semantic Web, seen from a computational linguistics perspective. The language focus is on Zone S of the Bantu language family, with the particular focus on two languages – a Sotho (Sesotho sa Leboa\textsuperscript{12}) and an Nguni (isiZulu\textsuperscript{13}) language – as representatives of the initial mapping, and one language (Sesotho sa Leboa) in the final mapping and mapping validation. The convention for the rest of the research will be to use the common English forms of the endonyms – Northern Sotho and Zulu respectively.

The upper ontology included in the scope of this dissertation is SUMO. Lexical ontologies used as reference criteria are the EuroWordNet Top Ontology, the Global WordNet Core Concepts and the BalkaNet Core Concepts. The ontology comparison was only done for nominal concepts as there is no theoretical framework available for doing this mapping with adjectives or verbs. There are many valid linguistic questions about the accuracy and usefulness of CBOLD and

\textsuperscript{12} Sesotho sa Leboa is an autonym and endonym of Northern Sotho.
\textsuperscript{13} isiZulu is an autonym and endonym of Zulu.
Proto-Bantu projects because reconstructions, as done through the BLR project, are not based on written, historical records as in the case of Middle-Eastern, Asian or European languages (Marten, 2006). These are not examined in this study. The assumption is made that, even if inaccurate or methodologically questionable (Fleisch, 2008), they are the result of work that has been meticulously undertaken by highly respected historical linguists over many decades. The reconstructions are accepted to be broadly representative of Bantu language concepts, and as such, are useful to answer this research question.

There are many philosophical questions about the definition and use of concept hierarchies and upper ontologies, from Aristotle (Ackrill, 1963), through Liebniz (Loemker, 1976) to modern debates. There are also many questions about whether ontology in philosophy itself is representative of African thought or not (Eze, 1998; Oruka, 1990). This dissertation deliberately avoids these philosophical debates. The assumption is made here that, since SUMO is a used, applied and accepted technology itself within Computer Science, and is significant in its use in the Semantic Web architecture, that there is intrinsic value in examining, as a research question, its claim of universality – computationally and not necessarily philosophically.

1.5 Style conventions

The typographical styles used by the two main sources used for this research are slightly different. Whereas Maho (2001) uses í, ñ, á and ñ, the BLR3 project uses í, ng, j and ny respectively. An attempt has been made to conform to BLR3 (Bastin et al., 2005). The en-GB-oed spelling style convention has been
followed for suffixes. Words of Greek and recently coined Latin etymology are terminated by -ize and those of French and classical Latin or Romance etymology by -ise. Hence, lexicalized, but generalised. When in doubt, -ize has been used.

1.6 Significance of the contribution

This study considers a novel approach of applying concept mapping to the African languages. Besides answering questions on universality of upper ontologies, this also provides new insights into the two current approaches to building WordNets and their usefulness - examining the research of Vossen (2007b), Pala and Wong (2001) and Ordan and Wintner (2007) from a new perspective. This should provide empirical validation of whether the correct decision was made by the African WordNet Project to start from first principles (Griesel and Bosch, 2013, 2014; Morapa et al., 2007; University of South Africa, 2008). This is achieved through constructing a prototype WordNet that focuses only on the concepts in this research. The term African language WordNets will be used as the more general term to include this prototype and any future African language WordNets not already incorporated in the African Language WordNet Project 14. The African WordNet Project, which started in 2007, is ongoing and the African WordNets for Zulu, Xhosa, Northern Sotho, Tswana and Venda have not yet been released. In this research, the question of how to make ontology comparisons is explored and the existing ontology mapping method of Xue et al. (2009) is used in a new application. This will also validate the usefulness of this ontology comparison

14. The term African WordNet Project will be exclusively used to refer to the established project detailed in Chapter 3.
approach. One publication has emanated from this research (Anderson et al., 2010), with the abstract reflected in Appendix D.

1.7 Structure of the dissertation

The structure of the dissertation is to first examine the Semantic Web Technologies and Standards as background context. This is foundational to understanding the notion behind upper ontologies (Chapter 2). Following that, there is an examination of how computational linguistics relates to Semantic Web standards. This is important in order to understand the rôle of linguistics in defining core concepts. In order to accomplish this, the research is then placed in the context of WordNet and the Bantu Lexical Reconstruction project (Chapter 3). Since mapping of linguistic concepts to upper ontologies is core to this study, the next portion of the research examines a methodology for ontology comparison (Chapter 4). The methodology for the research approach is then documented (Chapter 5), and finally, results are presented and conclusions drawn from the results (Chapter 6). Finally, the dissertation is concluded with a re-examination of the research questions and their answers (Chapter 7).

The appendices provide additional data that would detract from discussion in the text of the dissertation, but enhances the discussion related to the detail behind answering the research question. The supporting word lists are presented in Appendix A. This includes the original BLR3 Bantu word list used at the start of research, the attested word list of lexicalized concepts, and a final quality-assured word list that was used for final results and the conclusions. Sample usage of the results of this thesis in the form of the Semantic Web Resource Description
Framework (RDF), and the Semantic Web language used for upper ontology definition - the Web Ontology Language (OWL) - are shown in Appendix B. WordNet RDF and SUMO OWL examples are provided. Whereas the calculation method for ontology comparison is presented in Chapter 4, the detail resulting data behind the results of these calculations is shown in Appendix C.

A diagram summarising the structure of this dissertation is shown in Figure 1.2.
Figure 1.2: Structure of dissertation
CHAPTER 2

Semantic Web architecture

The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

The Semantic Web, Berners-Lee et al. (2001, p. 28)

The research question is whether a given natural language family’s core concepts are currently included in an existing, accepted upper ontology. In this chapter the Semantic Web technology architecture is examined to place ontologies and upper ontologies in context in the Semantic Web. In Section 2.1, the concept of the Semantic Web is introduced as a layered architecture. Section 2.2 examines the Semantic Web architecture by systemically examining the original requirement of the extension of the Web, the standards required to define the Semantic Web and the foundational layers of the Semantic Web.

In Section 2.3, the foundation is now extended to the significant new aspects of the Semantic Web architecture with a specific focus on formal ontologies and
how they relate to the Semantic Web. In Section 2.4, the top layers of the Semantic Web architecture are then briefly addressed for the sake of completeness. The goals of the Semantic Web architecture are summarized in Section 2.5.

2.1 Semantic Web layered architecture

The Semantic Web as touted by Tim Berners-Lee in various sources (Berners-Lee, 2000, 2005; Berners-Lee et al., 2001) was a dream to extend the Web to a new generation of technology to include more structure to make it more machine-readable, and not just human-readable as the original Web based on html (World Wide Web Consortium, 2001a,b, 2013)\(^1\). In order to achieve this Semantic Web, there are stringent meta-data requirements like the indexing of information and data, the adoption of meta-data definitions, standard taxonomies and ontologies, linkages between meta-data and the standards for machine readability. This would include

- the definition of services in a form that enables a computer to understand the functionalities that the services provide,

- the machine’s ability to ‘discover’ services, and

- the ability of automated agents to function ‘intelligently’ on the Web.

\(^1\) The use of a new generation of Web architectural standards is distinct from the Web 2.0 terminology (DiNucci, 1999) – Web applications with a focus on social media – and Web 3.0 – Web applications (Smart et al., 2007). Web 2.0 and Web 3.0 are referred to as part of the new generation of the Web because they make use of the Semantic Web architectural components.
All of these functions require the need to define and designate resources and their descriptions. This is done through a variety of World Wide Web Consortium (W3C) standards (World Wide Web Consortium, 2006).

A key to understanding the Semantic Web architecture is the computational understanding of meaning – how the meaning of data is represented computationally. Computationally, the meaning of data is represented using meta-data mark-up. A hierarchy of meaning is established by the different levels or layers of meta-data (Geroimenko, 2013). Linguistically, meaning is derived from a lexicon which specifies the meaning of words, combined with a set of semantic rules for establishing relations between words (Matthews, 2007). In computational linguistics, the lexicon and semantic rules are available computationally. Therefore meaning in the Semantic Web is usually established through either lexical ontologies or, alternatively, through a combination of lexicons and ontologies (McCrae et al., 2012, 2010, 2011; Protaziuk et al., 2012). These lexical ontologies, or combinations of lexicons and ontologies do not only provide the computational lexicon, but the ontology additionally provides the ability to specify semantic rules. Hyponymy, or the class/sub-class relationship, is an example of a word relationship that is a semantic rule. This understanding of semantics (or meaning) becomes foundational to the mapping already done from WordNet/s to SUMO and the proposed mapping in this dissertation from African language core concepts to SUMO (Pease, 2015). The mapping between the source and target concepts must be semantic alignment (equivalence) or semantic linkage (subsumption).

Berners-Lee produced 4 different versions of his architecture for the Semantic Web, all as a layered architecture (Gerber, 2006). For this research the focus is on
the common aspects of all 4 versions, already all depicted in the original version. A recent graphical representation of these layers is shown in Figure 2.1 (Nowack, 2009). In all of the versions its architecture is layered as follows (Berners-Lee, 1998; Gerber, 2006; Nowack, 2009; Some More Individual (Semantic Web Ontologies), 2011):

1. Layer 1 – The Web Platform: Unicode, URI/IRI and HTTP
2. Layer 2 – The Syntax: Namespaces, XML and XML Schema
3. Layer 3 – Knowledge Representation: RDF
4. Layer 4 – Semantics and Rules: RDF Schema and Other Ontology Vocabularies
5. Layer 5 – Logic/Logic Framework

6. Layer 6 – Proof

7. Layer 7 – Trust


2.2 The foundational layers

*Meta-data* provides structure to data by placing specific data in the context of a specific design or specification. Meta-data thus enables machine readability and is essential to the architectural foundations of the Web and the Semantic Web. It is also fundamental to the foundational layers, and a key aspect of all the layers, of the Semantic Web. *Mark-up* languages are used in all the layers of the Semantic Web and are therefore also similarly fundamental to the foundational layers. The layers of the Semantic Web provide ever more complexity to the meta-data as the layers progress to higher layers. For example, in the layers provided in Figure 2.1 above, XML provides a basis of meta-data extended by additional meta-data for XML Schema. Similarly RDF Schema extends the meta-data capabilities provided by RDF. A meta-data standard for referencing resources forms the first layer of the Semantic Web Architecture. Once the referencing architecture was defined, the main requirement was for one common format for the content of a resource, defined through a mark-up language constituting the second layer. The third and fourth layers provide the standards using this meta-data and mark-up to define ontologies. In short, Layer 1 powers the Web on the internet, Layer 2
addresses syntactic interoperability and Layers 3 and 4 lay the foundations for semantic interoperability.

2.2.1 Layer 1 – The Web platform

The first foundational layer, Layer 1, is defined as a combination of foundational standards to define the Web platform. These include an encoding standard (Unicode), a means of referencing resources (URI/IRI) and a transport protocol (HTTP). The Unicode Standard is the universal character encoding standard used for the representation of text for computer processing and provides a consistent way of encoding multilingual plain text (Davis et al., 2014a; Unicode Consortium, 2014a,b). Unicode is not just a new standard for the representation of text strings but specifically provides mechanisms to deal with strings that are natural language specific. For example, in many writing systems, a graphical unit is considered to be a single letter and may be represented in Unicode by a sequence of more than one coded character. A sequence of multiple coded characters that makes a single user-perceived character is termed a grapheme cluster (Davis, 2014). Major cases of this phenomenon include:

- Letters with applied diacritical or vowel marks (combining character sequences) as in ṭ in molló (“fire”) in Northern Sotho
- Language-dependent digraphs, such as ṓ in Northern Sotho, or ps in Tswana and Northern Sotho

Unicode is also used for string comparison that supports the multilingual Web. Comparison by binary code-point order (how a computer orders the script used)
rarely yields linguistically-correct results (how a declarative dictionary or lexicon
would order the script used).

No (dictionary or lexicon) user expects a sorting by code for characters which is
what the previous, non-language specific ASCII and EBCDIC standards,
specified.

\[ \text{E} < \text{S} < \text{Z} < \text{e} < \text{s} < \text{z} < \text{Š} < \text{ê} \]

Comparison is not simple for encoding systems because it is natural language
dependent. Typically accent differences become relevant only if there are no letter
differences and case differences become relevant only if there are no accent or let-
ter differences. To establish a framework for confronting these complexities, the
Unicode Collation Algorithm (UCA) (Davis et al., 2014b) specifies a comparison
for Unicode strings, now in terms of language specific sequencing, is termed col-
lation. Collation is the term used by Unicode for determining the sorting order of
strings of characters. It also provides for a default neutral ordering, for example:

\[ \text{e} < \text{ê} < \text{Š} < \text{z} < \text{Z} \]

Individual languages require tailoring of this foundation as in the Common
Local Data Repository (CLDR) (CLDR - Unicode Common Local Data Reposi-
tory Project, 2014). Tailorings establish equivalences among characters that are
used in language-sensitive searching and matching. For example, a tailoring
of Unicode for Northern Sotho would require that the character š occurs after
the letter s and the digraph ſš follows the digraph fs as in, amongst others, the

---

2. Also refer to §5.16 of The Unicode Standard (Davis et al., 2014a; Unicode Consortium,
2014a,b), UTN #9 (Davis, 2003) and UTS #10 (UCA) (Davis et al., 2014b) for further
information.
Therefore the Unicode standard is part of the answer to the multilingualism key challenge for the development of ontologies, which was highlighted previously. As already introduced, this multilingual challenge re-enforces the requirement of an upper ontology being open and, in particular, universal.

Everything on the Web can, and is, referenced as a resource. The resource can be any object of information, including a text document, video, picture, sound, page or a concept. This principle, that anything in the broadest, universal sense of anything, on the Web, should be identified uniquely by an opaque string of universal characters, is core to the universality of the Web (Berners-Lee, 2010; Leuf, 2005). References have standards defined to represent the reference, and these standards ensure a uniform representation of references. The mentioned resources can be uniquely identified using a uniform resource identifier URI, named using a uniform resource name (URN) and located using a uniform resource locator URL (Berners-Lee et al., 1998; Daigle et al., 2002). The standards ensure uniformity: in a URI the identification is uniform, in a URN the name is uniform and in a URL the linkage is uniform.

A named resource, or URN, by definition, also has an identifier, or URI. The standards proposal RFC 3986, proposed that, instead of separate standards for a URN and a URI, both the name and identifier are represented by a URI (Internet Engineering Task Force et al., 2009). In other words, prior to the year 2009 and RFC 3986, there was always a clear distinction between a URN and URI, but, subsequent to that standard, the URI standard is now also used to refer to both URIs and URNs.
The most common URIs in the original Web were those that referenced resources that could be located, or addressed and retrieved (Passin, 2004). A URL, is for example http://www.pansalb.org.za which represents the Web location of the Pan South African Language Board. A URI example, on the other hand, can refer to abstract resources, such as a scientific theory or the human concept of a bee or a guinea fowl, with a detailed description provided through RDF (Internet Engineering Task Force et al., 2009).

The aim of the Semantic Web is machine-readability, as opposed to just human-readability. Moreover, the higher layers, those above the Semantics and Rules Layer 4 of the Semantic Web (see Figure 2.1), also aim for machine-understandability or semantic computing. Therefore there is, additionally, besides the requirement for just resource referencing, a requirement to link the resources defined by URIs intelligently, in order to computationally reason about relationships.

URIs in Layer 1 provide the foundation for this intelligent linking which is accomplished in the higher layers. For example, in Listing 2.1, there are two URIs that provide information that is machine-readable, namely that the class bee is defined as a class in SUMO (line 11), and that its definition as bee is accessible via a URL at http://www.ontologyportal.org/SUMO.owl (line 17). Furthermore, it also provides information that an image of the concept bee can be found at the URL http://upload.wikimedia.org/wikipedia/commons/5/51-/Apis_mellifera_bi.jpg (line 14).

Listing 2.1: SUMO Bee class

```
1 @prefix : <http://www.ontologyportal.org/SUMO.owl#> .
2 @prefix wn: <http://www.ontologyportal.org/WordNet.owl#> .
3 @prefix owl: <http://www.w3.org/2002/07/owl#> .
```
In the architecture of the Semantic Web the focus is therefore concentrated on data, whereas it fell on documents in the original Web. One of the original motivations for the Semantic Web was to unlock the value of data in databases and the data in free text for machine interpretation and processing. Therefore, great potential for the reuse of this data is now possible, due to the fact that all languages use URIs as identifiers. This allows “things” defined in one natural or constructed language to refer to “things” defined in another natural or constructed language and preserving equivalence regardless of the defining language. The use of URIs allows a language to leverage the persistence, identity and equivalence in this uniform way (Geroimenko, 2013). This means that the concept of bee introduced above can be made a persistent concept that can be reused across different ontologies, different lexical mappings from different languages, yet still keep its unique identifier.

A graph (in computer science and mathematics) is a representation of objects
where some of the objects (nodes) can be linked (edges). The URI linkage can be represented by a graph - where the graph nodes are the resources and the linkages are edges in the graph. The importance of URIs\(^3\) to the Semantic Web Architecture is that URIs enable the first step in answering the Semantic Web question “When is a node in one graph the same node as a node in another graph?” (Allemang and Hendler, 2011). This comparison between nodes, which is conducted to answer the research question, will be further explored in Chapter 4.

In order to implement and use URIs practically, a small set of commands has been defined for the Web using standards for the predominant Web protocol - hypertext transfer protocol (HTTP) as a transport protocol to access these locatable resources (Internet Engineering Task Force, 2009). The HTTP protocol is specifically designed to use a small set of commands. These commands are universally understood by Web servers, clients (like browsers), intermediate components like caches and intelligent agents (Passin, 2004). With these commands, there is no question about what is being requested on the Web network and there is also, deliberately, no visibility into how the server fulfils the request on the network (Passin, 2004).

The protocol enables the linkage to the Semantic Web infrastructure. A URI merely refers to a resource through a reference, but that reference in the URI can be dereferenced. Dereferencing means using the information in the URI to locate its actual location on the Semantic Web infrastructure. The dereferencing succeeds if the protocol establishes an actual location on the Web, meaning that

\[^3\] The importance of URIs also applies to the importance of OWL 2 IRIs introduced subsequent to OWL 1 by OWL 2.
there is a URL for that URI. So, whereas the URI enables modelling on the Semantic Web, the URL enables participation in the Semantic Web infrastructure, through the use of the protocol (Allemang and Hendler, 2011).

For example, the HTTP protocol allows us to search for the concept bee on the Semantic Web Search engine - Falcons (http://ws.nju.edu.cn/falcons/conceptsearch/index.jsp) (Cheng et al., 2008). The results of the search can then be accessed via a client (in this case a browser), to show the ontological definition of bee.

2.2.2 Layer 2 – The syntax

The Web was originally designed with the principle that there would be many proprietary formats, and the hypertext protocol (HTTP) was designed as a negotiation connectivity mechanism, or transport, between client and server, as described above. In this original Web architecture, HTML was used as the dominant Web mark-up language. The limitations, proprietary nature and extensions of HTML gave rise to the requirements for a revision and development of a new standard that would be extensible without allowing proprietary changes. XML, in The Syntax Layer 2 of the architecture, is the outcome of this revision. XML is a language for data communication. XML is a plain text document containing both data and meta-data, but with no formatting information, unlike HTML\(^4\) (Bray et al., 2008; Geroimenko, 2013). XML is expressed with much of the same notation as HTML, but differs from the architecture of an HTML document. HTML contains data and formatting information, but lacks the extensible meta-data of

---

\(^4\) HTML in this context are the versions prior to version 5. HTML version 5 is based on elements of the XML specification (Berjon et al., 2014).
XML (Berners-Lee and Connolly, 1993; Geroimenko, 2013). For example, Listing 2.2 illustrates an XML representation of the meta-data for the concept bee. The comment for the concept bee is included between XML tags that start and end the comment shown by rdfs:comment, and the data describes a bee.

An example of the meta-data is the attribute xml:lang="en", which informs us that the enclosed data description is in English.

Listing 2.2: Bee Synset

```xml
<rdf:RDF
   xmlns="http://www.ontologyportal.org/WordNet.owl#"
   xmlns:rdf="http://www.w3.org/1999/02/rdf-syntax-ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:owl="http://www.w3.org/2002/07/owl#"
   rdf:about="WordNet">
  <owl:Ontology rdf:about="WordNet">
    <owl:Thing rdf:ID="WN30-102206656">
      ...<rdfs:comment xml:lang="en">any of numerous hairy-bodied insects including social and solitary species</rdfs:comment>
      ...
    </owl:Thing>
  </owl:Ontology>
</rdf:RDF>
```

Similarly, Listing 2.5 shows the representation of the word bee in OWL for WordNet in English –

```xml
<http://www.adampease.org/OP/WordNet.owl>
```

5. RDF and OWL, which are used in this example, are explained in more detail later. RDF is part of the knowledge representation layer in the Semantic Web architecture. OWL, as a Web ontology language will be introduced in the semantic and rules layer of the architecture. The focus in this section is on XML specifically. RDF can be represented in different formats, termed serializations. This RDF example is an XML serialization of RDF.

6. The comment is an RDF Schema comment illustrated by the namespace included at the start of the XML document.
A similar representation could be done for a different language, say Northern Sotho –

\[ < \text{http://www.adampease.org/OP/WordNet.owl} \]

\[ \#WN30Word - bee > rdfs:label "nose"@nso; \]

where nso is the standardized meta-data to indicate that the language being used is Northern Sotho, and therefore the data – nose\(^7\) – is the Northern Sotho word nose, or translated into English, the word bee (and not the English word nose for the appendage on a human face used for breathing and smelling).

So, although HTML was based on SGML, XML has proved to be a better design, primarily because of its extensibility and its inherent design for internationalisation. XML includes, in its definition, a reference to normative and non-normative references. The normative references are required and include key aspects required for human language use and for internationalisation. They are key to understanding the XML standard and its implementation (Geroimenko, 2013). The normative references are UCS, Unicode support, IETF language tag and IANA character set names. An Internationalisation Tag Set meta-data standard provides additions to tags within XML to enhance the internationalisation capabilities of XML (Savourel et al., 2013).

For example, the normative references are included in the first line of the WordNet XML representation in Listing 2.3. The IETF language tag was illustrated in Listing 2.2 above.

---

\(^7\) The scientific orthography for this is n̄ose but it could be written as nose in the practical orthography.
In addition, the optional non-normative references in the XML specification include UTF support, URI support, MIME types, SGML, country and language codes and HyTime. These non-normative references aid in understanding the design of XML and are used by attribute values in XML (Geroimenko, 2013). Since the normative and non-normative references enable XML, as a standard, to have better support for internationalisation and natural language than HTML, they therefore support the universality goal of the Semantic Web architecture.

2.3 The core layers

2.3.1 Layer 3 – Knowledge representation structure

The core new element of the Semantic Web, which makes it different to the original Web, is the Resource Description Framework (RDF)\(^8\). Phrased differently, RDF is the subsequent standard in the higher knowledge representation layer of the architecture that distinguishes the Semantic Web from the human-readable Web. All the other significant aspects of the Semantic Web, which further distinguish it from the original Web, are layered above it in Figure 2.1.

RDF, in Layer 3, is a data model which represents data, and therefore knowledge, as node-and-arc labeled directed graphs termed a triple. Each triple is

\(^8\) RDF was first defined through a series of standards as version 1 (Beckett, 2004; Guha and Brickley, 2004; Hayes, 2004; Klyne and Carroll, 2004) and then revised as version 1.1 (Cyganiak et al., 2014; Hayes and Patel-Schneider, 2014).
an assertion. In an RDF triple each subject and predicate is a URI and objects
are either a URI or a so-called literal. The initial use of RDF is to define basic
assertions and to encode logical facts or axioms. To identify the resources, RDF
uses URI from Layer 1.

An XML serialization example of RDF, termed RDF/XML (Gandon and
Schreiber, 2014), was shown in Listing 2.2. Another common serialization for-
matt for RDF is the Terse RDF Triple Language (TURTLE) (Carothers and
Prud’hommeaux, 2014). An example of the RDF triple for the English gloss
of the noun bee in TURTLE is shown in Listing 2.4 line 1.

Listing 2.4: RDF TURTLE gloss example

```
<http://wordnet-rdf.princeton.edu/wn31/bee-n@1-n> <http://wordnet-rdf.princeton.edu/ontology#gloss> "any_of_numerous_hairy-bodied_insects_including_social_and_solitary_species"@eng .


```

Listing 2.4 line 2 uses a reference to the RDF Schema standard to represent
the knowledge that the label for bee is identified in WordNet by the identifier
102209508-n. In the TURTLE standard a literal can be associated with a natural
language. Literals may be given a language suffix. Languages then are indicated
by appending the simple literal with @ and the IETF natural language tag. So in
the example the @eng represents that the literal is in International English.

That identifier 102209508-n allows us to obtain further knowledge on the
the concept bee, namely that it has a translation for a specific Chinese language
written standard form – Zhōngwén (中文), from the IETF language tag zho
represented by the Unicode character \u8702 and shown in Listing 2.4 line 3. A
graph of the examples above relating to bee is presented as Figure 2.2.
2.3.2 Layer 4 – Semantics and rules

In section 1.1 an ontology was defined as:

a formal, explicit specification of a shared conceptualization (Guarino et al., 2009, p. 2).

Whereas an ontology consists of a set of concepts, axioms, and relationships that describe a domain of interest (Colomb and Dampney, 2005), an upper ontology is limited to concepts that are “meta, generic, abstract and philosophical, and therefore are general enough to address (at a high level) a broad range of domain areas” (Niles and Pease, 2001). Concepts specific to given domains are not included; however, an upper ontology standard provides a structure and a set of general concepts upon which domain ontologies (e.g. lexical/linguistic, medical, financial, engineering, etc.) can be constructed (Niles and Pease, 2001).

Furthermore, a domain ontology establishes the things that a system (human or machine) can talk and reason about (Passin, 2004). It requires a classification system (also previously called a taxonomy) as its base. Classification can be by enumeration (the extension), definition (the intention), classes, sub-classes and instances, sets, names, identifiers or properties (Over et al., 2005).

Regardless of whether a domain ontology or upper ontology is defined, the Semantic Web architecture remains the same – the layer required to specify semantics and rules is one layer of architectural standards. The ontologies within this Semantics and Rules Layer 4 of the Semantic Web Architecture in Figure 2.1 provide more powerful schema concepts useful in inference, such as an inverse or transitive relationships. Furthermore, certain properties, when known, allow an agent navigating the Semantic Web to map different identifiers (URIs/IRIs)
Figure 2.2: Graph example of RDF
which, in fact, are have a bearing on the same concept, through ontologies (Over et al., 2005). For example, in SKOS, a Semantic Web vocabulary for representing thesauri, taxonomies and other structured, controlled vocabularies (Miles et al., 2005), the skos:altlabel and skos:narrower RDF properties are used to specify synonyms or near synonyms, or narrower concepts respectively (Isaac and Summers, 2009).

Ontologies can also be constructed by using RDF-based vocabularies of languages such as RDF Schema and OWL 2. One way to represent semantics and rules is through RDF Schema. RDF Schema (Resource Description Framework Schema), is a set of classes with certain properties using the RDF knowledge representation standard (Guha and Brickley, 2014). It provides basic elements for the description of ontologies (Rusher, 2003).

OWL 2 is another family of languages used to construct ontologies (Grau et al., 2008; Krötzsch et al., 2012; Patel-Schneider and Motik, 2012; Patel-Schneider et al., 2012a,b; Schneider, 2012; World Wide Web Consortium, 2006). OWL 2 is an extension of the Web Ontology Language OWL 1 family, where OWL 1 is a subset of OWL 2 – all ontologies created in OWL 1 can be read and understood by any application that understands the OWL 2 equivalent version (Yu, 2011). OWL 2 is therefore a more descriptive language family than OWL 1. OWL is used to refer to the complete family of web ontology languages, where the 2004 specifications relate to OWL 1, and the 2009 specifications refer to OWL 2. Within OWL 2, OWL 2 Full is the most descriptive language.

Listing 2.5 illustrates the representation of the word bee in OWL for WordNet.

Listing 2.5: WordNet Bee synset

```
prefix: <http://www.ontologyportal.org/WordNet.owl#>
```
@prefix owl: <http://www.w3.org/2002/07/owl#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix xml: <http://www.w3.org/XML/1998/namespace>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@base <http://www.adampease.org/OP/WordNet>.

<rdf:type owl:Ontology>


"Produced on date: Mon May 10 00:59:29 PDT 2010"@en.

@prefix owl: AnnotationProperty.
@prefix rdfs: AnnotationProperty.
@prefix rdfs: ObjectProperty.
@prefix word: AnnotationProperty.

@prefix hypernym: rdf:type owl:AnnotationProperty.
@prefix hyponym: rdf:type owl:AnnotationProperty.
@prefix member: rdf:type owl:AnnotationProperty.
@prefix member:owl:AnnotationProperty.
@prefix member:meronym: rdf:type owl:AnnotationProperty.
@prefix senseKey: rdf:type owl:AnnotationProperty.
@prefix synset: rdf:type owl:AnnotationProperty.
@prefix word: rdf:type owl:AnnotationProperty.

@prefix http://www.adampease.org/OP/WordNet#hypernym: rdf:type owl:ObjectProperty;
@prefix http://www.adampease.org/OP/WordNet#hyponym: rdf:type owl:ObjectProperty;
@prefix http://www.adampease.org/OP/WordNet#member: rdf:type owl:ObjectProperty;
@prefix http://www.adampease.org/OP/WordNet#senseKey: rdf:type owl:ObjectProperty;
@prefix http://www.adampease.org/OP/WordNet#synset: rdf:type owl:ObjectProperty.

rdfs:label "hyponym"@en;
rdfs:range <http://www.adampease.org/OP/WordNet.owl#Synset>;

rdfs:label "hypernym"@en;
rdfs:range <http://www.adampease.org/OP/WordNet.owl#Synset>;

rdfs:label "member-holonym"@en;
rdfs:range <http://www.adampease.org/OP/WordNet.owl#Synset>;

rdfs:label "sense_key"@en;
rdfs:comment "A relation between word and a particular sense of the word."@en;
rdfs:domain <http://www.adampease.org/OP/WordNet.owl#Word> ;

rdfs:domain <http://www.adampease.org/OP/WordNet.owl#synset> rdf:type owl:ObjectProperty ;
  rdfs:label "synset"@en ;
  rdfs:comment "A relation between a particular word and the synset in which it appears."@en ;
  rdfs:range <http://www.adampease.org/OP/WordNet.owl@Synset> ;

rdfs:domain <http://www.adampease.org/OP/WordNet.owl#word> rdf:type owl:ObjectProperty ;
  rdfs:label "word"@en ;
  rdfs:comment "A relation between a WordNet synset and a word which is a member of the synset."@en ;
  rdfs:range <http://www.adampease.org/OP/WordNet.owl@Synset> ;

<http://www.adampease.org/OP/WordNet.owl#WN30−102206856> rdf:type owl:NamedIndividual ,
  owl:NamedIndividual .
owl:Thing ;
rdfs:label "bee" ;
rdfs:comment "any of numerous hairy-bodied insects including social and solitary species"@en ;
:hypernym <http://www.adampease.org/OP/WordNet.owl#WN30−102206270> ;
:member−holonym <http://www.adampease.org/OP/WordNet.owl#WN30−102206624> ;
:hyponym <http://www.adampease.org/OP/WordNet.owl#WN30−102207179> ,
  <http://www.adampease.org/OP/WordNet.owl#WN30−102208260> ,
  <http://www.adampease.org/OP/WordNet.owl#WN30−102209354> ,
  <http://www.adampease.org/OP/WordNet.owl#WN30−102209664> ,
  <http://www.adampease.org/OP/WordNet.owl#WN30−102210427> ,
  <http://www.adampease.org/OP/WordNet.owl#WN30−102210921> ,
  <http://www.adampease.org/OP/WordNet.owl#WN30−102211444> ,
  <http://www.adampease.org/OP/WordNet.owl#WN30−102211627> ,
  <http://www.adampease.org/OP/WordNet.owl#WN30−102211896> ;
:word <http://www.adampease.org/OP/WordNet.owl#WN30Word−bee> ,
  owl:NamedIndividual ,
  owl:Thing ;
rdfs:label "bee"@en ;
rdfs:comment "The English word "bee"."@en ;
:senseKey <http://www.adampease.org/OP/WordNet.owl#WN30WordSense−bee>NN> ,
  <http://www.adampease.org/OP/WordNet.owl#WN30WordSense−bee>NN> ;
The first lines provide the prefixes. This allows URIs to be abbreviated by using TURTLE’s @prefix directive that allows declaring a short prefix name for a long prefix of repeated URIs later in the OWL example. Should a requirement be to understand the meaning of owl:NamedIndividual, the IRI reference for owl can be navigated by a machine to obtain the additional information.

The W3C OWL 2 recommendation explains that the Semantic Web is a vision for the future of the Web in which information is given explicit meaning, making it easier for machines to automatically process and integrate information available on the Web. The Semantic Web Architecture therefore is designed so that OWL 2 builds on both XML’s ability to define customised tagging schemes and RDF’s flexible approach to representing data (McGuinness et al., 2004). OWL 2 also makes the act of defining an ontology simpler. In the example note the use of the class Thing in Listing 2.5 line 96. There are two pre-defined classes in OWL 2, owl:Thing and owl:Nothing where these classes are the set of individuals and the empty set respectively.

Listing 2.5 lines 102-107 represent the concept of the first sense of the word bee in OWL for WordNet. Listing 2.5 lines 75-92 represents the noun synset for
bee in OWL 2 for WordNet with its English definition. Notice the linkage between concepts that an ontology language like OWL 2 provides in simple references to words, hypernyms, member-holonyms and hyponyms in lines 80-92. These are all pre-defined in the ontology as Object Properties. For example, in the ontology for WordNet the hyponym property is defined once as in Listing 2.5 lines 28-32.

Listing 2.6 represents the classes and their relationships for WordNet in OWL.

Listing 2.6: WordNet synset Class example

```xml
<pre>  : <http://www.w3.org/2006/03/wm/wm20/schema/> .
<pre>  owl: <http://www.w3.org/2002/07/owl#> .
<pre>  rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
<pre>  xsd: <http://www.w3.org/2001/XMLSchema#> .
<pre>  rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
<pre>  wn20schema: <http://www.w3.org/2006/03/wm/wm20/schema/> .
<pre>  base <http://www.w3.org/2006/03/wm/wm20/schema/> .
<pre>  rdf:type owl:Ontology .
<pre>  wn20schema:lexicalForm rdfs:comment "A datatype relation between Word and its lexical form." @en-us .
<pre>  wn20schema:gloss rdfs:comment "It specifies the gloss for a synset." @en-us .
<pre>  wn20schema:sense rdfs:type owl:ObjectProperty ;
  rdfs:comment "A relation added here to link words and word senses explicitly in the WordNet db, it is implicit in the synset record." @en-us ;
  rdfs:domain wn20schema:Word ;
  rdfs:range wn20schema:WordSense ;
  owl:inverseOf wn20schema:word .
<pre>  wn20schema:Word rdf:type owl:Class ;
  rdfs:subClassOf { rdfs:type owl:Restriction ;
  owl:onProperty wn20schema:sense ;
  owl:someValuesFrom wn20schema:WordSense } ;
  owl:disjointWith wn20schema:WordSense ;
<pre>  wn20schema:word rdf:type owl:ObjectProperty ;
  rdfs:comment "A relation added here to link word senses and word senses explicitly in the WordNet db, it is implicit in the synset record." @en-us ;
  rdfs:range wn20schema:Word ;
  rdfs:domain wn20schema:WordSense .
</pre>
```
wn20schema:WordSense rdf:type owl:Class ;
    rdfs:subClassOf [ rdf:type owl:Restriction ;
        owl:onProperty wn20schema:word ;
        owl:allValuesFrom wn20schema:Word
    ] ,
    [ rdf:type owl:Restriction ;
        owl:onProperty wn20schema:inSynset ;
        owl:cardinality "1"^^xsd:nonNegativeInteger
    ] ,
    [ rdf:type owl:Restriction ;
        owl:onProperty wn20schema:word ;
        owl:someValuesFrom wn20schema:Word
    ] ;
    rdfs:comment """A meaning of a word in WordNet. Each sense of a word is in exactly one synset."
""".

wn20schema:Synset rdf:type owl:Class .
    rdfs:subClassOf owl:Thing .
    rdfs:subClassOf [ rdf:type owl:Restriction ;
        owl:onProperty wn20schema:containsWordSense ;
        owl:someValuesFrom wn20schema:WordSense
    ] ;
    rdfs:comment """A synonym set; a set of words that are interchangeable in some context."
""" .

wn20schema:inSynset rdf:type owl:ObjectProperty ;
    rdfs:comment """A relation added here to link word senses and synsets explicitly (in the WordNet DB, it is implicit in the sense tag record)****
""" .
    rdfs:range wn20schema:Synset ;
    rdfs:domain wn20schema:WordSense .

wn20schema:classifies rdf:type owl:ObjectProperty ;
    rdfs:domain wn20schema:NounSynset ;
    rdfs:range wn20schema:Synset .

wn20schema:NounWordSense rdf:type owl:Class ;
    rdfs:subClassOf wn20schema:WordSense ;
    [ rdf:type owl:Restriction ;
        owl:onProperty wn20schema:inSynset ;
        owl:allValuesFrom wn20schema:NounSynset
    ] ,
    [ rdf:type owl:Restriction ;
        owl:onProperty wn20schema:inSynset ;
        owl:someValuesFrom wn20schema:NounSynset
    ] ;
    rdfs:comment """A synonym set; a set of words that are interchangeable in some context."
""" .
Listing 2.6 lines 67-81 represent the noun word sense class for WordNet in OWL 2. Listing 2.6 lines 63-65 represent the relationship between a NounSynset and a synset for WordNet in OWL 2, defining that the first synset has been classified as a member of the class represented by the second synset. Note that Noun synset is therefore a class belonging to all WordNet synsets.

Listing 2.6 also represents the properties for WordNet in OWL. Line 12 represents the lexical form property of a word as a data type for WordNet in OWL. Line 13 represents the gloss property of a synset as a data type for WordNet in OWL.

The Web platform, Layer 1, originally used Unicode as a standard for all data, except the URIs themselves, formed from strings using a subset of ASCII. URIs were used in OWL 1 to identify classes, ontologies, and other ontology elements. This non-use of Unicode for URI was inconsistent and limiting, particularly with respect to the multilingual challenge of the Semantic Web. Therefore OWL 2 introduced and uses Internationalized Resource Identifiers (IRIs) in RFC3987 for identifying ontologies and their elements instead (Wallace and Golbreich, 2012, §2.6.3).
Protége is an open-source platform with a suite of tools to construct domain models and knowledge-based applications with ontologies (Noy et al., 2001). Protége implements a “rich set of knowledge-modelling structures and actions that support the creation, visualisation and manipulation of ontologies in various representation formats” (Noy et al., 2006). Protége can be customised to provide domain support for creating knowledge models and can be for entering ontological data (Noy et al., 2006, 2001).

The Protége OWL editor enables one to build ontologies for the Semantic Web in OWL 2. According to the documentation, an OWL 2 ontology designed in Protége, could include descriptions of classes, properties and their instances. Protége can then use the OWL 2 formal semantics for inference – these are the facts not literally present in the ontology, but entailed by the semantics. These entailments may be based on a single document or multiple, distributed documents that have been combined using the defined OWL mechanisms in Protége (Noy et al., 2006, 2001). Protége was used in this research to access and navigate the upper ontology SUMO and the WordNet representation in OWL (van Assem et al., 2006).

SUMO was briefly introduced as the upper ontology used in this research in section 1.1. SUMO is an open source formal ontology and consists of approximately 1 000 terms and 4 000 axioms. First described by Niles and Pease, the goal was to develop a standard upper ontology that will promote data interoperability, information search and retrieval, automated inferencing, and natural language processing (Niles and Pease, 2001). The SUMO has subsequently been translated into various other representation formats, but the development language was a variant of KIF (a language supporting the first-order predicate calculus). It
covers areas of knowledge such as temporal and spatial representation, units and measures, processes, events, actions, and obligations. SUMO has been “mapped by hand (Niles and Pease, 2003) to the entire WordNet lexicon of approximately 100 000 noun, verb, adjective and adverb word senses, which not only acts as a check on coverage and completeness, but also provides a basis for application to natural language understanding” (Reed and Pease, 2015).

Listing 2.7 represents an example of WordNet represented in SUMO in entirety. The example chosen is one of an equivalence relation (see lines 18 and 27)⁹.

Listing 2.7: SUMO Bee Class

```
@prefix : <http://www.ontologyportal.org/SUMO.owl#> .
@prefix wn: <http://www.ontologyportal.org/WordNet.owl#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf-schema: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdfs:comment: <http://www.w3.org/2000/01/rdf-schema#comment> .
@prefix rdfs:label: <http://www.w3.org/2000/01/rdf-schema#label> .
@prefix rdfs:subClassOf: <http://www.w3.org/2000/01/rdf-schema#subClassOf> .
@prefix xml:base: <http://www.adampease.org/OP/SUMO#> .
@prefix xml:externalImage: <http://www.adampease.org/OP/SUMO#externalImage> .
@prefix xml:mds: <http://www.adampease.org/OP/SUMO#mds> .
@prefix xml:profile: <http://www.adampease.org/OP/SUMO#profile> .
@prefix xml:representation: <http://www.adampease.org/OP/SUMO#representation> .
@prefix xml:sections: <http://www.adampease.org/OP/SUMO#sections> .
@prefix xml:terms: <http://www.adampease.org/OP/SUMO#terms> .
@prefix xml:version: <http://www.adampease.org/OP/SUMO#version> .
@prefix xml:xmlns: <http://www.adampease.org/OP/SUMO#xmlns> .
@prefix xml:xmlns:xsi: <http://www.adampease.org/OP/SUMO#xmlns:xsi> .
@prefix xml:xs: <http://www.w3.org/2001/XMLSchema#> .
@prefix xml:xs:complexType: <http://www.w3.org/2001/XMLSchema#complexType> .
@prefix xml:xs:element: <http://www.w3.org/2001/XMLSchema#element> .
@prefix xml:xs:extension: <http://www.w3.org/2001/XMLSchema#extension> .
@prefix xml:xs:group: <http://www.w3.org/2001/XMLSchema#group> .
@prefix xml:xs:import: <http://www.w3.org/2001/XMLSchema#import> .
@prefix xml:xs:list: <http://www.w3.org/2001/XMLSchema#list> .
@prefix xml:xs:lt: <http://www.w3.org/2001/XMLSchema#lt> .
@prefix xml:xs:maxExclusive: <http://www.w3.org/2001/XMLSchema#maxExclusive> .
@prefix xml:xs:maxInclusive: <http://www.w3.org/2001/XMLSchema#maxInclusive> .
@prefix xml:xs:maxLength: <http://www.w3.org/2001/XMLSchema#maxLength> .
@prefix xml:xs:minExclusive: <http://www.w3.org/2001/XMLSchema#minExclusive> .
@prefix xml:xs:minInclusive: <http://www.w3.org/2001/XMLSchema#minInclusive> .
@prefix xml:xs:minLength: <http://www.w3.org/2001/XMLSchema#minLength> .
@prefix xml:xs:pattern: <http://www.w3.org/2001/XMLSchema#pattern> .
@prefix xml:xs:positiveInteger: <http://www.w3.org/2001/XMLSchema#positiveInteger> .
@prefix xml:xs:restriction: <http://www.w3.org/2001/XMLSchema#restriction> .
@prefix xml:xs:simpleType: <http://www.w3.org/2001/XMLSchema#simpleType> .
@prefix xml:xs:sequence: <http://www.w3.org/2001/XMLSchema#sequence> .
@prefix xml:xs:union: <http://www.w3.org/2001/XMLSchema#union> .
@prefix xml:xs:unionMember: <http://www.w3.org/2001/XMLSchema#unionMember> .
@prefix xml:xs:whiteSpace: <http://www.w3.org/2001/XMLSchema#WhiteSpace> .

<http://www.adampease.org/OP/SUMO#Bee rdf:type owl:Class ;
    rdfs:label: "bee"@en;
    rdfs:subClassOf <http://www.adampease.org/OP/SUMO#Insect> ;
    :externalImage "http://www.adampease.org/Articulate/

9. For the complete context of this example please refer to Appendix B.1.1.4.
Recall that an upper ontology is meant to bridge ontologies, so an example in the listing above is the reference of the upper ontology concept to the Mid-Level-Ontology MILO through the axiom `rdfs:resource="sumo:#axiom976385803Mid-level-ontology.kif"`. MILO is an ontology that was developed as a bridge between the abstract content of the SUMO and the rich detail of the various domain ontologies (Niles and Pease, 2001). In the MILO ontology a Bee is simply defined as a subclass of Insect\(^{10}\). As a mid-level ontology, MILO provides more information for logical interpretation than the upper ontology, including as an example that Honey is an Animal Substance produced by a Bee through a Physiological Process. This additional information is provided as logic rules through a version of the Knowledge Interchange Format (KIF), used to define SUMO and MILO (Pease, 2004). It is this mapping from upper ontologies to domain ontologies that provide the machine-readable inference capabilities of the Semantic Web architecture.

\(^{10}\) To see further details on MILO please refer to http://ontolog.cim3.net/file/resource/ontology/MILO/Mid-level-ontology.txt
The example in Listing 2.5 also illustrates the use, in an upper ontology, of equivalence to WordNet concepts/synsets (line 98) and subsumption of WordNet concepts/synsets (lines 101-103) as constructs. As discussed in Chapter 1 this formal definition of semantic interoperability is key to mapping and comparing ontologies. This is particularly important in the context of the African languages for the purpose of answering the research question and will be elaborated further in Chapter 4.

The mapping of upper ontologies to mid-level ontologies for further definition is very important in the context of natural language. Previous research has shown that there is sparsity of concepts in the vocabularies of natural language within upper ontologies. There is an argument that all natural languages should be able to inform the upper levels of an ontology since one would assume that natural languages have an “essential agreement about how the world is categorized, simply because the distinctions seem to be so fundamental and so basic to our biologically based, and therefore presumably universal, cognitive processes and perception of the world” (Guarino et al., 2009, p. 279). However, research shows that natural languages concentrate the richest and most commonly used parts of their vocabulary in the middle of the lexical hierarchy in a lexical ontology, which would be the focus of a mid-level ontology rather than an upper ontology (Guarino et al., 2009; Murphy and Lassaline, 1997). Therefore it is the mid-level ontologies, such as MILO, that “maximize both informativeness and distinctiveness” (Guarino et al., 2009, p. 279). Figure 2.3 illustrates how the general concepts in SUMO classes are mapped through MILO sub-classes to domain ontology classes (OntologyPortal, 2014).

It has been shown that one cannot build a good upper ontology merely by
looking at the relevant vocabulary of one, or even several, natural languages. Furthermore, there are extensive criticisms of the use of the top level of WordNet as an upper ontology. The upper levels of a lexical ontology are shown to produce a poor upper ontology (Gangemi et al., 2001; Guarino et al., 2009). However, SUMO is defined as an upper ontology. Subsequently a mapping has been done from WordNet to SUMO in order to compare the results and the representation of the top level concepts of WordNet in SUMO. This mapping does serve to confirm SUMO as a good upper ontology (Pease, 2005). Moreover, the comprehensive mapping of the upper ontology to the mid-level ontologies addresses the criticism introduced above.

It has also been shown that one can start with an upper ontology to produce a new lexical ontology, in a different natural language, and that this is beneficial. One of the significant aspects of SUMO, in this research, is that SUMO has been used in the construction of a WordNet (Arabic WordNet in particular (Pease, 2005)) and there is an existing mapping from Princeton WordNet to SUMO (Niles and Pease, 2003). A last, significant aspect of SUMO for this research is that it is the upper ontology, defined in OWL, that is used to answer the main research question.

2.4 The top layers of the Semantic Web architecture

This research focuses particularly on Layer 4 – ontologies representing the semantics and rules – and the layers below in Figure 2.1 that form the foundations
of the ontology implementation for the Semantic Web architecture. The higher layers are not directly pertinent to the research questions. However, for completion of the description of the Semantic Web architecture, these layers will be briefly mentioned in their context.

2.4.1 Layers 5, 6 and 7 – Logic, proof and trust

Layer 5 represents Logic and Logic Frameworks and these are important to enable inference on the Semantic Web from the knowledge presented in ontologies (Gerber, 2006, p. 112). A common logical framework is Description Logics (Grau et al., 2008, p. 335). Description Logics is a family of formal knowledge representation languages that models concepts, roles and individuals, and their relationships. The OWL 2 language family provides three increasingly expressive sub-languages as OWL species: OWL Lite, OWL DL and OWL Full. OWL DL is a species in the OWL 2 family to support Description Logics (McGuinness et al., 2004; Welty and McGuinness, 2004)\textsuperscript{11}. Logical reasoning and proof can be utilised to determine whether the data and data sets are consistent and correct. Logic can also be used to infer conclusions that are not explicitly stated, but are required, or consistent with, existing and known data sets on the Web (World Wide Web Consortium, 2006).

At the very top of the Semantic Web architecture is trust. The architecture allows that, once reasoning and proof are possible, one can determine a distributed version of trust based on the knowledge beneath (World Wide Web Consortium, 2006). Layer 6 (Proof) and Layer 7 (Trust) (together with the ver-

\textsuperscript{11} Refer to the references provided for further information.
tical layers) in Figure 2.1 respectively provide validity, trust levels and security (including identity) to the foundational layers. Layer 6 and 7 are considered to be unattainable at present for the Semantic Web (Patel-Schneider and Fensel, 2002).

2.5 Goals of the Semantic Web architecture

The Web, and by implication the Semantic Web, has been designed with specific architectural goals: to be scalable and open. It is also by design incomplete and inconsistent. Prior to looking at the goals of the Semantic Web architecture, a brief summary of the inherited architectural goals of the original Web is provided.

The original Web has been designed with two main architectural design goals in mind. One is that it is intentionally distributed and de-centralised. Secondly, each transaction on the Web contains all the information necessary to fulfil a request. These jointly allow the Web to grow and scale to any size (Passin, 2004). The Web is open. This means that Web sites and all resources available (URIs) can be added freely and without any central control. The assignment of domain names does, however, require central authority but domain names do not restrict the creation of Web servers and the information the servers provide. These highly scalable and open architectural goals allow the Web to grow. A Falcons\textsuperscript{12} search will return different but probably more results on each subsequent search for the concept bee as the Semantic Web grows (Ding et al., 2004, 2005).

The additional goals for the Semantic Web are interoperability and creating an evolvable or extensible technology (Passin, 2004). The Web is incomplete.\textsuperscript{12} http://ws.nju.edu.cn/falcons/conceptsearch/index.jsp
This means there is no guarantee that every URI will work, or that all possible information is available (Passin, 2004). The Web can be inconsistent. Information on the Web will never be fully consistent as different resources and statements by the sources providing the information can be conflicting. Software, and particularly the portions of the Semantic Web dealing with logic and reasoning, makes provision for change, potential inconsistency and incompleteness (d’Amato et al., 2012; Ma et al., 2014; Maarala et al., 2014; Passin, 2004).

Ontologies allow us to make logical deductions from information on the Web, even if some of the information is inconsistent or incomplete by design. In an examination of the concept bee the Semantic Web allows for retrieval of the logical fact that bee is a sub-class of insect, which would allow some logical deductions, even if the actual full definition of bee was not available on the Semantic Web.

This concludes the description of the Semantic Web Architecture. To provide more direct context towards answering the main research question, the following chapter examines lexical ontologies in particular.
Figure 2.3: The relationship between SUMO and mid-level ontologies
CHAPTER 3

Lexical core concepts and lexical ontologies

Words are ... battered relics of past ages often containing within them indelible records capable of intelligent interpretation.

John Herschel (Desmond and Moore, 1991, p. 215)

3.1 Introduction

The broader context of the Semantic Web Architecture, its foundational architecture, was introduced in Chapter 2 to place upper ontologies in context. The research question is whether a given natural language family’s core concepts are currently included in an existing, accepted upper ontology. This chapter examines natural language family core concepts by describing the more detailed context of lexical ontologies designed for natural language. Specifically the first sections,
from Section 3.2 to Section 3.5, are about lexical ontologies, their significance in natural language processing and computational linguistics, and their use as a foundation to defining natural language core concepts. The subsequent sections of this chapter, from Section 3.6 to Section 3.7, address previous research on African language core concepts and the establishment of the African WordNet Project. The last section, Section 3.8, considers the relationship between lexical ontologies and work done in specific language families in the context of SUMO. This is accomplished by considering the existing approaches to determining core concepts linguistically, and then mapping these to existing upper ontologies.

### 3.2 Semantic concepts in linguistics

Semantics in linguistics is the study of meaning, particularly the relationship between the morphemes that constitute words and their meaning. The meaning of a lexical item as distinguished from other meanings, such as in a dictionary, is called a “sense” (Matthews, 2007). Sense relations refer to the relation between lexical items or senses. **Antonymy** refers to the sense relation between lexical units that have opposite meanings. For example, *long* has, as an adjective, an opposite meaning to *short*. **Hyponymy** is a sense relation where the meaning of the first lexical unit is included in that of the second in a more general way. For example, *guinea fowl* is a hyponym of *fowl* and *bee* is a hyponym of *insect*. **Synonymy** refers to sense relations between lexical units where the meaning is similar or the same. For example *Ixodida* and *tick* mean the same thing and are therefore considered synonyms. Typically replacing a lexical unit with its synonymous counterpart will not change the logical truth condition of a sentence,
and hence change logical facts in the context of the Semantic Web. Lastly, meronymy refers to part-whole sense relations. For example, eye and tongue are different parts of a head.

### 3.3 Lexical ontologies

Lexical ontologies have been developed for reasons other than the Semantic Web, but are finding extensive application within the Semantic Web, particularly for upper ontology definition and confirmation.

The importance of lexical ontologies for ontology development has been highlighted as part of what now is referred to as the “ontology learning layer cake” as illustrated in Figure 3.1 (Buitelaar et al., 2005, p. 2).

The choice of concepts at level 3 of the cake in a lexical ontology is based on linguistic criteria instead of pure logical criteria (Farrar, 2003). His example is that, whereas we would categorise an animal in a formal ontology to include zebras, newts, and cows, in a lexical (he terms it a “linguistic”) ontology, an animal might not include certain individuals that we objectively know are animals. Examples would be ‘holy’ animals, ‘unclean’ animals, or animals that are marked linguistically. Animals that fall into different proto-Bantu noun classes might

---

1. *Meronymy* is the part-whole semantic relation used in linguistics and lexical ontologies. The related part-whole conceptual data structure in computer science and formal ontologies is termed *meronomy*, which like taxonomy refers in ontologies to a complex data structure built on the hyponymy lexical relation (World Wide Web Consortium, 2003).
2. For examples of how these antonymy, hyponymy, synonymy and meronymy relations are formalized in ontologies refer to Listing 3.1.
3. Int is *Intension*, Ext is *Extension* and Lex is *Lexical Realisation.*
∀x, y(sufferFrom(x, y) → ill(x))

cur(dom:DOCTOR, range:DISEASE)
is_a(DOCTOR, PERSON)

DISEASE:=⟨Int, Ext, Lex⟩

{disease, illness, bolwetsi}
disease, illness, hospital

Figure 3.1: Ontology learning layer cake

be categorised differently in a lexical ontology. For example, nocturnal animals and animals associated with spirits generally fall into a different Bantu noun class to most other animals and humans themselves belong in their own class. Putting a different root into another class can change lexical sense, i.e. -lwane is the root for animal in Classes 7 and 8 in Zulu, but is a derogatory term for a human in the human Classes 1 and 2. So even though in a scientific ontology a human might be an animal, the correct hyponym that categorises it as an animal would need to be chosen in a lexical ontology represented in Zulu. Therefore, although lexical ontologies do not claim to represent the object or cognitive world like formal ontologies, lexical ontologies are quite useful as processing resources (Farrar, 2003); for text understanding (Bateman, 1990; Farrar, 1991; Henschel and Bateman, 1994); for machine translation (Hovy and Nirenburg, 1992); and for common-sense reasoning (Dahlgren et al., 1989; Hobbs et al., 1987; Nirenburg et al., 1987).

Based on the formal definition of ontology in the context of the Semantic Web (provided in Chapters 1 and 2) the meaning of lexical ontology is often
less formal than the Semantic Web ontology definition and serves a different purpose. A lexical ontology exactly reflects the relationships between lexicalized words and expressions in a language (Vossen, 2007a, p. 9). In order to align lexical ontologies closer to formal ontologies, or to align lexical ontologies across different languages, there is a need to either:

- ignore levels that are lexicalized but not relevant for the purpose of an ontology or

- introduce artificial levels (for hyponymy or meronymy) that are not lexicalized in that specific lexical ontology (Vossen, 2007a,b).

So for example spoon in a formal ontology might be a “hand tool” where the concept of a tool used by the hand is not lexicalized in that language but introduced as an artificial level (Vossen, 2007a, p. 8). Similarly the grouping of spoons under tableware or silverware might be relevant in a lexical ontology based on Germanic languages such as English that have the suffix -ware, but could be ignored in a formal ontology where the properties of a spoon could be inferred: container; artefact; hand tool; object; made of metal or plastic; for eating, pouring or cooking (Vossen, 2007a, p. 8). It has been shown that a great deal of work would be required to adapt a lexical ontology such as WordNet into a formal ontology (Oltramari et al., 2002).

Even though the concepts and constructs in a lexical ontology are less formal than a Semantic Web ontology, a lexical ontology can be modelled and constructed using Semantic Web languages, frameworks and models. Just as mark-up was foundational to the layers of the Semantic Web, so lexical mark-up is foundational to lexical ontologies. Lexical Mark-up Framework (LMF), or ISO
24613:2008, is the international standard for lexical mark-up used in the founda-
tional layers of the Semantic Web architecture (Francopoulo et al., 2007, 2006;
International Organisation for Standardization, 2008). WordNetLMF is an LMF
format for WordNet (Soria et al., 2009). It is the standard used by the EU KY-
OTO Project: Knowledge Yielding Ontologies for Transition-based Organisation.
The goal of KYOTO is to make

knowledge shareable between communities of people, culture, lan-
guages and computers, by assigning meaning to text and giving text
to meaning (European Union, 2007).

Lemon is a formal model for defining lexical ontologies and is also used
for the integration of lexical ontologies through RDF within the Semantic
Web architecture (Eckle-Kohler et al., 2014; McCrae et al., 2012, 2010, 2011;
Protaziuk et al., 2012). It is based on LMF but extends the LMF formal model
to provide native integration of lexica with domain ontologies (Buitelaar et al.,
2013; Fiorelli et al., 2015). WordNet has been remodelled in the Lemon format as
LemonWordNet (Eckle-Kohler et al., 2014; McCrae et al., 2011; Open Linguistics
Working Group, 2014). For an example of how Lemon has been used in the
context of RDF refer to the example shown in Figure 2.2.

4. Note that in 1.1 it was decided to use the word mapping as the standard term for what is
variously defined as integration or linkage between ontologies in this dissertation. The terms
linkage and integration are used by sources in the context of lemon.
5. Refer to Eckle-Kohler et al. (2014) for further detail on the progress of the linkage of lexical
ontologies through lemon.
3.4 WordNet base concepts

WordNet describes itself (Fellbaum, 1998) as a large lexical database of English, developed under the direction of George A. Miller. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. This results in a network of meaningfully related words and concepts.

WordNet (Miller et al., 1990) is a combination of cognitive (or conceptual) (Miller, 1995) and lexical ontology (Fellbaum, 1998) and is based on a taxonomic structure based on hyponyms as core but also based on the concepts of synonyms, meronyms and antonyms. Taxonomy is the organisation of ideas and objects into categories and subcategories (Morville, 2005). A hyponym taxonomy in this case is a directed acyclic graph (DAG) that is specifically a tree in graph theory. For a lexical ontology the concept of entity is the traditional source node in the noun DAG, and all the other nodes have an indegree of 1. The first developed WordNet structure was for Princeton WordNet (United States English). It has been subsequently developed in other languages. The combined language project is called Global WordNet Project.

The Global WordNet Project has defined synsets (sets of synonymous word meanings) that are most important in 3 or up to 4 WordNets for different languages (English, Spanish, Dutch and Italian), the so-called Base Concepts. The Base Concepts (Vossen, 1998a) are the major building blocks on which the other word meanings in the WordNets depend. The importance of synsets is based on two criteria: the high number of relations with other synsets and a high position
in the hierarchy. This approach is similar to that often used in the construction of upper ontologies. Concepts that have high agreement between domain or mid-level ontologies and have high positions in their hierarchy (large outdegree or outreach) are the concepts chosen for inclusion in an upper ontology (Reed and Pease, 2015).

EuroWordNet (an extension of the key 4 European WordNets that included more languages) was developed with a shared set of so-called Common Base Concepts (CBCs) which were classified using a common shared semantic framework. These CBCs were chosen as the most significant meanings in the local European WordNets (Vossen et al., 1998c). The BalkaNet project extended the list by including Greek, Romanian, Serbian, Turkish and Bulgarian in a larger set of synsets and upgraded\(^6\) the mapping of the CBCs to Princeton WordNet 2.0\(^7\). The Balkanet project also divided these CBCs into 3 levels based on most significant meanings. These are referred to in this dissertation as BCS1, BCS2 and BCS3 respectively.

There is a reduced set of 164 CBCs that occur in 3 or more WordNets as important meanings for the Global WordNet Project. The Global WordNet Project further defined a lexical ontology of 71 Base Types (a reduction of the 164 CBCs). 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

---

6. The mappings had previously been to an earlier version of Princeton WordNet - hence the term “upgrade”.
7. IndoWordNet, like EuroWordNet and BalkaNet is also a multilingual WordNet project that has defined its own master interlingual synset (Redkar et al., 2015).
Base Types are a minimalized list of fundamental concepts. These Base Types (the semantic primitives or taxonomy top nodes) play a key application rôle in large-scale semantic networks like the Semantic Web (Vossen, 2007a,b).

### 3.5 Qualia rôles

Hierarchical structures, like hyponymy based taxonomies represented as DAGs, can be very complex and encode multiple hyponymy relations. Such a hierarchical structure can be populated with features that can be tested against a corpus to verify its quality (Vossen, 1998a). Qualia information can be additional information provided to each synset to provide a rôle related to the hierarchical structure (Mendes and Chaves, 2001). For example, tool is “an implement used in the practice of a vocation” and tool, as a separate concept, is “the means whereby something is accomplished” (the WordNet gloss for the second tool concept is “science has given us new tools to fight disease”). An important aspect is shared by both synsets – both are means to an end or have a telic rôle, i.e. a specified purpose and function (Mendes and Chaves, 2001).

In EuroWordNet the rôle relation is usually related to telicity but it could also cover other aspects of semantic entailment such as agent, patient or result (Mendes and Chaves, 2001). All of these rôles are collectively termed *qualia rôles*. They are important because the CBCs are categorised into these Aristotelian qualia rôles for classifying concrete entities (Agentive for the Aristotle origin rôle, Form for the formal rôle, Composition for the constitutional Role and Function for the telic rôle) (Calzolari et al., 2013; Vossen, 1998a). Composition is further categorized into Substance and Object, and Substance itself further
This dissertation will examine which of the qualia rôles are predominant in the African concepts that are regarded as core.

3.6 African language concepts

The 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 (Bosch, 2007). One of the treasures that emerged from these studies was the establishment of a broad taxonomy of all the African languages mainly through German researchers (Bleek, 1851, 1862, 1869; Meinhold, 1932), Guthrie (1948) and the linguistics department of Oxford University, Belgian research (Meeussen, 1956; Meeussen and Rodegem, 1969) and others. This research for a common lexical base and reconstructed forms for the all the African languages mirrored the original studies into Indo-European languages that attempted to find a reconstructed base for the European languages.

Towards the end of 1986 the HSRC (Human Sciences Research Council) commissioned the LEXINET investigation in order to determine the extent to which computer processing of language abroad might be relevant to South Africa, and to formulate proposals for possible local developments (Bosch, 2007; Morris, 1988). The investigation was divided into seven sub-areas, of which the so-called TEXTNET entailed the investigation into computer processing of language data.
In the ensuing report published in 1988, it was noted that in general there was very little progress in this field in South Africa at the time, especially in comparison to the pace at which NLP was developing abroad. The African WordNet Project gave new impetus to the requirements for contributing to NLP by developing either new base concepts or producing a mapping to Global WordNet base concepts. Significant progress has been made in these areas by the African WordNet Project (Griesel and Bosch, 2013, 2014; Madonsela et al., 2016; Mojapelo, 2016; University of South Africa, 2011, 2013, 2014). The aim of the African WordNet 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 the BalkaNet (Bosch, 2007).

Linking the African language WordNets to one another is strategic. Since much of the international work around WordNet and SUMO has been connected to interlingual indices and upper ontologies, this is also a goal of the Global WordNet Project (Bond et al., 2016; Vossen, 2007b). There are already over 40 different language WordNets, and the establishment of interlingual indices and ontologies would make cross-linguistic information retrieval and question answering possible, and significantly aid machine translation (Fellbaum and Vossen, 2012; Horák and Rambousek, 2010; Peters et al., 1998; Pianta et al., 2002).

In the linguistics of the Bantu languages, there have been projects over the last 50 years aimed at aligning the natural language core concepts of the Bantu languages. The two main approaches originally have been those of Comparative Bantu and Proto-Bantu (Fleisch, 2008). The Comparative On-line Bantu Dictionary (CBOLD) project has taken the initial linguistic comparative Bantu and Proto-Bantu approach and attempted to unify and extend it (Bostoen and
The CBOLD project was initiated in 1994 by Larry Hyman and John Lowe and was aimed at producing a lexicographic database in Berkeley to support and enhance the theoretical, descriptive, and historical linguistic study of the languages of the Bantu language family. CBOLD includes a list of reconstructed Proto-Bantu roots (based on the Comparative Bantu tables of Guthrie (1948) and the Bantu Lexical Reconstruction (BLR) list of (Meeussen and Rodegem, 1969)), thousands of additional reconstructed regional roots called Bantu Lexical Reconstructions 2 (BLR2) (based on the current work of scholars in Tervuren and elsewhere), and reflexes of these roots for a substantial subset of more than 500 daughter languages. The Tervuren Museum’s Linguistics Sections continued work and updated the original BLR list from (Meeussen and Rodegem, 1969). They combined it with the Guthrie research to produce an electronic database called BLR2. It was meant to be the follow-up of Meeussen’s original manuscript (Bostoen and Bastin, 2016; Schadeberg, 2002). A newer version of BLR2, called BLR3 was released in 2002 (Bastin et al., 2005; Schadeberg, 2002). The main enhancement from BLR2 to BLR3 was the data representation (Bostoen and Bastin, 2016).

Of these roots used by BLR3, the CBOLD project has selected 10 000 BLR3 reconstructions that represent so-called main entries of which there are 1 400. These main entries are referred to as basic reconstructed etymons. These 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 3.2 (Zone S is the region containing all the Southern African Bantu languages).

The reason for the choice of Zone A and Zone S is that these two zones...
are geographically maximally removed and hence it is of great significance if the same proto-form occurs in both (Maho, 2009). This emphasises the generality and the hierarchical importance level of a concept. This produces 375 roots. Maho (2001) also isolated all main entries that have modern reflexes in at least 14 zones (231 roots). The two lists produce a core collection of 407 roots.

Concerns have been expressed regarding the use of proto-language in the Bantu language context and the agreement of the unity within the Bantu languages, as well as the challenges to describe the disagreements on the nature of this unity (Marten, 2006). As mentioned in Section 1.4 these concerns are primarily based on the lack of written historical records for the Bantu languages.

The challenge in the last century that led to the compilation of BLR3 was the creation of lists of cognate linguistic items in the absence of written historical evidence. The scholars involved used the principles of historical linguistics and language reconstruction to find cognates that on the surface may seem unrelated due to phonological changes over time. Diachronic semantics and semantic reconstruction have received far less attention within Bantu historical linguistics (Bostoen and Bastin, 2016) than in other languages. Fleisch (2008) gives a detailed historical overview and summary of the reconstruction of lexical meaning in Bantu. Unlike sound change, semantic change is not necessarily unidirectional but could be multi-directional and cyclic (Bostoen and Bastin, 2016). Bostoen (2001) gives a detailed and specific Bantu case study involving these sort of semantic shifts. He cites an example in which oil palm, palm oil, palm nut, and blood are associated. It is shown that it is difficult to determine which of these was the original meaning of the BLR3 entry and in which direction it evolved semantically (Bostoen, 2001). As mentioned above, the particular challenge is
the lack of written historical records for the Bantu languages, and hence much of this semantic research remains purely theoretical.

### 3.7 African WordNet construction

Since an approach to interlingual mapping is important to lexical ontology design (Fellbaum and Vossen, 2012; Horák and Rambousek, 2010; Peters et al., 1998; Pianta et al., 2002), the approach for the design of the African language WordNets and their interlingual index is significant. In the construction of the Hebrew WordNet, Ordan and Wintner (2007) discuss two approaches for constructing WordNets – either construction from scratch followed by alignment, as proposed by EuroWordNet (Vossen, 1998a) (the merge approach); alternatively, there is strict alignment with Princeton WordNet as the base. The latter approach is based on the assumption that those concepts are universally shared (the expand approach). This second approach is that proposed by MultiWordNet (Pianta et al., 2002). The latter approach involves the potential risk that the resulting hierarchy will be influenced by Princeton WordNet. Ordan and Wintner (2007) propose that the expand approach is still a better approach for languages poor in resources.

The first approach is where a WordNet for each language is built from first principles, and aligning is done once complete, using an Interlingual Index (ILI). Examples of this merge approach are the Chinese (Wong and Pala, 2002), Russian (Balkova et al., 2004), Tartar (Galieva et al., 2014), Dutch, Italian and Swedish (Viberg et al., 2002) WordNets. An interesting alternative to the merge approach, in order to address the traditional labour and time intensity of Word-
Net creations, was done for Onto.PT. Onto.PT is a WordNet-like lexical ontology for Portuguese. It was created using an automated approach from existing Portuguese lexical resources (Gonçalo Oliveira and Gomes, 2014). Obviously, this alternative merge approach is only applicable to relatively resource-rich languages. This alternative merge approach has been termed the ECO approach since it focuses on Extraction, Clustering and Ontologising (Gonçalo Oliveira and Gomes, 2014, p. 377).

The second approach is where the WordNets are aligned as strictly as possible to the American-English version of Princeton WordNet (PWN), under the assumption that most of the concepts are universally shared. This approach involves a potential risk, namely that the resulting WordNet may be influenced by the structure of Princeton WordNet. This risk could be offset by devising a methodology to cope with it (Ordan and Wintner, 2007). Examples of the expansion approach already utilised for lesser resourced languages include Hungarian (Miháltz and Prószéký, 2004), Finnish (Lindén and Niemi, 2014), Serbian (Štoković et al., 2014), Croatian (Šojat and Srebačić, 2014), Persian (Rouhizadeh et al., 2008), Gujarati (Bhensdadia et al., 2010), Marathi, Sanskrit, Bodo and Telugu (Bhattacharyya, 2010), Basque (Alegria et al., 2011; Pociello et al., 2011), Indonesian (Putra et al., 2008) and Thai (Thongsup et al., 2009).

A similar argument for the two different WordNet construction approaches is also proposed by Vossen – what he terms the expand and merge approaches (Vossen, 2007a). 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” (Vossen, 2007a) and compatible with Princeton WordNet. This allows for the exploitation
of many resources already linked to Princeton WordNet. For instance SUMO, WordNet domains and selection restriction from the British National Corpus are resource examples that can be exploited using this approach. The disadvantage is that it will be biased by design.

In the *merge approach*, an independent WordNet is created in another language which is then aligned with the Princeton WordNet by generating the appropriate translations. This approach has the disadvantage of being 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 (Vossen, 2007a). It is also “typically slower” (Bhattacharyya, 2010, p. 2). It is also argued that in the merge approach there is the distracting influence of another language, due to the lexicographer encountering cultural and regional specific concepts of the source language (Bhattacharyya, 2010).

Benjamins et al. (2002) 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 (Eckle-Kohler et al., 2014; 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; Schwitter, 2005), and the challenges of trying to align natural language core concepts and lexical ontologies with the upper ontologies required for inference on the Semantic Web (Eckle-Kohler et al., 2014; Gangemi, 2004). The challenges of the implementation of HLT within the Bantu language domain influences resource development for the African languages (Bosch et al., 2006; Griesel and Bosch, 2013, 2014).
All of these challenges highlight the importance of the correct approach to an interlingual index for the African languages. The importance of examining previously defined core concepts in projects like BLR3, in concert with how they can be mapped to existing Global WordNet BCs is that it will inform which approach provides the best benefit or addresses the multilingual challenges best. They should provide evidence for answering sub-research question 5: What will a new structure of core concepts from an African linguistic base look like and how can it be compared to existing structures?

3.8 WordNet concepts and top lexical ontologies

WordNet was developed prior to the advent of the Semantic Web and its ontologies. What is the relation, therefore, between WordNet and the Semantic Web architecture and standards? The first WordNet structure developed was, as described above, for Princeton WordNet (US English), and although technically WordNet refers to all WordNets in the Global WordNet Project, it often directly refers to Princeton WordNet and US English as language in particular - a form of synecdoche or *totum pro parte*.

WordNet is considered to be one of the most important resources available to researchers in computational linguistics, text analysis, and many related areas. While its original design was inspired by psycho-linguistic and computational theories of human lexical memory, Princeton WordNet has been ported to the Semantic Web languages of RDF and OWL (van Assem et al., 2006) and Prince-
ton WordNet 3.0 is defined for use with SUMO. The DOLCE group has also ported EuroWordNet to the DOLCE ontology, called the OntoWordNet Project, but it uses an older version of WordNet (1.6), and aims to align only the upper levels of WordNet (Gangemi et al., 2003).

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. As mentioned, the BCs are supposed to be the natural language core concepts that play the most important rôle in the various WordNets of different languages.

Subsequent to the EuroWordNet Project, which started the drive towards the Global WordNet, there has been significant developments in constructing ontologies related to WordNets for other Indo-European languages. BalkaNet (Balkanet, 2001), Romanian WordNet (Tufiş et al., 2013) and Slovene WordNet (Fišer, 2009) also developed a mapping to a top ontology. IndoWordNet had plans to construct linkage to an ontology (Bhattacharyya, 2010; Boem et al., 2013; Redkar et al., 2015), and FarsNet has already linked Farsi to SUMO (Taheri and Shamsfard, 2011).

Also, as already mentioned, there are different approaches to designing top ontologies and interlingual indices. Some of the different applications of these approaches, particularly to the usage of BCs in those languages that fall outside the Indo-European family, are the Arabic WordNet (Black et al., 2006), Hebrew WordNet (Ordan and Wintner, 2007) and Chinese WordNet (Huang et al., 2004; Lee et al., 2009; Wong and Pala, 2002).
The WordNet “Top Ontology” refers to 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, 1998a; Vossen et al., 1998c). The 64 Top Ontology concepts are based on the fundamental semantic distinctions used in various semantic theories and paradigms forming a hierarchy of language-independent concepts that reflect the distinctions between, for example, object and subject or dynamic and static (Vossen, 1998b; Vossen et al., 1998a). They have explicitly been defined in terms of hyponymy and opposition (for example, animate and inanimate) relations (Vossen et al., 1998b). Much of the international work around WordNet and SUMO has been connected to interlingual indices (ILIs) and WordNet Top (lexical) Ontologies (Niles and Pease, 2003) or WordNet and OWL (van Assem et al., 2006).

The relationships between synsets defined in WordNet have been formalised in Semantic Web ontologies. Listing 3.1 represents the noun meronymy (lines 17-21), ”classified by usage” (lines 12-15), noun holonymy, the inverse of meronymy (lines 23-26), hyponymy for nouns and verbs (lines 28-33) and antonymy for all word classes (lines 35-39) as properties for WordNet in OWL. It can be seen that hyponomy is a transitive property in line 29 and antonymy is defined as a formalised symmetric property on line 36.

Listing 3.1: WordNet synset relations

```xml
1  @prefix : <http://www.w3.org/2006/03/wn/wn20/schema/> .
2  @prefix owl: <http://www.w3.org/2002/07/owl#> .
3  @prefix rdf: <http://www.w3.org/1999/02/rdf#> .
4  @prefix xml: <http://www.w3.org/XML/1998/namespace> .
5  @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
6  @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
7  @prefix wn20schema: <http://www.w3.org/2006/03/wn20/schema/> .
8  @base <http://www.w3.org/2006/03/wn/wn20/schema/> .
```
As introduced in Section 3.4, concepts that have high agreement between ontologies and have high positions in their hierarchy (large outdegree or outreach) are the concepts chosen for inclusion in an upper ontology. These criteria for inclusion align with the Global WordNet Project goal for specifying Base Concepts. These criteria will be examined further when I compare the results of the study with upper ontologies and with Global WordNet interlingual mapping and
In the EuroWordNet Top Ontology, three types of entities are distinguished at the first level of the Top Ontology (Vossen et al., 1998c).

1. 1st Order – any concrete entity publicly perceivable by the senses and located at any point in time, in a three-dimensional space, e.g. individual persons, animals and more or less discrete physical objects and physical substances. They are always denoted by (concrete) nouns.

2. 2nd Order – any static situation (property, relation) or dynamic situation, which cannot be grasped, heard, seen, felt as an independent physical thing. They occur or take place rather than exist, e.g. continue, occur, apply, and also events, processes, states-of-affairs or situations that can be located in time belong here. They can be expressed by nouns, verbs and adjectives.

3. 3rd Order – unobservable propositions which exist independently of time and space. They can be true or false rather than real. They can be asserted or denied, remembered or forgotten, e.g. ideas, thoughts, theories, plans, hypotheses, reasons, and they are always expressed by (abstract) nouns.

For EuroWordnet, these criteria have independently been applied to 4 different detailed language WordNets (UK English, Spanish, Dutch and Italian). By providing clear definitions or features for the Base Types in EuroWordNet (refer to section 3.4), the Global WordNet Project 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.
Of interest in this research is the analyses of Top Ontologies and natural language core concepts in those languages that fall outside the Indo-European family. The challenges of a multilingual WordNet catering for all languages has been highlighted. The challenges highlighted include the previous mapping through interlingua based on natural language which had been done for the European languages and the alternative option of mapping through a formal ontology has been proposed more recently (Fellbaum and Vossen, 2012).

These last two chapters have detailed the broader context of the Semantic Web architecture and the more detailed context of existing work on lexical ontologies, specifically related to WordNets and the African WordNet Project. How can a lexical ontology be compared to another lexical ontology or to any other existing ontology? The next chapter examines the notion of ontology comparison.
Figure 3.2: Bantu language zones in Sub-Saharan Africa
Part II

Research design and implementation
CHAPTER 4

Ontology comparison

He [Platon Karataev] ... did not understand and could not grasp the meaning of words apart from their context.

War and Peace, Tolstoy (2009)

4.1 Introduction

The previous chapters introduced the concepts in this dissertation, documented the research questions and provided an overview of the literature as a contextualisation for this research. They provided the context to this research in terms of the Semantic Web Architecture and the use of ontologies and upper ontologies as a key layer in that architecture. The domain focus of linguistics and the African language research in this area were introduced.

0. The following chapter is based on the work described in Xue et al. (2009) and to a large extent is almost verbatim for the descriptions and formulae.
The following chapters provide the details of the research design and implementation. Key to the design of the research is the approach to answering research question 5 in table 1.1: what will a new structure of core concepts, from an African linguistic base, look like and how can it be compared to existing structures? In order to validate the mapping needed to answer this question, an accepted approach for validation needs to be chosen. This chapter provides the design approach to the ontology comparison method used in this research. Initially, a motivation is provided as to why tree comparisons are useful for and adequate as an ontology comparison approach, followed by the actual calculation formulae in section 4.2. The limitations of these calculations are provided in section 4.3. Finally, the principles of how the results might relate to lexical ontologies are illustrated in section 4.4.

4.2 Ontology comparison

In its full generality, an ontology is a conceptual graph or a semantic net and is not a tree (Sowa, 1984). However, among others, two specific species of the Web Ontology Language family, OWL 2 DL and OWL 2 Full, were designed with a desire to “provide practically useful knowledge modeling primitives while ensuring decidability of reasoning” (Motik et al., 2008). This reduces the complexity so that the model of any class expressions in OWL DL does indeed ensure a tree model (Magka et al., 2012; Motik et al., 2008). It is this model that forms the basis of the remainder of the material in this chapter where knowledge in WordNet is provided in tree-like structures.

Although ontologies are more complex than trees, in WordNet trees are suf-
icient to describe the complexity of its relations, particularly with nouns. In the WordNet structures, nouns generally share a common root\(^1\), while verb structures have a variety of roots. An example of the hyponym trees for the nouns *bee* and *sangoma* are shown in Figures 4.1 and 4.2 respectively. Examples of the hyponymy trees for the verbs *roast* and *bite* are shown in Figures 4.3 and 4.4 respectively.

Figure 4.1 represents the relationship between *bee:1* and its hypernym tree to its root node *entity:1*. It starts with the hyponym *drone:1* which is a type of *bee:1*. Following this, the navigation downwards in this view of the tree in Figure 4.1 to illustrate that bee has the hypernym of *hymenopterous insect:1*, which has a hypernym of *insect:1*, which has a hypernym of *arthropod:1*, and so on. Eventually all *living thing:1* are hyponyms of *object:1* which has, as its hypernym, *entity:1*, the root node in this tree structure.

Figure 4.2 similarly represents the relationship between *sangoma:1* and its hypernym tree to the same root node *entity:1*. It starts with the synset *sangoma:1* which is a type of *therapist:1*. Eventually it also has, as its ultimate hypernym, *entity*, the root node in this tree structure. Therefore these two examples illustrate how the node *entity:1* is a common root node for nouns in the hypernym tree structure.

Figure 4.3 represents the tree structure for the verb synset *roast:1* which in this case has, as a verb, the root node *change:2*. Figure 4.4 in turn represents the tree structure for the verb synset *bite:2*. It has the root node *cause to be perceived:1*. Therefore the tree structure for verbs can have different root nodes.

Similar to the model of defining OWL languages that are less descriptive but

---

\(^1\) “All noun hierarchies ultimately go up the root node entity.” (Leung et al., 2013, p. 665)
Figure 4.1: Hyponymy tree for the noun bee
Figure 4.2: Hyponymy tree for the noun *sangoma*
Figure 4.3: Hyponymy tree for the verb *roast*
Figure 4.4: Hyponymy tree for the verb \textit{bite}
more useful for specific application, when looking at these WordNet structures from the limited perspective of hyponymy, or a less descriptive WordNet synset relation approach, the concept relations produce a non-cyclic tree structure that is more useful for comparison purposes. Tree structures\(^2\), since they do not contain cycles, make comparison simpler than cyclic graph structures. Therefore, for many practical purposes of knowledge representation using ontologies, a tree structure is a useful and an adequate model for comparison and is the commonly used form for representing concept structures in a domain (Xue et al., 2009, p. 1767).

The similarity measures for ontology comparison can be divided into general groups: lexical measures (string distances), structural measures (taxonomic similarities) and combinations of these, often termed semantic measures (Banerjee et al., 2010; Grover et al., 2010, 2011; Jiang et al., 2014; Ngo and Bellahsene, 2012). Lexical measures use mappings that have similar names or descriptions across ontologies. Structural measures focus on the adjacent nodes in the ontology graphs. Semantic measures rely on information distance between the nodes being compared (Bennett et al., 1998; Chen et al., 2009; Vitányi et al., 2009). However, the field of ontology comparison is a broad and growing field and there are many detailed discussions on the methods for comparing ontologies as trees (Choi et al., 2006; Wang et al., 2010). A detailed discussion of these falls outside the scope of this dissertation. For this dissertation the focus will be on the method of Xue et al. (2009), as expanded on in Xue (2010).

The sections below document the costs described in Xue et al. (2009) which are used as a basis for comparison in all the calculations and results of this

\(^2\) Trees are types of graphs.
research. All of this information is directly taken from Xue et al. (2009) for use in calculating the costs. Comments about the usage of these costs are included in Chapter 7 (Conclusion and Future Work). In the conclusion (Chapter 7) there are comments on the usage of these costs.

4.2.1 Concept tree

In order to use the tree similarity measure of Xue et al. (2009), the following definitions are necessary:

An unordered and unlabelled concept tree is the six-tuple

\[ T = (V, E, L^V, \text{root}(T), D, M) \]

where

- \( V \) is a finite set of nodes
- \( E \) is a set of edges satisfying that \( E \subset V \times V \) which implies an irreflexive and antisymmetric relationship between nodes
- \( L^V \) is a set of terms for concepts used as node labels
- \( \text{root}(T) \in V \) is the root of the tree
- \( D \) is the discourse domain
- \( M \) is the injective mapping from \( V \to L^V \). A mapping from node \( v \) to label \( l \) is written as the tuple \((v, l) \in M\).

If \((u, v) \in E\) then \( u \) is a parent of \( v \) defined as \( \text{parent}(v) \) and \( v \) is a child of \( u \) defined as \( \text{child}(u) \).
The set of all children of node $u$ are denoted as $C(u)$. For two nodes $u_1, u_2 \in V$ if $(u_1, u_2) \in E^*$, then $u_1$ is an ancestor of $u_2$ and $u_2$ is a descendant of $u_1$.

### 4.2.2 Conceptual similarity measures

The conceptual similarity measure $S_{L^{V_1},L^{V_2}}$ is the set of mappings from two term sets $L^{V_1}, L^{V_2}$ used in different concept trees to $R$, i.e. $S_{L^{V_1},L^{V_2}} : L^{V_1} \times L^{V_2} \rightarrow R$. $R$ has a range of $(0, 1]$. $S_{L^{V_1},L^{V_2}}$ is for $l_1 \in L^{V_1}$ and $l_2 \in L^{V_2}$:

- **semantically reflexive**: here $S_{L^{V_1},L^{V_2}}(l_1, l_2) = 1$
- **symmetric**: here $S_{L^{V_1},L^{V_2}}(l_1, l_2) = S_{L^{V_1},L^{V_2}}(l_2, l_1)$

$w = s(l_1, l_2)$ refers to the number value of conceptual similarity from two trees $T_1$ and $T_2$. The larger the value of $w$ the closer the two concepts are and $w = 1$ means identical concepts (synonymy of the concepts). For for $l_1 \in L^{V_1}$ and $l_2 \in L^{V_2}$, if there is no definition for $l_1$ and $l_2$ in the measure, then $l_1$ and $l_2$ are disjoint concepts.

![Figure 4.5: Deleting a node](image)
4.2.3 Tree operations: deletion

If

\[ v \neq \text{root}(T), \]

then

\[ V' = V - v, \]

\[ E' = E - \{(u, v) \mid u = \text{parent}(v)\} - \{(v, v_c) \mid v_c \in C(v)\} + \]
\[ \{(u, v_c) \mid u = \text{parent}(v) \land v_c \in C(v)\}, \]

\[ L^{V'} = L^V - M(v) \]

and

\[ M' = M - \{(v, M(v))\}. \]

If \( v = \text{root}(T) \) then \( v \) cannot be deleted. Deleting one node effectively means eliminating the node from the tree and then making its children nodes new direct children nodes of its parent node. Deleting a node is therefore not the same as deleting a sub-tree.

If the node to be deleted is the root, then the result is no longer a tree. In a concept tree, the root is usually a very general node like “object”, or “entity:1” in WordNet nominal trees or the class “owl:Thing” in OWL. For this reason, Xue proposes a rule to restrict the deletion of the root node (Xue et al., 2009, p. 1771). Deletion of a node is represented in Figure 4.5. There are no examples of node deletion with reference to the data used in this research.

4.2.4 Tree operations: insertion

\[ V' = V + v, \]
\[ E' = E + \{(u, v)\} + \{(v, u_c) | u_c \in C'(u)\} - \{(u, u_c) | u_c \in C'(u)\}, \]

\[ L^{V'} = L^V + \{l_v\} \]

and

\[ M' = M + \{(v, l_v)\} \]

where \( l_v \) is the term assigned to the new node \( v \), and \( C'(u) \subseteq C(u) \) meaning that some children nodes of \( u \) are changed to be children nodes of the new node \( v \). The elements contained within \( C'(u) \) are determined by the context when performing the editing operations. Insertion of a node is represented in Figure 4.6, while insertion of a node in terms of the data used in this research is represented in Figure 4.7. For a more detailed discussion on this synset refer to Section 6.3.1.

![Figure 4.6: Inserting a node](image)
4.2.5 Tree operations: re-labelling

Re-labelling $v$ with label $l_v$ is to assign $v$ a new label $l_v$ and to keep positions of all nodes unchanged.

$$L' = L - l_v + l'_v$$

and

$$M' = M - (v, l_v) + (v, l'_v)$$

where $l_v$ is the new label assigned to $v$. Re-labelling of a node is represented in Figure 4.8, while re-labelling of a node in terms of the data used in this research is represented in Figure 4.9. For a more detailed discussion of this synset refer to Section 6.3.5.

4.2.6 Tree operations: movement

This is a new operation introduced by Xue that is not normally covered in classical tree editing operation sets. In a pure, structured tree, a move operation could be achieved by deleting a node in the source tree and then inserting it correctly.
Figure 4.8: Re-labelling a node

Figure 4.9: Example of node re-labelling
in the target tree. However, in a concept tree, the assumption cannot be made that the node destined for insertion does not already exist in the target. If it does already exist, then an insertion of a duplicate node would violate the notion of a concept tree. The moving operation moves the node from its original position in the source to the chosen position in the target.

\[ V' = V, \]

\[ E' = E + \{(u, v)\} + \{(v, u_c) | u_c \in C'(u)\} + \{\text{parent}(v), v, u_c | v_c \in C(v)\} - \{\text{parent}(v), v\} - \{(v, v_c) | v_c \in C(v)\} - \{(u, u_c) | u_c \in C'(u)\} \]

where \( C'(u) \subseteq C(u) \), implying some children of node \( u \) will be changed to children of node \( v \) based on the operation context.

### 4.2.6.1 Transformation costs

Each transformation operation \( \text{Op} \) on tree \( T \) is mapped to a real number that defines the transformation cost and is denoted as \( \gamma(\text{Op}) \).

If \( \text{OP} = \text{Op}_1, \text{Op}_2, \ldots, \text{Op}_k \) is a transformation sequence, then the transformation cost of the sequence is defined as

\[ \gamma(\text{OP}) = \sum_{i=1}^{\text{|OP|}} \gamma(\text{Op}_i) \]

If \( \text{OP} \) is a transformation sequence mapping a tree \( T_1 \) to tree \( T_2 \), then the transformation cost from \( T_1 \) to \( T_2 \) is

\[ \gamma(T_1 \rightarrow T_2) = \min\{\gamma(\text{OP})\} \]

The similarity index of two trees \( T_1 \) and \( T_2 \) is
\[ \gamma(T_1, T_2) = \min\{\gamma(T_1 \rightarrow T_2), \gamma(T_2 \rightarrow T_1)\} \]

The similarity of two individual concepts needs to be estimated by domain experts (Xue et al., 2009, p. 1767). The concept -ngaka is translated into English as “witch-doctor, doctor, medical practitioner, surgeon” (Ziervogel and Mokgokong, 1985) and a domain expert would give these different conceptual similarity measures values to node mappings of the WordNet synsets sangoma:1, doctor:1, witch doctor:1, herbalist:1, surgeon:1, medical practitioner:1 and others. Where the meaning is exactly the same, or synonymous, the similarity degree would be 1 – say the mapping of -ngaka to sangoma:1. Sangoma in this case is a borrowing of a Zulu term into English that maps to the Northern Sotho concept. However, if the meaning does not always refer to the same thing, a similarity degree can be assigned, say 0.9, to mean that in around 90% of occasions the two concepts are describing the same group – say the mapping of -ngaka to surgeon:1.

The following are requirements for determining the transformation cost:

- **height**\((T)\) is a function calculating the height of tree \(T\).

- **depth**\((v)\) is a function calculating the height of node \(v\).

\[ \text{depth}(\text{root}(T)) = 1 \]

and

\[ \text{depth}(\text{root}(T)) > 1 \]

iff \(v\) is not the root.
• $|D(v)|$ is the number of descendants of node $v$ including direct children and indirect offspring. If $v$ is a leaf node then $D(v) = \phi$ and $|D(v)| = 0$.

• $s$ is the conceptual similarity measure between two labels $l_1$ and $l_2$ where $s \in [0, 1]$.

The transformation cost then is

$$
\gamma_{T_1 \rightarrow T_2}(OP) = \min \left\{ \sum_{i \in D} \gamma(\text{delete}(i)) + \sum_{i \in I} \gamma(\text{insert}_u(i)) + \sum_{i \in M} \gamma(\text{move}(i)) + \sum_{i \in R} \gamma(\text{relabel}(i)) \right\}
$$

where

$$
\gamma(\text{delete}(v)) = \frac{\text{height}(T) - \text{depth}(v) + 1 + |D(v)|}{|V|}
$$

where $v$ is a non-root node, and

$$
\gamma(\text{insert}_u(v)) = \frac{\text{height}(T) - \text{depth}(u) + 1 + |D(v)|}{|V|}
$$

$$
\gamma(\text{move}(v)) = \frac{(\gamma(\text{delete}(v)) + \gamma(\text{insert}_u(v))) \times (|V| - 2)}{2|V|}
$$

$$
\gamma(\text{relabel}_{l_1 \rightarrow l_2}(v)) = (\gamma(\text{delete}(v)) + \gamma(\text{insert}_{\text{parent}(v)}(v))) \times (1 - s)
$$

The time complexity of computing the transformation cost is $O(n^4)$.

In the above definitions, the worst case is an insertion operation for all nodes at the second level and the best case is when every operation is re-labelling.

Princeton WordNet is accessible using Prolog (Witzig, 2003). It was therefore possible to use Prolog predicates to calculate the height and descendants of any given node.
4.3 Limitations of calculations

The mechanism above was only used for calculation of the noun stems and the ontology comparison of two trees representing the nominal concepts. Verb and adjective root positions are only compared qualitatively in this dissertation.

4.4 Comparison principles

In this chapter an existing method of ontology comparison was introduced in order to illustrate how it is applied to the research data. The use of the method highlights additional future areas of research that might be pertinent to ontology comparison, specifically in the context of upper ontologies.

For an upper ontology, as opposed to a domain ontology, a number of conceptual guidelines have emerged. SUMO upper ontology concepts should, in relation to WordNet synsets, satisfy the following:

- have a large outdegree;
- be “high up” in the tree – that implies a large outreach and a low inreach, or equivalently, the graph theory levels function should be low;
- not be a sink node;
- have a short path length from the root, relative to the maximum path length in the WordNet structure;
- delineate a component in some form, and preferably a strong component rather than a weak component.
An outcome of considering these specific list items could be the identification of more accurate measures for comparing trees where a node can also be given a level of significance, say in movement or deletion, that gains more significance in the case of upper ontologies.
CHAPTER 5

Ontology mapping approach

\[ \ldots \text{est non verbum e verbo sed sensum exprimere de sensu (I express not the word for the word but the sense for the sense).} \]

Patrologia Latina, Jerome (PL 1877: XXII, 571) (Migne and Hamman, 1859)

5.1 Introduction

This chapter continues documenting the approach taken in this study in the context of the Semantic Web, the usage of upper ontologies and the application to African language WordNets. The approach describes how a natural language core concept hierarchy is defined using existing African language research in conjunction with methods proposed in the African WordNet Project. This is done by construction of a specific African language WordNet prototype focusing on core concept hierarchies. The goal of the approach is to prepare the data for the ontology comparison described in Chapter 4 in order to determine whether there
will be a significant difference in the natural language core concept mapping when starting from an African language base. This approach aids in answering the research question about whether the original mapping from WordNet to SUMO, that is from one linguistic base only, provides representativeness and comprehensiveness as applicable to other languages, particularly in other language families and specifically to the African language families. The approach also provides a methodology to answer the research question about whether the language used as a basis to for the upper ontology definition affects the concepts that are regarded as broad and comprehensive enough for inclusion in the upper ontology.

5.2 Methodological approach

The *modus operandi* was as follows: the 1 400 main entries from the CBOLD BLR3 list of 10 000 suggested Proto-Bantu reconstructions were utilised as the theoretical base, and then further reduced to the subset proposed by Maho for Zone A and S languages. Figure 5.1 illustrates the main search window for BLR3. A number of criteria can be used for the search entry. These include:

- English or French equivalent word or concept,
- BLR3 ID,
- tone,
- proto-Bantu root,
- grammatical part of speech,
- noun class,
Figure 5.1: BLR3 Search Entry

- zone and wider regions within the language domain and the total number of zones and regions, and
- consonant and vowel slots and vowel tones.

Figure 5.2 shows a result for a search for a guinea fowl. The result shows a
Number of details:

- the main reference identifier
- the proto-Bantu root
- the tone pattern
- the noun classes where the root occurs
- the English and French translations
- the regional distribution
- the zonal distribution, and
- a coloured dot reflecting the reliability code (Bostoen and Bastin, 2016):
  - main reconstructions in yellow as in Figure 5.2,
  - derived reconstructions in green,
  - variant reconstructions in purple,
  - compound reconstructions in blue,
  - inclusive reconstructions (that were previously proposed but are now included in one of the above types) in gray, and
  - refused reconstructions in red.

The main entries have been categorized (Maho, 2001) to isolate all main entries that have modern reflexes in Zone A and Zone S (Zone S is the region containing all the Southern African Bantu languages) (375 roots). Maho also
isolated all main entries that have modern reflexes in at least 14 zones (231 roots). The two lists produce a core collection of 407 lexical roots.

Of these Maho determined which main entries have modern reflexes with a claimed total zone-spread covering at least 14 of a total of 16 zones, yielding 231 roots. These were then further reduced to roots that have zone spread across all 16 zones and are therefore also in Zone S\(^1\), where equivalent modern reflexes can be found in Northern Sotho and Zulu (with reference to the predominant local dictionary for each). Northern Sotho and Zulu are representative of two significant different large groups within Zone S\(^2\).

My methodology has involved taking the mentioned 407 roots and only using those that occur in all 16 zones. This produced a list of 99 roots. These roots were then analysed to establish if they have modern reflexes in the Comprehensive Northern Sotho Dictionary (Ziervogel and Mokgokong, 1985). The exercise yields a list of 80 potential candidates. These 80 were mapped to their Princeton WordNet equivalents if they existed or marked if no mappings were found. Verification of the candidate concepts, that is the quality assurance of those concepts, was done by two individuals, both of whom are Northern Sotho linguists. One of them is familiar with Northern Sotho linguistics as a mother-tongue speaker and the other is a Northern Sotho phoneticist, who is also familiar with research in BLR3.

---

1. Refer to Figure 3.2
2. The examples and results given here are shown for Northern Sotho only, since its lexicalization has been verified and quality assured.
5.3 Quality assurance

Once the mappings were verified \(^3\) and the phonetic mapping to the BLR3 was quality assured \(^4\), this final list was reduced to 67 roots. Roots of which the status is doubted (the difference between the original 80 and the final 67 concepts or roots) were sent to an international BLR expert for possible additional inclusion in the final result table, but the 67 final roots are used in the results for all calculations \(^5\). If any of the 67 roots did not match a main entry in the BLR3 list, a variant of such a root was used if one existed \(^6\). For example, the Northern Sotho root for two – -bêdi – is a closer match to variant 190 in the BLR3 list than to the main entry 36, which is also the entry in Maho’s list \((\text{Maho, 2001})\).

Once the quality assurance review was done, a re-examination of every mapping was performed to ensure that no comments from the quality assurance feedback affected the existing mapping. This resultant table is termed the *quality assured word list* in this dissertation. The quality assurance is based on the veracity as acknowledged by three independent experts.

\(^3\) The initial mapping was checked by the researcher.
\(^4\) Secondary phonetic correspondence of the BLR3 entries to the Northern Sotho realisation was checked by two local experts in the field of phonetics.
\(^5\) The feedback by the international BLR expert was that the key data had been reviewed and he confirmed that the Bantu language data was presented, interpreted and used correctly, and no changes or additions were recommended \((\text{Maho, 2012})\).
\(^6\) BLR3 entries are recorded as a main entry or as a variant.
5.4 Meta-data documentation

The 67 roots were then added to WordNet using DEBVisDic (Horáček et al., 2008, 2006; Horáček and Rambousek, 2010), software produced by the BalkaNet team to define, manage and map WordNets (Bukatović et al., 2010). Where there were direct mappings to Princeton WordNet, the ILI for the word was used as a linkage. In this case the word sense is the standard representation of the lexical root. For nouns, this would be the singular class of the word. See Figure 5.3 for an example of the synset bee (nôse). For verbs this would be the present tense un-extended verb. Where there is a one-to-one correspondence between the primary Comprehensive Northern Sotho Dictionary sense of the word and WordNet, the mapping was made (Ziervogel and Mokgokong, 1985). If the mapping was to the incorrect level of the tree in terms of definition, then the tree was adjusted in Northern Sotho WordNet. Where a word did not exist in Princeton Wordnet, it was added to the Northern Sotho WordNet structure without the ILI relationship.

All words that are in the list were marked as being part of the *African WordNet Core Set 1* in the African language WordNet prototype. All additional words required to complete the WordNet tree to the top level of the hierarchy were added as *African WordNet Core Set 2*. The principles used for the mapping were the ILI, EuroWordNet base concept methodology and the existing SUMO mapping as a form of verification. All words were grouped according to the part of speech they represent as proposed in *Maho (2001)* and the part of speech as attested in the Comprehensive Northern Sotho Dictionary (Ziervogel and Mokgokong, 1985). The results of this prototype are available as a resource on the
5.5 SUMO mapping confirmation

All of the SUMO mappings for the words were documented. If a word in African Core Set 2 was not a Northern Sotho root, the actual root was added in its correct place as being part of Core Set 2, or the derivative relationship encoded to that word. The final result for each word is an XML entry conforming to the DEBVisDic XML standard for WordNet (Bukatović et al., 2010). Any ontological relationship gaps in WordNet and SUMO were noted and any patterns in the mapping from Northern Sotho WordNet to SUMO were noted. These are discussed in the subsequent chapter (Chapter 6).

The main technologies used were

- DEBVisDic (see Figure 5.3) – a tool built using XML and Berkeley Database technologies for constructing WordNets, mapping to core concept sets and for documenting interlingual relationships, and

- Protegé (see Figure 5.4) – an ontological design environment for examination and comparison.

In Figure 5.3, the structure of the entry is exemplified in the XML format used by DEBVisDic:

- the STAMP tag records the author, the date and the time stamp of the entry,
Figure 5.3: DEBVisDic
• the *ILR* tag records internal language sense relationships using the ILI reference in the XML attribute,

• the *SUMO* tag records the SUMO concept and the attribute records the mapping type (in this example there is equivalence),

• the *ID* tag records the ILI if it exists, otherwise a unique ID,

• the *SYNONYM* tag records the synonyms in the synset,

• the *DEF* tag records the definition (not used in this research since the CNSD is not a defining dictionary but a multilingual dictionary),

• the *SNOTE* tag records notes, in this case the stem is recorded for nouns as well as the English and Afrikaans entries from the CNSD,

• the *VERSION* tag records the version of DEBVisDic used during definition,

• the *BCS* tag records the nature of the lexeme’s status as a core concept, 1 is used for all the African language core concepts,

• the *DOMAIN* tag records the domain of the noun,

• the *NL* tag records whether this is a lexicalized or non-lexicalized entry in the language, and

• the *POS* tag records the part of speech.

In Figure 5.4, the structure of the entry is exemplified in the ontology class as shown in the Protegé ontology editor software. The top left-hand side shows the class in its hierarchy. In this case it is shown that in SUMO a bee is an
organism, that is an animal, that is an invertebrate, that is an arthropod and that is an insect. On the top right-hand side of the figure the following relevant class annotations are represented:

- the *label* annotation records the label for the class, and in this case the original SUMO class has an English label but in my research I have added a Northern Sotho label to the class as well,

- the *isDefinedBy* annotation records where this OWL class is defined,

- the *axiom* annotation records the MILO axiom,

- the *comment* annotation provides the comment in the specific language,

- the *equivalenceRelation* annotation documents that this has an equivalence relation to a WordNet identifier or ILI,

- the *externalImage* annotation links to a URI of an image for this concept or class,

- the *subsumingRelation* annotation shows other links to WordNet identifiers or ILIs that are subsumed by this class.
5.6 Applying ontology comparison

An ontological tree comparison measure has been proposed for measuring the similarity of concept trees as discussed in Chapter 4 (Xue et al., 2009). Their definitions have been reused 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 take into account the position and conceptual similarities of the affected nodes that must be considered in a comparison process. They achieve this by defining four distinct tree transformation operations, each of which has a different transformation cost. Of interest are the insert, move and relabelling operations. The reason for using these costs is that at the completion of all the research one could determine a final transformation cost. The final cost will measure the transformation of the resultant Bantu language core concept tree to the corresponding Global WordNet Base Concept tree. This could also be applied to the mapping from the African language Northern Sotho WordNet prototype to SUMO.

Before the calculation of the measure in Xue et al. (2009) can be executed a number of steps to prepare the data need to be completed. This is termed pre-processing. The algorithm as described by Xue (2010) was used as shown in two separate phases. Algorithm 1 represents the algorithm for the initial pre-processing steps. Once this is completed then it is followed by the required transforming phase of the algorithm presented in Algorithm 2. The pre-processing includes finding the nodes that need to be deleted and inserted. The transform-
Figure 5.4: Protegé
ing phase applies an exhaustive method for trying every possible transformation sequence to obtain the minimal cost. The transforming phase has a backup and restore operation to ensure a common starting point each time a new operation sequence is tried to determine the minimal cost. This algorithm can similarly also be used to compute the cost of transforming $T_2$ into $T_1$, so that the similarity index of $T_1$ and $T_2$ can be determined (Xue, 2010).

\begin{algorithm}
\begin{algorithmic}[1]
\State \textbf{Input:} Tree $T_1$ and $T_2$; Concept similarity measure set $S_{L_{V_1},L_{V_2}}$
\State \textbf{Output:} Sets of nodes to be deleted, $D$, and inserted, $I$
\State \textbf{begin}
\State \hspace{1em} $D = \emptyset$
\State \hspace{1em} \textbf{foreach} node $u$ in $V_1$ \textbf{do}
\State \hspace{2em} \textbf{if} not exists any $l$ in $L_{V_2}$ such that $M_1(u) = l$ \textbf{then}
\State \hspace{3em} \textbf{if} not exists any $s(M_1(u), l)$ in $S_{L_{V_1},L_{V_2}}$ \textbf{then}
\State \hspace{4em} add $u$ into $D$;
\State \hspace{2em} \textbf{end}
\State \hspace{1em} \textbf{end}
\State \hspace{1em} $I = \emptyset$
\State \hspace{1em} \textbf{foreach} node $v$ in $V_2$ \textbf{do}
\State \hspace{2em} \textbf{if} not exists any $l$ in $L_{V_1}$ such that $M_2(v) = l$ \textbf{then}
\State \hspace{3em} \textbf{if} not exists any $s(l, M_2(v))$ in $S_{L_{V_1},L_{V_2}}$ \textbf{then}
\State \hspace{4em} add $v$ into $I$;
\State \hspace{2em} \textbf{end}
\State \hspace{1em} \textbf{end}
\State \hspace{1em} \textbf{return} $D$ and $I$
\State \textbf{end}
\end{algorithmic}
\end{algorithm}

\textbf{Algorithm 1:} The transformation pre-processing phase
5.7 Methodological questions

As discussed in Section 1.4 there are a number of questions that could arise from the methodological approach, but these have been excluded from the scope of this research for various reasons:

1. Ontological comparison using the method of Xue (2010); Xue et al. (2009) was only performed in respect of nominal concepts. No theoretical framework for performing it in respect of adjectives or verbs exists.

2. The questions on the linguistic accuracy and usefulness of BLR3 are not discussed, but the concepts are used since they have been shown to be broadly representative of the Bantu language concepts.

3. Questions about the applicability of current ontology approaches in philosophy itself to African thought, are not treated. It is assumed that it would be worthwhile to examine if upper ontologies are universally representative from a computational perspective.
**Input:** Tree $T_1$ and $T_2$; $D, I$; Concept similarity measure set $S_{L_{V_1}, L_{V_2}}$

**Output:** $\gamma(T_1 \rightarrow T_2)$

1. begin
2. find all permutations composed by elements in $D \cup I$ and store in $P$;
3. $\text{transformCost} = +\infty$;
4. foreach permutation $p$ in $P$ do
   5. backup $T_1$ and $T_2$;
   6. $\text{editCost} = 0$;
   7. foreach element $u$ in $p$ do
      8. perform deletion (if $u \in D$) or insertion (if $u \in I$) on $u$ if applicable;
      9. $\text{editCost} = \text{editCost} + (\gamma(\text{delete}(u)) \text{ or } \gamma(\text{insert}(u)))$;
   10. end
   11. foreach $u$ in $V_1$ but not in $p$ do
      12. /* handle the nodes to be moved */
      13. if exists $l$ in $L_{V_2}$ such that $M_1(u) = l$ or exists any $s(M_1(u), l)$ in $S_{L_{V_1}, L_{V_2}}$ then
         14. if $M_1(\text{parent}(u)) \neq M_2(\text{parent}(M_2 - 1(l)))$ and not exists any $s(M_1(\text{parent}(u)), l)$ in $S_{L_{V_1}, L_{V_2}}$ then
             15. perform moving on $u$;
             16. $\text{editCost} = \text{editCost} + \gamma(\text{move}(u))$;
         17. end
      18. end
   19. end
   20. foreach $u$ in $V_1$ but not in $p$ do
      21. /* handle the nodes to be re-labelled */
      22. if exists $l$ in $L_{V_2}$ such that exists any $s(M_1(u), l)$ in $S_{L_{V_1}, L_{V_2}}$ then
         23. perform re-labelling on $u$;
         24. $\text{editCost} = \text{editCost} + \gamma(\text{relabel}(u))$;
      25. end
   26. end
   27. $\text{transformCost} = \min(\text{transformCost}, \text{editCost})$;
   28. restore $T_1$ and $T_2$;
   29. return $\text{transformCost}$
30. end

**Algorithm 2:** The transformation cost computing phase
Part III

Contribution and conclusion
CHAPTER 6

Results

And that ... is why nothing in Nature is quite regular. There are always exceptions. A good average uniformity, but not complete.

Lewis (1943)

6.1 Introduction

The research questions, the context and the approach to the research have now been concluded. Whereas the first few chapters provided the answers to the first research questions on the state of the art, the following chapters provide the results of the research and the final answers to the research questions.

This chapter documents the results of the research. The mapping of core concepts to upper ontologies has been applied to the Bantu languages - a new African linguistic base. The approach to this mapping was discussed in Chapter 5. The results cannot be presented without choosing a method of ontological
comparison. The approach to the comparison was presented in Chapter 4. In the results presented here two research questions are explored in detail: is the mapping comprehensive and does the mapping indicate whether SUMO is universally representative?

6.2 Final word list

This final resultant word list of 67 is shown in Table 6.1. For details of how this list is derived refer to the descriptions in Chapter 5 and the description and tables in Appendix A. Main Ref refers to the reference for the main entry as described in Section 5.3.

Table 6.1: BLR roots and meanings

<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-bûł</td>
<td>5841</td>
<td>‘bad’</td>
</tr>
<tr>
<td>-bá-</td>
<td>4</td>
<td>‘to dwell; to be; to become’</td>
</tr>
<tr>
<td>-báb-</td>
<td>5</td>
<td>‘to be bitter; to be smart; to itch; to be sharp; to sting; to hurt’</td>
</tr>
<tr>
<td>-bâdî</td>
<td>36</td>
<td>‘two’</td>
</tr>
<tr>
<td>-bòd-</td>
<td>253</td>
<td>‘to be rotten’</td>
</tr>
<tr>
<td>-bûdà</td>
<td>368</td>
<td>‘rain’</td>
</tr>
<tr>
<td>-bûà</td>
<td>282</td>
<td>‘dog’</td>
</tr>
<tr>
<td>-dí</td>
<td>944</td>
<td>‘to eat’</td>
</tr>
</tbody>
</table>

Continued on next page

1. Research questions 3 and 4 in section 1.2.
Table 6.1 – continued from previous page

<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-dímì</td>
<td>973</td>
<td>‘tongue; language; flame’</td>
</tr>
<tr>
<td>-dúm-</td>
<td>1181</td>
<td>‘to bite’</td>
</tr>
<tr>
<td>-dì</td>
<td>940</td>
<td>‘to be’</td>
</tr>
<tr>
<td>-dúd-</td>
<td>959</td>
<td>‘to weep; to shout; to wail’</td>
</tr>
<tr>
<td>-dá</td>
<td>780</td>
<td>‘louse’</td>
</tr>
<tr>
<td>-dài</td>
<td>3705</td>
<td>‘long’</td>
</tr>
<tr>
<td>-dèdù</td>
<td>897</td>
<td>‘beard; chin’</td>
</tr>
<tr>
<td>-dibà</td>
<td>1025</td>
<td>‘pool; pond; deep water; well’</td>
</tr>
<tr>
<td>-dúad-</td>
<td>1234</td>
<td>‘to wear’</td>
</tr>
<tr>
<td>-gí</td>
<td>1368</td>
<td>‘egg’</td>
</tr>
<tr>
<td>-gàngà</td>
<td>1332</td>
<td>‘medicine man’</td>
</tr>
<tr>
<td>-gèd-</td>
<td>1345</td>
<td>‘to try’</td>
</tr>
<tr>
<td>-gènd-</td>
<td>1362</td>
<td>‘to walk; to travel’</td>
</tr>
<tr>
<td>-gùdùbè</td>
<td>1494</td>
<td>‘pig’</td>
</tr>
<tr>
<td>-kúd-</td>
<td>1997</td>
<td>‘to grow up’</td>
</tr>
<tr>
<td>-kúmì</td>
<td>2027</td>
<td>‘ten’</td>
</tr>
<tr>
<td>-kúnì</td>
<td>2042</td>
<td>‘firewood’</td>
</tr>
<tr>
<td>-kádà</td>
<td>1662</td>
<td>‘ember; charcoal’</td>
</tr>
<tr>
<td>-kádàng-</td>
<td>1665</td>
<td>‘to fry, to roast’</td>
</tr>
<tr>
<td>-kángà</td>
<td>1720</td>
<td>‘guinea fowl’</td>
</tr>
</tbody>
</table>

Continued on next page
Table 6.1 – continued from previous page

<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-kúdà</td>
<td>1793</td>
<td>‘tail’</td>
</tr>
<tr>
<td>-kókó</td>
<td>1904</td>
<td>‘chicken’</td>
</tr>
<tr>
<td>-kómb-</td>
<td>1916</td>
<td>‘to scrape; to dig; to lick with finger’</td>
</tr>
<tr>
<td>-kót-</td>
<td>7350</td>
<td>‘to stoop; to be bent’</td>
</tr>
<tr>
<td>-kú-</td>
<td>2089</td>
<td>‘to die’</td>
</tr>
<tr>
<td>-kúm-</td>
<td>2113</td>
<td>‘to be honoured; to be rich’</td>
</tr>
<tr>
<td>-kúpá</td>
<td>2071</td>
<td>‘tick; insect’</td>
</tr>
<tr>
<td>-jádà</td>
<td>1558</td>
<td>‘finger-nail, toe-nail, claw’</td>
</tr>
<tr>
<td>-jádà</td>
<td>1555</td>
<td>‘hunger; famine’</td>
</tr>
<tr>
<td>-jákà</td>
<td>3169</td>
<td>‘year; cultivation season; harvest’</td>
</tr>
<tr>
<td>-jánà</td>
<td>3203</td>
<td>‘child’</td>
</tr>
<tr>
<td>-jánuk-</td>
<td>3206</td>
<td>‘to spread to dry in the sun; to spread out’</td>
</tr>
<tr>
<td>-jéd-</td>
<td>3273</td>
<td>‘to shine; to be clear; to be ripe; to be favourable’</td>
</tr>
<tr>
<td>-jícò</td>
<td>3405</td>
<td>‘eye’</td>
</tr>
<tr>
<td>-jídà</td>
<td>1593</td>
<td>‘path’</td>
</tr>
<tr>
<td>-jíkì</td>
<td>3350</td>
<td>‘bee’</td>
</tr>
<tr>
<td>-jíkì</td>
<td>3442</td>
<td>‘smoke’</td>
</tr>
<tr>
<td>-jíkút-</td>
<td>3445</td>
<td>‘to be satiated’</td>
</tr>
<tr>
<td>-júmb-</td>
<td>3361</td>
<td>‘to sing; to dance’</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-jíñà</td>
<td>3464</td>
<td>‘name’</td>
</tr>
<tr>
<td>-jíŋí</td>
<td>3485</td>
<td>‘many, much’</td>
</tr>
<tr>
<td>-jípí</td>
<td>3495</td>
<td>‘short’</td>
</tr>
<tr>
<td>-jókà</td>
<td>3536</td>
<td>‘snake; intestinal worm’</td>
</tr>
<tr>
<td>-jót-</td>
<td>3579</td>
<td>‘to warm oneself’</td>
</tr>
<tr>
<td>-nà</td>
<td>3674</td>
<td>‘with; and’</td>
</tr>
<tr>
<td>-ncè</td>
<td>500</td>
<td>‘all’</td>
</tr>
<tr>
<td>-ntū</td>
<td>4807</td>
<td>‘some (entity); any’</td>
</tr>
<tr>
<td>-nyàmà</td>
<td>3180</td>
<td>‘animal; meat’</td>
</tr>
<tr>
<td>-nyó-</td>
<td>7047</td>
<td>‘to drink’</td>
</tr>
<tr>
<td>-pá-</td>
<td>2344</td>
<td>‘to give’</td>
</tr>
<tr>
<td>-pácà</td>
<td>2348</td>
<td>‘twin’</td>
</tr>
<tr>
<td>-pàp-</td>
<td>2407</td>
<td>‘to flap wings; to flutter’</td>
</tr>
<tr>
<td>-pèp</td>
<td>2463</td>
<td>‘to blow as wind; to winnow; to smoke tobacco; to breathe’</td>
</tr>
<tr>
<td>-pí</td>
<td>2491</td>
<td>‘to be burnt; to be hot; to be cooked; to be ripe; to ferment; to be red’</td>
</tr>
<tr>
<td>-pód-</td>
<td>2589</td>
<td>‘to be cold; to cool down; to be quiet’</td>
</tr>
<tr>
<td>-túng-</td>
<td>3081</td>
<td>‘to put through; to thread on string; to plait; to sew; to tie up; to build; to close in’</td>
</tr>
</tbody>
</table>

Continued on next page
Table 6.1 – continued from previous page

<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-tátù</td>
<td>2811</td>
<td>‘three’</td>
</tr>
<tr>
<td>-tí</td>
<td>2881</td>
<td>‘tree stick’</td>
</tr>
<tr>
<td>-túd-</td>
<td>3101</td>
<td>‘to hammer; to forge’</td>
</tr>
</tbody>
</table>

6.3 Qualitative comparison results

The final quality assured concept list described in Section 5.3 was analysed. A subset of this Bantu concept list is shown in Table 6.1 with a sample shown in Table 6.2. In Table 6.2 the heading *Proto-Bantu* refers to the original root concept that has been attested in all 16 Bantu languages zones, including Zones A and S. It has been verified that such roots have local Northern Sotho lexicalizations. The *BLR3 reference* is the number for the proto-Bantu root in the CBOLD project. The *attested meaning* is the meaning provided by Maho (2005). The POS indicates the part of speech of the proto-Bantu root. The *WordNet sense* is the English Princeton WordNet closest equivalent mapped via the ILI. The *tree operation* indicates the base operation required to calculate the ontological similarity measurement. *Word* is the noun stem, verb root or adjectival root 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 set specifically because there are
different list groupings in BCS\textsuperscript{2}. Set membership in the BCS includes being a member in the Global WordNet Core Concept list (Vossen and Fellbaum, 2014a). The \textit{SUMO domain} is the mapping of the concept to SUMO as provided via the ILI link to Princeton WordNet. The \textit{SUMO operation} indicates the WordNet mapping operation to SUMO and the \textit{SUMO node} indicates the mapped node.

\textsuperscript{2} Refer to section 3.4 for further detail on the BalkaNet Core Set.
<table>
<thead>
<tr>
<th>Proto-Bantu</th>
<th>BLR3 Ref</th>
<th>Attested and/or reconstructed meaning</th>
<th>POS</th>
<th>WordNet sense</th>
<th>Tree Operation</th>
<th>Word or Stem</th>
<th>Core Set</th>
<th>SUMO Domain</th>
<th>SUMO Operation</th>
<th>SUMO Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>-jánà</td>
<td>3203</td>
<td>‘child’</td>
<td>n</td>
<td>Child:2</td>
<td>re-labelling</td>
<td>ngwana</td>
<td>1</td>
<td>person</td>
<td>+</td>
<td>Human</td>
</tr>
<tr>
<td>-jólà</td>
<td>3536</td>
<td>‘snake; intestinal worm’</td>
<td>n</td>
<td>Snake:1</td>
<td>re-labelling</td>
<td>noga</td>
<td>3</td>
<td>zoology</td>
<td>=</td>
<td>Snake</td>
</tr>
<tr>
<td>-jíkù</td>
<td>3350</td>
<td>‘bee’</td>
<td>n</td>
<td>Bee:1</td>
<td>re-labelling</td>
<td>nose</td>
<td>2</td>
<td>entomology</td>
<td>=</td>
<td>Bee</td>
</tr>
<tr>
<td>-ntù</td>
<td>4807</td>
<td>‘some (entity); any’</td>
<td>n</td>
<td>Person:1</td>
<td>re-labelling</td>
<td>motha</td>
<td>1</td>
<td>biology</td>
<td>=</td>
<td>Human</td>
</tr>
<tr>
<td>-jíngí</td>
<td>3485</td>
<td>‘many, much’</td>
<td>adj</td>
<td>Many:1</td>
<td>re-labelling</td>
<td>-ntši</td>
<td>None</td>
<td>factotum</td>
<td>=</td>
<td>Subjective</td>
</tr>
<tr>
<td>-nyó-</td>
<td>7047</td>
<td>‘to drink’</td>
<td>v</td>
<td>Drink:1</td>
<td>re-labelling</td>
<td>-nwa</td>
<td>1</td>
<td>alimentation</td>
<td>=</td>
<td>Beverage</td>
</tr>
<tr>
<td>-jó-</td>
<td>3579</td>
<td>‘to warm oneself’</td>
<td>v</td>
<td>Bask:2</td>
<td>re-labelling</td>
<td>-ora</td>
<td>None</td>
<td>factotum</td>
<td>+</td>
<td>Process</td>
</tr>
</tbody>
</table>

Table 6.2: Sample BLR roots and meanings
A verb sample list is shown in Table 6.3, adjectival roots in Table 6.4 and nouns in Table 6.5. Variant Ref refers to the reference for the variant entry, if it exists, as described in Section 5.3. POS refers to the part of speech as defined in the Comprehensive Northern Sotho Dictionary (Ziervogel and Mokgokong, 1985). WordNet Sense refers to the synset that corresponds in Princeton WordNet to the root.

Table 6.3: BLR verb roots and meanings

<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Variant Ref</th>
<th>Attested and/or reconstructed meaning</th>
<th>POS</th>
<th>WordNet Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>-jánk-</td>
<td>3206</td>
<td></td>
<td>‘to spread to dry in the sun; to spread out’</td>
<td>v</td>
<td>Air:1</td>
</tr>
<tr>
<td>-bá-</td>
<td>4</td>
<td></td>
<td>‘to dwell; to be; to become’</td>
<td>v</td>
<td>Be:1</td>
</tr>
<tr>
<td>-báb-</td>
<td>5</td>
<td></td>
<td>‘to be bitter; to be smart; to itch; to be sharp; to sting; to hurt’</td>
<td>v</td>
<td>Bitter:1(^3)</td>
</tr>
</tbody>
</table>

Continued on next page

---

3. This is mapped to the verb sense in WordNet which is to be bitter. In the Bantu languages
Table 6.3 – continued from previous page

<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Variant Ref</th>
<th>Attested and/or reconstructed meaning</th>
<th>POS</th>
<th>Wordnet Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>-júmb-</td>
<td>3361</td>
<td>244</td>
<td>'to sing; to dance’</td>
<td>v</td>
<td>Dance:1</td>
</tr>
<tr>
<td>-bòd-</td>
<td>253</td>
<td></td>
<td>'to be rotten’</td>
<td>v</td>
<td>Rotten:3</td>
</tr>
<tr>
<td>-dí</td>
<td>940</td>
<td></td>
<td>'to be’</td>
<td>v</td>
<td>Do:1</td>
</tr>
<tr>
<td>-gènd-</td>
<td>1362</td>
<td></td>
<td>'to walk; to travel’</td>
<td>v</td>
<td>Walk:1</td>
</tr>
<tr>
<td>-pá-</td>
<td>2344</td>
<td></td>
<td>'to give’</td>
<td>v</td>
<td>Give:3</td>
</tr>
<tr>
<td>-pép</td>
<td>2463</td>
<td></td>
<td>'to blow as wind; to winnow; to smoke tobacco; to breathe’</td>
<td>v</td>
<td>Winnow:1</td>
</tr>
</tbody>
</table>

Continued on next page

such concepts, although sometimes adjectives in English, are expressed in a verbal structure. Therefore the part of speech is a verb.
Table 6.3 – continued from previous page

<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Variant Ref</th>
<th>Attested and/or reconstructed meaning</th>
<th>POS</th>
<th>Wordnet Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>-pí</td>
<td>2491</td>
<td></td>
<td>'to be burnt; to be hot; to be cooked; to be ripe; to ferment; to be red'</td>
<td>v</td>
<td>Heat:1</td>
</tr>
<tr>
<td>-pód-</td>
<td>2589</td>
<td></td>
<td>'to be cold; to cool down; to be quiet'</td>
<td>v</td>
<td>Cool:1</td>
</tr>
<tr>
<td>-kádàng-</td>
<td>1665 1680</td>
<td></td>
<td>'to fry, to roast'</td>
<td>v</td>
<td>Roast:1</td>
</tr>
<tr>
<td>-kúd-</td>
<td>1997</td>
<td></td>
<td>'to grow up'</td>
<td>v</td>
<td>Grow:2</td>
</tr>
<tr>
<td>-kúm-</td>
<td>2113</td>
<td></td>
<td>'to be honoured; to be rich'</td>
<td>v</td>
<td>Enrich:1</td>
</tr>
<tr>
<td>-köt-</td>
<td>7350 1961</td>
<td></td>
<td>'to stoop; to be bent'</td>
<td>v</td>
<td>Stoop:1</td>
</tr>
<tr>
<td>-kú-</td>
<td>2089</td>
<td></td>
<td>'to die'</td>
<td>v</td>
<td>Die:1</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Variant Ref</th>
<th>Attested and/or re-constructed meaning</th>
<th>POS</th>
<th>Wordnet Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>-dí</td>
<td>944</td>
<td></td>
<td>‘to eat’</td>
<td>v</td>
<td>Eat:1</td>
</tr>
<tr>
<td>-kómbo-</td>
<td>1916</td>
<td></td>
<td>‘to scrape; to dig; to lick with finger’</td>
<td>v</td>
<td>Dig:1</td>
</tr>
<tr>
<td>-díð-</td>
<td>959</td>
<td></td>
<td>‘to weep; to shout; to wail’</td>
<td>v</td>
<td>Cry:2</td>
</tr>
<tr>
<td>-díým-</td>
<td>1181</td>
<td></td>
<td>‘to bite’</td>
<td>v</td>
<td>Bite:2</td>
</tr>
<tr>
<td>-nýo-</td>
<td>7047</td>
<td></td>
<td>‘to drink’</td>
<td>v</td>
<td>Drink:1</td>
</tr>
<tr>
<td>-jóþ-</td>
<td>3579</td>
<td></td>
<td>‘to warm oneself’</td>
<td>v</td>
<td>Bask:2</td>
</tr>
<tr>
<td>-páþ-</td>
<td>2407</td>
<td></td>
<td>‘to flap wings; to flutter’</td>
<td>v</td>
<td>Flutter:3</td>
</tr>
<tr>
<td>-túþ-</td>
<td>3101</td>
<td></td>
<td>‘to hammer; to forge’</td>
<td>v</td>
<td>Smelt:1</td>
</tr>
<tr>
<td>-díýad-</td>
<td>1234</td>
<td></td>
<td>‘to wear’</td>
<td>v</td>
<td>Carry:2</td>
</tr>
<tr>
<td>-jíkúþ-</td>
<td>3445</td>
<td></td>
<td>‘to be satiated’</td>
<td>v</td>
<td>Appease:2</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Variant Ref</th>
<th>Attested and/or re-constructed meaning</th>
<th>POS</th>
<th>Wordnet Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>-gèd-</td>
<td>1345</td>
<td>'to try'</td>
<td>v</td>
<td>Try:1</td>
<td></td>
</tr>
<tr>
<td>-túng-</td>
<td>3081</td>
<td>'to put through; to thread on string; to plait; to sew; to tie up; to build; to close in'</td>
<td>v</td>
<td>Plait:1</td>
<td></td>
</tr>
<tr>
<td>-nà</td>
<td>3674</td>
<td>'with; and'</td>
<td>v</td>
<td>Attach To:1</td>
<td></td>
</tr>
<tr>
<td>-jéd-</td>
<td>3273</td>
<td>'to shine; to be clear; to be ripe; to be favourable'</td>
<td>v</td>
<td>Twinkle:1</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.3 – continued from previous page
Table 6.4: BLR adjective roots and meanings

<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Variant Ref</th>
<th>Attested and/or reconstructed meaning</th>
<th>POS</th>
<th>WordNet Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>-bû</td>
<td>5841</td>
<td></td>
<td>‘bad’</td>
<td>adj</td>
<td>Bad:1</td>
</tr>
<tr>
<td>-bàdî</td>
<td>36</td>
<td>190</td>
<td>‘two’</td>
<td>adj</td>
<td>Two:1</td>
</tr>
<tr>
<td>-jîpî</td>
<td>3495</td>
<td>2133</td>
<td>‘short’</td>
<td>adj</td>
<td>Short:2</td>
</tr>
<tr>
<td>-jîngî</td>
<td>3485</td>
<td></td>
<td>‘many, much’</td>
<td>adj</td>
<td>Many:1</td>
</tr>
<tr>
<td>-ncè</td>
<td>500</td>
<td>499</td>
<td>‘all’</td>
<td>adj</td>
<td>Whole:1</td>
</tr>
<tr>
<td>-tâtû</td>
<td>2811</td>
<td></td>
<td>‘three’</td>
<td>adj</td>
<td>Three:1</td>
</tr>
<tr>
<td>-kûmî</td>
<td>2027</td>
<td></td>
<td>‘ten’</td>
<td>adj</td>
<td>Ten:1</td>
</tr>
<tr>
<td>-dàî</td>
<td>3705</td>
<td></td>
<td>‘long’</td>
<td>adj</td>
<td>Long:1</td>
</tr>
</tbody>
</table>
Table 6.5: BLR noun stems and meanings

<table>
<thead>
<tr>
<th>Stem</th>
<th>Main Ref</th>
<th>Variant Ref</th>
<th>Attested and/or reconstructed meaning</th>
<th>POS</th>
<th>Wordnet Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>-dibà</td>
<td>1025</td>
<td></td>
<td>'pool; pond; deep water; well'</td>
<td>n</td>
<td>Pool:2</td>
</tr>
<tr>
<td>-gì</td>
<td>1368</td>
<td></td>
<td>'egg'</td>
<td>n</td>
<td>Egg:2</td>
</tr>
<tr>
<td>-pácà</td>
<td>2348</td>
<td></td>
<td>'twin'</td>
<td>n</td>
<td>Twin:1</td>
</tr>
<tr>
<td>-kádà</td>
<td>1662</td>
<td></td>
<td>'ember; charcoal'</td>
<td>n</td>
<td>Ember:1</td>
</tr>
<tr>
<td>-kùni</td>
<td>2042</td>
<td></td>
<td>'firewood'</td>
<td>n</td>
<td>Firewood:1</td>
</tr>
<tr>
<td>-jìcò</td>
<td>3405</td>
<td></td>
<td>'eye'</td>
<td>n</td>
<td>Eye:1</td>
</tr>
<tr>
<td>-jìnà</td>
<td>3464</td>
<td></td>
<td>'name'</td>
<td>n</td>
<td>Name:1</td>
</tr>
<tr>
<td>-kàngà</td>
<td>1720</td>
<td></td>
<td>'guinea fowl'</td>
<td>n</td>
<td>Numida meleagris:1</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Stem</th>
<th>Main Ref</th>
<th>Variant Ref</th>
<th>Attested and/or reconstructed meaning</th>
<th>POS</th>
<th>WordNet Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>-kúpá</td>
<td>2071</td>
<td>‘tick; insect’</td>
<td>n</td>
<td>Tick:2</td>
<td></td>
</tr>
<tr>
<td>-kókó</td>
<td>1904</td>
<td>‘chicken’</td>
<td>n</td>
<td>Poultry:2</td>
<td></td>
</tr>
<tr>
<td>-dèdù</td>
<td>897</td>
<td>‘beard; chin’</td>
<td>n</td>
<td>Beard:1</td>
<td></td>
</tr>
<tr>
<td>-dímì</td>
<td>973</td>
<td>‘tongue; language; flame’</td>
<td>n</td>
<td>Tongue:1</td>
<td></td>
</tr>
<tr>
<td>-gùdùbè</td>
<td>1494</td>
<td>‘pig’</td>
<td>n</td>
<td>Pig:1</td>
<td></td>
</tr>
<tr>
<td>-búà</td>
<td>282</td>
<td>‘dog’</td>
<td>n</td>
<td>Dog:1</td>
<td></td>
</tr>
<tr>
<td>-jádà</td>
<td>1558</td>
<td>‘finger-nail, toe-nail, claw’</td>
<td>n</td>
<td>Unguis:1</td>
<td></td>
</tr>
<tr>
<td>-nyàmà</td>
<td>3180</td>
<td>‘animal; meat’</td>
<td>n</td>
<td>Meat:1</td>
<td></td>
</tr>
<tr>
<td>-gàngà</td>
<td>1332</td>
<td>‘medicine man’</td>
<td>n</td>
<td>Sangoma:1</td>
<td></td>
</tr>
</tbody>
</table>

Continued on next page
Table 6.5 – continued from previous page

<table>
<thead>
<tr>
<th>Stem</th>
<th>Main Ref</th>
<th>Variant Ref</th>
<th>Attested and/or reconstructed meaning</th>
<th>POS</th>
<th>WordNet Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>-jákà</td>
<td>3169</td>
<td></td>
<td>'year; cultivation season; harvest'</td>
<td>n</td>
<td>Year:2</td>
</tr>
<tr>
<td>-jánà</td>
<td>3203</td>
<td></td>
<td>'child'</td>
<td>n</td>
<td>Child:2</td>
</tr>
<tr>
<td>-jókà</td>
<td>3536</td>
<td></td>
<td>'snake; intestinal worm'</td>
<td>n</td>
<td>Snake:1</td>
</tr>
<tr>
<td>-jíkì</td>
<td>3350</td>
<td>1622</td>
<td>'bee'</td>
<td>n</td>
<td>Bee:1</td>
</tr>
<tr>
<td>-ntù</td>
<td>4807</td>
<td></td>
<td>'some (entity); any'</td>
<td>n</td>
<td>Person:1</td>
</tr>
<tr>
<td>-búdà</td>
<td>368</td>
<td></td>
<td>'rain'</td>
<td>n</td>
<td>Rain:1</td>
</tr>
<tr>
<td>-tì</td>
<td>2881</td>
<td></td>
<td>'tree stick'</td>
<td>n</td>
<td>Branch:2</td>
</tr>
<tr>
<td>-kúdà</td>
<td>1793</td>
<td></td>
<td>'tail'</td>
<td>n</td>
<td>Tail:1</td>
</tr>
<tr>
<td>-jíkì</td>
<td>3442</td>
<td></td>
<td>'smoke'</td>
<td>n</td>
<td>Smoke:1</td>
</tr>
</tbody>
</table>

4. Note that although this variant was found it is indicated as a refused reconstruction by BLR3

Continued on next page

137
<table>
<thead>
<tr>
<th>Stem</th>
<th>Main Ref</th>
<th>Variant Ref</th>
<th>Attested and/or reconstructed meaning</th>
<th>POS</th>
<th>WordNet Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>-dá</td>
<td>780</td>
<td></td>
<td>‘louse’</td>
<td>n</td>
<td>Louse:1</td>
</tr>
<tr>
<td>-jàdà</td>
<td>1555</td>
<td></td>
<td>‘hunger; famine’</td>
<td>n</td>
<td>Hunger:1</td>
</tr>
<tr>
<td>-jìdà</td>
<td>1593</td>
<td></td>
<td>‘path’</td>
<td>n</td>
<td>Path:1</td>
</tr>
</tbody>
</table>
6.3.1 Sense mapping with WordNet

The majority (62 or 93%) of the 67 concepts map to an English Princeton WordNet concept that has already been defined. Mapping means that the major sense of the word (the first listed sense of the word) in at least the 2 most authoritative dictionaries (Kriel et al., 2003; Ziervogel and Mokgokong, 1985) in a lexicalized form (Northern Sotho) has one-to-one synonymy with a Princeton WordNet sense. For example, ‘-dibâ’, which BLR3 represents as the noun for ‘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 root ‘-jánik-’, 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 ‘-jíngú’, which BLR3 represents as ‘many, much’ and which is lexicalized in Northern Sotho as ntšhi:1, maps to the Princeton WordNet adjective sense many:1. This one-to-one mapping is referred to as “re-labelling” in the context of ontological comparison measure as described in Chapter 4.

If we consider more complicated sense mappings (the remaining 7%), then there are 3 other potential scenarios - insert, move, and combinations of insert and move. This is as a result of either the concept not having become fully lexicalized in English (insert), or the Northern Sotho sense, when compared to the English equivalent sense, does not align with the current position of that English sense in Princeton WordNet (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 *fê-fera*: and described by the Comprehensive Northern Sotho Dictionary as the primary sense: ‘winnow (stamped corn is shaken in a lesêlô until the chaff lies on top)’ (Ziervogel and Mokgokong, 1985). 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:* which has the sense *Numida coronata, crowned guinea-fowl* in the Comprehensive Northern Sotho Dictionary (Ziervogel and Mokgokong, 1985). The complexity is because this should be inserted as a 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 synset hypernym tree *bird*, with hyponym *gallinaceous bird* and further hyponym *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). These authors have shown that for animals there is ontological confusion in WordNet between types (landfowl versus waterfowl) and rôles (domestic fowl versus gamefowl).

Apart from the 67 quality assured concepts, there were other concepts that were inserted into the African language WordNet prototype 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, and is distinct from the primary sense of partridge in English. 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 the use of these examples is not to detract from the broader fit to the BCs, but merely to highlight that there will be obvious divergence for Africa 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) “-kúpá”, 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 the ILI, but has the hypernym structure Parasitiformes:1/kgofa:2, kgofa:1 rather than the current Princeton WordNet hypernym structure arachnid:1, acarine:1, tick:2

The sense mappings 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.

### 6.3.2 Mapping of BLR3 with Balkanet common synsets

The mapping of the Bantu language concepts in this research (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 as well as the Zones in between) to the Global WordNet BCs is not as good as the individual word sense mapping 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 adjectival root 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.

6.3.3 Mapping of BLR3 with Global 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.

5. Refer to Section 3.4 for a description of BCS
The 5000 Balkanet Common Synsets include all the original EuroWordNet and Global WordNet BCs. The mismatch of 49% of the concepts mentioned in Section 6.3.2 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.

### 6.3.4 Top Ontology comparison

EuroWordNet defined 3 different order entity types for the Top Ontology (refer to Section 3.8). In summary these are:

1. **1st Order** – any concrete entity publicly perceivable by the senses and located at any point in time, in a three-dimensional space.

2. **2nd Order** – any static situation (property, relation) or dynamic situation, which cannot be grasped, heard, seen, felt as an independent physical thing. They occur or take place rather than exist.

3. **3rd Order** – unobservable propositions which exist independently of time and space. They can be true or false rather than real. They can be asserted or denied, remembered or forgotten.

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 6.1).

Section 1.2 emphasizes that, to answer the research question, it is critical to investigate the state of the art of mappings from other, specifically non-Indo-
European, language family concepts, to upper ontology concepts. Significant work in regard to mapping to the WordNet Top Ontology has been done with Chinese WordNet. This lack of mappings to 3rd order entities, which is a result of this work, 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, 2001). The large amount of mappings to 1st order entities in this research similarly corresponds to the previous results on linkage of Chinese BCs to WordNet (Pala and Wong, 2001). For illustrative purposes I represent some of my results mapped to Chinese radicals, which might be useful for further research. Table 6.6 shows the amount of mappings from the original Proto-Bantu word list in Table A.1 to the Kángxī Chinese Radicals shown by standard number and actual radical. Pīnyīn is the official phonetic system for transcribing the Mandarin pronunciations of Chinese characters into the Latin alphabet in mainland China, Taiwan and Singapore.
Table 6.6: BLR3 to Kāngxī radical mapping

<table>
<thead>
<tr>
<th>Latex</th>
<th>BLR3 Main Ref</th>
<th>Pinyin</th>
<th>Kāngxī Radical Number</th>
<th>Kāngxī Radical</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-kúmb-</td>
<td>2120</td>
<td>piě</td>
<td>4</td>
<td>丿</td>
<td>‘to bend’</td>
</tr>
<tr>
<td>-bàdí</td>
<td>36</td>
<td>èr</td>
<td>7</td>
<td>二</td>
<td>‘two’</td>
</tr>
<tr>
<td>-ntū</td>
<td>4807</td>
<td>rén</td>
<td>9</td>
<td>人</td>
<td>‘some entity; any’</td>
</tr>
<tr>
<td>-kúmì</td>
<td>2027</td>
<td>shí</td>
<td>24</td>
<td>十</td>
<td>‘ten’</td>
</tr>
<tr>
<td>-cí</td>
<td>562</td>
<td>tǔ</td>
<td>32</td>
<td>土</td>
<td>‘ground; country; underneath’</td>
</tr>
<tr>
<td>-kádí</td>
<td>1674</td>
<td>nˇü</td>
<td>38</td>
<td>女</td>
<td>‘woman; wife’</td>
</tr>
<tr>
<td>-jánà</td>
<td>3203</td>
<td>zǐ</td>
<td>39</td>
<td>子</td>
<td>‘child’</td>
</tr>
<tr>
<td>-jéné</td>
<td>3296</td>
<td>jǐ</td>
<td>49</td>
<td>己</td>
<td>‘self; same’</td>
</tr>
<tr>
<td>-jípí</td>
<td>3495</td>
<td>yǎo</td>
<td>52</td>
<td>ㄠ</td>
<td>‘short’</td>
</tr>
<tr>
<td>-táà</td>
<td>9207</td>
<td>gōng</td>
<td>57</td>
<td>弓</td>
<td>‘bow’</td>
</tr>
<tr>
<td>-dèdú</td>
<td>897</td>
<td>shān</td>
<td>59</td>
<td>𠍇</td>
<td>‘beard; chin’</td>
</tr>
<tr>
<td>-tí</td>
<td>2881</td>
<td>zhī</td>
<td>65</td>
<td>支</td>
<td>‘tree stick’</td>
</tr>
<tr>
<td>-júbà</td>
<td>1614</td>
<td>rì</td>
<td>72</td>
<td>日</td>
<td>‘sun’</td>
</tr>
<tr>
<td>-kú-</td>
<td>2089</td>
<td>dāi</td>
<td>78</td>
<td>死</td>
<td>‘to die’</td>
</tr>
<tr>
<td>-pép</td>
<td>2463</td>
<td>qì</td>
<td>84</td>
<td>气</td>
<td>‘to blow as wind; to winnow; to smoke tobacco; to breathe’</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Latex</th>
<th>BLR3 Main Ref</th>
<th>Pinyin</th>
<th>Kângxî Radical Number</th>
<th>Kângxî Radical</th>
<th>Attested and/or re-constructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-jìjì</td>
<td>3433</td>
<td>shuǐ</td>
<td>85</td>
<td>水</td>
<td>`water’</td>
</tr>
<tr>
<td>-jádà</td>
<td>1558</td>
<td>zhǎo</td>
<td>87</td>
<td>爪</td>
<td>`finger-nail, toe-nail, claw’</td>
</tr>
<tr>
<td>-kùnì</td>
<td>2042</td>
<td>qiáng</td>
<td>90</td>
<td>爿</td>
<td>`firewood’</td>
</tr>
<tr>
<td>-gòmbè</td>
<td>1434</td>
<td>niú</td>
<td>93</td>
<td>牛</td>
<td>`cattle’</td>
</tr>
<tr>
<td>-búà</td>
<td>282</td>
<td>quàn</td>
<td>94</td>
<td>犬</td>
<td>`dog’</td>
</tr>
<tr>
<td>-jíd-</td>
<td>6142</td>
<td>xuán</td>
<td>95</td>
<td>玄</td>
<td>`to get dark; to get black’</td>
</tr>
<tr>
<td>-jícò</td>
<td>3405</td>
<td>mù</td>
<td>109</td>
<td>目</td>
<td>`eye’</td>
</tr>
<tr>
<td>-nyàmà</td>
<td>3180</td>
<td>ròu</td>
<td>130</td>
<td>肉</td>
<td>`animal; meat’</td>
</tr>
<tr>
<td>-dímì</td>
<td>973</td>
<td>shé</td>
<td>135</td>
<td>舌</td>
<td>`tongue; language; flame’</td>
</tr>
<tr>
<td>-játò</td>
<td>3252</td>
<td>zhōu</td>
<td>137</td>
<td>舟</td>
<td>`canoe’</td>
</tr>
<tr>
<td>-kúpá</td>
<td>2071</td>
<td>huǐ</td>
<td>142</td>
<td>虫</td>
<td>`tick; insect’</td>
</tr>
<tr>
<td>-dì</td>
<td>940</td>
<td>行</td>
<td>144</td>
<td>行</td>
<td>`to be’</td>
</tr>
<tr>
<td>-kíngó-</td>
<td>1845</td>
<td>yán</td>
<td>149</td>
<td>言</td>
<td>`neck; nape; voice’</td>
</tr>
<tr>
<td>-gùdòbè</td>
<td>1494</td>
<td>shǐ</td>
<td>152</td>
<td>葬</td>
<td>`pig’</td>
</tr>
</tbody>
</table>
Table 6.6 – continued from previous page

<table>
<thead>
<tr>
<th>Latex</th>
<th>BLR3 Main Ref</th>
<th>Pinyin</th>
<th>Kāngxī Radical Number</th>
<th>Kāngxī Radical</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-pǐ</td>
<td>2491</td>
<td>chì</td>
<td>155</td>
<td>赤</td>
<td>‘to be burnt; to be hot; to be cooked; to be ripe; to ferment; to be red’</td>
</tr>
<tr>
<td>-báb-</td>
<td>5</td>
<td>xīn</td>
<td>160</td>
<td>辛</td>
<td>‘to be bitter; to be smart; to itch; to be sharp; to sting; to hurt’</td>
</tr>
<tr>
<td>-ké</td>
<td>7986</td>
<td>chén</td>
<td>161</td>
<td>辰</td>
<td>‘dawn’</td>
</tr>
<tr>
<td>-gènd-</td>
<td>1362</td>
<td>chuò</td>
<td>162</td>
<td>走</td>
<td>‘to walk; to travel’</td>
</tr>
<tr>
<td>-dài</td>
<td>3705</td>
<td>cháng</td>
<td>168</td>
<td>長</td>
<td>‘long’</td>
</tr>
<tr>
<td>-búdà</td>
<td>368</td>
<td>yǔ</td>
<td>173</td>
<td>雨</td>
<td>‘rain’</td>
</tr>
<tr>
<td>-dí</td>
<td>944</td>
<td>shí</td>
<td>184</td>
<td>食</td>
<td>‘to eat’</td>
</tr>
<tr>
<td>-tứè</td>
<td>3023</td>
<td>shǒu</td>
<td>185</td>
<td>首</td>
<td>‘head’</td>
</tr>
<tr>
<td>-kúpà</td>
<td>2132</td>
<td>gǔ</td>
<td>188</td>
<td>骨</td>
<td>‘bone’</td>
</tr>
</tbody>
</table>

Section 3.5 introduced the significance of qualia rôles. In terms of qualia rôles within the Top Ontology, my results show that the majority rôles mapped are Physical, Dynamic, BoundedEvent, Object and Agentitive. For the comprehensive list, refer to Table 6.7. The Physical qualia rôle has the largest proportion mapped, and is more than double the Dynamic qualia rôle which is next on the list. For my noun example of bee:1, it has the qualia rôles Object and Animal.
Roast:1, as a verb example, has the qualia rôles UnboundedEvent, Agentive, Physical, Condition and Purpose.

Table 6.7: Bantu Concept mapping to Top Ontology qualia rôles

<table>
<thead>
<tr>
<th>Qualia rôle</th>
<th>Bantu concepts mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical</td>
<td>25</td>
</tr>
<tr>
<td>Dynamic</td>
<td>12</td>
</tr>
<tr>
<td>BoundedEvent</td>
<td>11</td>
</tr>
<tr>
<td>Object</td>
<td>11</td>
</tr>
<tr>
<td>Agentive</td>
<td>10</td>
</tr>
<tr>
<td>Animal</td>
<td>9</td>
</tr>
<tr>
<td>Condition</td>
<td>8</td>
</tr>
<tr>
<td>Location</td>
<td>8</td>
</tr>
<tr>
<td>Quantity</td>
<td>8</td>
</tr>
<tr>
<td>UnboundedEvent</td>
<td>8</td>
</tr>
<tr>
<td>Cause</td>
<td>7</td>
</tr>
<tr>
<td>Experience</td>
<td>7</td>
</tr>
<tr>
<td>Part</td>
<td>7</td>
</tr>
<tr>
<td>Purpose</td>
<td>6</td>
</tr>
<tr>
<td>Living</td>
<td>5</td>
</tr>
<tr>
<td>Property</td>
<td>5</td>
</tr>
<tr>
<td>Static</td>
<td>5</td>
</tr>
</tbody>
</table>

*Continued on next page*
Table 6.7 – continued from previous page

<table>
<thead>
<tr>
<th>Qualia rôle</th>
<th>Bantu concepts mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>4</td>
</tr>
<tr>
<td>Phenomenal</td>
<td>4</td>
</tr>
<tr>
<td>Solid</td>
<td>4</td>
</tr>
<tr>
<td>Comestible</td>
<td>3</td>
</tr>
<tr>
<td>Relation</td>
<td>3</td>
</tr>
<tr>
<td>Social</td>
<td>3</td>
</tr>
<tr>
<td>Usage</td>
<td>3</td>
</tr>
<tr>
<td>Existence</td>
<td>2</td>
</tr>
<tr>
<td>Manner</td>
<td>2</td>
</tr>
<tr>
<td>Natural</td>
<td>2</td>
</tr>
<tr>
<td>Artifact</td>
<td>1</td>
</tr>
<tr>
<td>Covering</td>
<td>1</td>
</tr>
<tr>
<td>Language</td>
<td>1</td>
</tr>
<tr>
<td>Representation</td>
<td>1</td>
</tr>
<tr>
<td>Liquid</td>
<td>1</td>
</tr>
<tr>
<td>Mental</td>
<td>1</td>
</tr>
<tr>
<td>Place</td>
<td>1</td>
</tr>
<tr>
<td>Plant</td>
<td>1</td>
</tr>
<tr>
<td>Possession</td>
<td>1</td>
</tr>
<tr>
<td>Substance</td>
<td>1</td>
</tr>
</tbody>
</table>

*Continued on next page*
6.3.5 **Upper Ontology comparison**

Of the 33 Bantu Language concepts not mapped to BCs in Global WordNet, the majority have a hypernym relationship to SUMO (not synonymy to a SUMO node, but subsumption). SUMO categorizes concepts into domains. Concepts that have no specific domain are put into the domain *factotum* (*Kozareva et al., 2007*, p. 334). Of the 67 roots in the list of Bantu language concepts mapped to SUMO, more than 30 concepts map to the factotum domain. Following the factotum domain, the number of concepts covering other domains, by decreasing number, is:

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.

I used the mapping methodology proposed by *Niles and Pease (2003)* to accomplish the mapping to SUMO of the Bantu language concepts. They propose

<table>
<thead>
<tr>
<th>Qualia rôle</th>
<th>Bantu concepts mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>1</td>
</tr>
</tbody>
</table>
three possible relations of interest: synonymy, hypernymy and instantiation. Synonymy is where there is a clear direct relation. This synonymy (from WordNet terminology) or equivalence relation (from ontology terminology) is represented by Niles and Pease (2003) using the symbol \(=\). For example nose:1 in Northern Sotho is synonymous to bee:1 in Princeton WordNet which is synonymous, or equivalent, to Bee in SUMO. 36 of the Bantu language concepts are mapped to SUMO via synonymy. There is no concept in SUMO that is equivalent to the Princeton pig:1 or kolobê in Northern Sotho. SUMO does have the concept Hoofed Mammal which was mapped by Niles and Pease (2003) through considering the hypernym relation in Princeton WordNet. Since ungulate:1, hoofed mammal:1 in Princeton WordNet is synonymous with Hoofed Mammal in SUMO, its hyponym pig:1 is directly mapped to SUMO as subsumption through this hypernym relation. This hypernymy (from WordNet terminology) or subsumption relation (from ontology terminology) is represented using the symbol \(+\). 28 of the concepts are linked via a hypernym to a SUMO node. The third relation is instantiation. Three of the concepts have neither equivalence in meaning in SUMO nor subsumption in meaning. All of these are numbers that are adjectival concepts mapped to the SUMO Positive Integer node\(^6\). This instantiation relation indicates that the numbers (two:1, three:1 and ten:1) are members of the class denoted by the SUMO concept Positive Integer. It is represented using the symbol \(\oplus\). The only concept that mapped to Princeton WordNet, for which Princeton WordNet does not have an existing SUMO mapping, is sangoma:1. As a solution for my research the same methodology used by Niles and Pease (2003) was applied by considering the WordNet hypernym therapist:1 and its

\(^6\) Refer to Listing B.14 for the detail.
SUMO mapping (TherapeuticProcess in domain medicine), thus also regarding this relation for *sangoma:1* as subsumption with SUMO.

SUMO contains a hierarchy of classes. The topmost class is *Entity*. The *Entity* class is specialised into the *Physical* and *Abstract* subclasses. The *Physical* class is further specialised into the *Object* and *Process* subclasses. A comprehensive description of the top classes in SUMO is provided by Breitman et al. (2007, p.187). In terms of the top classes in SUMO, the attribute class and the process class are the best represented in their sub-classes for the 67 concepts. Between physical and abstract concept classes, the physical class is better represented. Within the physical class, of the 4 types of object sub-classes, 3 are well represented. All of the process sub-classes are represented by concepts. The abstract class is not as well represented. Figure 6.2 illustrates the subsumption of the Bantu core concepts in these SUMO top level classes. The dotted nodes reflect classes not covering any core concepts. For deeper sub-class levels, these are just summarised by the number of nodes.

### 6.4 Quantitative ontology comparison

The costs described in Xue et al. (2009) were used as a basis for comparison in respect of all the calculations. The details of the results are shown in Appendix C. The totals for the transformation costs are shown in Table C.1. In order to arrive at the results in Table C.1, calculations were required on each node in the tree. These calculations per node are illustrated in Table C.2.

The Similarity Index \(\gamma(T_1, T_2) = \min\{\gamma(T_1 \rightarrow T_2), \gamma(T_2 \rightarrow T_1)\}\) is 1.36 between African WordNet Base Concepts in the African language WordNet pro-
totype and the equivalent subset of Princeton WordNet mapped to SUMO. Note that, in obtaining this measure, there are no delete operations. Insertion is always much more costly if there are more descendants for a given node, that is $|D(v)|$ is large in comparison with the average value of $|D(v)|$ for nodes in the tree. Insertion is also more costly if the node is closer in path length to the root of the tree, that is $\text{depth}(v)$ is small in comparison to $\text{height}(T)$. Movement is less costly if there are more descendants for a given node, that is $|D(v)|$ is large in comparison to the average value of $|D(v)|$ for nodes in the tree, than is the cost for insertion. Movement is also less costly for a node than insertion in relation to $\text{depth}(v)$. The Similarity Index of 1.36 can be compared qualitatively to previous examples of tree comparison – Figure 1 in Xue et al. (2009, p. 1768) and Figure 2 in Banerjee et al. (2010, p. 586). By qualitative comparison, the figure of 1.36 signifies a slight difference but not significant difference.
Figure 6.2: Bantu Base Concept subsumption in SUMO
CHAPTER 7

Conclusion and future work

Of course it is the words on the page that lead one to the ideas, but paradoxically keeping one’s trust in the words after one has found the ideas that they stand for amounts to a knee-jerk preference for letter over spirit ...

Translator, Trader: An Essay on the Pleasantly Pervasive Paradoxes,

Sagan (2009)

7.1 Introduction

This dissertation covers the state of the art in upper ontologies, lexical ontologies, WordNet core concepts and the Bantu Lexical Reconstructions to date. It further investigates the mapping of core concepts taken from African languages to concepts already included in the Suggested Upper Merged Ontology (SUMO), and provides results on the nature of the possible mapping, qualitatively and quantitatively.
Qualitatively, results are provided for the mapping between the final BLR3 core concepts (detailed in Appendix A) and:

1. Princeton WordNet,
2. BalkaNet Common Synsets,
3. Global WordNet Base Concepts,
4. EuroWordNet Top Ontology and
5. SUMO.

Quantitatively, the BLR3 core concepts were mapped from an African language WordNet prototype to SUMO via the WordNet mapping to SUMO and an ontology comparison that was done using the existing approach of Xue et al. (2009). The quantitative result showed no real marked difference between the placing of the concepts and their existence in an upper ontology, thus lending support to answer the primary research question – that the human language chosen to define the core concepts in an upper ontology such as SUMO has no effect on the universal and comprehensive nature of the upper ontology.

Both qualitative and quantitative results were provided in Chapter 6. This chapter examines those results in the context of the research questions posed in Chapter 1 and provides a conclusion to this research and dissertation.
<table>
<thead>
<tr>
<th>Number</th>
<th>Main Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Are the core concepts from a proposed natural language family currently included in an existing, accepted upper ontology?</td>
</tr>
<tr>
<td>1a</td>
<td>Is every one of these core concepts equivalent to or subsumed by a concept in a defined upper ontology?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number</th>
<th>Research Sub-Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>What is the state of the art of the natural language core concept definition in WordNets?</td>
</tr>
<tr>
<td>3</td>
<td>What is the state of the art of the upper ontology usage in the context of these natural language core concepts?</td>
</tr>
<tr>
<td>4</td>
<td>How do existing mappings of non-Indo-European language family core concepts to upper ontologies compare to that of Princeton WordNet?</td>
</tr>
<tr>
<td>5</td>
<td>What will a new structure of core concepts, from a novel African linguistic base, look like and how can it be compared to existing structures?</td>
</tr>
</tbody>
</table>

Table 7.1: Research questions

7.2 Answering the research questions

7.2.1 Research sub-questions

What is the state of the art of the natural language core concept definition in WordNets? The state of the art of linguistic core concepts in WordNet was described through the method of a literature review provided in Chapter 3.

What is the state of the art of the upper ontology usage in the context of
these natural language core concepts? The state of the art of upper ontologies and how ontologies could be compared was described through the method of a literature review provided in Chapters 2 and 4 respectively.

How do existing mappings of non-Indo-European language family core concepts to upper ontologies compare to that of Princeton WordNet? What will a new structure of core concepts, from an African linguistic base, look like and how can it be compared to existing structures? These two research sub-questions were examined through a mapping design and results provided in Chapter 6 with further detail in Appendices A, B and C.

7.2.2 Main research question

Are a given natural language family’s core concepts currently included in an existing, accepted upper ontology? This question was ultimately answered in the results provided in Chapter 6. Although there is not yet enough empirical research of the usage of the Similarity Index to draw a comprehensive quantitative comparison, the figure of 1.36 signifies a slight but not significant difference. Quantitatively, the Similarity Index $\gamma(T_1, T_2)$ is 1.36 between African language WordNet Base Concepts and the equivalent subset of Princeton WordNet mapped to SUMO, which is a measure of the mapping from an African language WordNet prototype to SUMO. Qualitatively, there are some differences but no glaring omissions in SUMO where a sensible mapping (at least subsumptively) could not be made.

Therefore, the research question can be effectively answered by producing a negative response to the primary research question: the human language chosen
to define core concepts in SUMO has no observably, significant effect on the universal and comprehensive nature of an upper ontology such as SUMO.

7.3 Reflection

This research has highlighted a number of key issues. The construction from scratch approach, followed by mapping, was used on a subset of core concepts, strategically chosen (as already recognized core concepts in the Bantu languages) by multiple linguists over many years of research. The aim was to produce an informed approach to mapping to WordNet and upper ontologies, and determine whether there are significant differences. It is clear from the results that mapping at the word sense level is good (93% fit), but mapping between the BC set proposed by the Global WordNet Project and the BC set is not good (only half fit).

The 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 mapping with Princeton WordNet as the base. Within Africa, this strict mapping was used for the construction of the Afrikaans WordNet (Kotzé, 2008). This made sense as the core concepts in Afrikaans would probably closely map to the core concepts in Dutch, which was used as one of the input languages to the Global WordNet BCs. The advantages of the strict mapping approach used with the Afrikaans language is that bootstrapping is made easier, and automation can be utilised to advantage with a less resourced language – the advantages proposed by Ordan and Wintner (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 Base Concepts.

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, a hybrid approach is proposed 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 mapping with Princeton WordNet (first the expand and then the merge approach (Vossen, 2007a)). 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, and 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 mapping to Global WordNet of the BCs was "better" at the top levels for verbs, and "better" at the lower levels for nouns. The Global WordNet BC requirement for a concept to occupy a "high position in the semantic hierarchy" implies the importance of verb roots in African languages will need to be considered. For African languages, it might be appropriate to focus on the verb structure first in terms of BCs.

The result in terms of mapping to 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:1 match well, but it is not immediately clear why bee:1 (a Global WordNet Base Concept and a Bantu
language core concept) has synonymy with SUMO, but *tick:*1 (only a Bantu language core concept) does not. Should they be part of SUMO or rather part of a domain specific ontology? This highlights the potential discrepancy of what is included in an upper ontology and what is included in a mid-level ontology.

Consider the verb examples of *heat:*1 and *cool:*1. Both words exist as BCs in Global WordNet BCs and in the 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 mismapping between Princeton WordNet and SUMO, or if mapped correctly would produce a different logical interpretation of OWL and RDF results for these concepts. Logical discrepancies can result from this – mapping of one concept via synonymy and the opposite concept by a sub-class relationship.

This research has produced peripheral resource artefacts that are useful for further research in general. In particular, the open available natural language core concept base for the Northern Sotho WordNet is available as a result of the African language WordNet prototype, developed as part of this study. These development results are available as part of the DEBVisDic project

1. The results of this prototype are available as a resource on the DEBVisDic server at https://abulafia.fi.muni.cz:9001/editor hosted by Masaryk University (Rambousek and Horáček, 2016).
language WordNets in the African WordNet Project if required.

The list chosen as a subset, discussed in this research, is the quality assured 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 additional languages have been added, 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 mapped easily to upper ontology concepts that are claimed to be universally shared. Additionally, mapping of concepts to SUMO occurs independent of the part of speech associated with a concept in any language. This underscores the fact that mapping is executed on a concept rather than a lexical basis.

7.4 Recommendations

7.4.1 Policy and practice

It was shown that use of the Global WordNet BCs as a starting point for building African language WordNets will not necessarily be a good idea for the Bantu lan-
guages. This research strengthens the proposed argument that a *hybrid* approach is used for constructing the future African language WordNets. The base of any African language WordNet should be constructed from scratch by mother tongue speakers, with bootstrapping only applied subsequently to the bases’ completion. In other words, the initial frame of reference for African language concepts should be African and not borrowed. These recommendations could influence the policy and practices of the various language teams in the in building future WordNets for African languages.

### 7.4.2 Evaluation

The goal of this dissertation was to address a research challenge: the assumption that the universality of the upper ontology is preserved for the concepts realized in other languages, particularly in other language families. The results addressed this challenge for the African languages specifically. They could be strengthened by looking at other language families as well. The method of using an existing tree comparison approach worked well and provided a practical method for comparison. These comparison results could be strengthened by similar studies extending this to concepts from other language families. The systems used in assisting the research were proven research technologies of WordNet (and its ancillary tools), Protegé and DEBVisDic. They were well documented and there were no serious hurdles in using them in the context required. These tools all were fit for purpose to meet this research’s objectives. The data artifacts produced by them are all available in standards-based formats and can be used by others to compare and improve the confidence in the results of this research.
7.4.3 Future research

Further research could be done on the African language WordNet prototype, produced in this study. This could extend the mapping to SUMO beyond these 67 concepts to determine more conclusive results than those in this research. This would be of benefit for any other further research in the use of the Similarity Index for ontology comparisons, so that a more robust mechanism could be used for ontology comparisons going forward. This would also develop a broader base of empirical data on ontological comparison for additional research.

7.4.4 Further development work

These core concepts in the online WordNet can be used through the ILI for the benefit of the related African language WordNets in the African WordNet Project. Once all the African language WordNets are completed as a first version, it would be useful to revisit the approach used to define Base Concepts in Global WordNet, reapply these, and subsequently, reexamine the SUMO mappings produced in this research.

This research can also be used as a basis for further research in multilingual WordNet applications that rely on mappings to SUMO, the ILI or to Princeton English WordNet. For example, the KYOTO project has extended the use of Global WordNets to practical applications, using word sense disambiguation through WordNet graphs to determine personalised PageRanks (Haveliwala, 2002) for a word sense in a specified context (Soroa et al., 2010). An advantage to this technique being used in other language families was the graph alignment between more than one language, and, in particular, the alignment to Princeton English
WordNet. Furthermore, DEBVisDic has been used as a tool for applications of XML WordNet structures within the KYOTO project (Horák and Rambousek, 2010). Therefore, the results produced in this research (a lexical ontology base in DEBVisDic) could be used for further exploration of personalised PageRank to other language family WordNets, such as those in the African WordNet Project.
REFERENCES


Bleek, W. H. I. (1862). *A Comparative Grammar of South African Lan-


177


182


Guthrie, M. (1948). *The classification of the Bantu languages*. Published for the International Institute of African Languages and Cultures by the Ox-


Kriel, T., Prinsloo, D., and Sathekge, B. (2003). *Popular Northern Sotho Dic-


Nowack, B. (2009). The Semantic Web - Not a piece of cake .... [On-
line; accessed 2 August 2014], \url{http://bnode.org/blog/2009/07/08/the-semantic-web-not-a-piece-of-cake}.


197


Pease, A. (2005). Construction of Arabic WordNet in Parallel with an On-


sight project. [Online; accessed 2 August 2014], http://metaverseroadmap.org/.


Part IV

Additional information
Word and concept lists

The linguistic department of Oxford University originally categorized Proto-Bantu roots (Guthrie, 1948). The Comparative On-line Bantu Dictionary (CBOLD) project took this initial linguistic unification work and extended it (Schadeberg, 2002). 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., 2005) that represent so called main entries of which there are 1400. The main entries have been further categorized by Maho (2001, 2005) to isolate all main entries that have modern reflexes in Zone A and Zone S.

This produces 375 roots. Maho (2001) also isolated all main entries which have modern reflexes in at least 14 zones (231 roots). The two lists produce a core collection of 407 roots.
A.1 Original word list

My methodology in this research has involved taking these 407 roots and only using those roots that occur in all 16 zones. This produces a list of 99 roots.

This list of 99 is shown in Table A.1.

<table>
<thead>
<tr>
<th>Root</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-bů́</td>
<td>'bad'</td>
</tr>
<tr>
<td>-bá-</td>
<td>'to dwell; to be; to become'</td>
</tr>
<tr>
<td>-bá̄b-</td>
<td>'to be bitter; to be smart; to itch; to be sharp; to sting; to hurt'</td>
</tr>
<tr>
<td>-bá̄d́</td>
<td>'two'</td>
</tr>
<tr>
<td>-bò̄d-</td>
<td>'to be rotten'</td>
</tr>
<tr>
<td>-bú̄dā</td>
<td>'rain'</td>
</tr>
<tr>
<td>-bú̄̄à</td>
<td>'dog'</td>
</tr>
<tr>
<td>-cí</td>
<td>'ground; country; underneath'</td>
</tr>
<tr>
<td>-cè̄ŋgà</td>
<td>'sand; sandy ground'</td>
</tr>
<tr>
<td>-cò̄k-</td>
<td>'to poke in; to put in; to prick with a point; to hide'</td>
</tr>
<tr>
<td>-cò̄ng-</td>
<td>'to sharpen to a point'</td>
</tr>
<tr>
<td>-dí̄</td>
<td>'to eat'</td>
</tr>
<tr>
<td>-dí̄mî</td>
<td>'tongue; language; flame'</td>
</tr>
<tr>
<td>-dí̄m-</td>
<td>'to bite'</td>
</tr>
<tr>
<td>-dí̄</td>
<td>'to be'</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Root</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-díd-</td>
<td>‘to weep; to shout; to wail’</td>
</tr>
<tr>
<td>-dá</td>
<td>‘louse’</td>
</tr>
<tr>
<td>-dàì</td>
<td>‘long’</td>
</tr>
<tr>
<td>-dèdù</td>
<td>‘beard; chin’</td>
</tr>
<tr>
<td>-démà-</td>
<td>‘invalid; physical disability’</td>
</tr>
<tr>
<td>-dì</td>
<td>‘root; fibre’</td>
</tr>
<tr>
<td>-dibà</td>
<td>‘pool; pond; deep water; well’</td>
</tr>
<tr>
<td>-dó-</td>
<td>‘to sleep’</td>
</tr>
<tr>
<td>-dònìg-</td>
<td>‘to heap up; to arrange; to pack up’</td>
</tr>
<tr>
<td>-dùad-</td>
<td>‘to wear’</td>
</tr>
<tr>
<td>-gì</td>
<td>‘egg’</td>
</tr>
<tr>
<td>-gùdò</td>
<td>‘yesterday; day before yesterday; evening’</td>
</tr>
<tr>
<td>-gàngà</td>
<td>‘medicine man’</td>
</tr>
<tr>
<td>-gèd-</td>
<td>‘to try’</td>
</tr>
<tr>
<td>-gènd-</td>
<td>‘to walk; to travel’</td>
</tr>
<tr>
<td>-gòmbè</td>
<td>‘cattle’</td>
</tr>
<tr>
<td>-gùdùbè</td>
<td>‘pig’</td>
</tr>
<tr>
<td>-kúd-</td>
<td>‘to grow up’</td>
</tr>
<tr>
<td>-kúmi</td>
<td>‘ten’</td>
</tr>
<tr>
<td>-kúni</td>
<td>‘firewood’</td>
</tr>
<tr>
<td>-kàdà</td>
<td>‘ember; charcoal’</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Root</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-kádàng-</td>
<td>‘to fry, to roast’</td>
</tr>
<tr>
<td>-kádí</td>
<td>‘woman; wife’</td>
</tr>
<tr>
<td>-kángà</td>
<td>‘guinea fowl’</td>
</tr>
<tr>
<td>-ké</td>
<td>‘dawn’</td>
</tr>
<tr>
<td>-kídà</td>
<td>‘tail’</td>
</tr>
<tr>
<td>-kingó-</td>
<td>‘neck; nape; voice’</td>
</tr>
<tr>
<td>-kókó</td>
<td>‘chicken’</td>
</tr>
<tr>
<td>-kómb-</td>
<td>‘to scrape; to dig; to lick with finger’</td>
</tr>
<tr>
<td>-köt-</td>
<td>‘to stoop; to be bent’</td>
</tr>
<tr>
<td>-kú-</td>
<td>‘to die’</td>
</tr>
<tr>
<td>-kúm-</td>
<td>‘to be honoured; to be rich’</td>
</tr>
<tr>
<td>-kúmb-</td>
<td>‘to bend’</td>
</tr>
<tr>
<td>-kúpà</td>
<td>‘bone’</td>
</tr>
<tr>
<td>-kúpá</td>
<td>‘tick; insect’</td>
</tr>
<tr>
<td>-júbà</td>
<td>‘sun’</td>
</tr>
<tr>
<td>-júà</td>
<td>‘thing; bead; iron’</td>
</tr>
<tr>
<td>-jüng-</td>
<td>‘to sift’</td>
</tr>
<tr>
<td>-jádà</td>
<td>‘finger-nail, toe-nail, claw’</td>
</tr>
<tr>
<td>-jádà</td>
<td>‘hunger; famine’</td>
</tr>
<tr>
<td>-jákà</td>
<td>‘year; cultivation season; harvest’</td>
</tr>
<tr>
<td>-jánà</td>
<td>‘child’</td>
</tr>
<tr>
<td>Root</td>
<td>Attested and/or reconstructed meaning</td>
</tr>
<tr>
<td>--------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>-jánk-</td>
<td>‘to spread to dry in the sun; to spread out’</td>
</tr>
<tr>
<td>-játò</td>
<td>‘canoe’</td>
</tr>
<tr>
<td>-jéd-</td>
<td>‘to shine; to be clear; to be ripe; to be favourable’</td>
</tr>
<tr>
<td>-jéné</td>
<td>‘self; same’</td>
</tr>
<tr>
<td>-jib-</td>
<td>‘to steal’</td>
</tr>
<tr>
<td>-jícò</td>
<td>‘eye’</td>
</tr>
<tr>
<td>-jid-</td>
<td>‘to get dark; to get black’</td>
</tr>
<tr>
<td>-jidà</td>
<td>‘path’</td>
</tr>
<tr>
<td>-jík-</td>
<td>‘to come or go down’</td>
</tr>
<tr>
<td>-jíkì</td>
<td>‘bee’</td>
</tr>
<tr>
<td>-jíkì</td>
<td>‘smoke’</td>
</tr>
<tr>
<td>-jikùt-</td>
<td>‘to be satiated’</td>
</tr>
<tr>
<td>-jijad</td>
<td>‘to be full’</td>
</tr>
<tr>
<td>-jìjì</td>
<td>‘water’</td>
</tr>
<tr>
<td>-júmb-</td>
<td>‘to sing; to dance’</td>
</tr>
<tr>
<td>-jinà</td>
<td>‘name’</td>
</tr>
<tr>
<td>-jíngí</td>
<td>‘many, much’</td>
</tr>
<tr>
<td>-jípí</td>
<td>‘short’</td>
</tr>
<tr>
<td>-jójà</td>
<td>‘fear’</td>
</tr>
<tr>
<td>-jókà</td>
<td>‘snake; intestinal worm’</td>
</tr>
<tr>
<td>-jót-</td>
<td>‘to warm oneself’</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Root</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-mìd-</td>
<td>‘to blow nose’</td>
</tr>
<tr>
<td>-nà</td>
<td>‘with; and’</td>
</tr>
<tr>
<td>-ncè</td>
<td>‘all’</td>
</tr>
<tr>
<td>-ntù</td>
<td>‘some entity; any’</td>
</tr>
<tr>
<td>-nyàmà</td>
<td>‘animal; meat’</td>
</tr>
<tr>
<td>-nyó-</td>
<td>‘to drink’</td>
</tr>
<tr>
<td>-pá-</td>
<td>‘to give’</td>
</tr>
<tr>
<td>-pácà</td>
<td>‘twin’</td>
</tr>
<tr>
<td>-pàp-</td>
<td>‘to flap wings; to flutter’</td>
</tr>
<tr>
<td>-pép</td>
<td>‘to blow as wind; to winnow; to smoke tobacco; to breathe’</td>
</tr>
<tr>
<td>-pí</td>
<td>‘to be burnt; to be hot; to be cooked; to be ripe; to ferment; to be red’</td>
</tr>
<tr>
<td>-pód-</td>
<td>‘to be cold; to cool down; to be quiet’</td>
</tr>
<tr>
<td>-púd-</td>
<td>‘to froth over’</td>
</tr>
<tr>
<td>-túng-</td>
<td>‘to put through; to thread on string; to plait; to sew; to tie up; to build; to close in’</td>
</tr>
<tr>
<td>-tùè</td>
<td>‘head’</td>
</tr>
<tr>
<td>-táà</td>
<td>‘bow’</td>
</tr>
<tr>
<td>-tátù</td>
<td>‘three’</td>
</tr>
<tr>
<td>-tì</td>
<td>‘tree stick’</td>
</tr>
<tr>
<td>-tíg-</td>
<td>‘to leave behind’</td>
</tr>
</tbody>
</table>

Continued on next page
Table A.1 – continued from previous page

<table>
<thead>
<tr>
<th>Root</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-tó-</td>
<td>‘to stamp; to pound; to bite’</td>
</tr>
<tr>
<td>-túd-</td>
<td>‘to hammer; to forge’</td>
</tr>
</tbody>
</table>

A.2 Attested word list

These roots were then analysed for potential modern reflexes in Northern Sotho based on the listing in the Comprehensive Northern Sotho Dictionary Groot Noord-Sotho Woordeboek (1985), producing a list of 84 potential candidates.

This list of 84 is shown in Table A.2.

Table A.2: Attested word list

<table>
<thead>
<tr>
<th>Root</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-bū́</td>
<td>‘bad’</td>
</tr>
<tr>
<td>-bá-</td>
<td>‘to dwell; to be; to become’</td>
</tr>
<tr>
<td>-báb-</td>
<td>‘to be bitter; to be smart; to itch; to be sharp; to sting; to hurt’</td>
</tr>
<tr>
<td>-bádí</td>
<td>‘two’</td>
</tr>
<tr>
<td>-bōd-</td>
<td>‘to be rotten’</td>
</tr>
<tr>
<td>-búdá</td>
<td>‘rain’</td>
</tr>
<tr>
<td>-búá́</td>
<td>‘dog’</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Root</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-còng-</td>
<td>‘to sharpen to a point’</td>
</tr>
<tr>
<td>-dí</td>
<td>‘to eat’</td>
</tr>
<tr>
<td>-dímì</td>
<td>‘tongue; language; flame’</td>
</tr>
<tr>
<td>-dúm-</td>
<td>‘to bite’</td>
</tr>
<tr>
<td>-dì</td>
<td>‘to be’</td>
</tr>
<tr>
<td>-dídid-</td>
<td>‘to weep; to shout; to wail’</td>
</tr>
<tr>
<td>-díá</td>
<td>‘louse’</td>
</tr>
<tr>
<td>-dài</td>
<td>‘long’</td>
</tr>
<tr>
<td>-dèdù</td>
<td>‘beard; chin’</td>
</tr>
<tr>
<td>-dì</td>
<td>‘root; fibre’</td>
</tr>
<tr>
<td>-dibà</td>
<td>‘pool; pond; deep water; well’</td>
</tr>
<tr>
<td>-dúad-</td>
<td>‘to wear’</td>
</tr>
<tr>
<td>-gí</td>
<td>‘egg’</td>
</tr>
<tr>
<td>-gàngà</td>
<td>‘medicine man’</td>
</tr>
<tr>
<td>-gèd-</td>
<td>‘to try’</td>
</tr>
<tr>
<td>-gènd-</td>
<td>‘to walk; to travel’</td>
</tr>
<tr>
<td>-gòmbè</td>
<td>‘cattle’</td>
</tr>
<tr>
<td>-gùdübè</td>
<td>‘pig’</td>
</tr>
<tr>
<td>-kúd-</td>
<td>‘to grow up’</td>
</tr>
<tr>
<td>-kúmì</td>
<td>‘ten’</td>
</tr>
<tr>
<td>-kúnì</td>
<td>‘firewood’</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Root</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-kádá</td>
<td>‘ember; charcoal’</td>
</tr>
<tr>
<td>-kádàng-</td>
<td>‘to fry, to roast’</td>
</tr>
<tr>
<td>-kádí</td>
<td>‘woman; wife’</td>
</tr>
<tr>
<td>-kángà</td>
<td>‘guinea fowl’</td>
</tr>
<tr>
<td>-ké</td>
<td>‘dawn’</td>
</tr>
<tr>
<td>-kídà</td>
<td>‘tail’</td>
</tr>
<tr>
<td>-kókó</td>
<td>‘chicken’</td>
</tr>
<tr>
<td>-kómb-</td>
<td>‘to scrape; to dig; to lick with finger’</td>
</tr>
<tr>
<td>-kòt-</td>
<td>‘to stoop; to be bent’</td>
</tr>
<tr>
<td>-kú-</td>
<td>‘to die’</td>
</tr>
<tr>
<td>-kúm-</td>
<td>‘to be honoured; to be rich’</td>
</tr>
<tr>
<td>-kúmb-</td>
<td>‘to bend’</td>
</tr>
<tr>
<td>-kúpá</td>
<td>‘bone’</td>
</tr>
<tr>
<td>-kúpá</td>
<td>‘tick; insect’</td>
</tr>
<tr>
<td>-júng-</td>
<td>‘to sift’</td>
</tr>
<tr>
<td>-jádá</td>
<td>‘finger-nail, toe-nail, claw’</td>
</tr>
<tr>
<td>-jádá</td>
<td>‘hunger; famine’</td>
</tr>
<tr>
<td>-jákà</td>
<td>‘year; cultivation season; harvest’</td>
</tr>
<tr>
<td>-jánà</td>
<td>‘child’</td>
</tr>
<tr>
<td>-jánúk-</td>
<td>‘to spread to dry in the sun; to spread out’</td>
</tr>
<tr>
<td>-jéd-</td>
<td>‘to shine; to be clear; to be ripe; to be favourable’</td>
</tr>
<tr>
<td>Root</td>
<td>Attested and/or reconstructed meaning</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------------------------</td>
</tr>
<tr>
<td>-jéné</td>
<td>‘self; same’</td>
</tr>
<tr>
<td>-jicò</td>
<td>‘eye’</td>
</tr>
<tr>
<td>-jíd-</td>
<td>‘to get dark; to get black’</td>
</tr>
<tr>
<td>-jùdà</td>
<td>‘path’</td>
</tr>
<tr>
<td>-jík-</td>
<td>‘to come or go down’</td>
</tr>
<tr>
<td>-jíkùt-</td>
<td>‘to be satiated’</td>
</tr>
<tr>
<td>-jíjad</td>
<td>‘to be full’</td>
</tr>
<tr>
<td>-júmb-</td>
<td>‘to sing; to dance’</td>
</tr>
<tr>
<td>-jínà</td>
<td>‘name’</td>
</tr>
<tr>
<td>-jíngí</td>
<td>‘many, much’</td>
</tr>
<tr>
<td>-jípí</td>
<td>‘short’</td>
</tr>
<tr>
<td>-jógà</td>
<td>‘fear’</td>
</tr>
<tr>
<td>-jókà</td>
<td>‘snake; intestinal worm’</td>
</tr>
<tr>
<td>-jót-</td>
<td>‘to warm oneself’</td>
</tr>
<tr>
<td>-mid-</td>
<td>‘to blow nose’</td>
</tr>
<tr>
<td>-nà</td>
<td>‘with; and’</td>
</tr>
<tr>
<td>-ncè</td>
<td>‘all’</td>
</tr>
<tr>
<td>-ntù</td>
<td>‘some entity; any’</td>
</tr>
<tr>
<td>-nyàmà</td>
<td>‘animal; meat’</td>
</tr>
</tbody>
</table>

Continued on next page
Table A.2 – continued from previous page

<table>
<thead>
<tr>
<th>Root</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-nyó-</td>
<td>‘to drink’</td>
</tr>
<tr>
<td>-pá-</td>
<td>‘to give’</td>
</tr>
<tr>
<td>-pácà</td>
<td>‘twin’</td>
</tr>
<tr>
<td>-pàp-</td>
<td>‘to flap wings; to flutter’</td>
</tr>
<tr>
<td>-pép</td>
<td>‘to blow as wind; to winnow; to smoke tobacco; to breathe’</td>
</tr>
<tr>
<td>-pí</td>
<td>‘to be burnt; to be hot; to be cooked; to be ripe; to ferment; to be red’</td>
</tr>
<tr>
<td>-pód-</td>
<td>‘to be cold; to cool down; to be quiet’</td>
</tr>
<tr>
<td>-púd-</td>
<td>‘to froth over’</td>
</tr>
<tr>
<td>-túng-</td>
<td>‘to put through; to thread on string; to plait; to sew; to tie up; to build; to close in’</td>
</tr>
<tr>
<td>-táà</td>
<td>‘bow’</td>
</tr>
<tr>
<td>-tátù</td>
<td>‘three’</td>
</tr>
<tr>
<td>-tí</td>
<td>‘tree stick’</td>
</tr>
<tr>
<td>-tíg-</td>
<td>‘to leave behind’</td>
</tr>
<tr>
<td>-túd-</td>
<td>‘to hammer; to forge’</td>
</tr>
</tbody>
</table>

A.3 Quality assured word list

The potential candidate list of 84 roots was quality assured by two external linguists to produce a further subset of 67 roots.
Within the 67, if the main entry did not occur, but its variant did, then the variant was used.

This list of 67 is shown in Table A.3.

Table A.3: Quality assured word list

<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-bû</td>
<td>5841</td>
<td>‘bad’</td>
</tr>
<tr>
<td>-bá-</td>
<td>4</td>
<td>‘to dwell; to be; to become’</td>
</tr>
<tr>
<td>-báb-</td>
<td>5</td>
<td>‘to be bitter; to be smart; to itch; to be sharp; to sting; to hurt’</td>
</tr>
<tr>
<td>-bàdî</td>
<td>36</td>
<td>‘two’</td>
</tr>
<tr>
<td>-bòd-</td>
<td>253</td>
<td>‘to be rotten’</td>
</tr>
<tr>
<td>-búdâ</td>
<td>368</td>
<td>‘rain’</td>
</tr>
<tr>
<td>-bùa</td>
<td>282</td>
<td>‘dog’</td>
</tr>
<tr>
<td>-dí</td>
<td>944</td>
<td>‘to eat’</td>
</tr>
<tr>
<td>-dímì</td>
<td>973</td>
<td>‘tongue; language; flame’</td>
</tr>
<tr>
<td>-dím-</td>
<td>1181</td>
<td>‘to bite’</td>
</tr>
<tr>
<td>-dí</td>
<td>940</td>
<td>‘to be’</td>
</tr>
<tr>
<td>-díd-</td>
<td>959</td>
<td>‘to weep; to shout; to wail’</td>
</tr>
<tr>
<td>-dá</td>
<td>780</td>
<td>‘louse’</td>
</tr>
<tr>
<td>-dài</td>
<td>3705</td>
<td>‘long’</td>
</tr>
<tr>
<td>-dèdù</td>
<td>897</td>
<td>‘beard; chin’</td>
</tr>
<tr>
<td>-díbà</td>
<td>1025</td>
<td>‘pool; pond; deep water; well’</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-dúad-</td>
<td>1234</td>
<td>‘to wear’</td>
</tr>
<tr>
<td>-gí</td>
<td>1368</td>
<td>‘egg’</td>
</tr>
<tr>
<td>-gàngà</td>
<td>1332</td>
<td>‘medicine man’</td>
</tr>
<tr>
<td>-gèd-</td>
<td>1345</td>
<td>‘to try’</td>
</tr>
<tr>
<td>-gènd-</td>
<td>1362</td>
<td>‘to walk; to travel’</td>
</tr>
<tr>
<td>-gùdùbè</td>
<td>1494</td>
<td>‘pig’</td>
</tr>
<tr>
<td>-kúd-</td>
<td>1997</td>
<td>‘to grow up’</td>
</tr>
<tr>
<td>-kúmì</td>
<td>2027</td>
<td>‘ten’</td>
</tr>
<tr>
<td>-kúñì</td>
<td>2042</td>
<td>‘firewood’</td>
</tr>
<tr>
<td>-kádà</td>
<td>1662</td>
<td>‘ember; charcoal’</td>
</tr>
<tr>
<td>-kádàng-</td>
<td>1665</td>
<td>‘to fry, to roast’</td>
</tr>
<tr>
<td>-kàngà</td>
<td>1720</td>
<td>‘guinea fowl’</td>
</tr>
<tr>
<td>-kídà</td>
<td>1793</td>
<td>‘tail’</td>
</tr>
<tr>
<td>-kókó</td>
<td>1904</td>
<td>‘chicken’</td>
</tr>
<tr>
<td>-kób-</td>
<td>1916</td>
<td>‘to scrape; to dig; to lick with finger’</td>
</tr>
<tr>
<td>-kót-</td>
<td>7350</td>
<td>‘to stoop; to be bent’</td>
</tr>
<tr>
<td>-kú-</td>
<td>2089</td>
<td>‘to die’</td>
</tr>
<tr>
<td>-kúm-</td>
<td>2113</td>
<td>‘to be honoured; to be rich’</td>
</tr>
<tr>
<td>-kúpá</td>
<td>2071</td>
<td>‘tick; insect’</td>
</tr>
<tr>
<td>-jádà</td>
<td>1558</td>
<td>‘finger-nail, toe-nail, claw’</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-jàdà</td>
<td>1555</td>
<td>‘hunger; famine’</td>
</tr>
<tr>
<td>-jákà</td>
<td>3169</td>
<td>‘year; cultivation season; harvest’</td>
</tr>
<tr>
<td>-jánà</td>
<td>3203</td>
<td>‘child’</td>
</tr>
<tr>
<td>-jánk-</td>
<td>3206</td>
<td>‘to spread to dry in the sun; to spread out’</td>
</tr>
<tr>
<td>-jéd-</td>
<td>3273</td>
<td>‘to shine; to be clear; to be ripe; to be favourable’</td>
</tr>
<tr>
<td>-jícò</td>
<td>3405</td>
<td>‘eye’</td>
</tr>
<tr>
<td>-jídà</td>
<td>1593</td>
<td>‘path’</td>
</tr>
<tr>
<td>-jíkì</td>
<td>3350</td>
<td>‘bee’</td>
</tr>
<tr>
<td>-jíkì</td>
<td>3442</td>
<td>‘smoke’</td>
</tr>
<tr>
<td>-jíkùt-</td>
<td>3445</td>
<td>‘to be satiated’</td>
</tr>
<tr>
<td>-jùmb-</td>
<td>3361</td>
<td>‘to sing; to dance’</td>
</tr>
<tr>
<td>-jínà</td>
<td>3464</td>
<td>‘name’</td>
</tr>
<tr>
<td>-jíngí</td>
<td>3485</td>
<td>‘many, much’</td>
</tr>
<tr>
<td>-jípí</td>
<td>3495</td>
<td>‘short’</td>
</tr>
<tr>
<td>-jókà</td>
<td>3536</td>
<td>‘snake; intestinal worm’</td>
</tr>
<tr>
<td>-jót-</td>
<td>3579</td>
<td>‘to warm oneself’</td>
</tr>
<tr>
<td>-nà</td>
<td>3674</td>
<td>‘with; and’</td>
</tr>
<tr>
<td>-ncè</td>
<td>500</td>
<td>‘all’</td>
</tr>
<tr>
<td>-ntù</td>
<td>4807</td>
<td>‘some (entity); any’</td>
</tr>
</tbody>
</table>

Continued on next page
### Table A.3 – continued from previous page

<table>
<thead>
<tr>
<th>Root</th>
<th>Main Ref</th>
<th>Attested and/or reconstructed meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-nyām̀a</td>
<td>3180</td>
<td>‘animal; meat’</td>
</tr>
<tr>
<td>-nyó-</td>
<td>7047</td>
<td>‘to drink’</td>
</tr>
<tr>
<td>-pá-</td>
<td>2344</td>
<td>‘to give’</td>
</tr>
<tr>
<td>-pácà</td>
<td>2348</td>
<td>‘twin’</td>
</tr>
<tr>
<td>-pàp-</td>
<td>2407</td>
<td>‘to flap wings; to flutter’</td>
</tr>
<tr>
<td>-pèp</td>
<td>2463</td>
<td>‘to blow as wind; to winnow; to smoke tobacco; to breathe’</td>
</tr>
<tr>
<td>-pì</td>
<td>2491</td>
<td>‘to be burnt; to be hot; to be cooked; to be ripe; to ferment; to be red’</td>
</tr>
<tr>
<td>-pòd-</td>
<td>2589</td>
<td>‘to be cold; to cool down; to be quiet’</td>
</tr>
<tr>
<td>-túng-</td>
<td>3081</td>
<td>‘to put through; to thread on string; to plait; to sew; to tie up; to build; to close in’</td>
</tr>
<tr>
<td>-tátù</td>
<td>2811</td>
<td>‘three’</td>
</tr>
<tr>
<td>-tí</td>
<td>2881</td>
<td>‘tree stick’</td>
</tr>
<tr>
<td>-túd-</td>
<td>3101</td>
<td>‘to hammer; to forge’</td>
</tr>
</tbody>
</table>

### A.4 Variant BLR3 list

The term *variant* has a particular meaning in BLR3:

These are reconstructions which are considered to descend from
another, more basic (MAIN) etymon. They show some variation in form for which no regular sound correspondence is known. Reconstructions derived from such variant forms are also classified as variants.

(Bastin et al., 2005)

BLR3 makes reference to derivatives:

These are reconstructions which are derived from a basic (MAIN) etymon. The derivation can be by affixation (e.g., verb extensions, nominalization, change of noun class, reduplication) or by semantic shift. In some cases, the decision which item is basic and which is derived, is somewhat arbitrary.

(Bastin et al., 2005)

In the subsequent table, the use of the word variant applies to actual BLR3 variants and derivations.

We can consider the term distribution which means that reconstructions may have derived variant and included forms in the different zones. These forms, once verified, have also been assigned identification numbers.

The list of variants is shown in Table A.4

<table>
<thead>
<tr>
<th>Proto-Bantu form</th>
<th>BLR3 Main Ref</th>
<th>Variant</th>
<th>Variant Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>-bàdî</td>
<td>36</td>
<td>-bîdî</td>
<td>190</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Proto-Bantu form</th>
<th>BLR3 Main Ref</th>
<th>Variant</th>
<th>Variant Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>-kádàng-</td>
<td>1665</td>
<td>-kàdùng-</td>
<td>1680</td>
</tr>
<tr>
<td>-kòt-</td>
<td>7350</td>
<td>-kòtam</td>
<td>1961</td>
</tr>
<tr>
<td>-kúpá</td>
<td>2071</td>
<td>-gúpá</td>
<td>1516 ¹</td>
</tr>
<tr>
<td>-jìkì</td>
<td>3350</td>
<td>-jìkì</td>
<td>6225 ²</td>
</tr>
<tr>
<td>-jìngí</td>
<td>3485</td>
<td>-nyìngí</td>
<td>2329</td>
</tr>
<tr>
<td>-jìpí</td>
<td>3495</td>
<td>-kúpí</td>
<td>2133</td>
</tr>
<tr>
<td>-ncè</td>
<td>500</td>
<td>-cè</td>
<td>499</td>
</tr>
<tr>
<td>-pèp</td>
<td>2463</td>
<td>-pèpud</td>
<td>1469</td>
</tr>
</tbody>
</table>

1. Refused by BLR3
2. Derived through semantic shift: honey from bee
The RDF and OWL files were accessed from the First Public Working Draft produced by the WordNet Task Force of the Semantic Web Best Practices and Deployment Working Group, part of the W3C Semantic Web Activity. The files represented use the format provided in RDF/XML and converted to TURTLE.
B.1 Sample WordNet RDF results

B.1.1 Nouns

B.1.1.1 Sangoma

The example below can be generated using the following URL: http://wordnet-rdf.princeton.edu/wn31/110569926-n.ttl.

Listing B.1: The synset for Sangoma

```xml
@prefix lemon: <http://lemon-model.net/lemon#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix wordnet-ontology: <http://wordnet-rdf.princeton.edu/ontology#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://wordnet-rdf.princeton.edu/wn31/110569926-n> a wordnet-ontology:Synset ;
  rdfs:label "sangoma"@eng ;
  wordnet-ontology:gloss "a traditional Zulu healer and respected elder"@eng ;
  wordnet-ontology:hypernym <http://wordnet-rdf.princeton.edu/wn31/110726882-n> ;
  wordnet-ontology:lexical_domain wordnet-ontology:noun.person ;
  wordnet-ontology:part_of_speech wordnet-ontology:noun ;
  wordnet-ontology:synset_member <http://wordnet-rdf.princeton.edu/wn31/sangoma-n> ;
  wordnet-ontology:translation "sangoma"@fin ;
  owl:sameAs <http://lemon-model.net/lexica/uby/wm/WN_Synset_57158>,
       <http://www.w3.org/2006/03/wn/wn20/instances/synset-sangoma-noun> .
```

B.1.1.2 Entity

The example below can be generated using the following URL: http://wordnet-rdf.princeton.edu/wn31/100001740-n.ttl.

Listing B.2: The synset for Entity

```xml
@prefix lemon: <http://lemon-model.net/lemon#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
```

232
B.1.1.3 Numida

Listing B.3: The synset for Numida
B.1.4 Bee

Listing B.4: The synset for Bee

```
Prefix lemon: <http://lemon-model.net/lemon#>
Prefix owl: <http://www.w3.org/2002/07/owl#>
Prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
Prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
Prefix wordnet-ontology: <http://wordnet-rdf.princeton.edu/ontology#>
Prefix xsd: <http://www.w3.org/2001/XMLSchema#>

<http://wordnet-rdf.princeton.edu/wn31/102209508> a wordnet-ontology:Synset;
  rdfs:label "bee"@en;
  wordnet-ontology:gloss "any of numerous hairy-bodied insects including social and solitary species"@en;
  wordnet-ontology:hypernym <http://wordnet-rdf.princeton.edu/wn31/102208922>;
  wordnet-ontology:hyponym <http://wordnet-rdf.princeton.edu/wn31/102210893>;
  <http://wordnet-rdf.princeton.edu/wn31/102210932>;
  <http://wordnet-rdf.princeton.edu/wn31/102211027>;
  <http://wordnet-rdf.princeton.edu/wn31/102211032>;
  <http://wordnet-rdf.princeton.edu/wn31/102211033>;
  wordnet-ontology:lexical_domain wordnet-ontology:noun.animal;
  wordnet-ontology:member_meronym <http://wordnet-rdf.princeton.edu/wn31/102209831>;
  wordnet-ontology:part_of_speech wordnet-ontology:noun;
  wordnet-ontology:synset_member <http://wordnet-rdf.princeton.edu/wn31/break>;  
  owl:sameAs <http://lemon-model.net/lexical/uby/wn/WN_Synset_11541>,
  <http://www.w3.org/2006/03/wn20/instances/synset-bee-noun-1>.
```

B.1.2 Verbs

B.1.2.1 Dance

http://wordnet-rdf.princeton.edu/wn31/201712535-v.ttl

Listing B.5: The synset for Dance

```
Prefix lemon: <http://lemon-model.net/lemon#>
```
<prefix owl: <http://www.w3.org/2002/07/owl#> .  
<prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .  
<prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .  

<http://wordnet-rdf.princeton.edu/wn31/201712535—from a wordnet-ontology:Synset ;  
   rdfs:label "dance" @eng ,  
   "trip the light fantastic" @eng ,  
   "trip the light fantastic toe" @eng ;  
wordnet-ontology:domain_category <http://wordnet-rdf.princeton.edu/wn31/201712401—from> ;  
wordnet-ontology:hyponym <http://wordnet-rdf.princeton.edu/wn31/201835473—from> ;  
wordnet-ontology:hyponym <http://wordnet-rdf.princeton.edu/wn31/201712401—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201713640—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201713790—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201713907—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201714049—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201755353—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201759233—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201899256—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201899376—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201899512—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201899605—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201899750—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900000—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900011—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900112—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900206—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900228—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900477—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900650—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900760—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900874—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900988—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201901090—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900196—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900199—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201900180—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201901399—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201901417—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201901482—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201901576—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201901670—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201901772—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201901878—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201902025—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201902174—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201902762—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201902886—from> ,  
<http://wordnet-rdf.princeton.edu/wn31/201903151—from> ,
B.1.2.2 Carry

http://wordnet-rdf.princeton.edu/wn31/20272977-v.ttl

Listing B.6: The synset for Carry

```xml
<prefix lemon:="http://lemon-model.net/lemon#">.
prefix rdf:="http://www.w3.org/1999/02/22-rdf-syntax-ns#">.

<http://wordnet-rdf.princeton.edu/wn31/20272977-v> a wordnet-ontology:Synset ;
	rdfs:label "carry" @eng ,
	"pack" @eng ,
	"take" @eng ;
wordnet-ontology: gloss "have with oneself; have on one's person" @eng ;
wordnet-ontology: hypernym <http://wordnet-rdf.princeton.edu/wn31/202636270-v> ;
wordnet-ontology: lexical_domain wordnet-ontology: verb.stative ;
wordnet-ontology: part_of speech wordnet-ontology: verb ;
wordnet-ontology: sample "I always carry money" @eng ,
"She always takes an umbrella" @eng ,
"She packs a gun when she goes into the mountains" @eng ;
wordnet-ontology: synset_member <http://wordnet-rdf.princeton.edu/wn31/carry-v> ,
<http://wordnet-rdf.princeton.edu/wn31/pack-v> ,
<http://wordnet-rdf.princeton.edu/wn31/take-v> ;
wordnet-ontology: verb_group <http://wordnet-rdf.princeton.edu/wn31/202642600-v> ;
owl:sameAs <http://lemon-model.net/lexica/sb/wn/WT/006277> ,
```
B.1.2.3 Winnow

http://wordnet-rdf.princeton.edu/wn31/201463566-v.ttl

Listing B.7: The synset for Winnow

```xml
@prefix lemon: <http://lemon-model.net/lemon#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix wordnet-ontology: <http://wordnet-rdf.princeton.edu/ontology#> .
@prefix rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix wordnet-ontology: <http://wordnet-rdf.princeton.edu/ontology#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .

<http://wordnet-rdf.princeton.edu/wn31/201463566-v> a wordnet-ontology:Synset ;
  rdfs:label "winnow"@eng ;
  wordnet-ontology: gloss "separate the chaff from grain by using air currents"@eng ;
  wordnet-ontology:hypernym <http://wordnet-rdf.princeton.edu/wn31/201462658-vc> ;
  wordnet-ontology:lexical_domain wordnet-ontology:verb .contact ;
  wordnet-ontology:part_of_speech wordnet-ontology:verb ;
  wordnet-ontology:sample "She stood there winnowing grain all day in the field"@eng ;
  wordnet-ontology:synset_member <http://wordnet-rdf.princeton.edu/wn31/winnow-w> ;
  owl:sameAs <http://lemon-model.net/lexica/uby/wm/WN_Synset_89365> ,
```

B.1.3 Adjectives

B.1.3.1 Bad

http://wordnet-rdf.princeton.edu/wn31/301129296-a.ttl

Listing B.8: The synset for Bad

```xml
@prefix lemon: <http://lemon-model.net/lemon#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix wordnet-ontology: <http://wordnet-rdf.princeton.edu/ontology#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .

<http://wordnet-rdf.princeton.edu/wn31/301129296-a> a wordnet-ontology:Synset ;
  rdfs:label "bad"@eng ;
```
B.1.3.2 Many

http://wordnet-rdf.princeton.edu/wn31/301555990-a.ttl
B.2 Sample SUMO results

The following extracts of information are taken from the WordNet OWL representation (Semantic Web Best Practices and Deployment Working Group, 2001) using the Standard Upper Merged Ontology (Niles and Pease, 2001; Niles et al.,
B.2.1 Nouns

B.2.1.1 Bee

Listing B.10: The Bee Class

```xml
@prefix opwn: <http://www.ontologyportal.org/WordNet.owl#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sumo: <http://www.ontologyportal.org/SUMO.owl#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

@Bee a owl:Class ;
  rdfs:label "bee"@en ;
  sumo:axiom <sumo:axiom976385803Mid-level-ontology.klf> ;
  sumo:equivalenceRelation opwn:WN30−102208656 ;
  sumo:externalImage "http://upload.wikimedia.org/wikipedia/commons/5/51/Apis_mellifera_bj.jpg"ˆˆxsd:anyURI ,
  "http://www.adamease.org/Articulate/sumopictures/pictures/animals/bugs/bee/bee.png"ˆˆxsd:anyURI ;
  sumo:subsumingRelation opwn:WN30−102208280 ,
  opwn:WN30−102210427 ,
  opwn:WN30−102210921 ;
  rdfs:isDefinedBy <http://www.ontologyportal.org/SUMO.owl> ;
  rdfs:subClassOf <sumo:Insect> ;
  owl:comment "A hairy Insect, some species of which produce honey and/or sting."@en .
```

B.2.1.2 Tongue

Listing B.11: The Tongue Class

```xml
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sumo: <http://www.ontologyportal.org/SUMO.owl#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```
B.2.2 Verbs

B.2.2.1 Weeping

Listing B.12: The Weeping Class
B.2.2.2 Giving

Listing B.13: The Giving Class

```xml
<sumo:Giving> a owl:Class ;
  rdfs:label """"giving""""@en ;
  sumo:axiom <sumo:sumo : axiom -1649778498 Military.kif> ,
  <sumo:sumo : axiom -2024056464Mid-level-ontology.kif> ,
  <sumo:sumo : axiom -2139476126Merge.kif> ,
  <sumo:sumo : axiom -421464319Merge.kif> ,
  <sumo:sumo : axiom 134628666FinancialOntology.kif> ,
  <sumo:sumo : axiom 1688421466Merge.kif> ,
  <sumo:sumo : axiom 695217888Merge.kif> ;
  sumo:equivalenceRelation <opwn:WN30-101086081> ,
  <opwn:WN30-200878636> ,
  <opwn:WN30-202199590> ;
  sumo:externalImage """"http://upload.wikimedia.org/wikipedia/commons/0/01/Gift_giving_ceremony.jpg"
  """"xsd:anyURI ,
  """"http://upload.wikimedia.org/wikipedia/commons/2/29/Morgan_giving_lecture.png"
  """"xsd:anyURI ,
  """"http://upload.wikimedia.org/wikipedia/commons/6/62/Gifts_xmas.jpg"
  """"xsd:anyURI ,
  """"http://upload.wikimedia.org/wikipedia/commons/e/cd/Mumbai-street-kids.jpg"
  """"xsd:anyURI ,
  """"http://upload.wikimedia.org/wikipedia/commons/f/f2/Fairbanks_Robin_Hood_giving_Marian_a_dagger.jpg"
  """"xsd:anyURI ,
  """"xsd:anyURI ;
  sumo:instanceRelation <opwn:WN30-1113266690> ;
  sumo:subsumingRelation <opwn:WN30-100097348> ,
  <opwn:WN30-100246659> ,
  <opwn:WN30-100205543> ,
  <opwn:WN30-100212808> ,
  <opwn:WN30-100213052> ,
  <opwn:WN30-100213186> ,
  <opwn:WN30-100213343> ,
  <opwn:WN30-100213482> ,
  <opwn:WN30-100260881> ,
  <opwn:WN30-100318035> ,
  <opwn:WN30-100318391> ,
  <opwn:WN30-101060530> ,
  <opwn:WN30-101083350> ,
  <opwn:WN30-101083645> ,
```

242
B.2.3 Adjectives

B.2.3.1 Two

The concept Two is subsumed under the concept of PositiveInteger in SUMO.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<sumo:PositiveInteger> a owl:Class ;
  rdfs:label "positive integer"@en ;
  sumo:axiom <sumo:axiom − 1010125721 Merge.kif>,
  <sumo:axiom − 1030169404 Merge.kif>,
  <sumo:axiom1647430257Merge.kif>,
  <sumo:axiom473174484Merge.kif>,
  <sumo:axiom642933629Merge.kif> ;
  sumo:equivalenceRelation <opwn:WN30−113728367>,
  <opwn:WN30−302186338>,
  <opwn:WN30−302186470> ;
  sumo:instanceRelation <opwn:WN30−113742573>,
  <opwn:WN30−113744044>,
  <opwn:WN30−113744521>,
  <opwn:WN30−113744722>,
  <opwn:WN30−113744916>,
  <opwn:WN30−113745086>,
  <opwn:WN30−113745270>,
  <opwn:WN30−113746512>,
  <opwn:WN30−113746672>,
  <opwn:WN30−113746785>,
  <opwn:WN30−113747199>,
  <opwn:WN30−113747448>,
  <opwn:WN30−113747469>,
  <opwn:WN30−113747606>,
  <opwn:WN30−113747725>,
  <opwn:WN30−113747865>,
  <opwn:WN30−113747989>,
  <opwn:WN30−113748128>,
  <opwn:WN30−113748246>,
  <opwn:WN30−113748367>,
  <opwn:WN30−113748493>,
  <opwn:WN30−113748622>,
  <opwn:WN30−113748763>,
  <opwn:WN30−113748890>,
  <opwn:WN30−113749017>,
  <opwn:WN30−113749146>,
  <opwn:WN30−113749278>,
  <opwn:WN30−113749407>,
  <opwn:WN30−113749527>,
  <opwn:WN30−113749644>,
  <opwn:WN30−113749778>,
  <opwn:WN30−113749894>,
  <opwn:WN30−113750164>,
  <opwn:WN30−113750297>,
  <opwn:WN30−113750415>,
  <opwn:WN30−113750504>,
  <opwn:WN30−113750574>,
  <opwn:WN30−113750712>,
\textless \text{opwn} : WN30 \text{−} 113750844\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113751158\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113751295\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113751404\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113751533\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113751829\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113752033\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113752443\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113752679\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113752911\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113753067\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113753274\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113753894\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113776432\textgreater ,
\textless \text{opwn} : WN30 \text{−} 302824825\textgreater ,
\textless \text{opwn} : WN30 \text{−} 302854257\textgreater ,
\textless \text{opwn} : WN30 \text{−} 400257784\textgreater ,
\textless \text{opwn} : WN30 \text{−} 400344500\textgreater ,
\textless \text{opwn} : WN30 \text{−} 400344659\textgreater ,
\textless \text{opwn} : WN30 \text{−} 400410317\textgreater ,
\textless \text{opwn} : WN30 \text{−} 400450382\textgreater ,
\textless \text{opwn} : WN30 \text{−} 400455508\textgreater ,
\textless \text{opwn} : WN30 \text{−} 400476680\textgreater ;
\textless \text{sumo:subsumingRelation} \text{opwn} : WN30 \text{−} 108272652\textgreater ,
\textless \text{opwn} : WN30 \text{−} 108272774\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113336368\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113342398\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113597588\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113597794\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113598408\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113598556\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113598715\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113598960\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113599114\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113599348\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113744304\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113745420\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113746449\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113747114\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113750033\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113751036\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113752372\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113753430\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113753585\textgreater ,
\textless \text{opwn} : WN30 \text{−} 113753740\textgreater ;
\textless \text{opwn} : WN30 \text{−} 113779804\textgreater ,
\textless \text{opwn} : WN30 \text{−} 400083541\textgreater ,
\textless \text{opwn} : WN30 \text{−} 400083666\textgreater ;
`rdfs:isDefinedBy <http://www.ontologyportal.org/SUMO.owl> ;
rdfs:subClassOf <sumo:NonnegativeInteger>,
<sumo:PositiveRealNumber> ;
owl:comment "An Integer that is greater than zero." @en .`
APPENDIX C

Ontology comparison calculations

C.1 Final word list comparison values

The costs described in Xue et al. (2009) were used as a basis for comparison in all the calculations. The totals for the transformation costs ($\gamma(\text{OP})$) where it is calculated as $\sum_{i=1}^{\text{|OP|}} \gamma(\text{Op}_i)$ and the calculations leading to the similarity index ($\gamma_{T_1 \rightarrow T_2}(\text{OP})$) are shown in Table C.1.

The first column contains the formula for the value being calculated, the second column contains the transformation cost from the first tree to the second tree and the third column contains the transformation cost from the second tree to the first tree.
Table C.1: Transformation cost and similarity index

<table>
<thead>
<tr>
<th>Value</th>
<th>$\gamma_{T_1 \rightarrow T_2}$</th>
<th>$\gamma_{T_2 \rightarrow T_1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sum_{i=1}^{\text{OP}} \gamma(\text{Op}_1)$</td>
<td>1.36</td>
<td>1.44</td>
</tr>
<tr>
<td>$\sum_{i \in \text{D}} \gamma(\text{delete}(i))$</td>
<td>0</td>
<td>0.85</td>
</tr>
<tr>
<td>$\sum_{i \in \text{D}} \gamma(\text{insert}(i))$</td>
<td>0.79</td>
<td>0</td>
</tr>
<tr>
<td>$\sum_{i \in \text{D}} \gamma(\text{move}(i))$</td>
<td>0.56</td>
<td>0.59</td>
</tr>
<tr>
<td>$\sum_{i \in \text{D}} \gamma(\text{relabel}(i))$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\sum_{i \in \text{D}} \gamma(\text{delete}(i)) + \sum_{i \in I} \gamma(\text{insert}<em>u(i)) + \sum</em>{i \in \text{M}} \gamma(\text{move}(i)) + \sum_{i \in R} \gamma(\text{relabel}(i))$</td>
<td>1.36</td>
<td>1.44</td>
</tr>
<tr>
<td>$\gamma_{T_1 \rightarrow T_2} (\text{OP})$</td>
<td></td>
<td>1.36</td>
</tr>
</tbody>
</table>

C.2 Calculation details

In order to arrive at the results in Table C.1, calculations were required on each node in the tree. These calculations per node are illustrated in Table C.2.

The columns contain respectively the operation type, the WordNet reference sense, the depth of the node, the descendants of the node, the deletion cost, the insertion cost, the movement cost, the relabelling cost and the transformation cost.

In the calculations:

- $D$ is the discourse domain
- $M$ is the injective mapping from $V \rightarrow L^V$
- $I$ is the set of nodes to be inserted into $T_1$
- $R$ is the set of nodes to be re-labelled
Table C.2: Tree cost calculations

<table>
<thead>
<tr>
<th>Op</th>
<th>Sense</th>
<th>depth(v)</th>
<th></th>
<th>D(v)</th>
<th>delete(v)</th>
<th>insert(v)</th>
<th>move(v)</th>
<th>relabel(v)</th>
<th>(\gamma(Op_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>relabelling</td>
<td>Air:1</td>
<td>4</td>
<td>1</td>
<td>0.18</td>
<td>0.16</td>
<td>0.17</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Be:1</td>
<td>0</td>
<td>252</td>
<td>3.99</td>
<td>3.97</td>
<td>3.86</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Bitter:1</td>
<td>2</td>
<td>1</td>
<td>0.21</td>
<td>0.19</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Bad:1</td>
<td>1</td>
<td>0</td>
<td>0.21</td>
<td>0.19</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Two:1</td>
<td>1</td>
<td>0</td>
<td>0.21</td>
<td>0.19</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Dance:1</td>
<td>1</td>
<td>4</td>
<td>0.27</td>
<td>0.25</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Rotten:3</td>
<td>1</td>
<td>0</td>
<td>0.21</td>
<td>0.19</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Pool:2</td>
<td>5</td>
<td>7</td>
<td>0.25</td>
<td>0.24</td>
<td>0.24</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Do:1</td>
<td>0</td>
<td>5</td>
<td>0.3</td>
<td>0.28</td>
<td>0.28</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Egg:2</td>
<td>6</td>
<td>1</td>
<td>0.15</td>
<td>0.13</td>
<td>0.14</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Walk:1</td>
<td>1</td>
<td>54</td>
<td>1.01</td>
<td>1</td>
<td>0.98</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Give:3</td>
<td>1</td>
<td>348</td>
<td>5.4</td>
<td>5.39</td>
<td>5.23</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Twin:1</td>
<td>9</td>
<td>4</td>
<td>0.15</td>
<td>0.13</td>
<td>0.14</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>insert</td>
<td>Winnow:1</td>
<td>4</td>
<td>1</td>
<td>0.18</td>
<td>0.16</td>
<td>0.17</td>
<td>0.34</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Heat:1</td>
<td>1</td>
<td>11</td>
<td>0.37</td>
<td>0.36</td>
<td>0.35</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Cool:1</td>
<td>1</td>
<td>4</td>
<td>0.27</td>
<td>0.25</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Roast:1</td>
<td>3</td>
<td>2</td>
<td>0.21</td>
<td>0.19</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Ember:1</td>
<td>5</td>
<td>1</td>
<td>0.16</td>
<td>0.15</td>
<td>0.15</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Grow:2</td>
<td>3</td>
<td>70</td>
<td>1.22</td>
<td>1.21</td>
<td>1.18</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>insert</td>
<td>Firewood:1</td>
<td>5</td>
<td>6</td>
<td>0.24</td>
<td>0.22</td>
<td>0.22</td>
<td>0.46</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Enrich:1</td>
<td>2</td>
<td>7</td>
<td>0.3</td>
<td>0.28</td>
<td>0.28</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Stoop:1</td>
<td>3</td>
<td>3</td>
<td>0.22</td>
<td>0.21</td>
<td>0.21</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Die:1</td>
<td>2</td>
<td>10</td>
<td>0.34</td>
<td>0.33</td>
<td>0.33</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Eye:1</td>
<td>7</td>
<td>8</td>
<td>0.24</td>
<td>0.22</td>
<td>0.22</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Name:1</td>
<td>5</td>
<td>70</td>
<td>1.19</td>
<td>1.18</td>
<td>1.15</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Eat:1</td>
<td>2</td>
<td>19</td>
<td>0.48</td>
<td>0.46</td>
<td>0.46</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>relabelling</td>
<td>Dig:1</td>
<td>1</td>
<td>8</td>
<td>0.33</td>
<td>0.31</td>
<td>0.31</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>insert and</td>
<td>'Numida</td>
<td>1</td>
<td>0</td>
<td>0.21</td>
<td>0.19</td>
<td>0.2</td>
<td>0.4</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>move</td>
<td>melea-</td>
<td>gris':1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>move</td>
<td>Tick:2</td>
<td>11</td>
<td>12</td>
<td>0.24</td>
<td>0.22</td>
<td>0.22</td>
<td>0.46</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>move</td>
<td>Poultry:2</td>
<td>7</td>
<td>16</td>
<td>0.36</td>
<td>0.34</td>
<td>0.34</td>
<td>0.7</td>
<td>0.34</td>
<td></td>
</tr>
</tbody>
</table>

Continued on next page
| Op         | Sense | depth(v) | $|D(v)|$ | delete(v) | insert(v) | move(v) | relabel(v) | $\gamma_{Op_i}$ |
|------------|-------|----------|-------|----------|-----------|---------|------------|-----------------|
| relabelling| Short:2| 1        | 0     | 0.21     | 0.19      | 0.2     | 0          | 0               |
| relabelling| Beard:1| 9        | 8     | 0.21     | 0.19      | 0.2     | 0          | 0               |
| relabelling| Tongue:1| 6        | 1     | 0.15     | 0.13      | 0.14    | 0          | 0               |
| relabelling| Cry:2  | 1        | 6     | 0.3      | 0.28      | 0.28    | 0          | 0               |
| relabelling| Pig:1  | 14       | 2     | 0.04     | 0.03      | 0.04    | 0          | 0               |
| relabelling| Bite:2 | 2        | 2     | 0.22     | 0.21      | 0.21    | 0          | 0               |
| relabelling| Dog:1  | 13       | 190   | 2.87     | 2.85      | 2.77    | 0          | 0               |
| relabelling| Unguis:1| 6       | 10    | 0.28     | 0.27      | 0.27    | 0          | 0               |
| relabelling| Meat:1 | 5        | 198   | 3.1      | 3.09      | 3       | 0          | 0               |
| relabelling| Sangoma:1| 9      | 1     | 0.1      | 0.09      | 0.09    | 0          | 0               |
| relabelling| Year:2 | 5        | 2     | 0.18     | 0.16      | 0.17    | 0          | 0               |
| relabelling| Child:2| 9        | 31    | 0.55     | 0.54      | 0.53    | 0          | 0               |
| relabelling| Snake:1| 11       | 115   | 1.78     | 1.76      | 1.72    | 0          | 0               |
| relabelling| Bee:1  | 11       | 15    | 0.28     | 0.27      | 0.27    | 0          | 0               |
| relabelling| Person:1| 6       | 6979  | 104.3    | 104.28    | 101.18  | 0          | 0               |
| relabelling| Many:1 | 1        | 0     | 0.21     | 0.19      | 0.2     | 0          | 0               |
| relabelling| Drink:1| 1        | 10    | 0.36     | 0.34      | 0.34    | 0          | 0               |
| relabelling| Whole:1| 1        | 0     | 0.21     | 0.19      | 0.2     | 0          | 0               |
| relabelling| Bask:2 | 1        | 1     | 0.22     | 0.21      | 0.21    | 0          | 0               |
| relabelling| Flutter:3| 2    | 2     | 0.21     | 0.19      | 0.2     | 0          | 0               |
| relabelling| Rain:1 | 9        | 9     | 0.22     | 0.21      | 0.21    | 0          | 0               |
| relabelling| Three:1| 1        | 0     | 0.21     | 0.19      | 0.2     | 0          | 0               |
| relabelling| Smelt:1| 1        | 1     | 0.22     | 0.21      | 0.21    | 0          | 0               |
| insert     | Carry:2| 1        | 1     | 0.22     | 0.21      | 0.21    | 0.43       | 0.21            |
| relabelling| Branch:2| 8      | 10    | 0.25     | 0.24      | 0.24    | 0          | 0               |
| relabelling| Tail:1 | 6        | 10    | 0.28     | 0.27      | 0.27    | 0          | 0               |
| relabelling| Smoke:1| 8        | 3     | 0.15     | 0.13      | 0.14    | 0          | 0               |
| relabelling| Ten:1  | 1        | 0     | 0.21     | 0.19      | 0.2     | 0          | 0               |
| relabelling| Louse:1| 10       | 5     | 0.15     | 0.13      | 0.14    | 0          | 0               |
| relabelling| Long:1 | 1        | 0     | 0.21     | 0.19      | 0.2     | 0          | 0               |
| relabelling| Hunger:1| 7     | 6     | 0.21     | 0.19      | 0.2     | 0          | 0               |
| relabelling| Path:1 | 7        | 6     | 0.21     | 0.19      | 0.2     | 0          | 0               |
| relabelling| Appease:2| 4    | 1     | 0.18     | 0.16      | 0.17    | 0          | 0               |
| Op          | Sense | depth(v) | |D(v)| | delete(v) | | insert(v) | | move(v) | | relabel(v) | | γ(Op) |
|------------|-------|----------|--------|-----|--------|--------|--------|--------|--------|--------|--------|
| relabelling | Try:1 | 1        | 22     | 0.54 | 0.52  | 0.51   | 0      | 0      |
| relabelling | Plait:1 | 3       | 1      | 0.19 | 0.18  | 0.18   | 0      | 0      |
| relabelling | With:1 | 1        | 0      | 0.21 | 0.19  | 0.2    | 0      | 0      |
| relabelling | Twinkle:1 | 2      | 2      | 0.22 | 0.21  | 0.21   | 0      | 0      |
Abstract of publication

The following is an abstract of a publication (Anderson et al., 2010) that resulted from the core research for this dissertation referenced in Section 1.6.

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

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?
APPENDIX E

Bantu Base Concept
subsumption in SUMO

Figure E.1: Bantu Base Concept subsumption in SUMO: refers and class
Figure E.2: Bantu Base Concept subsumption in SUMO:physical processes

Figure E.3: Bantu Base Concept subsumption in SUMO:physical objects
Figure E.4: Bantu Base Concept subsumption in SUMO:abstract