Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment

Sonja Bosch, Thomas Eckart, Bettina Klimek, Dirk Goldhahn, Uwe Quasthoff
Department of African Languages, University of South Africa;
Natural Language Processing Group, Agile Knowledge Engineering and Semantic Web Group, Leipzig University
boschse@unisa.ac.za, {teckart, klimek, dgoldhahn, quasthoff}@informatik.uni-leipzig.de

Abstract

The South African linguistic landscape is characterised by multilingualism and the influence between their eleven official and some local languages. Unfortunately, for most of the languages the amount and quality of available lexicographical data is suboptimal, even though its availability is essential for all educational institutions and for the development of state-of-the-art language technology. In this paper we present a new source of lexicographical data for Xhosa, a language spoken by more than eight million speakers. For its utilisation in a multilingual and federated environment it is modelled using a dedicated OWL ontology for Bantu languages and possesses all features that are currently considered integral for the promotion of resource reuse as well as long-term usage. In the future, the introduced ontology may be used for other Bantu languages as well and may ease their combination to achieve more extensive, multilingual data stocks.

Keywords: Xhosa, lexicography, research infrastructures, linked data

1. Introduction

A basic requirement for the language processing capability for any language is the availability of lexicographical data, ideally open source data which is often hard to find for less resourced languages. This includes many members of the Bantu language family. For enhancing the usability of this kind of data this paper presents a Bantu Language Model for describing lexicographical data in RDF. Furthermore, it presents a new resource of lexicographical data for the Xhosa language based on this new model. The data to be presented in this model is a representative sample of raw data for a Xhosa-English dictionary, containing approximately 6,800 lexical entries. In its final state, the data set should contain approximately 10,000 lexical entries. Whereas the available data enables us to use Xhosa as the language of instantiation, the method and model are extensible and applicable to many other Bantu languages, in particular those belonging to the same group.

Together with this paper, both the Bantu Language Model and the current state of the lexicographical data set are freely available for download and for querying via a dedicated SPARQL endpoint. It should be noted that the research reported on in this paper is work in progress. Its final version will be provided via SADiLaR, the South African Centre for Digital Language Resources. Moreover, the current version of the data is already available via CLARIN-D (see section 4.).

The remainder of this paper is structured as follows: Section 2 describes the origin of the Xhosa language material and explains essential features of the Xhosa language with a focus on its morphology. Section 3 explains the new Bantu Language Model that is based on the established MMoOn ontology[1]. Section 4 gives detailed information about the structure of the Xhosa RDF data set using a concrete example. Furthermore, information about its current extent are provided. Section 5 demonstrates the relationship of the described work in the context of federated research infrastructures and how they can simplify access to and enhance usability of modern lexicographical data. The paper closes with a short summary and an outlook to planned further work.

2. The Xhosa Source Data

The data used for this case is based on Xhosa [xho][2], one of the official languages of South Africa belonging to the so-called Bantu language family. It is spoken predominantly in the Eastern Cape and Western Cape regions. There are approximately 8.1 million Xhosa speakers[3], adding up to about 16% of the South African population. Xhosa, as member of the Nguni language group, shares many linguistic features with other Nguni languages, which include Zulu [zul], Swati [ssw], Southern Ndebele [nbl] and Northern Ndebele [nbe][4]. Xhosa, like the other Bantu languages, is structurally agglutinating and is therefore characterised by words usually consisting of more than one morpheme. Each morpheme corresponds to a single lexical meaning or grammatical function. This particular Xhosa lexicographical data set is accompanied by English translations and was compiled and made available for purposes of further developing Xhosa language resources[5]. The process involved digitisation into CSV tables and various iterations of quality control in order to make the

[1] http://mmoon.org/core/

[2] Each language is followed by its ISO 639-3 code http://www.loc.gov/standards/iso639-2/php/code_list.php in order to distinguish one language from other languages with the same or similar names and to identify the names of cross-border languages.

[3] http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf

[4] The names of the Nguni languages in the languages themselves are respectively: isiXhosa, isiZulu, Siswati and isiNdebele.

[5] Bilingual (Xhosa-English) word lists were compiled by JA Louw after his retirement with the intention of documenting Xhosa words and expanding existing bilingual Xhosa dictionaries by means of among others botanical, animal and bird names, grammar terms, modern forms etc., as well as lexicalisations of verbs with extensions.

data reusable and shareable. In this paper, we concentrate on nouns and verbs. The excerpt of the lexicographical data set is a representative sample of Xhosa nouns and verbs. Nouns of all possible regular and irregular combinations of noun classes, and verbs with a variety of verbal extensions (leading to lexicalisations in meaning) are represented. Nouns are listed alphabetically according to noun stems, followed by the POS, the surface form of the singular and plural class prefixes (if applicable) as well as the number(s) of the class prefixes, and finally the English translations, e.g.

Noun stem	*POS*	*Class pref sg*	*Class no.*
phathi	noun	um	1
Class pref pl	*Class no.*	*English translation*	
aba	2	superintendent	

Verbs are listed alphabetically according to verb stem, i.e. the basic verb root followed by the inflection suffix -a, or sometimes -i, e.g.

Verb stem	*POS*	*English translation*
mi	verb	be standing
tyalisa	verb	help to plant

The lexicographic data is by no means based on corpus frequencies of nouns and verb stems as for instance the Oxford School Dictionary (De Schryver, 2014) but rather on complementation of existing, established dictionaries.

2.1. Xhosa Morphology

The noun is made up of two main parts, namely a noun prefix and a noun stem. All nouns are assigned to a particular class, as reflected in the class prefix. For practical and comparative purposes, noun classes have been given numbers by scholars working in the field of Bantu linguistics. Although 23 such classes have been reconstructed in Proto-Bantu, most Bantu languages have fewer than 20 classes (Nurse and Philippson, 2003). In Xhosa for instance, the class numbers end with class 17, while classes 12 and 13 do not occur at all (Pahl, 1967). It should be added that class 15 represents the infinitive class, while class 16 is no longer used productively to form nouns in Xhosa, but rather has an adverbial significance. Each noun class is characterised by a distinct prefix, which also includes a pre-prefix, and a particular singular/plural pairing with uneven numbers signifying singular and even numbers signifying plural. These class prefixes may show agreement with other constituents in a sentence. The only class pair with specific semantic contents is class 1/2 which contains personal nouns only. This does not, however, mean that all personal nouns occur in this class pair. For the rest of the noun classes, semantic arbitrariness is observed, although certain semantic generalisations do occur, e.g. classes 9 and 10 are generally referred to as the "animal classes" since they contain many animal names, but also many other miscellaneous terms. Noun stems may also be suffixed with morphemes such those indicating diminutive, augmentative, derogatory or feminine modifications to the basic meaning of the noun. A verb consists of a series of prefixes and suffixes that are built around a basic verb root carrying the basic meaning. A final inflection suffix completes the verb stem, to which pre-stem inflection is added in the form of, for example, the following morphemes: subject agreement, object agreement, negation, tense and aspect. Verbal suffixes may include morphemes such as: negation and derivational extension. The verb therefore carries much information and is pivotal in the sentence.

2.2. Discussion of the Xhosa Data

In the Xhosa data set under discussion, noun stems are separated from their class prefixes as is the case in traditional Bantu language dictionaries. In each instance the class prefix modifies the meaning of the basic noun stem e.g.

balo (isi 7; izi 8) "arithmetic"
balo (u 11) "census"

Although noun stems can be sub-divided into a root plus suffixes, any suffixes that occur, e.g. the feminine suffix *–kazi* and the diminutive suffix *–ana*, are not identified separately in our data. This is illustrated in the following examples where the modification of the basic meaning only appears in the English translations:

caka (isi 7; izi 8) "servant"
cakakazi (isi 7; izi 8) "servant girl"
cakazana (isi 7; izi 8) "young servant girl"

Noun class pairs normally signify singular/plural that correspond to the odd and even class numbers respectively, e.g.

khwenyana (um 1; aba 2) "son-in-law"
kroti (i 5; ama 6) "hero"

There are exceptions, however, for instance the singular class 11 takes its plural in class 10 (instead of 12, which does not exist in Xhosa), e.g.

diza (u 11; iin 10) "straw"

Also, the distinction between singular and plural does not apply to nouns that denote, for example, mass or abstract concepts, as in the case of:

bisi (u 11) "milk"
ophu (um 3) "vapour"

The following examples demonstrate that phonetically and phonologically conditioned allomorphs of class prefixes 1/2 (um-/aba- versus um-/ab-); 7/8 (isi-/izi- versus is-/iz-) and 11 (u- versus ulu-) appear in the data, e.g. in the case of vowel initial noun stems or monosyllabic noun stems (Kosch, 2006):

biki (um 1; aba 2) "reporter" vs. ongi (um 1; ab 2) "nurse"
kolo (isi 7; izi 8) "school" vs. enzo (is 7; iz 8) "deed, act"
patho (u 11) "school" vs. bi (ulu 11)

"misfortune, calamity"

These examples, therefore, demonstrate that for some noun classes more than one prefix member exists, resulting in allomorphs that occur in complementary distribution.
Verb stems are listed according to their infinitive form minus the infinitive prefix, i.e. the basic verb root followed by the inflection suffix *–a*. In some few cases, the final suffix presents as *–i* or *-e*. The latter only occurs in the case of stative verbs such as *–krekrelele* "stand in line". In the data there is no morphological differentiation between basic verb stems and verb stems with suffixed extension morphemes. The modification of the basic meaning of the verb stem, however, appears in the English translation, as in:

tenda "entertain"
tendana "entertain one another"
tendeka "be able to be entertained"
tendela "entertain at or for"
tendisa "help to entertain"

3. The Bantu Language Model

For the representation of the tabular Xhosa dictionary data and their translations we chose to convert the data into the RDF (Resource Description Framework) format. The mapping of the source data to RDF, however, requires a specific vocabulary which can be some existing or newly created ontology. While the lexicon model for ontologies (Lemon) (McCrae et al., 2011) was designed to represent lexical language data, its usage has been proven to be problematic for Bantu languages (Chavula and Keet, 2014). This is mainly due to the lack of the conceptualisation for morphological language data. Even though the Lemon model evolved to become a W3C recommendation published as the OntoLex-Lemon model that is split into five specified modules[6] (McCrae et al., 2017), the necessary modelling of morphological data has not been worked into this refined model.
Therefore, we created the Bantu Language Model[7] (in short BantuLM) as illustrated in Figure 1. This ontology is fully based on the reuse of and alignment to already existing vocabularies[8]. The largest part is based on the Multilingual Morpheme Core Ontology (MMoOn Core)[9] because it provides fine-grained classes and properties for representing morphological data and, moreover, already shares a considerable amount of overlap to the ontolex module for lexical data (Klimek, 2017). By taking the Xhosa verb and noun source data as an orientation point we identified three major linguistic subdomains that were to be modelled: 1) lexicographic data which is based on the OntoLex lime module[10] and MMoOn Core, 2) morphological data which is solely based on MMoOn Core, and 3) translational data which is based on the OntoLex vartrans module[11]. Despite the best practice recommendation to make direct reuse of existing vocabularies if they appropriately fit the modelling domain in question, a different approach of vocabulary reuse has been taken. In order to represent the BantuLM under a single namespace, all classes and properties have been newly created, however, corresponding to the reused external vocabularies. I.e. all classes that are based on MMoOn Core have identical labels and are aligned by usage of the `rdfs:subClassOf` object property in accordance to the creation procedure for MMoOn Core-based data sets. Otherwise, all classes that are based on the ontolex and lime vocabulary are interconnected with their derived counterparts via the `owl:equivalentClass` object property. The equivalence of all object properties within the BantuLM vocabulary is created with the `owl:equivalentProperty` property. Consequently, all definitions of the classes and properties need to be obtained from the interconnected original vocabularies. This poses, however, no disadvantage since the naming of the classes and properties is quite self-explanatory. While this kind of duplication of vocabularies is rather unusual it is formally valid in terms of ontology creation. This modelling of the BantuLM vocabulary has been chosen in preference of user-friendliness given that the data creators are mainly linguists that have only little or no expertise in creating language resources in the RDF format or within the Linked Data framework. It is assumed that a vocabulary that is applicable to all Bantu languages is easier to use and query for non-experts if it is built on a single namespace instead of a variety of vocabularies that need to be studied before they can be actually used for language data representation.
To conclude, the BantuLM is an aggregation of those classes and properties from the mentioned vocabularies that are necessary or useful to represent not only the Xhosa source data but also other Bantu languages in general, e.g. we had no data for the class `blm:Wordform`, but other Bantu language resources might well have and can then use this class accordingly. In contrast to the reused models the BantuLM is a language-specific model and, hence, specified for its affiliation to the Bantu language family. That means in particular, that grammatical meanings such as wordclass, number or nominal classifier are newly created and consequently specific to and shared by all Bantu language resources that will be based on the BantuLM ontology.
For the creation of the Xhosa RDF inventory data set the BantuLM proved not only to be fully suitable but also contributed to an explicit semantic interrelation between the lexical and morphological elements which is rather implicit in the tabular source data[12].

[6] `https://www.w3.org/community/ontolex/wiki/Final_Model_Specification`

[7] The ontology is available under the URI: `http://mmoon.org/bnt/schema/bantulm/`.

[8] Please consult the ontology URI for more information on how to use the ontology for creating other Bantu language data.

[9] Cf. `http://mmoon.org/` and `http://mmoon.org/core/` for more information.

[10] `http://www.w3.org/ns/lemon/lime#`

Model for lexical and morphological data of Bantu languages.

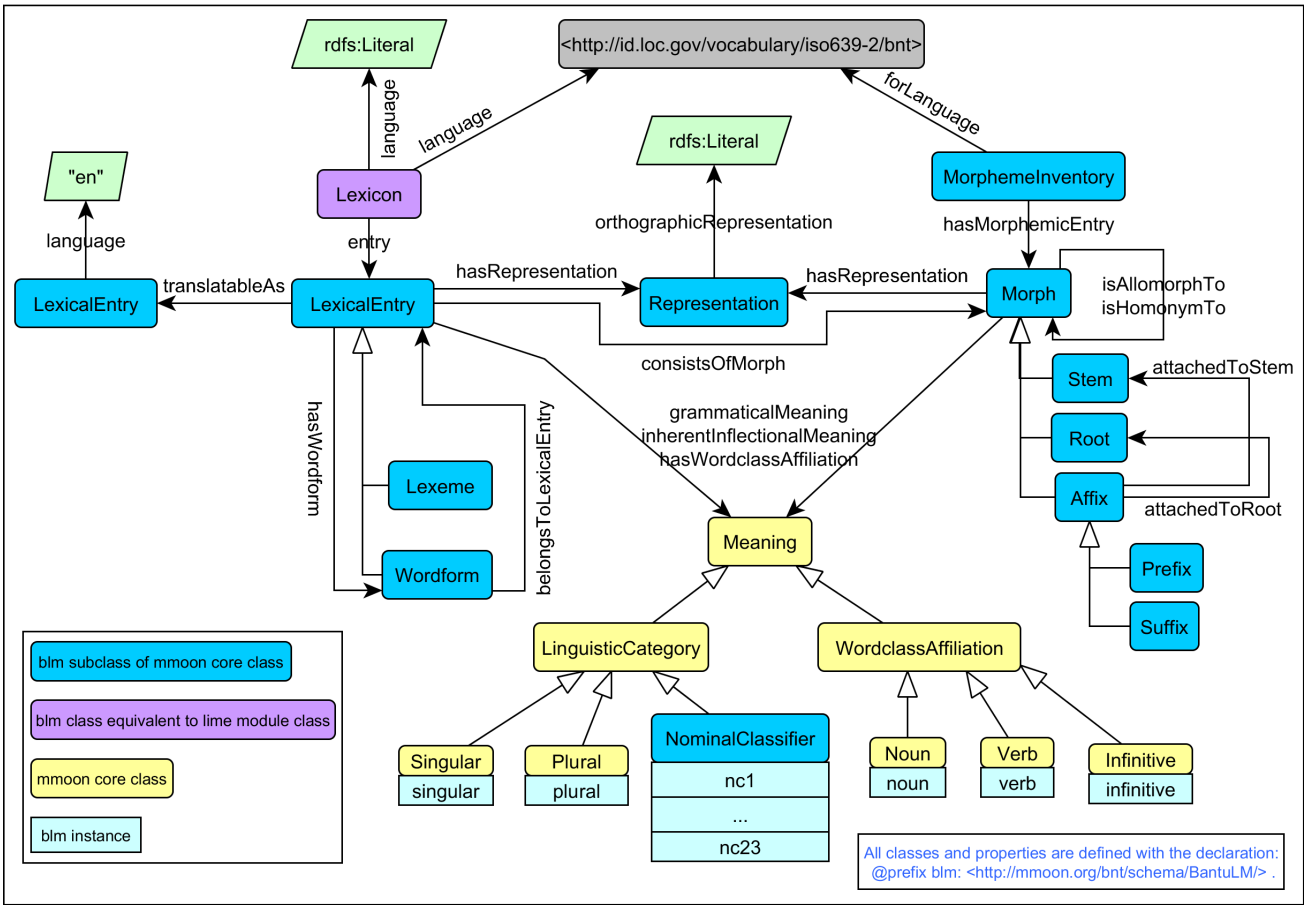

Figure 1: Ontology for the Bantu Language Model.

4. The Xhosa RDF Data set

The creation of the BantuLM ontology enabled the conversion from the Xhosa tabular source data into the Xhosa RDF data graph without any data loss. Necessary meta data is explicitly stated within the data set declaring information such as the data set creator, version and the underlying license.

In addition to the source data, the ontology-based representation of the Xhosa language data allowed for an explication of indirectly contained linguistic information. This is exemplified in Figure 2 which illustrates the graph representation of the lexical and morphological data.

With regard to the lexical data it can be seen that unique lexeme resources, like `xho_inv:lexeme_umbiki_n`[13] and `xho_inv:lexeme_ababiki_n`[14], have been created which were formerly separated as root and affix entries within the tabular data. As for the morphological data, the relationship that holds between affixes could be further specified by making use of the two object properties `blm:isAllomorphTo` and `blm:isHomonymTo`. That is, Figure 2 shows that the prefixes *aba-* and *ab-* are allomorphs to each other since they share the same meaning (noun class 2 and plural) but differ in their orthographic representation. Not illustrated, but included in the data set, are the homonymous relations that hold between affixes that share the same orthographic and/or phonological representation but differ in meaning. Such detailed linguistic information might be very useful for linguistic research investigating Bantu noun class systems.

Next to this internal enrichment of the tabular source data, the Xhosa RDF data set has been also externally enriched by linking the English translations, e.g. `xho_inv:trans_reporter_n` to lexical entries of the WordNet RDF data set[15] (McCrae et al., 2014). The object property `owl:sameAs` has been used to automatically create appropriate links. The full equivalence between the Xhosa RDF and WordNet RDF lexical resources is assured because only those lexemes have been interlinked that consisted of exactly one and the same word and also agreed in their part of speech. Figure 2 shows an example linking

[11] `http://www.w3.org/ns/lemon/vartrans#`

[12] To examine the increased expressivity, please compare an example of the source and RDF data here: `http://mmoon.org/lrec2018figures/`

[13] `http://rdf.corpora.uni-leipzig.de/resources/xho/inventory/lexeme_umbiki_n`

[14] `http://rdf.corpora.uni-leipzig.de/resources/xho/inventory/lexeme_ababiki_n`

[15] Please cf. `http://wordnet-rdf.princeton.edu/about` for more information. The data set can be found here: `http://wordnet-rdf.princeton.edu/static/wordnet.nt.gz`.

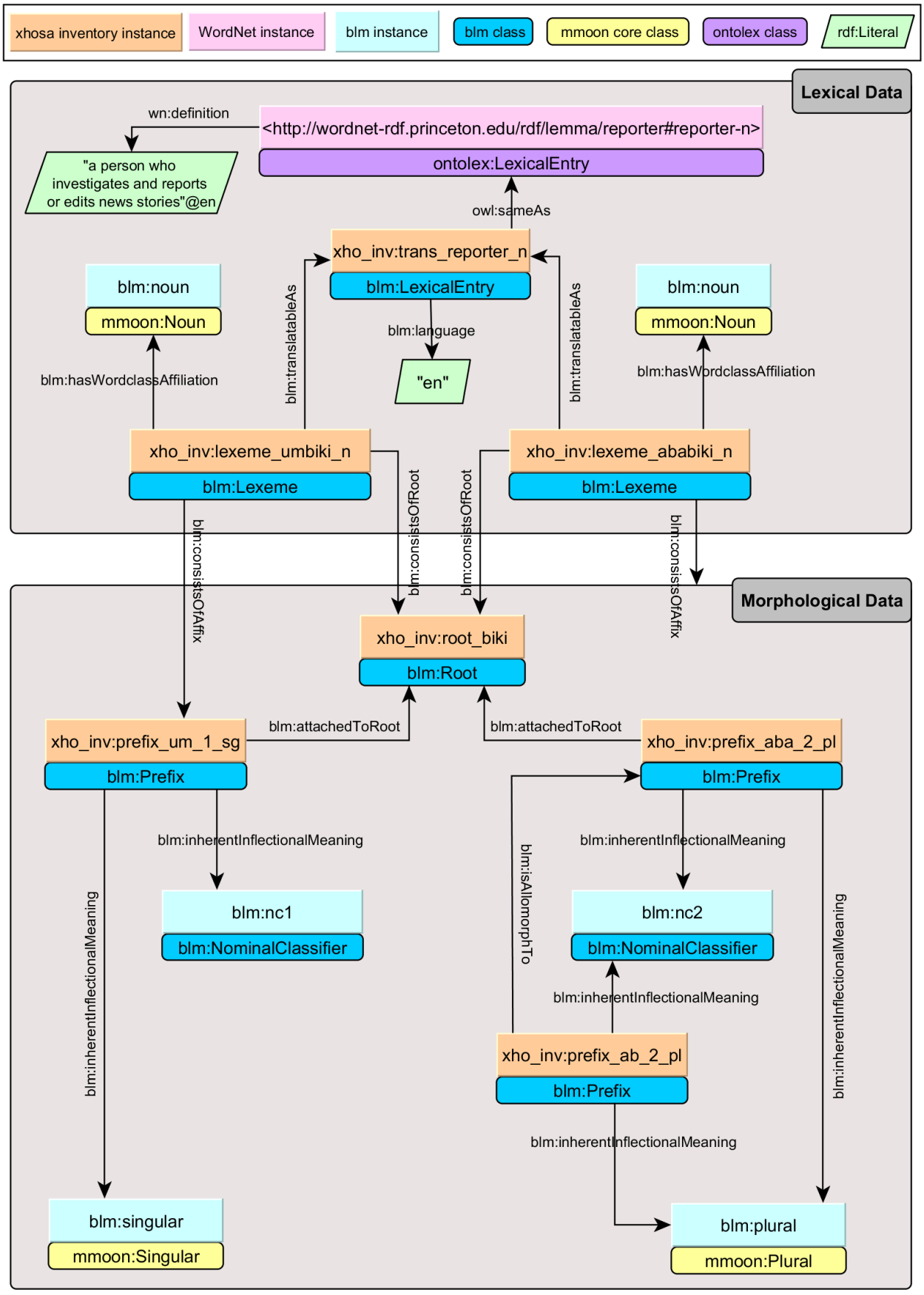

Figure 2: Excerpt from the Xhosa RDF data graph.

for the English translation *reporter* of the Xhosa nouns *umbiki* (singular) and *ababiki* (plural) to the corresponding WordNet lexical entry. Further, it can be seen that this WordNet entry ultimately[16] leads to a sense definition of the lexeme *reporter*. As a result, the interlinking of the Xhosa English translations with the WordNet RDF lexical entries, consequently, leads to an enrichment of the Xhosa noun and verb lexemes with corresponding lexical senses. Senses or sense definitions have not been part of the source data but are now accessible for all Xhosa lexemes whose translations are linked to WordNet and can be obtained by traversing through the interconnected data graph. While this enrichment with lexical senses already leads to a more coherent lexical data set for Xhosa, the linking to WordNet RDF entails an additional value in the context of the multilingual Bantu language landscape. Provided that more Bantu language data sets will be similarly converted into RDF and interlinked with WordNet, an interconnection of different Bantu language data sets could be realised by using the WordNet RDF as the pivot data basis for a multilingual Bantu language data graph.

Finally, the Xhosa RDF data set has been validated by using the RDF Unit[17] (Kontokostas et al., 2014) which conducts syntactic and semantic data quality tests of RDF data, which have been all passed by the Xhosa RDF data set.

In summary, the presented Xhosa RDF data set generates an added-value in comparison to its underlying tabular source data due to the successful internal and external data enrichment just explained. The Xhosa RDF data set in its current state contains 4,014 noun and 2,763 verb lexemes, 66 affixes as well as 2,818 links from the English translations to WordNet RDF. The Xhosa RDF data set is available within the LLOD Cloud and also accessible here: `https://github.com/MMoOn-Project/OpenBantu/blob/master/xho/inventory/ob_xho.ttl`. Moreover, the SPARQL endpoint provided at the URL `http://rdf.corpora.uni-leipzig.de/sparql` enables the querying of the data set to obtain deeper insights into the Xhosa language data.

5. Lexicographical Infrastructures in a Federated Environment

Despite strong efforts and significant progress towards open access to linguistic resources over the last years, many languages still lack those resources or their uncomplicated availability for larger user groups. Therefore, the presented work should not only be seen as another building block for a more complete landscape of linguistic resources, but in the context of federated and distributed infrastructures in a sometimes complex political and administrative environment.

Many countries with heterogeneous linguistic environments have decided to promote joint efforts for documenting their native languages for the benefit of education — primary, secondary, and academic — or the promotion of language technology, which currently is often only available for a highly resourced subset. This is especially problematic in a larger context where rights on relevant resources are held by different institutions with a varying degree of openness and each providing their own proprietary access interfaces.

As a consequence of this rather typical situation many large-scale infrastructures in the field of linguistic resources promote the usage of service-oriented architectures (SOAs) that provide data and services via standardised Web interfaces and data models. One of the benefits of this approach is that data can still be hosted by the publishing institution — being the main authority for the specific resource — and still allow access for the broader (or academic) public while promoting use and re-use in an active research environment.

In the South African context, the recently established Centre for Digital Language Resources[18] (SADiLaR) is a new research infrastructure with a focus on the creation, management and distribution of digital language resources of all official languages of the country. The ultimate aim is to provide a central repository for reusable language resources as well as applicable software tools that will be made freely available for research purposes (cf. Roux (2016)).

In the European context, CLARIN-D (cf. Hinrichs and Krauwer (2014)) is a long-term digital research infrastructure for language resources in the Humanities and Social Sciences. This includes language data bases, highly interoperable language technology tools as well as web-based language processing services. Researchers and students of Humanities and Social Sciences can use resources and technologies easily and in a standard way, without having to deal with technical complexities. The CLARIN-D infrastructure is built upon a network of centres, each of which with its own established competence and international reputation. For the time being, the described resource is hosted via CLARIN-D's infrastructure.

In our work we utilize this approach of making data available based on a standardised data model, i.e. the MMoOn Core ontology as the main basis of the BantuLM ontology, that has already proven to be adequate for describing morphological and lexical data (Klimek et al., 2016) and that is especially suitable to be used for other members of the Bantu language family as well.

The strict separation of data model, technical interface and end-user applications in a service-oriented environment opens the data for innovative applications. Among others, this is especially relevant for the field of meta-lexicography in the context of a multilingual environment. Besides the benefit of combining resources hosted and administered in different locations by different institutions, a SOA is a suitable backbone for enhancing usability with the major aim of addressing and reaching new user groups. This can be established by creating specific portals for different target audiences with varying and partially incompatible requirements. The specific demand may range from looking up simple words for language learners

[16] Please note, that there are several nodes in the WordNet RDF graph between the lexical entries and the sense definitions which are omitted in the Figure.

[17] `http://aksw.org/Projects/RDFUnit.html`

[18] http://www.nwu.ac.za/sites/www.nwu.ac.za/files/files/p-ctext/documents/Graphics.RMA.Newsletter.1.0.3.LvdB.2016-11-23.pdf

to concrete usage examples for dictionary enrichment or highly specific information of different linguistic fields for academic studies. Naturally, aspects such as necessary functions, form and content aspects and intended use are playing a vital role here (Gouws et al., 2007).

6. Summary and Outlook

The presentation of a new Xhosa lexicographical resource for a multilingual federated environment is an example for the transformation of isolated and unpublished dictionary data to the digital age. However, the data set used to develop the BantuLM ontology is only a snapshot of a resource in development. Currently, more lexemes are curated and quality assurance methods will be used to improve the already available data constantly. The publication date of the final data set is expected to be within the next 15 months.

The Bantu Language Model described in this paper can be used for many more languages. Dictionary data is available in a variety of formats, see, for instance, *http://www.cbold.ish-lyon.cnrs.fr/Dico.asp* with dictionaries for about 70 Bantu languages with 5,000 to 10,000 entries per dictionary.

A next logical step is the construction of a user interface to use this data as an actual online dictionary. For comfortable dictionary look-up an additional morphological analysis would be helpful. Again, a unified approach for many Bantu languages seems possible here. As most existing dictionaries translate to English or French, the transitive connection of several dictionaries can be used to interconnect different Bantu languages and allow their combination to a joined "virtual" resource for the whole language family in the future.

Acknowledgment The various phases of research activities related to this paper were funded by grants from the: H2020 EU projects ALIGNED (GA-644055); Smart Data Web BMWi project (GA-01MD15010B); BMBF project CLARIN-D (01UG1620C); South African Centre for Digital Language Resources (SADiLaR); Erasmus+ Programme; Scientific eLexicography for Africa project; and the South African National Research Foundation. The late JA Louw is acknowledged for making the Xhosa data available for purposes of further developing Xhosa language resources.

7. Bibliographical References

Chavula, C. and Keet, C. (2014). Is lemon Sufficient for Building Multilingual Ontologies for Bantu Languages? In *OWLED*, pages 61–72.

De Schryver, G.-M. (2014). Oxford school dictionary: Xhosa-english. Cape Town. Oxford University Press Southern Africa.

Gouws, R. H., Heid, U., Schweickard, W., and Wiegand, H. E. (2007). Dictionaries. an international encyclopedia of lexicography. supplementary volume: Recent developments with special focus on computational lexicography. an outline of the project. *Edited by Fredric FM Dolezal et al.*, page 262.

Hinrichs, E. and Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531, May.

Klimek, B., Arndt, N., Krause, S., and Arndt, T. (2016). Creating Linked Data Morphological Language Resources with MMoOn – The Hebrew Morpheme Inventory. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016), 23-28 May 2016, Slovenia, Portoroz*.

Klimek, B. (2017). Proposing an OntoLex-MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model*, pages 68–73.

Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., and Zaveri, A. (2014). Test-driven evaluation of linked data quality. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 747–758, New York, NY, USA. ACM.

Kosch, I. M. (2006). Topics in morphology in the african language context. Unisa Press.

McCrae, J., Spohr, D., and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *Extended Semantic Web Conference*, pages 245–259, Berlin, Heidelberg. Springer.

McCrae, J., Fellbaum, C., and Cimiano, P. (2014). Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The ontolex-lemon model: Development and applications. In *Electronic lexicography in the 21st century: Lexicography from Scratch. Proceedings of the eLex 2017 conference.*

Nurse, D. and Philippson, G. (2003). *The Bantu languages*. London: Routledge.

Pahl, H. (1967). *isiXhosa*. Johannesburg: Educum.

Roux, J. C. (2016). South African Centre for Digital Language Resources. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016), 23-28 May 2016, Slovenia, Portoroz*.