Enriching and Increasing the Usability of Lexicographical Data for Less-Resourced Languages

Dirk Goldhahn
Natural Language Processing Group, University of Leipzig, Germany
Saxon Academy of Sciences and Humanities, Leipzig, Germany
goldhahn@saw-leipzig.de

Thomas Eckart
Natural Language Processing Group, University of Leipzig, Germany
Saxon Academy of Sciences and Humanities, Leipzig, Germany
teckart@informatik.uni-leipzig.de

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

Abstract

This paper presents a use case for enriching lexicographical data for less-resourced languages employing the CLARIN infrastructure. Newly prepared lexicographical data sets for under-resourced Bantu languages spoken in southern regions of the African continent form the basis of the presented work. These datasets have been made digitally available using well-established standards of the Linguistic Linked Open Data (LLOD) community. To overcome the insufficient amount of freely available reference material, a crowdsourcing web portal for collecting textual data for less-resourced languages has been created and incorporated into the CLARIN infrastructure. Using this portal, the number of available text resources for the respective languages was significantly increased in a community effort. The collected content is used to enrich lexicographical data with real-world samples to increase the usability of the entire resource.

1 Introduction

The availability of contemporary text material is a prerequisite for a variety of applications and research scenarios, especially including studying recent developments in language use. Projects such as *An Crúbadán*[1] offer text freely available on the web to enable such studies for languages with small numbers of speakers. Resources in *An Crúbadán* are typically added manually, have a limited scope but are available for over 2,000 languages.

To expand the amount of available textual data, crawling and processing web content is now a standard procedure to acquire those needed resources. As a positive consequence of the vast amount of available online content, preselection of highly specific material is now possible for many languages and allows examination of all sorts of linguistic phenomena for specific domains and genres. Automatically generating valuable data sources from online resources requires specific means of text acquisition and pre-processing of the gathered material. Subsequently, different systems that simplify the crawling and processing of web pages for end users were developed and are in active use. One of the most popular services is the SketchEngine (Kilgarriff et al., 2014) which has a focus on lexicography.

On the other hand, for many less-resourced languages, significant amounts of material are now available for the first time. Using standards of the growing Linguistic Linked Open Data (LLOD) community,

[1] http://crubadan.org/

Dirk Goldhahn, Thomas Eckart and Sonja Bosch 2020. Enriching and Increasing the Usability of Lexicographical Data for Less-Resourced Languages. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 23–32.

connecting – so far isolated – datasets of all kinds, helps creating substantial resources for these languages in a federated infrastructure. One of the benefits of these interconnections is the ability of building bridges between lexicographical entries and concrete, real-world usage examples. The resulting resources have a high value for all kinds of use cases and user groups, including being essential for first- and second-language acquisition.

This paper focuses on a specific use case, facilitated by the federated research infrastructure CLARIN (Hinrichs and Krauwer, 2014), in which newly created lexicographical datasets for some Bantu languages were enriched using a new web crawling portal that focuses on the acquisition of text material for less-resourced languages.

2 Crawling Under-Resourced Languages

The situation concerning the availability of digital language resources is satisfactory only for a small number of languages. Even for most of the languages with more than one million speakers, no reasonably sized textual resources or tools like POS taggers are available. This points to a widespread need for digital language resources for many languages of the world. Therefore, the CURL (**C**rawling **U**nder-**R**esourced **L**anguages) web portal[2] for corpus collection with a focus on languages with more than one million speakers, has been initiated (Goldhahn et al., 2016) as a service in the CLARIN infrastructure. It relies on native speakers with knowledge of web pages in their respective language. The initiative gives interested scholars and language enthusiasts the opportunity to contribute to corpus creation or extension by simply entering a URL into a web interface.

In the backend, Heritrix (Mohr et al., 2004), the crawler of the Internet Archive[3], is used in combination with a well-established corpus processing chain that was adapted to append newly added web pages to continuously growing corpora. This enables us to collect larger corpora for under-resourced languages by means of a community effort. These corpora are made publicly available within the CLARIN infrastructure, both directly in the created portal and as part of the Leipzig Corpora Collection project (Goldhahn et al., 2012). Figure 1 gives an overview of the general structure of the application.

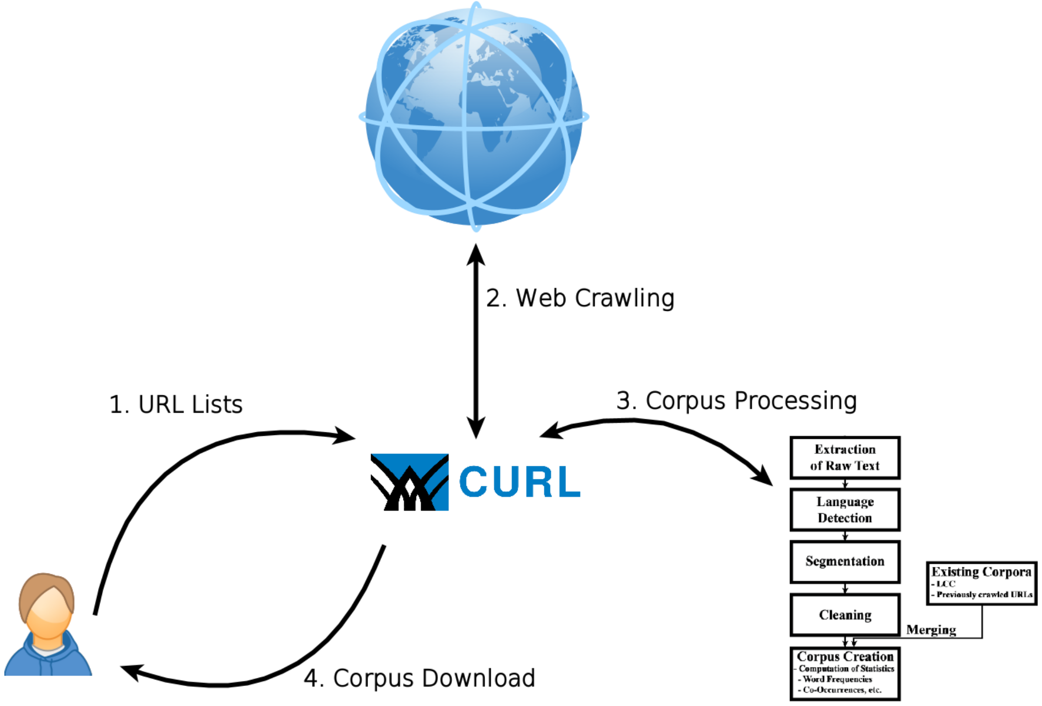


Figure 1: Schematic overview of the CURL portal

[2] https://curl.corpora.uni-leipzig.de/
[3] https://archive.org

The data provided do not only include the crawled text material with information about their source of origin and date of crawling; in addition, statistical analysis of word frequencies and word relationships based on statistical word co-occurrences is also generated and included in the download files. This kind of information has proven to be a valuable resource when working with language material, where manually created and evaluated resources on word semantics like thesauri are hardly available.

Since its establishment, the web portal has helped creating initial text resources for several languages and also facilitated the expansion of available text collections. All in all, more than 10,000 URLs were submitted to the system in 127 crawling jobs for 62 languages. Concrete efforts for two languages belonging to the Bantu language family - Xhosa (ISO 639-3:xho) and Kalanga (ISO 639-3:kck)[4] - will be described in the following sections.

3 Bantu Languages

The Bantu languages are a family of languages spoken in sub-Saharan Africa with around 240 million speakers spread across 27 African countries (Nurse and Philippson, 2003:1). The exact number of Bantu languages cannot easily be determined because it is often difficult to draw the line between a language and a dialect. However, taking this into consideration, Nurse and Philippson's (2003:3) estimate is about 300 Bantu languages two of which are under discussion, namely Xhosa and Kalanga. Xhosa, with approximately 8.1 million speakers, is spoken predominantly in the Eastern Cape and Western Cape regions of South Africa, and is classified as a member of the Nguni group of languages (Nurse and Philippson, 2003:649). The cross-border language Kalanga is spoken in eastern Botswana and western Zimbabwe and has a total of 338,000 users[5]. It is classified as a member of the larger Shona group of languages (Nurse and Philippson, 2003:609).

Many linguistic features are shared by these two languages. Morphologically, they are of an agglutinating nature with verbs exhibiting an intricate set of affixes, while nouns are assigned to classes by means of so-called class prefixes. The number of class prefixes per language varies, but in general Bantu languages have between 12 and 20 class prefixes. Each class is characterised by a distinct prefix, a particular singular/plural pairing and agreement that branches out to other word categories, namely verbs (subject and object markers), adjectives, possessives and so on. Exceptions to the singular/plural rule also occur, for example 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 a number at all.

4 Dictionary Data

Like most Bantu languages, Xhosa and Kalanga 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, available sources for lexicographical data for Bantu languages used in this publication are described in more detail.

4.1 Xhosa Dictionary

Xhosa lexical data were 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 (Bosch et al., 2018). The data were compiled with the intention of documenting Xhosa words and expanding existing bilingual Xhosa dictionaries by means of – among others – botanical names, animal names, grammar terms, modern forms and so on, 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.

[4] The speakers of the two languages use the names isiXhosa and Ikalanga.

[5] https://www.ethnologue.com/language/kck

In its current state, the data set contains approximately 6,800 lexical entries and is already published in the CLARIN infrastructure[6] and available via a dedicated web portal[7]. Table 1 gives a short summary of its current inventory.

Dataset feature	Value
Number of noun lexemes	4020
Number of verb lexemes	2763
Number of noun classes	15
Number of English translations	7807

Table 1: Characteristics of the Xhosa dataset (as of 2020-01-12)

However, the compilation process is not completed yet and the data are still subject to quality assurance measures. The final dataset is expected to contain approximately 10,000 lexical entries and will also be available in the context of CLARIN via the South African Centre for Digital Language Resources (SADiLaR[8]).

4.2 Kalanga Dictionary

Lexicographical data for the Kalanga language was extracted from the Comparative Bantu OnLine Dictionary (CBOLD[9]). The project started in 1994 to create open source lexicographical data for Bantu languages. The amount and range of available data, and its quality and format vary from dictionary to dictionary. The CBOLD dictionary for Kalanga was created in 1994 by Joyce Mathangwane and is provided as a plain text file. 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. The resulting data set is currently used primarily for testing different approaches of dictionary alignment (Eckart et al., 2019). However, it also shows the high relevance of openly available data as a starting point for building up structured data sets – including their extension and improvement – in cases where no comparable resources are available. In the case of the CBOLD project, this includes material for dozens of less-resourced languages.

4.3 Bantu Language Model

The lexical resources introduced were transformed into a unified schema to simplify all relevant data enrichment and quality assurance procedures and to form a basis for future applications and user interfaces. The Bantu Language Model (BLM) (Bosch et al., 2018) is an ontology of the Linguistic Linked Open Data (LLOD) community that ensures semantic and structural interoperability. The BLM is based on the MMoOn ontology (Klimek, 2017) and 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.

A summary of the chosen data model is depicted in Figure 2. In the context of Bantu languages, both the nominal classifier system – as a (language family) specific *LinguisticCategory* – and the support of their rich morpheme structure and morph-related predicates (*isAllomorphTo*, *isHomonymTo*) are of particular relevance.

The benefit of using an LOD-based format is the simplicity to enrich existing datasets with additional information. In the context of this contribution, the focus lies on connecting lexicographical data based on full forms with real-world samples. The resulting interconnected resources are helpful for a variety

[6] https://hdl.handle.net/11022/0000-0007-C655-A

[7] https://rdf.corpora.uni-leipzig.de

[8] https://www.sadilar.org

[9] http://www.cbold.ish-lyon.cnrs.fr

of user groups and topics, including support of first and second language acquisition, or as a general resource for all types of text-producing activities.

A more detailed discussion of the specific problems when representing these interconnections and the current state of helpful data models is given in the following section.

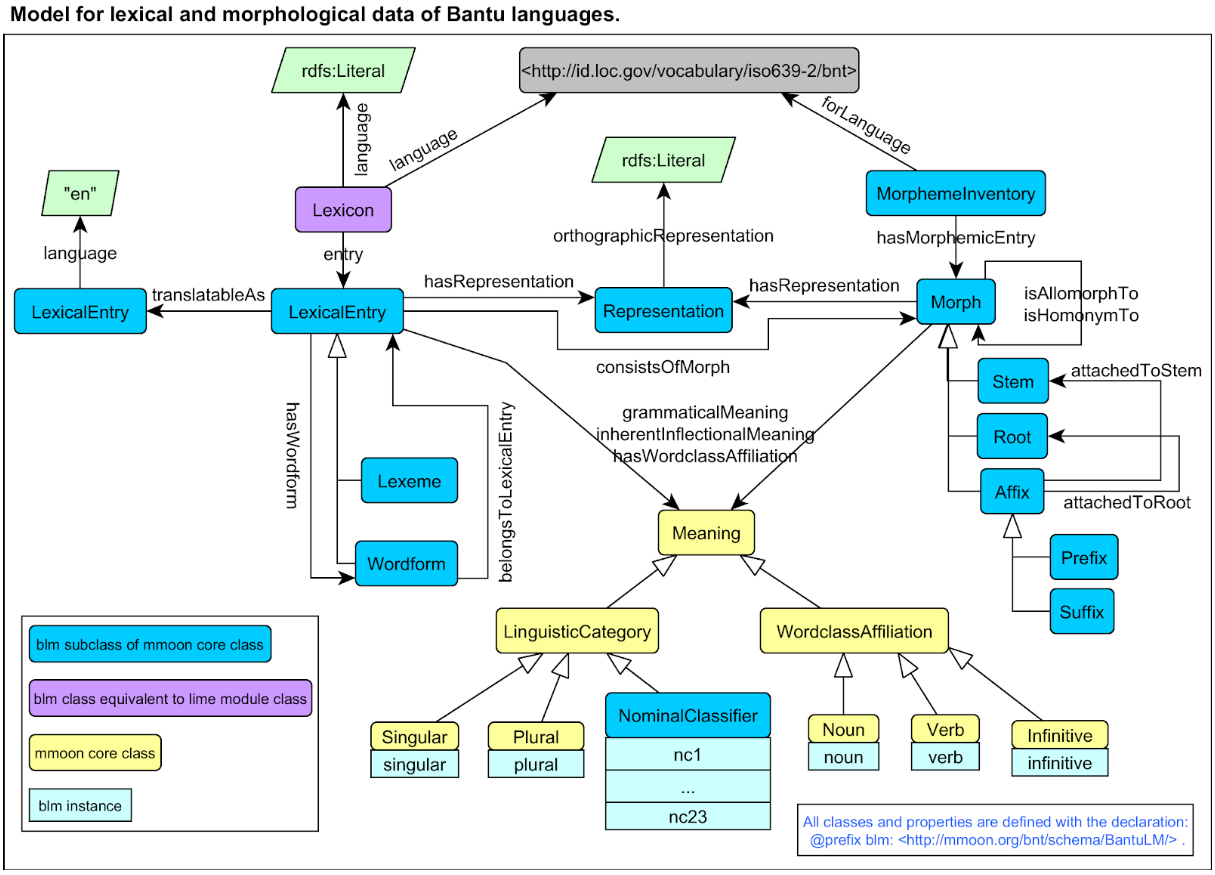


Figure 2: Summary of the Bantu Language Model

4.4 Representing Examples in Current Lexicographical Data Models

Besides standard lexicographical data types focusing on pragmatic, morphosyntactic and similar information with their established means of representation and publication, the growing utilisation of statistical analysis in the field extends the lexicographical focus significantly. Unfortunately, many established data models and formats (like ISO 24613:2008 LMF or the dictionary module of the TEI guidelines) allow the incorporation of these data types only in small parts, if at all.

Current standardisation endeavours try to reduce this problem by extending established standards with new modules or by extending established data models. One candidate currently under progress in the LLOD community is *OntoLex-FRAC*[10] (Frequency, Attestations and Corpus Data) as a new model of the established OntoLex ontology (McCrae et al., 2017). In the context of this paper, it seems to be a fit solution as it will both support referencing concrete usage examples and statistical results. However, the OntoLex-FRAC model is still under development. The incorporation of references to the described data sets based on included full forms will be started as soon as its standardisation is completed.

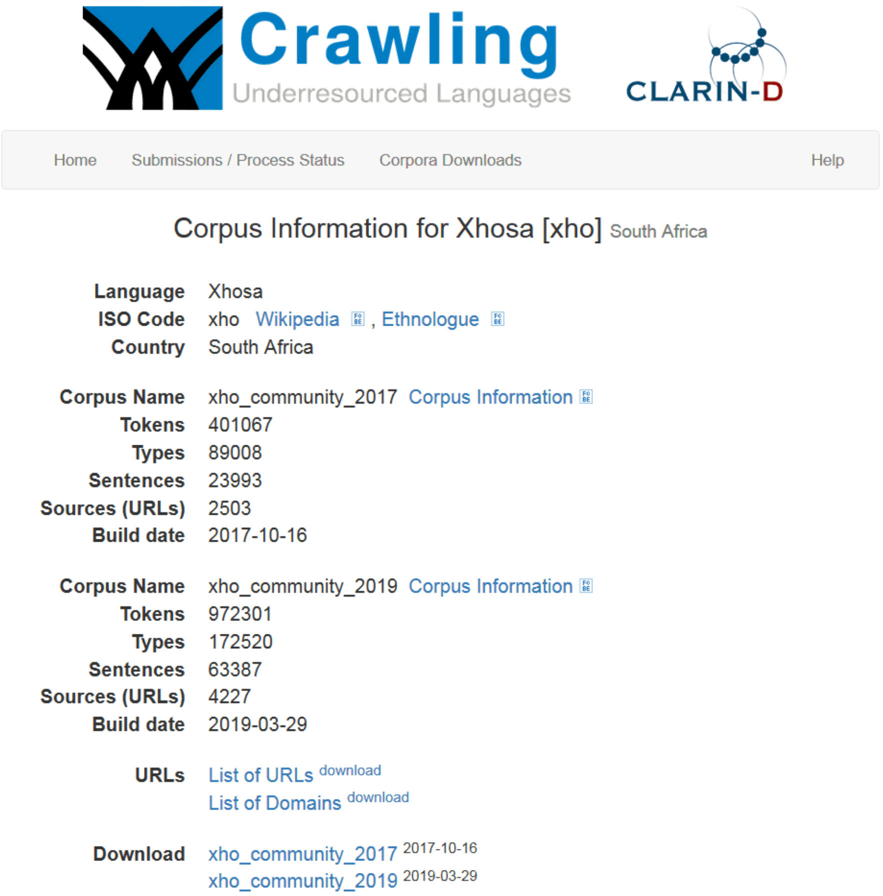
The inclusion of examples by external references makes good use of the CLARIN architecture that strongly relies on persistent identifiers and distributed resources. For example, corpora provided by the Leipzig Corpora Collection – including the data generated by the CURL portal – provide a granularity down to the sentence level (Boehlke et al., 2012), can be addressed using handles with part identifier, and are therefore easy to use for direct reference in a Linked Data environment.

[10] https://github.com/acoli-repo/ontolex-frac

5 Language Data

5.1 Collecting and Processing of Language Data

In a next step, the newly created lexical data were enriched with additional information: the availability of sample sentences proves to be valuable for users of lexical resources since they provide real-world usage examples of the lexical units. In the beginning of the project, available text data for the respective languages was very limited, both in the Leipzig Corpora Collection (LCC) and in other freely available digital resources. For Xhosa, fewer than 18,000 sentences could be found in the LCC. For Kalanga, the situation was even worse with only about 600 sentences. By advertising the initiative to some researchers in the respective communities, we were able to collect 180 seed URLs for Xhosa and one web domain for Kalanga. Crawling resulted in 45,585 additional unique sentences for Xhosa and 996 for Kalanga, increasing available resources significantly. Figure 3 depicts these text collecting efforts for Xhosa.



Corpus Information for Xhosa [xho] South Africa

Language	Xhosa
ISO Code	xho Wikipedia , Ethnologue
Country	South Africa
Corpus Name	xho_community_2017 Corpus Information
Tokens	401067
Types	89008
Sentences	23993
Sources (URLs)	2503
Build date	2017-10-16
Corpus Name	xho_community_2019 Corpus Information
Tokens	972301
Types	172520
Sentences	63387
Sources (URLs)	4227
Build date	2019-03-29
URLs	List of URLs download List of Domains download
Download	xho_community_2017 2017-10-16 xho_community_2019 2019-03-29

Figure 3: Overview of crawling activities in the CURL portal for Xhosa since 2017

The textual resources were processed to serve as a basis for assigning sample sentences to dictionary entries. For Xhosa, this resulted in sample sentences for about 25% of the lexical entries available. Since the coverage is significantly higher for more frequent words and these words are typically queried more often, this will result in a higher sample sentence coverage for actual queries. Additionally, the integration of tools for lemmatisation and morphological decomposition can increase the number even further.

The textual data are made available in the CLARIN infrastructure via the Leipzig repository. This allows for download of the data sets, for searching in the data via the Federated Content Search (FCS) or the local Leipzig Corpora Collection portal, for sustainably citing textual resources on sentence level and for further processing using web tools such as WebLicht. In a next step, extending the Bantu Language Model dataset is planned to allow for a direct linking of lexical entries and sample sentences using LLOD formats and therefore for easier integration, as depicted above.

Although extracting authentic examples from corpora is often a contentious issue, in particular for the purpose of language learners, there are opinions such as that of Frankenberg-Garcia (2012:290) that several corpus examples per lexeme can offer learners concentrated patterns of language that would encourage appropriate generalisations. Hanks (2012:431) is also of the opinion that corpora reflect language as communicative behaviour since “meanings reside only partly in individual words; meanings also reside in the phraseology (or constructions) in which words are used.”

5.2 Sample Results

In this section, identified sample sentences for lexical entries in Xhosa and Kalanga will be presented. For Xhosa, the lexemes *umfazi* and *abafazi* were chosen. Their lexical data is summarized in Table 2.

Lexeme	Prefix	Root	Gram. number	Class	POS	Gloss
umfazi	um	fazi	singular	1	noun	wife
abafazi	aba	fazi	plural	2	noun	wives

Table 2: Lexical information available for the lexemes *umfazi* and *abafazi*

For both lexical entries, sample sentences (with corresponding English translations[11]) can now be provided by matching lexemes with word forms occurring in the crawled text material.

Sample sentences for *umfazi*:

(1) *Abazali bam abagulayo kwaye bafuna ukutyelela US kwaye nzima kuhlala **umfazi** wam.*
“My parents are sick and want to visit the US and it is difficult for my **wife** to stay.”

(2) *Ingaba **umfazi** uzakuziva njani xa indoda yakhe ingaphangeli?*
“How would the **wife** feel if her husband was unemployed?”

(3) *Ndinabantwana abathathu kwaye **umfazi** wam akaphangeli.*
“I have three children and my **wife** is unemployed.”

(4) *‘Uza kugalela ntoni kuyo?’ kwabuza omnye **umfazi**.*
“‘What will you add to it?’ asked another **woman**.”

(5) *Baya efanayo baya oonyana bam xa ufune **umfazi**.*
“The same goes for my sons when you find a **wife**.”

Sample sentences for *abafazi*:

(6) *Ndiza kuqala bathi **abafazi** kufuneka uyeke nangegunya ngokwabo, kodwa nam ndiya umngeni amadoda ukwandisa ku nembasa & ukuqala imbeko **abafazi**.*
*“*I will start by saying that **women** need to give up the power themselves, but I also challenge men to increase in honor & start respecting **women**.”

(7) *Ngenxa yokuba kwakusele amadoda ambalwa kuloo mmandla, kwanyanzeleka ukuba **abafazi** baluthathele kubo uxanduva lokuloba.*
*“*With so few men left in the area, **women** had to take responsibility for their fishing.”

(8) *USankara ukwa ngumongameli wokuqala eAfrika ukuphuhlisa amalungelo **abafazi**; esithi le nto ingumfazi nayo ingumlingani na le nto iyindoda.*
*“*Sankara is also the first African president to promote **women**'s rights; saying that this woman is also a partner is this man.”

[11] The translations provided are automatically generated based on Google Translate results: https://translate.google.com

(9) *Iyomhla kuphela yaye wayezeke* ***abafazi*** *abamnyama ndatshata ngakumbi kuxanduva (umntwana) nolindelo yenkcubeko ngaphezu umdla ofanayo.*
*"*Only date and had married black **women** I married more to the (child) responsibility and cultural expectations than the same interest."

(10) *Beyonce and ezinye iimvumi ababhinqileyo' iingoma kukhokelela* ***abafazi*** *evakalisa ubuni babo yaye ke objectified.*
*"*Beyonce and other female singers' songs lead **women** to express their identity and are therefore objectified."

The identical procedure was applied to the Kalanga inventory. Lexical information for the sample lexeme *ikombo* can be found in Table 3.

Lexeme	**Prefix**	**Root**	**Tone**	**Class**	**POS**	**Gloss**
ikombo	i	kombo	HH	7	noun	navel

Table 3: Lexical information for the lexeme *ikombo*

Identified sample sentences for *ikombo* include the following[12]:

(11) *Atitjaka dwilila taka lingilila baka tatana bakajalo,* ***ikombo*** *tjibe tji shomoka mu liboko gwe mdala tjino wila mbeli kwe mtshana.*

(12) *Atitji gele towana shango ya pituka; mdala wabe mbeli, mtshana wabe iye ushule ne* ***ikombo****, kufiwa tate.*

(13) *Ebe atji nunga* ***ikombo*** *mtshana ndokubudza, ebe e shanduka e lingisana ne mdala.*

(14) *Koti zhulo tili gele kusi kwe mpani pa khisimusi tobona mdala e pinda aka tatamila mtshana elitsha* ***ikombo*** *to come nice.*

(15) *Mtshana a ka amuchila kwa ka nlingisana (****ikombo****).*

(16) *Mtshana alishule ne* ***ikombo*** *akabata; "andibilo mubudza tate ati muletje ndideelela."*

(17) *Mtshana ebe e ntumba ne* ***ikombo*** *atenti.*

(18) *Yaka bobola ikano ngina ka mai ne mabilo titjara alishule ne* ***ikombo*** *atenti.*

(19) *Yeela, tjakalila* ***ikombo*** *ilelo zhuba abona kuti ya, kwiba kuna zwibili.*

For nouns, sample sentences can be attributed easily since the lexical entries are usually available for the singular and/or plural form (see *umfazi* and *abafazi* in Xhosa). For verbs, the situation is more challenging. Lexemes as found in the dictionary do not appear unchanged in the sample sentences; affixation has to be taken into account.

By using tools such as a lemmatiser[13], morphological decomposer[14] and POS-tagger[15] for Xhosa, this gap between the lexicon and crawled full texts can be overcome. The usage of these tools will be investigated in the near future.

5.3 Lexical Ambiguity

Besides lemmatisation, aspects of disambiguation play a role when attributing sample sentences to dictionary entries. In this section, cases of homonymy and polysemy will be discussed. Currently these cases are handled manually. Applying methods of automatic sense disambiguation are limited due to the

[12] Due to missing support in Google Translate, English translations had to be omitted for Kalanga sample sentences and are left as future work.
[13] https://repo.sadilar.org/handle/20.500.12185/310
[14] https://repo.sadilar.org/handle/20.500.12185/311
[15] https://repo.sadilar.org/handle/20.500.12185/323

small number of available text samples.

An example of the sense relation homonymy is demonstrated by means of the noun *ithanga* (plural *amathanga*) which has two unrelated meanings, namely “pumpkin” (a type of vegetable) and “colony” (a geographical area politically controlled by a distant country). The following sentences from the crawled text material (with their English translations) illustrate disambiguation in context:

(20) *Ilizwe alisawulwa ngamandla* ***amathanga*** (colonial powers).
“The country is not ruled by colonial powers (literally – powers of **colonies**).”

(21) *Isitiya sakhe semifuno sasisoloko siyokozela zizinto ezimnandi zokutya, kodwa ngamanye amaxesha ayede adlulise ngobuninzi* ***amathanga,*** *umbona, iitapile nezinye iintlobo zemifuno.*
“His vegetable garden was always full of delicious things to eat, but sometimes even heaped large quantities of **pumpkins**, corn, potatoes and other vegetables.”

The sense relation polysemy, is illustrated in the case of the noun *inyanga* (plural *izinyanga*) which has two related meanings “moon” (astronomy) and “month” (time period). The two sense relations, as disambiguated in context, are illustrated in the following example sentences (accompanied by English translations) extracted from the crawled text material:

(22) *Ndibona* ***inyanga*** *iphuma ndiselapha.*
“I can see the **moon** rising from here.”

(23) *Ngobunye ubusuku,* ***inyanga*** *eyayikhanya yenza kwakho izithunzi edlelweni.*
“One night, a bright **moon** made shadows in the fields.”

(24) *Yayibubusuku obubanda kakhulu, kukhanyise luzizi* ***inyanga*** *eliceba.*
“It was a very cold night, with a clear **moon**light.”

(25) *Wahlala naye* ***inyanga*** *iphela.*
“And he stayed with him for a **month**.”

(26) *Ukuba umsebenzi unqunyanyisiwe okanye utshintshelwe kwenye indawo, umqeshi makenze konke okusemandleni akhe okanye aqukumbele ukuthethwa kwetyala ingadlulanga* ***inyanga*** *umsebenzi lowo enqunyanyisiwe okanye etshintshelwe kwenye indawo*.
“If an employee is suspended or transferred, the employer must do everything in his power or conclude a hearing within one **month** of the employee's termination or transfer.”

6 Conclusion and Further Work

This paper presented a use case for enriching lexicographical data for less-resourced languages with sample sentences. The basis was recently added resources and services of the CLARIN infrastructure such as the Xhosa lexicographical data based on the Bantu Language Model[16] and a portal for crawling under-resourced languages (CURL[17]). Results are made available via the CLARIN infrastructure to allow for wide applicability. They include text corpora for Xhosa[18] and Kalanga[19].

Future work will focus on deeper integration of the CURL portal into the CLARIN infrastructure. Advanced options for format conversion (e.g. TCF or plain text) are planned to be implemented. This will allow for direct processing of crawling results in environments such as WebLicht by employing the Language Resource Switchboard. Support for CLARIN’s private work space solution will increase usability even further. In addition, the RDF datasets will be extended to allow direct reference of sample sentences of which many are already integrated in CLARIN and available for reference using persistent identifiers. The integration of statistical outcomes in the data sets for enhanced usage scenarios is currently under investigation. This has the potential to enhance their utilisation in a mobile dictionary application that is presently under development.

[16] https://hdl.handle.net/11022/0000-0007-C655-A
[17] https://hdl.handle.net/11022/0000-0007-D369-5
[18] https://hdl.handle.net/11022/0000-0007-D396-1
[19] https://hdl.handle.net/11022/0000-0007-D395-2

References

Volker Boehlke, Torsten Compart and Thomas Eckart 2012. Building up a CLARIN resource center – Step 1: Providing metadata. In: Workshop on Describing Language Resources with Metadata at 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul 2012.

Sonja Bosch, Thomas Eckart, Bettina Klimek, Dirk Goldhahn and Uwe Quasthoff 2018. Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment, Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki (Japan) 2018.

Thomas Eckart, Sonja Bosch, Dirk Goldhahn, Uwe Quasthoff and Bettina Klimek 2019. Translation-based Dictionary Alignment for Under-resourced Bantu Languages, OpenAccess Series in Informatics (OASIcs), Vol. 70: Language Data and Knowledge LDK 2019.

Ana Frankenberg-Garcia 2012. Learners' Use of Corpus Examples. International Journal of Lexicography, Vol. 25 No. 3, pp. 273–296. doi:10.1093/ijl/ecs011

Dirk Goldhahn, Thomas Eckart and Uwe Quasthoff 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012) 2012.

Dirk Goldhahn, Maciej Sumalvico and Uwe Quasthoff 2016. Corpus Collection for Under-Resourced Languages with more than One Million Speakers, CCURL 2016 Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity 2016.

Patrick Hanks 2012. The Corpus Revolution in Lexicography. International Journal of Lexicography, Vol. 25 No. 4, pp. 398–436. doi:10.1093/ijl/ecs026

Erhard Hinrichs and Steven Krauwer 2014. The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014) 2014.

ISO 24613:2008 - Language resource management - Lexical markup framework (LMF). Iso.org.

Adam Kilgarriff, Vit Baisa, Jan Buta, Milos Jakubicek, Vojtech Kova, Jan Michelfeit, Pavel Rychly and Vit Suchomel 2014. The Sketch Engine: ten years on, Lexicography, pp. 7–36, Springer 2014.

Bettina Klimek 2017. Proposing an OntoLex-MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models, Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets 2017.

John P. McCrae, Julia Bosque Gil, Jordi Gràcia, Paul Buitelaar and Philipp Cimiano 2017. The OntoLex-Lemon Model: Development and Applications 2017.

Gordon Mohr, Michele Kimpton, Michael Stack and Igor Ranitovic 2004. Introduction to heritrix, an archival quality web crawler, Proceedings of the 4th International Web Archiving Workshop (IWAW'04) 2004.

Derek Nurse and Gérard Philippson 2003. The Bantu languages. London: Routledge.