Translation-Based Dictionary Alignment for Under-Resourced Bantu Languages

Thomas Eckart
Natural Language Processing Group, University of Leipzig, Germany
teckart@informatik.uni-leipzig.de

Sonja Bosch
Department of African Languages, University of South Africa, Pretoria, South Africa
Boschse@unisa.ac.za

Dirk Goldhahn
Natural Language Processing Group, University of Leipzig, Germany
dgoldhahn@informatik.uni-leipzig.de

Uwe Quasthoff
Natural Language Processing Group, University of Leipzig, Germany
quasthoff@informatik.uni-leipzig.de

Bettina Klimek
Institute of Computer Science, University of Leipzig, Germany
klimek@informatik.uni-leipzig.de

Abstract

Despite a large number of active speakers, most Bantu languages can be considered as under- or less-resourced languages. This includes especially the current situation of lexicographical data, which is highly unsatisfactory concerning the size, quality and consistency in format and provided information. Unfortunately, this does not only hold for the amount and quality of data for monolingual dictionaries, but also for their lack of interconnection to form a network of dictionaries. Current endeavours to promote the use of Bantu languages in primary and secondary education in countries like South Africa show the urgent need for high-quality digital dictionaries. This contribution describes a prototypical implementation for aligning Xhosa, Zimbabwean Ndebele and Kalanga language dictionaries based on their English translations using simple string matching techniques and via WordNet URIs. The RDF-based representation of the data using the Bantu Language Model (BLM) and – partial – references to the established WordNet dataset supported this process significantly.

2012 ACM Subject Classification Information systems → Resource Description Framework (RDF); Computing methodologies → Phonology / morphology; Information systems → Dictionaries

Keywords and phrases Cross-language dictionary alignment, Bantu languages, translation, linguistic linked data, under-resourced languages

Digital Object Identifier 10.4230/OASIcs.LDK.2019.17

Category Short Paper

1 Introduction

For less resourced languages, dictionary compilation is still a labour intensive task. The number of active speakers (typically between 1 and 10 million) and the number of available digital resources can be very limited: it is often difficult to collect even 100.000 sentences of raw text or get access to any enriched linguistic resources. The situation with freely available lexicographical resources is especially challenging. If available at all, the few resources are usually of questionable quality and consistency. These dictionaries are often scanned versions of dictionaries dating back a few decades. For the purpose of multilingual dictionary alignment, they often lack direct references to similar languages, but instead only

2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 17; pp. 17:1–17:11

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

provide inconsistent translations to European languages, like English or French. To the best knowledge of the authors no related work exists up to today that proposes a computational Linked Data-based method for aligning such multilingual fragmented and heterogeneous data for less-resourced languages. As such the presented investigation can be regarded as a promising step in building a homogeneous foundation that enables further enrichment and extension of the original data.

In this paper we will focus on examples of available dictionary sources from the Bantu language family. Many of these dictionaries have a similar, but not an identical structure: They provide word lists with varying grammatical information, translations to the target language English (or, sometimes, French), and some optional explanation in the target language. The aim of this paper is to transform this data into a unified RDF representation using the Bantu Language Model BLM. The availability of several dictionaries with different source languages, but a common target language allows the creation of aligned dictionaries using English (or French) as a pivot language. The aim of the paper is to use the Bantu Language Model to align lexical data for the three languages Ndebele [nde][1], Xhosa [xho], and Kalanga [kck] and to investigate methods which would be helpful for the generation of derived dictionaries. It will be demonstrated how the underlying graph model of the BLM enables the alignment task.

The resulting resources have the potential for a variety of use cases, like their application in all areas of language education. This is especially relevant for many Bantu languages, as their use in both primary and secondary education is currently promoted in numerous African countries, like for instance in the Republic of South Africa.

The remainder of this paper is structured as follows. Section 2 gives an overview of the Bantu language family and outlines the current situation of lexical language resources thereof. Additionally, the dictionary sources that have been used for alignment are presented. The Bantu Language Ontology as the shared modelling basis for the aligned Bantu language dictionaries is introduced in Section 3. The implementation and outcomes of the conducted RDF-based multilingual dictionary creation are then described in Section 4. Finally, a summary and prospect of future work will conclude this paper with Section 5.

2 The Bantu Language Family and Available Lexical Resources

The Bantu languages are a family of languages spoken in Sub-Saharan Africa. The total number of Bantu languages (depending on the distinction between language and dialect) is estimated at 440 to 680 distinct languages, with approximately 240 million speakers [9]. This language family represents a group of closely related languages which shows similarities in the fields of phonetics, phonology, morphology and syntax. A certain amount of common vocabulary is also involved.

The landscape of Bantu dictionary data is diverse and heterogeneous. The use of open and well-documented standards is a cornerstone for the long-term availability and reuse of existing resources, and their efficient retrieval. For example, lexicographical data for Xhosa was recently prepared and converted using a dedicated OWL ontology and is now available for all kinds of applications via standard retrieval mechanisms [1]. However, many other resources

[1] We refer to languages by their names as presented in the Ethnologue (`https://www.ethnologue.com`) and also indicate their particular ISO 639-3 codes: Xhosa [xho] is referred to as "isiXhosa", Ndebele [nde] as "isiNdebele" and Kalanga [kck] as "Kikalanga" by their respective speakers. It is important to differentiate between so-called Zimbabwe Ndebele [nde] spoken mainly in Zimbabwe, and Southern Ndebele [nbl] spoken in South Africa.

are already available in a heterogeneous digital format. One such valuable source is the Comparative Bantu OnLine Dictionary (CBOLD), which offers Bantu language dictionaries under an open licence, including data for Zimbabwean Ndebele [10] and Kalanga [7].

Two of the languages under discussion are cross-border languages. Kalanga is spoken in eastern Botswana and western Zimbabwe and has a total of 338,000 users[2]. While Kalanga is a minority language in Botswana with no official status [8, p.176], it is an officially recognised language in Zimbabwe[3]. Kalanga is classified as S16 in Guthrie's larger Shona group of languages (S10) [9, p.609]. Zimbabwean Ndebele is spoken by approximately 1.6 million people in Zimbabwe, Botswana and Zambia [4], and is also officially recognized in Zimbabwe. Xhosa, an official language in South Africa, has approximately 8.1 million speakers and is spoken predominantly in the Eastern Cape and Western Cape regions of the country. According to the new updated Guthrie classification of Bantu languages list [9, p.648], Zimbabwean Ndebele (S44) and Xhosa (S41) are classified as members of the Nguni group (S40).

These three languages all being members of the Bantu language family, in particular of the S group of languages, share many linguistic features – for instance, they are structurally agglutinating and are therefore characterised by words usually consisting of more than one morpheme. They adhere to the typical Bantu languages nominal classification system according to which nouns are categorised by prefixal morphemes. For analysis purposes, these prefixes have been sorted into classes and given numbers by scholars who have worked within the field of the Bantu language family. A total of 24 noun classes is recognized [9, p.108], but these are not all attested in any single Bantu language. Noun prefixes usually indicate number, whereby the uneven class numbers indicate singular and the corresponding even class numbers indicate plural. However, exceptions to this rule also occur, e.g. mass nouns such as "water" in so-called plural classes do not have a singular form; plurals of class 11 nouns are found in class 10, while a class such as 14 is usually not associated with number at all. Irregular pairing also occurs occasionally, e.g. classes 9/6:

Table 1 Ndebele (excerpt from Pelling's Ndebele dictionary, source: CBOLD).

Prefix	Noun stem	Lexeme	Sg./Pl.	POS	Gloss	Comments
in	simu	in-simu	in/ama	n.	(pl. ama-simu): field;	[classes 9/6]
u	suku	u-suku	ulu/izin	n.	day.	[classes 11/10]
ubu	thongo	ubu-thongo	ubu	n.	sleep.	[class 14]

Table 2 Kalanga (excerpt from Mathangwane's Kalanga dictionary, source: CBOLD).

Prefix	Noun stem	Tone	POS	Class	Gloss
	bhaisikili	LLHHH	n	9/6	bicycle
lu	nji	H	n	11/10	knitting needle; [...]; an injection needle
bu	nyambi	LH	n	14	neatness; skilfulness; cleverness

[2] https://www.ethnologue.com/language/kck
[3] Cf. https://www.constituteproject.org/constitution/Zimbabwe_2013.pdf

Table 3 Xhosa (excerpt from Louw's Xhosa data set).

Noun stem	POS	Sg. prefix	Class	Pl. prefix	Class	Gloss
khitshi	noun	i	9	ama	6	kitchen
phahla	noun	u	11	ii	10	roof
phuthuphuthu	noun	ubu	14			hastiness

It is notable that, in contrast to the other two languages under discussion, Kalanga has an additional class 21, employed to express the augmentative by means of the class 21 prefix *zhi-*, as illustrated in Table 4.

Table 4 Kalanga (excerpt from Mathangwane's Kalanga dictionary, source: CBOLD).

Prefix	Noun stem	Tone	POS	Class	Gloss	Comment
zhi	nyala	HL	n	21	thumb; big toe	(compare with: *chi-nyala*: a finger; a toe)
zhi	midza-mbila	LLLH	n	21	huge mamba snake	

In the Nguni language group, augmentation is usually indicated by means of a noun suffix which does not influence the noun class, as illustrated in the following Xhosa example:

um-thi (class 3) "tree" > um-thi-kazi (class 3) "big tree"

Like most Bantu languages, Zimbabwean Ndebele, Kalanga, and Xhosa are considered resource scarce languages, implying that linguistic resources such as large annotated corpora and machine-readable lexicons are not available. Moreover, academic and commercial interest in developing such resources is limited. In the following section, some of the available sources for lexicographical data for Bantu languages are described in more detail.

2.1 Comparative Bantu OnLine Dictionary

The Comparative Bantu OnLine Dictionary (CBOLD[4]) project started in 1994 to create a source for lexicographical data for Bantu languages. It is committed to open access principles as stated in the "Bantuists' Manifesto" [2]. Between 1994 and 2000, a large number of Bantu dictionaries were digitized by CBOLD and provided via the project Web page for external use and applications.

The amount and range of available data, and its quality vary from dictionary to dictionary. For many dictionaries, information about the respective Bantu noun classes and morphological structure is available. There is no interlinkage between lexical items of different dictionaries; an alignment is therefore not directly feasible. However, all datasets contain translations to either English or French.

Despite the completion of the project in the year 2000 with no further updates since, it is still one of the most comprehensive sources for lexicographical data of Bantu languages. The list of supported languages contains – among many others – Swahili, Zimbabwean Ndebele, Venda, and Kalanga.

The CBOLD dictionaries are provided in inconsistent data structures and schemata using a variety of file formats, including FileMaker databases, HyperCard[5], Microsoft Word documents and plain text files. Obviously, this schematic and technical heterogeneity can

[4] `http://www.cbold.ish-lyon.cnrs.fr/`
[5] A proprietary hypertext format created by Apple Inc. in the 1980s.

not be used as a basis for modern cross-dictionary alignment and inter-lingual applications. As a consequence, transformation and quality assurance measures are required to allow the active usage of this valuable lexical data source in the future.

In the following sections, two of the included CBOLD dictionaries (Kalanga and Ndebele) are described in more detail.

2.1.1 Ndebele Dictionary

The CBOLD dictionary for Zimbabwean Ndebele was compiled by James N. Pelling [10] in 1971. CBOLD provides the data as plain text file and a FileMaker database. The dictionary contains 5000 lexemes with information about the part of speech, prefix/stem structure for the nouns, translations to English and corresponding forms in the perfect passive.

For this submission, only nouns and verbs were considered. This includes 4632 of the provided lexemes (i.e. 92.6%). Table 5 shows an excerpt of the available data.

Table 5 Excerpt from Pelling's Ndebele dictionary (Source: CBOLD).

Prefix	Stem	Lexeme	Prefix	POS	Gloss	Perfect Passive
is	ayobe	is-ayobe	isi/izi	n	spider	
ama	ququ	ama-ququ	ama	n	bad smell, stench	
	cutha	-cutha		v.t.	pluck feathers	cuthwa
	finyeza	-finyeza		v.t.	shorten	finyezwa

2.1.2 Kalanga Dictionary

The CBOLD dictionary for Kalanga was created 1994 by Joyce Mathangwane [7]. CBOLD provides the data as plain text file and a FileMaker database. The dictionary contains 2960 lexemes with information about the part of speech, tone, noun classes and prefix/stem structure for the nouns. Additionally, English translations are provided.

For this submission, only nouns and verbs were considered. This includes 2796 of the provided lexemes (i.e. 94.5% of all). Table 6 shows an excerpt of the available data.

Table 6 Excerpt from Mathangwane's Kalanga dictionary (Source: CBOLD).

Prefix	Stem	Tone	POS	Class	Gloss
chi	ako	LL	n	7	corn head
m	bala	HH	n	3	colour
	anga	LL	v		freeze; congeal
	baka	HH	v		build; construct

2.2 Xhosa Dictionary

Since CBOLD dictionary data is not available for all Bantu languages, Xhosa data used for this publication was taken from a resource compiled by J.A. Louw (University of South Africa UNISA) which is available under a Creative Commons (CC) license. This Xhosa lexicographical data set consists of morphological information accompanied by English translations. It was created and made available by the authors for purposes of further developing Xhosa language resources [1]. The data were compiled with the intention of

documenting Xhosa words and expanding existing bilingual Xhosa dictionaries by means of among others botanical, animal names, grammar terms, modern forms etc., as well as lexicalisations of verbs with extensions. The publication process involved digitisation into CSV tables and several iterations of quality control in order to make the data reusable and shareable. Since this process has not yet been completed, we concentrate in this paper on two word classes for which extensive results already exist, namely 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, like shown in Table 7.

Table 7 Excerpt of nouns from the Xhosa dictionary.

Noun stem	POS	Class pref sg	Class no.	Class pref pl	Class no.	English translation
phathi	noun	um	1	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, like shown in Table 8.

Table 8 Excerpt of verbs from the Xhosa dictionary.

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 [3] but rather on complementation of existing, established dictionaries.

3 The Bantu Language Model

Aligning lexical content requires semantic and structural consistency between two or more language datasets. In the case of Bantu languages, as already explained, no shared structural basis for representing lexical data exists to date. The available digital resources are highly heterogeneous with regard to their size, content and format. In order to undertake any kind of alignment task these resources need to be transformed into a shared format first. While this can be done by using structured formats such as XML or entering and maintaining the lexical data in a database we decided to apply the Linked Data framework and reuse the Bantu Language Model (BLM)[6]. This model is an ontology that was introduced in Bosch et al. 2018 [1] in the RDF and OWL formats that ensure semantic and structural interoperability between all data that is described with it. An overview of the BLM is illustrated in Figure 1 which shows the underlying graph that integrates and unifies all data that is created based on the BLM. The BLM allows for the representation and interrelation of lexical, morphological and translational elements but also common grammatical meanings as well as noun class elements of Bantu languages. This is in accordance with the content

[6] The URL of the ontology is: `http://mmoon.org/bnt/schema/bantulm/`

that we found in existing tabular lexical data of various Bantu languages and with the three language datasets that were just described. More details on the underlying development and design decisions of the ontology are discussed in [1][7]. The applicability of this ontology has been proven by using it to create a Xhosa RDF dataset[8].

The choice of the BML as a suitable modelling basis that facilitates dictionary alignment is motivated by a number of aspects. First, this ontology is already specified for the peculiarities of Bantu languages, and above all, it was created together with Bantu language experts. In this way, semantic coherence between lexical elements is already ensured on the data representation level. Second, the Linked Data approach allows for the separate development of single language resources that can be later integrated and interrelated, if desired, within one unified graph due to the shared vocabulary. A third advantage is entailed in the possibility to not only interconnect various Bantu language datasets with each other but also extend the data with already existing other language resources, i.e. available English or French Wordnet RDF editions that are useful as a pivot language for identifying translations. What is more, the BLM ontology can be easily extended according to representational needs. It is not a fixed model but can be later on modified to include elements and relations that might be necessary for describing a more detailed language dataset. Finally, with regard to the practical aspect of transforming, editing, merging and analysing existing lexical Bantu resources, the compliance to the Linked Open Data framework is an additional decisive factor for the BLM, because various tools for enriching or analyzing RDF-based linguistic data already exist.

4 RDF-based Dictionary Alignment

4.1 Technical Implementation

All three dictionaries mentioned in section 2 were transformed into the RDF format by using the Bantu Language Model. The Xhosa RDF dataset could be reused directly, while for the Ndebele and Kalanga data transformation code was used, that had already generated the Xhosa RDF dataset[9]. As a result, links between English translation resources and their respective lexical WordNet resources have also been established within those two datasets. Due to missing data[10] or additional data[11], the implementation had to be adapted insignificantly. For example, temporary noun and number classes were introduced to the data set, that still have to be replaced by their correct classes during future quality assurance and enhancement procedures. Similar requirements exist for enhancing the quality of translations. For those procedures, the still ongoing work of double checking the Xhosa dataset by native speakers can be seen as a template.

The resulting RDF datasets were imported into a SPARQL endpoint[12] where they are publicly available and where future updates will also take place. All results included in the next subsection were extracted using SPARQL queries and are therefore easily reproducible.

[7] There, also the question why the OntoLex-Lemon model as widely accepted recommendation for representing lexical language data has not been used instead is answered and shall not be addressed in this publication again.

[8] The data is available here: `https://github.com/MMoOn-Project/OpenBantu/blob/master/xho/inventory/ob_xho.ttl/`

[9] The code will be available at the GitHub repository of the MMoOn project (`https://github.com/MMoOn-Project`) soon.

[10] This includes explicit noun class information for Ndebele or information about number for both Ndebele and Kalanga.

[11] Like information about tone for Kalanga.

[12] `https://rdf.corpora.uni-leipzig.de/sparql`

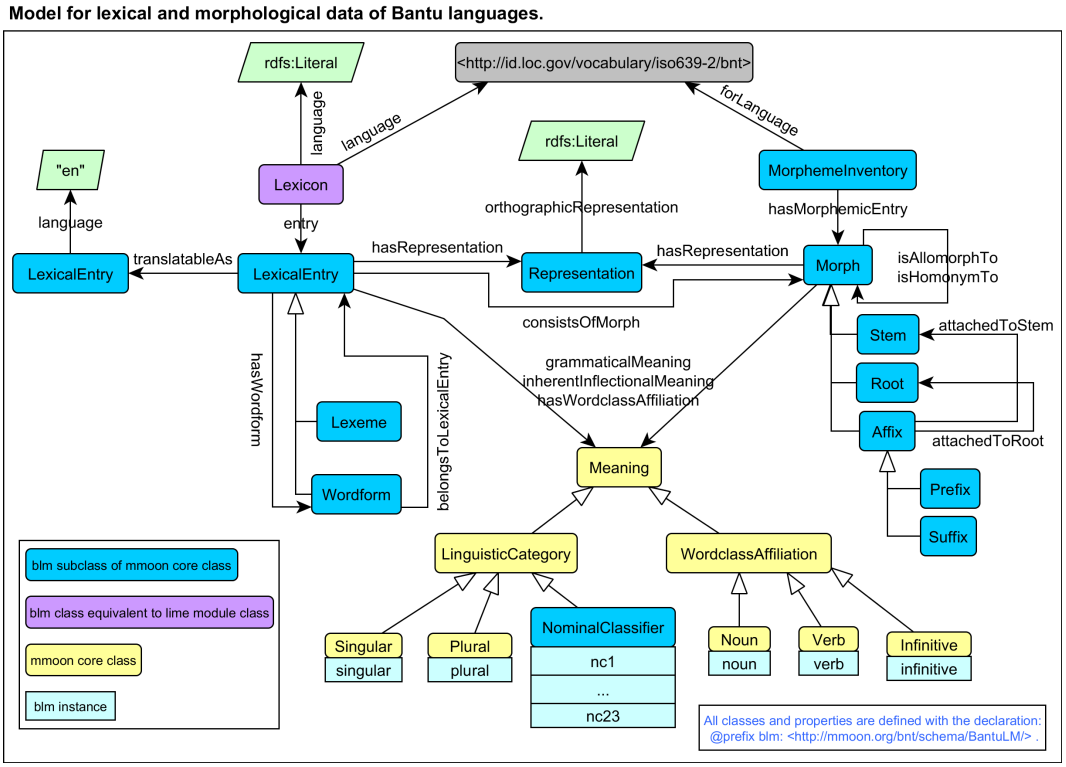

Figure 1 Ontology for the Bantu Language Model.

The actual alignments were not persisted in the endpoint, as quality assurance measures are not finished yet.

The underlying graph model of the RDF-based BLM ontology made the aggregation of the first results especially easy and is seen as a well-defined but still flexible backend for future, more user-friendly applications by the authors. First work on integrating the endpoint into an existing Web portal for lexicographical data has already shown some positive results.

4.2 Alignment Methods and Results

The identification of translations between lexical resources is already a challenging task if extensive data exists. In general, translation equivalences between two lexical entries are established if both entries share the same conceptual description, e.g. sense resources or definitions. For RDF-based datasets such an alignment between multiple dictionaries has been undertaken for the Apertium Bilingual Dictionaries [6]. While such an encompassing sense-based alignment is not feasible due to the outlined shortcomings of the source data for the three Bantu RDF language datasets under investigation, however, the demonstrated usage of pivot languages for aligning dictionaries with no direct translations was applicable for this case. Moreover, it should be noted that this contribution focuses on providing a foundation for the further enrichment of the aggregated data using a common data model. It presents work in progress and is seen by the authors as a first step towards the integration of more comparable languages. For this reason, a deeper evaluation of the results or the discussion of borderline cases was postponed to a later date.

Due to the underlying shared BLM vocabulary, semantic coherence between the lexical elements and translations between the three Bantu language RDF datasets is ensured. Loading all datasets into a single SPARQL endpoint, as done, then renders a unified data graph that

can be analysed and traversed along the nodes and edges across the three dictionaries. For the alignment only the lexeme, translation and WordNet resources could be used, since no sense definitions exist within the data. Provided with these resources we identified two methods for finding alignments. Similarly to the Apertium Bilingual Dictionaries we made use of the English translations contained in all three datasets as the pivot language interconnecting the Bantu dictionaries.

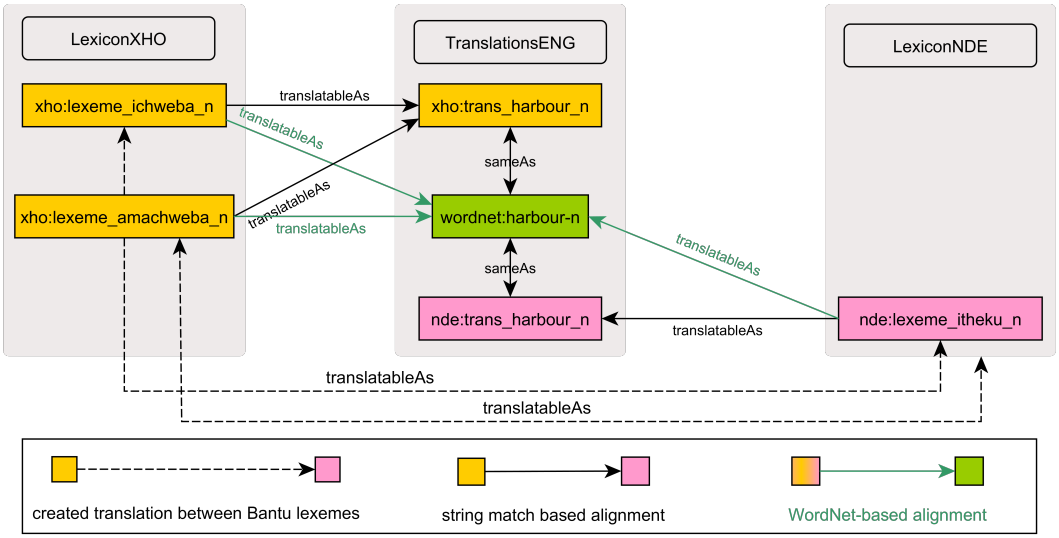

Figure 2 Example translation between lexemes in Xhosa and Ndebele in BLM RDF.

For the first method we aligned lexical entries based on the contained WordNet data [5], that is two lexical entitites are considered as translations if they point to the same WordNet resource. Since the English translation resources are interlinked with a WordNet resource via the `owl:sameAs` object property also a direct translation between a Bantu language lexeme and this WordNet resource can be inferred. The second method involves the identification of translations for which no shared WordNet resource exists. By conducting a simple string match between all translation resources across the three dictionary pairs, an alignment between lexical entries could be obtained whenever the strings of two English translation resources of different dictionaries were identical. Both methods are illustrated in Figure 2. As can be seen, the WordNet-based alignment contains the string match based alignment in that the WordNet links were also created based on string match with the English translation resources. While this seems to occur redundant we explicitly represent this method here because we regard the identification of translations by pointing to a single English dataset, which is the English WordNet in this case, as more accurate than the string match based alignment. In this special case for available Bantu language data the prospective creation of more BLM-based RDF dictionaries will result in a number of duplicate and ambiguous English translation strings without any further lexical information, e.g. `xho:trans_harbour_n` and `nde:trans_harbour_n`. Indeed, as the number of resulted translations in the three bilingual dictionary pairs in Table 9 show, there could be only one more translation for the Ndebele-Kalange and Xhosa-Kalanga dictionaries and just 67 translations for the Ndebele-Xhosa dictionary obtained via the string match based method in addition to the WordNet-based method.

Table 9 Available alignments for all dictionary pairs.

Dictionary pair	WordNet-based alignments	String-matching alignments
Ndebele, Xhosa	1541	1608
Ndebele, Kalanga	62	63
Xhosa, Kalanga	106	107

Consequently, we regard the WordNet-based method as more suitable for retrieving translations. Creating links from translations of single Bantu dictionaries to one shared and already existing dataset, such as the English RDF WordNet, facilitates the quality assessment of obtained alignments by language experts because WordNet also comes with definitions which can be used to ensure that the right translation has been found. Moreover, WordNet provides lexical entries with sense resources which could be used to arrive at more accurate sense-based translations in the future.

In addition to the bilingual translation data that was found, multilingual translations between all three dictionaries could be identified using the same methods (cf. Table 10). By that, it could be shown that analysing lexical data in the RDF format is very simple and efficient since every data point is interconnected and retrievable by traversing the graph. The quality of the established alignments with regard to their linguistic accuracy cannot be evaluated at this stage since it is future work to be done by language experts. Nevertheless, we judge the resulted numbers of obtained alignments across the bilingual dictionaries as promising. Taking into consideration that the strings of the English translation resources were the only available information usable as a comparative measure between Bantu language lexemes, the presented alignments can be considered as the closest one can get to bi- and multilingual translations for Bantu language data given the current state of the language data situation. What is more, the outcome of this translation-based dictionary alignment provides valuable additional data for the less-resourced Bantu languages that is easy to obtain and directly usable by language experts.

Table 10 Examples for aligned lexemes in all three source languages.

English	Xhosa	Kalanga	Ndebele
companion	iqabane	nkwinya	umngane
debt	isikweliti	nlandu	isikwilidi
doctor	ugqirha	nlapi	udokotela
image	umfanekiso	itshwantsho	isithombe
witch	igqwirha	nloyi	umthakathi

5 Conclusion

The presented prototypical implementation for aligning Xhosa, Zimbabwean Ndebele and Kalanga language dictionaries revealed typical problems of this task for less-resourced languages. While there is a need for aligned data, the available dictionaries are typically unsatisfactory concerning size, quality and consistency, which makes interconnecting them to form a network of dictionaries a challenging task. As in our case for three specific Bantu languages, data is rarely available in a schematic and technical homogeneous way. Transformation into a common model such as the BLM is, therefore, a first helpful step towards aligning datasets in a more straightforward fashion.

Missing reference data is another problematic aspect that has to be dealt with. Dictionaries as compiled by the CBOLD project have been compiled over decades and have only been assigned with loose and inconsistent translations to English or French instead of direct translations to other Bantu languages. By linking lexemes to concepts within WordNet, stable referencing of an external vocabulary can be ensured. This provides a common basis for linking with further dictionary data in the future.

The result of dictionary alignment is a relevant resource for fields such as teaching, where comprehensive dictionaries of high quality that may include references to external and even non-lexical data such as sample sentences or similar words are of fundamental importance. In the context of countries like South Africa, it becomes obvious that there is an urgent need for such data since mother-tongue education has gained popularity in recent years while the importance of international languages such as English is also incorporated into teaching concepts.

To allow for these use cases and an even wider applicability of dictionaries, the overall reliability and consistency of the data need to be assured. The presented systematic extraction and preparation of a shared integration model allows for collaborative approaches to quality assurance which can significantly boost the grade of the data.

Future work will include the incorporation of additional dictionaries based on the BLM and improving and extending their bilingual alignment. Further possibilities for expanding the alignment between dictionary entries in different languages needs to be considered. For the similarity of translations or descriptions in the pivot language English (or French), not only simple string similarities, but also similarities of the corresponding word embeddings can be used to link semantically similar lexemes.

Naturally, meaningful results can only be achieved with direct collaboration with language experts and native speakers. The systematic transformation and enrichment of public dictionaries like the ones provided by CBOLD have the potential to be an important starting point and a valuable resource for the Bantu language family.

References

1. Sonja Bosch, Thomas Eckart, Bettina Klimek, Dirk Goldhahn, and Uwe Quasthoff. Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki (Japan)*, 2018.
2. CBOLD. Bantuists' Manifesto. Website, 1996. Available: `http://www.cbold.ish-lyon.cnrs.fr/Docs/manifesto.html`; Accessed on 8 January 2019.
3. G.-M. De Schryver. *Oxford School Dictionary: Xhosa-English.* Oxford University Press Southern Africa, Cape Town, 2014.
4. Ethnologue. Ndebele. Website, 2019. Available: `https://www.ethnologue.com/language/nde`; Accessed on 8th January 2019.
5. Christiane Fellbaum, editor. *WordNet: an electronic lexical database.* MIT Press, 1998.
6. Jorge Gracia, Marta Villegas, Asuncion Gomez-Perez, and Nuria Bel. The apertium bilingual dictionaries on the web of data. *Semantic Web*, pages 231–240, 2018. `doi:10.3233/SW-170258`.
7. J. T. Mathangwane. *Kalanga.* Comparative Bantu OnLine Dictionary CBOLD, 1994. URL: `http://www.cbold.ish-lyon.cnrs.fr/Load.aspx?Langue=Kalanga&Type=Text&Fichier=Kalanga.Mathangwane1994.txt`.
8. J. T. Mathangwane. *Ikalanga 50 Years On: A Cross Border Language Against Tremendous Odds.* Botswana Notes and Records, 48, 2016.
9. Derek Nurse and Gérard Philippson. *The Bantu Languages.* Routledge, London, 2003.
10. J.N. Pelling. *A Practical Ndebele Dictionary.* Comparative Bantu OnLine Dictionary CBOLD, 1971. URL: `http://www.cbold.ish-lyon.cnrs.fr/Load.aspx?Langue=Ndebele&Type=Text&Fichier=Ndebele.Pelling.1971.txt`.