ALEXANDRU IOAN CUZA
UNIVERSITY of IAŞI

Proceedings of the Eighth Global WordNet Conference

Editors:

Verginica Barbu Mititelu, Corina Forăscu, Christiane Fellbaum, Piek Vossen

Bucharest, Romania, January 27-30, 2016

ISBN 978-973-0-20728-6

Semantics of body parts in African WordNet: a case of Northern Sotho

Mampaka Lydia Mojapelo

University of South Africa

Department of African Languages

mojapml@unisa.ac.za

Abstract

This paper presents a linguistic account of the lexical semantics of body parts in African WordNet, with special reference to Northern Sotho. It focuses on external human body parts synsets in Northern Sotho. The paper seeks to support the effectiveness of African WordNet as a resource for services such as in the healthcare and medical field in South Africa. It transpired from this exploration that there is either a one-to-one correspondence or some form of misalignment of lexicalisation with regard to the sample of examined synsets. The paper concludes by making suggestions on how African WordNet can deal with such semantic misalignments in order to improve its efficiency as a resource for the targeted purpose.

1 Introduction

African WordNet is a project that aims to build a lexical database for all indigenous official languages of South Africa, which will be linked to one another. It is modelled on Princeton WordNet[1] through the expand approach (Vossen, 1998). The approach was informed by experiences shared by earlier Wordnets such as BalkaNet, MultiWordNet, and other languages in the EuroWordNet, to name but a few. The expand approach takes synonym sets (synsets) from Princeton WordNet, with their relations, and convert them into the target language. The approach already lends the development of African Wordnets to the use of more than one language, that is, English and the target language concerned. African WordNet is further internally multilingual with five out of nine official African languages of South Africa that are currently part of the project. Northern Sotho (Sesotho sa Leboa)[2] is one of the languages involved. The premise in building African Wordnets is to model it on the Princeton structure while staying true to the African context.

Among the challenges that were encountered in the process of building African Wordnets was that some of the synsets extracted from Princeton for development of African WordNet did not make immediate sense for African languages and the African context, for a number of reasons. For example, among them are synsets for concepts that are geographically distant from the South African context, such as animal and plant species. This situation would result in non-lexicalised concepts. Some non-lexicalised concepts were left blank and for some it was decided that available linguistic resources would be used for coinage and borrowing. The envisaged convenience of African WordNet became clearer to the writer (a linguist, project translator or lexicographer) through other synsets of a more general nature that were easy to work with. One of the semantic domains that was considered generally applicable to any context was Anatomy, BodyPart. It was assumed that this kind of domain would have relatively fewer gaps compared to domains that are geographically or culturally more restricted. BodyPart also ranks ninth among the 50 most frequently suggested upper merged ontologies (SUMOs) in Princeton WordNet (PWN), as at 2014-03-11.

The downside of BodyPart was that the synsets extracted from Northern Sotho showed that none of the synsets done so far were aimed at the human anatomy. The SUMO_BodyPart consisted of words that were unrelated to humans, such as 'scale' (as in fish-scale), 'shell', 'paw', 'feather' and 'wool'. Other examples to illustrate unrelatedness to humans is that the senses of the word *seatla* 'hand' were limited to Domain_Transport, SUMO_Device

[1] http://wordnet.princeton.edu

[2] Cf. Guthrie's zone S30

and Domain_Factotum, SUMO_Constant Quantity, and denotation to parts of the human body did not feature. Similarly, the senses of *leoto* 'leg' were limited to Domain_Factotum, SUMO_Shape-Attribute and Domain_Zoology, SUMO_Mammal, which is a different synset from Domain_Anatomy, SUMO_BodyPart. This paper was premised on the understanding that, comparatively speaking, non-human body parts and other domains mentioned here may not demonstrate the immediate and direct societal impact of African WordNet to the extent that may be achieved with human body parts.

South Africa is a multilingual and multicultural country. According to the latest South African statistics (Statistics South Africa, 2012) on the use of home languages only 9,6% of the general population speak English as their home language (L1), while the majority speak the other ten official languages and their dialects as L1. The remainder (>90%) speak English either as a second, third or fourth language or not at all. Among this vast majority are healthcare workers, medical students and practitioners, as well as individuals and communities who should receive healthcare and medical services. Another issue is that studies incidental to most academic qualifications in South Africa are presented through the medium of English, which inevitably means that most students learn through a foreign medium. For some English schooling starts before they have duly mastered their L1. This apparent disadvantage is balanced by the foundation laid in English, which will give the student a significant headstart in his or her academic career, still with insufficient knowledge of his or her L1. L1 English speakers on the other hand are not motivated to learn other languages until they have completed their studies and happen to find themselves in an occupational environment where they have to adjust to a different language medium. It may therefore be useful to provide a multilingual platform for accessing domain lexicons on a level that is more than just a dictionary. Terminology lists and glossaries are being developed for various purposes in South Africa, including healthcare and medicine, but none of these is an African language Wordnet. African WordNet will not only provide definitions and contextual usages of words, but will be based on synsets. Synsets are sets of lexicalisations of a particular concept, and WordNet links them to other concepts through semantic relations such as hyponymy and meronymy, in the case of nouns. African WordNet will further link the languages spoken in the country to each other.

2 About the body parts lexicon in Northern Sotho

Since the available body-parts synsets in the Northern Sotho Wordnet were deemed not immediately useful for human healthcare and medicine purposes, the writer considered exploring external human body parts, which will later be followed by internal ones to complete the healthcare and medical intent. A list was drawn, verified and augmented against Northern Sotho Language Board (1988) as well as Ziervogel & Mokgokong (1975) and a paper in progress on verbs expressing physical pain. The list had Northern Sotho and English equivalents. Already when giving equivalents outside WordNet it emerged that there may be misalignment in the form of general-specific lexicalisation of senses. For example, Northern Sotho uses the same word for 'finger' and 'toe'. Unless the difference is readily apparent from the context a descriptive phrase is used for ease denotation. The question is: How big is the misalignment and how are we going to solve the problem linguistically? The sample used here is used as an index of misalignments, as well as possible solutions, for the rest of the development of the Northern Sotho Wordnet. The next step was to match the body parts on the list with English synsets.

3 Lexical entries in Northern Sotho Wordnet

In keeping with Princeton the lexical entries in African WordNet are guided by information such as part of speech (POS), domain, SUMO, definition, usage and the English ID. This paper focuses on the Northern Sotho nouns under the Domain_Anatomy, SUMO_BodyPart. According to the definition and usage provided in English as well as the ID, only body parts that are specifically human were picked out. Fellbaum (1998) contends that although the majority of lexicalised concepts are shared among languages, not every language will have words denoting concepts that are lexicalised in other languages. Therefore it is expected that some concepts may

be lexicalised in English and not in Northern Sotho, and *vice versa*. It is deemed necessary for this semantic domain to have as many lexicalised concepts as possible, given the envisaged use in the healthcare sector. The paper will also look into these semantic relations and ensure that the Northern Sotho synsets are presented in a manner that is not misconstrued.

Lexicalisation is defined as realisation of meaning in a single word or morpheme where words are already present in a language, as well as the addition of new words as new concepts enter the languages in due course. The addition of new words involves strategies of word formation such as compounding, derivation and borrowing. Another issue to lexicalisation is some level of acceptability among the speakers of a language, which will lead to general acceptability. The body-parts synsets in Northern Sotho reflect different types of lexicalisation, including addition of new words by the strategies mentioned above. There are also cases of non-lexicalisation which have yet to be resolved.

Although the expand approach has proved to be most expedient for new wordnets, lexicalisation challenges are inevitable for most of them. For example, in building the Konkani WordNet from Hindi WordNet (Walawalikar et. al 2010), which is a closely related language, some challenges were experienced. The challenges also involved the English source and they include linking errors, missing entries, definitions, concept misalignment and lexicalisation. The issue of culture-specificity is also reported as one of the causes of misalignment. In dealing with alignment in the Hebrew WordNet, which was also built on the expand approach; Ordan and Winter (2007) distinguish between contingent and systematic instances of non-equivalence. The two cases attest to the fact that lexicons of different languages mirror misalignments of both cultural and internal language structural nature.

Vincze and Almási (2014) also treat lexicalisation challenges encountered in dealing with the Hungarian WordNet. The intention of this paper is not to reinvent the wheel but to learn from others' experiences in the realisation that languages may be dissimilarly resourced, materially and structurally. Northern Sotho is a Bantu language of the Niger-Congo language family, which is agglutinating with productive morphology. Therefore one lexicalisation type or mechanism may prove to be more practical than another. For the purpose of this paper it is assumed that Northern Sotho may be differently resourced, given the object to explore how the project can try to solve extant misalignment challenges without losing the Princeton structure while remaining true to the African context, a manoeuvre requiring a certain amount of fineness.

4 Queries and results

To begin, the items on the list were queried from the English dictionary in DEBVisDic (WordNet editor and browser). Only sense 1 of SUMO_BodyPart under Domain_Anatomy was selected. The definitions, usages and synset IDs were used to obtain correct matches. General personal knowledge of Northern Sotho, as a mother tongue speaker, was complemented and verified against the Northern Sotho-English bilingual and Northern Sotho-English-Afrikaans trilingual dictionaries. The results gained from the queries confirmed some degree of misalignment between Northern Sotho and English. Clearly no comment is required on the one-to-one matches. The examples used here represent one-to-many and many-to-one mappings as well as lexicalisation gaps.

A sample of words representing 88 Northern Sotho concepts, with English equivalents, was used. The list is not exhaustive, but it is a fair representation of external human body parts. Also, not all possible connections have been indicated in the illustrations. While the initial focus was on external body parts, parts of the oral cavity were included as they are too close to the external facial body-parts and not as concealed as other internal body-parts. The English equivalents of the Northern Sotho words on the list were browsed and their IDs noted in order that their definitions and usages establish correct matches.

Queried senses in English (anatomy, human body part) were not found for the following words:

head

big hair

hair on arms and legs

protruding forehead

eye ridge

cheek

tongue

adam's apple

below the buttock (where the thigh starts)

back of hand

back

back of knee

foot

heel

When queried, the relevant senses of the words above could not be matched with the IDs found in DEBVisDic. A peculiar gap in English on human body parts relates to 'head', 'cheek', 'tongue', 'adam's apple', 'back', 'foot' and 'heel'. It is assumed that the rest of the words may be more physiologically or culturally relevant in Northern Sotho than in English. While it is still peculiar to some extent that 'back' was not found because physiologically, especially in the healthcare and medical context, the concept should have the same denotative significance in both languages, the gap was understood in the context of possible cultural dissimilarities. *Mokokotlo* 'back', as in the 'back part of the human torso', is one of the most recognisable lexical items in Northern Sotho due to what the concept represents. It is the part of the body that a baby or toddler is carried and strapped on for guaranteed safety and protection. In this context the back is culturally associated with care, nurturing, raising, acceptance and protection. The concept (and therefore the word) is culturally significant. With regard to *setšhitšhi* 'big hair' (not the same as 'long hair', which would be natural in the English lexicon) the gap in English is understood to be due to physiological difference.

Halliday et. al. (2004) explicate at length problems of cross-language mapping even for concepts that seem simple such as kinship terms. The examples of siblings and cousins between English and Australian Pitjantjatjara resonate with Northern Sotho and other Bantu languages. Therefore the issue of misalignment is not only a matter of lexical items, but of concepts as well.

The following diagrams provide reference for the current discussion. For every Northern Sotho lexical item, an English translation equivalent is provided. For combined connections, refer to appendix 1.

Diagram 1: Arm connections

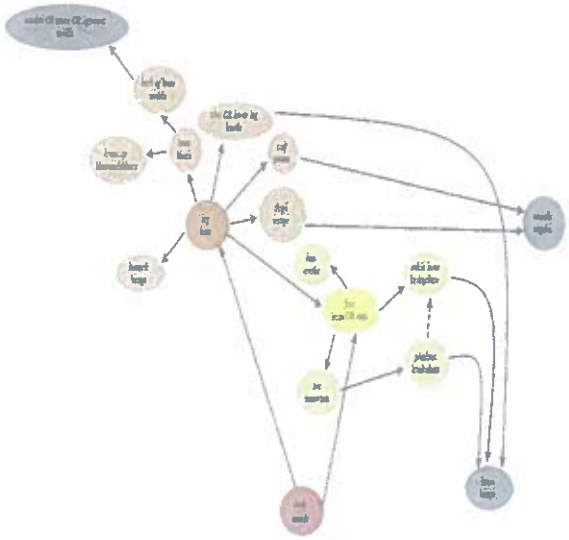

Diagram 2: Leg connections

Diagram 3: Torso connections

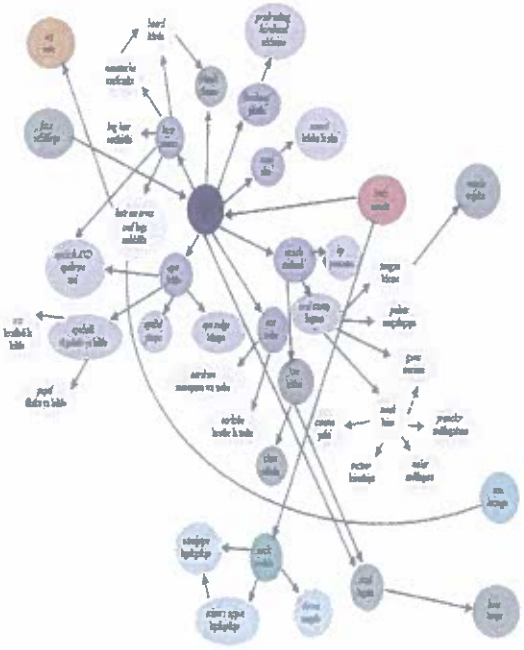

Diagram 4: Head connections

4.1 One-to-many and many-to-one

Two types of misalignment will be used for illustration here. There are cases of Northern Sotho lexicalisation of human body parts that mingle synonymy and meronymy, not in a confusing way though. In the context of WordNet words are synonymous if they express the same concept and can be interchanged in some contexts (Fellbaum 1998). Meronymy is explained by Croft and Cruse (2004) as a sense relation between meanings rather than between individual entities, that is when the meaning of one word is part of the meaning of another. The word for 'hand' in Northern Sotho is *seatla*. It expresses the same concept expressed by [POS: n ID: ENG 20-05246212-n BCS: 3], which is sense 1 of the Domain_ Anatomy, SUMO_Bodypart and defined in English as "the (prehensile) extremity of the superior limb". *Letsogo* is Northern Sotho for 'arm' [POS: n ID: ENG 20-05245410-n BCS: 3], Arm: 1, defined in English as "a human limb; technically part of the superior limb between the shoulder and the elbow but commonly used to refer to the whole superior limb". In Northern Sotho *letsogo* refers to the whole superior limb, which includes the hand. According to the definition provided above the common usage of the English 'arm' is the same as the Northern Sotho *letsogo*, but the technical usage is not. In Northern Sotho the word *letsogo* is also used to refer to *seatla* 'hand', but the whole limb is never called *seatla*. That is, while *seatla* 'hand' is a meronym of *letsogo* 'arm', the two are also synonymous. Similarly *leoto* 'leg' [ENG20-05242579-n] is used for both 'leg' and 'foot' while a separate specific word for 'foot' is *lenao*. These examples illustrate lexicalisation that reflects the occurrence of meronymy between lexical items that are also synonymous.

Another scenario relates to the case of *monwana* for both 'toe' [ENG20-05258265-n] and 'finger' [ENG20-05247839-n], and *ntši* for 'eyebrow' [ENG20-05007503-n] and 'eyelash' [ENG20-05008887-n]. In this case Northern Sotho uses one word to express separate concepts, or concepts that are viewed as separate in English. These two examples illustrate that the words *monwana* and *ntši* are used in Northern Sotho as hypernyms. Descriptive phrases 'of the foot' and 'of the hand' are used as hyponyms of *monwana* in cases where distinction is deemed necessary. A similar descriptive strategy is not used for *ntši*; it may also be cumbersome as both 'eyebrow' and 'eyelash' belong to the eye.

4.2 Possible non-lexicalisation in English

Another concept that is lexicalised in Northern Sotho but could not be found from querying the English in DEBVisDic is *nyaraga* (Mokgokong and Ziervogel 1975), also pronounced *nyarago*. The English trees relating to 'leg' and 'buttock' were examined as the concept is understood to be either a body part below the buttock or the uppermost back part of the leg. Its absence in the two trees pointed to possible non-lexicalisation.

The following section proposes possible linguistic means of catering for the misalignment issues mentioned above in African WordNet.

5 Handling misalignments

It is necessary to provide linguistic solutions to the misalignment challenges mentioned above. Vincze and Almási (2014) suggest a number of strategies for the Hungarian lexicalisation issues, namely to shorten the tree, flatten the tree, restructure the tree and lexicalize the concepts. They are also of the opinion that the merge approach would have alleviated some of the challenges. For Konkani Walawalikar et. al (2010) suggest, among others, that the target language synsets for which there were gaps in the source language could be used to fill the gaps, thereby strengthening the HWN. Ordan and Winter (2007) detail strategies for building Hebrew synsets, which include linking Hebrew word senses to related PWN sysnsets from Hebrew to English and from English to Hebrew. Lexical gaps from both sides are acknowledged and used to preserve and link semantic information.

This paper takes a linguistic view to addressing the challenges mentioned above, which relate to lexicalisation of the concepts. The first group of Northern Sotho words which could not be matched from English seem to be a matter of misses which can be addressed if probed further. The next situation concerns *seatla* 'hand' and *lenao* 'foot' which are meronyms of *letsogo* 'arm' and *leoto* 'leg', respectively, and proved to be synonymous as well. Therefore lexical items *seatla* and *letsogo* will be in the same synset while they are meronymically related as well. The same applied to *lenao* 'foot' and *leoto* 'leg'.

The next issue concerns *monwana* 'finger' and 'toe' and *ntši* 'eyelash' and 'eyebrow'. In the language synonyms for *monwana* are provided in the form of descriptive phrases to distinguish 'finger' and 'toe'. The descriptions *wa lenao* and *wa leoto* 'of the foot'; *wa seatla* and *wa letsogo* 'of the hand' are consistent with language usage and are not expected to pose any problems. The same solution cannot work in the case of *ntši* since eyebrow and eyelash are both 'of the eye'. Northern Sotho Language Board (1988) uses compounding as a strategy to distinguish the two. While they are both *ntši* the source coined *ntšikgolo* as additional lexicalisation for 'eyebrow'. The second component of the compound *-kgolo* (*-golo*) 'big' suggests that an eyebrow is dominant. The source was produced by a standardising body (Northern Sotho Language Board) which was obviously cognisant of the gaps in terms of lexicalisation. They probably considered either the overaching position of the eyebrow in relation to the eyelashes or the perceived amount of hair in both, to come up with a suggestion that an eyebrow is the main *ntši*. Another example of compounding from the same source is *khurumelakhuru* for 'kneecap'. *-khurumela* is a verb stem which means to close or to cover. *Khuru* is 'knee'. Therefore conceptualisation points to something that covers, closes off or protects the knee. Lexicalisation strategies such as these provide promising resources for African WordNet. What remains is whether or not such lexical items will filter down to everyday usage.

The last issue relates to the apparent English non-lexicalisation of concepts that are lexicalised in Northern Sotho, and *vice versa*. *Nyaraga* 'below the buttocks' is part of the Northern Sotho lexicon whose lexicalisation could not be ascertained in English. The English equivalent is provided in Northern Sotho dictionaries as a phrase. The English lexicalisation of the Northern Sotho *ntahle* 'back of hand' could also not be ascertained. Over and above being a body part, part of a hand, *ntahle* has an added connotation relating to slapping (backhand slap). That is, slapping someone with the inner part of a hand and the outer part of a hand would be reflected by the use of different lexical items. Such words need to be added as they represent concepts that are intertwined with the idiom of the language.

An expected scenario of the expand approach where English is the source language would obviously reveal Northern Sotho non-lexicalisation of concepts that are lexicalised in English. With regard to the domain under

discussion descriptive phrases are common, for example 'nose' is *nko* and 'nostril' is *lešoba la nko*, literally 'hole of nose'. 'Pubis' is *lerapo la pele la noka*, literally 'bone of front of waist'. Another lexicalisation mechanism that is productive in Bantu languages, which was nevertheless not observed in the current sample, is derivation. Affixes are used productively to form words from different word categories. Direct borrowing is also not evident in the current sample, but it is commonly used in the lexicalisation of technological concepts and specific disease names. From this sample an example of indirect borrowing is evident in coinage that resembles the English formations such as *khurumelakhuru* above and *moropana wa tsebe* literally 'small drum of ear' for 'eardrum'. Lexicalisation mechanisms that were employed for this sample hint at linguistic routes to follow in dealing with further development of human body parts.

6 Challenges

While the linguistic side of the project may prove exciting, there are challenges of an IT nature. The challenges include changes in the IT infrastructure at the hosting institutions, as well as problems with the DEBVisDic editor. Such challenges hamper the development of the wordnets, as they result in interrupted access to the server and inconsistent functionality of the editor. This becomes a challenge if one wants to browse and edit existing synsets, or add new synstes. Nonetheless, manual and semi-automatic data gathering methods are used so that when a permanent IT solution is reached there is enough linguistic data to fast-track the development of the wordnets.

7 Conclusion

The paper presented actual and possible scenarios that may pose challenges when developing the Northern Sotho Wordnet on Domain_Anatomy, SUMO_BodyPart. Human body parts are targeted in this paper due to their connection to human health care and medicine. Many speakers whose L1 is not Northern Sotho may benefit from the database as it will be linking Northern Sotho not only to English but to other South African indigenous languages as well. Not only were different types of lexical misalignment presented, but also lexicalisation mechanisms that are used in the language. While the mentioned mechanisms may be grammatically sound and fill lexicalisation gaps, the words also need to receive general acceptability to the point of being in reasonably high frequency used rather than merely existing.

It is envisaged that the proposed strategies will fill the gaps, and that inclusion of internal body parts and functions, as well as verbs of expressing physical pain will produce trees that mirror the language. It remains to be seen how far the translators in the project will go in utilising the lexicalisation strategies mentioned in this paper. To assist with acceptability and standardisation the synsets will also be shared with selected practitioners in the target field for comments.

Acknowledgement

Ms Marissa Griesel, for support with the illustrations

References

Christiane Fellbaum (Ed). 1998. *Wordnet: an electronic lexical database*. Cambridge, Mass: The MIT Press.

Dirk Ziervogel and Pothinus C. Mokgokong. 1975. *Pukuntšu ye kgolo ya Sesotho sa Leboa/ Comprehensive Northern Sotho dictionary/ Groot Noord-Sotho woordeboek*. Pretoria: J. L. Van Schaik.

M.A.K Halliday, Wolfgang Teubert, Colin Yallop and Anna Čermáková. 2004. *Lexicology and Corpus Linguistics: an introduction*. London: Continuum.

Noam Ordan and Shuly Wintner. 2007. Hebrew WordNet: a test case of aligning lexical databases across languages. *International Journal of Translation* 19(1):39-58.

Northern Sotho Language Board. 1988. *Sesotho sa Leboa Mareo le Mongwalo No. 4/ Northern Sotho Terminology and Orthography No. 4/ Noord-Sotho Terminologie en Spelreëls No. 4*. Pretoria: Government Printer.

Piek Vossen (ed.) 1998. *EuroWordNet:A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

Shantaram Walawalikar, Shilpa Desai, Ramdas Karmali, Sushant Naik, Damodar Ghanekar, Chandralekha D'Souza and Jyoti Pawar. 2010. Experiences in Building the Konkani WordNet Using the Expansion Approach. In *Proceedings of the Fifth Global WordNet Conference, January* 2010. Mumbai, India.

Statistics South Africa http://www.stassa.gov.za accessed on 19 August 2015

Verinika Vincze and Attila Almási (2014) In *Proceedings of the Seventh Global WordNet Conference,* January 2014. University of Tartu, Estonia.

William Croft and D. Alan Cruse 2004. *Cognitive linguistics*. Cambridge: University Press.