Considerations for a model for NCB noun classes in Wikidata

C. Maria Keet
Department of Computer Science
University of Cape Town
Cape Town, South Africa
mkeet@cs.uct.ac.za

Langa Khumalo
South African Centre for Digital
Language Resources (SADiLaR)
Potchefstroom, South Africa
Langa.Khumalo@nwu.ac.za

Zola Mahlaza
Department of Informatics
University of Pretoria
Pretoria, South Africa
z.mahlaza@up.ac.za

ABSTRACT

Abstract Wikipedia aims to document lexicographic information in Wikidata. A key stumbling block to realise this for Niger-Congo B languages, is the centrality of its noun class system that governs the rest of a sentence. At present, Wikidata has almost no data and information, and what is there is an unusable encoding for the NCB languages and the required tasks. In this abstract, we present some first steps in the direction of the creation of an inclusive model.

ACM Reference Format:
C. Maria Keet, Langa Khumalo, and Zola Mahlaza. 2022. Considerations for a model for NCB noun classes in Wikidata. In *Proceedings of Wiki Workshop 2022 (Wiki Workshop '22).* ACM, New York, NY, USA, 2 pages. .

1 INTRODUCTION

Abstract Wikipedia [14] requires lexicographic information in Wikidata, along with the natural language generation functions. The noun class (NC) system is emblematic for the up to 700 Niger-Congo B ('Bantu') languages, which are spoken in Sub-Saharan Africa by some 350 million people. Nouns with different categories of referents are classified into different NCs; the general idea with some examples are included in Table 1. It affects the morphosyntax of the rest of the sentence, which poses challenges in the computation of these NCB languages. For instance, consider verb conjugation and adjectives: *inja* 'dog' (NC 9), *umuntu* 'human' (NC 1), *-dla* 'eat', and *-de* 'tall': it is *inja ende iyadla* 'the tall dog eats' but *umuntu omude uyadla* 'the tall human eats' (differences underlined).

At present, the various resources, including Wikidata, contain very little information and in an unusable encoding for the NCB languages. General language models such as [5, 9] lack the required level of detail. There are ad hoc approaches, like for the isiZulu and related verbalisers [2, 8], models for recording the NC of nouns given a particular NC system [3] and then typically tailored to a particular NCB language or a family thereof, or for more or all part-of-speeches [4, 6, 13], and a model [1] that makes interoperability across the different linguistic categorisations and theories challenging as a result of its design. Thus, there is not one that will work across all languages and respects alternate NC systems that have been proposed over the years, nor the different names that some languages may give to their NCs. An example of such informal naming, is Kiswahili's "u-class"[1], which is problematic because there are NCs in other languages that have either different or more than one class starting with an *u-* as class prefix; e.g., for isiZulu, NCs 1a, 3a, and 11 have an *u-* prefix, whereas they are 11 and 14 for Kiswahili. Also the vernacular practice of lumping singular and plural classes together, like a 'class 1/2', is problematic, since the pairings can differ by language or one language has a pairing but not another; e.g., isiZulu has only 14 singular, but Chichewa has 14 singular that goes with 6 plural. Such shortcomings are, however, useful for casting them as requirements for a model.

Table 1: Semantic generalisation of the categories of objects the nouns refer to in the respective noun classes, with examples from isiZulu (1-11, 14, 15; South Africa), Chichewa (12,13,16-18; Malawi), Hunde (19; DRC), Runyankore (20,22; Uganda), and Luganda (23; Uganda). Other locatives in NCs 16-18 are, e.g., 'on top', and NC refinements exist, such as NCs 1a/2a, 3a, and 9a in isiZulu. (Source: adapted from [2].)

Noun Class	Description of nouns typically found in those classes	Examples
1 and 2	People and kinship	*umuntu*, *abantu* 'human(s)'
3 and 4	Plants, nature, and some parts of the body	*umfula*, *imifula* 'river(s)'
5 and 6	Fruits, liquids, some parts of the body, loan words, and paired things	*ikhala*, *amakhala* 'nose(s)'
7 and 8	Inanimate objects	*isihlalo*, *izihlalo* 'chair(s)'
9 and 10	Loan words, tools, and animals	*inja*, *izinja* 'dog(s)'
11 (with 10 pl.)	Long thin stringy objects, languages, and inanimate objects	*uthi*, *izinthi* 'stick(s)'
12, 13	Diminutives	*kagalimoto*, *timagalimoto* 'small car(s)'
14	Abstract concepts	*ubuhle* 'beauty'
15	Infinitive nouns	*ukucula* 'to sing'
16, 17, 18	Locative classes	*pamsika*, *kumsika*, *mumsika* '(round/at/in) the market'
19	Diminutives	*hyùndù* 'a little bit of porridge'
20, 21 and 22	Augmentative and pejorative	*ogusajja*, *agasajja* 'big ugly man (men)'
23	Locative class	*eka* 'at home'

2 SOME REQUIREMENTS FOR A MODEL

As a step toward designing a model for recording lexicographic information for nouns and their noun classes for NCB languages,

[1] https://www.wikidata.org/wiki/Q109671890

we collected a set of requirements that an inclusive model and its implementation ideally would cater for. This is based on our own research, relevant literature, and knowledge of such languages (mainly isiNdebele, isiXhosa, isiZulu, Runyankore, and Chichewa), and two premises. First, that whatever will be done is linguistically sound (cf. adhoccing) and, second, that it facilitates bootstrapping across similar languages. They are:

- Meinhof's system [10] (updated), since it is used by linguists for harmonisation and comparison across NCB languages, and it is also supportive of bootstrapping across languages.
- Meinhof's list receives updates over time, thanks to new insights and language development, so the model must allow for more classes, such as isiZulu's 1a, 2a, 3a and 9a.
- All NCB languages have a subset of those 23 NCs, which varies by language. It needs to be recorded which language has which subset of noun classes. This then also prevents mis-applications and, hence, avoiding dirty data.
- Indicate NC pairings for singular (sg.) and plural (pl.).
- Indicate mass nouns, which either do not have a pl. or no sg.
- Some languages have different sg./pl. pairings or no pl. for some sg.; e.g., 14/- in isiZulu, but 14/6 in Chichewa.
- The class names are the same (i.e., just numbering), but different languages and customs may have other names for them, too. Whether the interface should show only Meinhof numbering for linguistic precision or rather some 'vernacular' version or both, remains to be determined.
- A noun may be classified in a different noun class in a different language for the same word in, e.g., English; e.g., *ulendo* (14) 'journey' in Chichewa and *uhambo* (11) in isiZulu.
- The augment and prefix, or extended prefix when taken together, that is added to the stem may be the same or different for the different languages, they each may be an empty string, and the extended prefix is typically at most 4 characters.
- A word may have a noun class in one language but not in another, notwithstanding that they have the same meaning and behave the same grammatically; e.g., *eka* (23) 'at home', in Luganda and *ekhaya* in isiZulu. That is: a language may have let the noun class go in disuse, but not the grammatical feature that is implicitly or explicitly still there.
- Relate nouns in NC21 and 22 (considered secondary nouns) to the NC they are derived from through superimposition of the NC 21 and 22 prefix over prefixes of various other NCs [11]; e.g., *jiti* (21) 'giant tree' cf. *mti* (3) in Kiswahili.
- If an indication of the semantics of each noun class is going to be added, e.g., as a description of the noun class: this has slight variations across languages (see e.g., [7, 12] cf. Table 1).
- If other NCB NC systems are going to be entertained, record: (i) Name of the NC system, (ii) Possibly also additional properties for relating a noun to the NC, alike 'NC according to Meinhof', 'NC according to Doke' etc., (iii) Some notion of conversion between alternative NC systems, if known or feasible, and (iv) Criteria used for each class' membership so as to allow interoperability across the NC systems.
- The extant noun classes with their augments and prefixes may not be fully clear for a very under-resourced language, and so then nor the noun class a noun belongs to, from the viewpoint of linguistics. Also, for loan words and new words, it may take time to settle on which noun class the noun of the entity belongs to (e.g., [12]). Therefore, it should allow for gaps and modifications as to which NCs a language has (regarding Meinhof) and for multiple allocations for the same word into different noun classes.

Although we think we have been comprehensive in noting variability, it may require more features and capabilities due to NCB language features we are not aware of at present.

3 CLOSING REMARKS

It is possible to make one's life's work just on the noun class systems of NCB languages. As with any lexicographic resource, one has to forge ahead at some point rather than analysing languages further to collect ever more requirements. The alternative is to start with at least something, which runs the risk that changes down the line may be costly. Where the tipping point of the trade-off lies is unknown. The list provided in this abstract may not be feasible to implement fully, but we are nonetheless investigating the design of a comprehensive model that is extensible. A first concrete action for Wikidata would be to use Meinhof's system as default, as linguistic foundation and for bootstrapping. It could then also align with extant implemented functions [8] to take a first step toward realising Abstract Wikipedia for isiZulu.

ACKNOWLEDGMENTS

CMK was financially supported by the National Research Foundation (NRF) of South Africa (Grant Number 120852).

REFERENCES

[1] S. E. Bosch, T. Eckart, B. Klimek, D. Goldhahn, and U. Quasthoff. 2018. Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment. In *Proc. of LREC 2018*, N. Calzolari et al. (Eds.). ELRA.

[2] J. Byamugisha, C. M. Keet, and B. DeRenzi. 2018. Pluralizing Nouns across Agglutinating Bantu Languages. In *Proc. of COLING'18*. ACL, 2633–2643.

[3] C. Chavula and C. M. Keet. 2015. An Orchestration Framework for Linguistic Task Ontologies. In *Proc of MTSR'15 (CCIS, Vol. 544)*. Springer, 3–14.

[4] G. Faaß, S. E. Bosch, and R. H. Gouws. 2014. A general lexicographic model for a typological variety of dictionaries in African languages. *Lexikos* 24 (2014), 94–115.

[5] S. Farrar and D. T. Langendoen. 2003. A linguistic ontology for the semantic web. In *GLOT International (3, Vol. 7)*. 97–100.

[6] C. M. Keet and T. Chirema. 2016. A model for verbalising relations with roles in multiple languages. In *Proc. of EKAW'16 (LNAI, Vol. 10024)*, E. Blomqvist et al. (Eds.). Springer, 384–399.

[7] C. M. Keet and L. Khumalo. 2017. Toward a knowledge-to-text controlled natural language of isiZulu. *Language Resources and Evaluation* 51, 1 (2017), 131–157.

[8] C. M. Keet, M. Xakaza, and L. Khumalo. 2017. Verbalising OWL ontologies in isiZulu with Python. In *The Semantic Web: ESWC 2017 Satellite Events (LNCS, Vol. 10577)*, E. Blomqvist et al. (Eds.). Springer, 59–64.

[9] M. Kemps-Snijders, M. Windhouwer, P. Wittenburg, and S.E. Wright. 2009. ISOcat: Remodelling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies* 4, 4 (2009), 261–276.

[10] C. Meinhof. 1906. *Grundzüge einer Vergleichenden Grammatik der Bantusprachen.* Reimer, Berlin.

[11] C. Miti. 2006. *Comparative Bantu Phonology and Morphology.* CASAS, Cape Town.

[12] M. Ngcobo. 2013. Loan words classification in isiZulu: The need for a sociolinguistic approach. *Language Matters: Studies in the Languages of Africa* 44, 1 (2013), 21–38.

[13] E. Taljard, G. Faaß, and S. Bosch. 2015. Implementation of a Part-of-Speech Ontology: Morphemic Units of Bantu languages. *Nordic Journal of African Studies* 24, 2 (2015), 23–23.

[14] D. Vrandecic. 2018. Capturing Meaning: Toward an Abstract Wikipedia. In *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks (CEUR-WS, Vol. 2180)*, Marieke van Erp et al. (Eds.). CEUR-WS.org.