Ry/Rk-Lex: A Computational Lexicon for Runyankore and Rukiga Languages

David Sabiiti Bamutura
Chalmers University of Technology / Gothenburg, Sweden
Mbarara University of Science & Technology / Mbarara, Uganda
bamutra@chalmers.se | dbamutura@must.ac.ug

Abstract

Current research in computational linguistics and NLP requires the existence of language resources. Whereas these resources are available for only a few well-resourced languages, there are many languages that have been neglected. Among the neglected and / or under-resourced languages are Runyankore and Rukiga (henceforth referred to as *Ry/Rk*). In this paper, we report on *Ry/Rk-Lex*, a moderately large computational lexicon for Ry/Rk that we constructed from various existing data sources. Ry/Rk are two under-resourced Bantu languages with virtually no computational resources. About 9,400 lemmata have been entered so far. Ry/Rk-Lex has been enriched with syntactic and lexical semantic features, with the intent of providing a reference computational lexicon for Ry/Rk in other NLP (1) tasks such as: morphological analysis and generation; part of speech (POS) tagging; named entity recognition (NER); and (2) applications such as: spell and grammar checking; and cross-lingual information retrieval (CLIR). We have used Ry/Rk-Lex to dramatically increase the lexical coverage of previously developed computational resource grammars for Ry/Rk.

1 Introduction

Almost all computational linguistics and natural language processing (NLP) research areas require the use of computational language resources. However, such resources are available for a few well-resourced and "politically advantaged" languages of the world. As a result, most languages remain neglected. Recently, the NLP community has started to acknowledge that resources for under-resourced languages should also be given priority. Why? One reason being that as far as language typology is concerned, the few well-resourced languages do not represent the structural diversity of the remaining languages (Bender, 2013).

This study is a follow-up to a previous, but related study on the engineering of computational resource grammars for Runyankore and Rukiga (henceforth referred to as *Ry/Rk*) (Bamutura et al., 2020), using the Grammatical Framework (GF) and its Resource Grammar Library (Ranta, 2009a,b). In the previous study, a narrow-coverage lexicon of 167 lexical items was sufficient for grammar development. In order to both encourage wide use of the grammar (in real-life NLP applications) and fill the need for computational lexical language resources for Ry/Rk, it was necessary to develop a general-purpose lexicon. Consequently, we set out to create *Ry/Rk-Lex*, a computational lexical resource for Ry/Rk. Despite the challenges faced due to lack of substantial open source language resources for Ry/Rk, we have so far entered about 9,400 lemmata into Ry/Rk-Lex. Ry/Rk has been enriched with syntactic and lexical semantic features, with the intent of providing a reference computational lexicon for Ry/Rk that can be used in other NLP tasks and applications.

1.1 Runyankore and Rukiga Languages

Ry/Rk are two languages spoken by about 3.4 and 2.4 million people (Simons and Fennig, 2018) respectively. They belong to the *JE10* zone (Maho, 2009) of the Great Lakes, Narrow Bantu of the Niger-Congo language family. The native speakers of these languages are called Banyankore and Bakiga respectively. The two peoples hail from and / or live in the regions of Ankole and Kigezi — both located in South Western Uganda, East Africa.

Just like other Eastern Great Lakes Bantu languages, Ry/Rk are *mildly tonal* (Muzale, 1998), *highly agglutinating* with a *large noun class system* (Katushemererwe and Hanneforth, 2010; Byamugisha et al., 2016). They exhibit high incidences of *phonological conditioning* (Katushemererwe et al., 2020) that makes them complex to deal with computationally. The agglutinating nature, intricate concordial agreement system and phonological conditioning make it more difficult to model

and formalise the grammars for these languages using the symbolic approach. For details about the nominal and verbal morphology of these languages from the perspective of computational linguistics, the reader should see (Katushemererwe, 2013; Byamugisha, 2019; Bamutura et al., 2020; Katushemererwe et al., 2020).

1.2 Challenges of Creating Computational Lexica for Runyankore and Rukiga

Though Ry/Rk languages are spoken by a sizeable population they are under-resourced and have a limited presence on the web. When we consider the creation of computational language resources for these languages, four major problems stand out: (1) large amounts of language data must be collected manually by copy-typing which is time-consuming and error-prone; (2) refusal by publishers of books and dictionaries to allow their texts to be used as sources of these data; (3) lack of an easy to use and extensible modelling and storage format for computational lexicons for Bantu languages; and (4) lack of funds to procure copyrighted works for the extraction and processing of computational lexicons and other resources. These lexical resources are however very important for the success of other NLP (1) tasks such as: morphological analysis and generation; part of speech (POS) tagging; named entity recognition (NER); and (2) applications such as spell and grammar checking ; and cross-lingual information retrieval (CLIR).

1.3 Research Questions

This study was guided by the following research questions:

RQ.1 What are the existing linguistic data sources that can be used for the development of computational lexicons for Ry/Rk?

RQ.2 Out of the sources identified in RQ.1, which sources are suitable for use as a computational lexicon for Ry/Rk?

RQ.3 How can computational lexicons for Ry/Rk be extracted and modelled or structured in a simple, flexible and extensible manner?

The rest of the paper is structured as follows: Section 2 presents related work; Section 3 presents the data used for the study, its sources, how it was curated and processed; and Section 4 presents the findings in terms of how Ry/Rk-Lex was described i.e. how the different parts of speech were handled, the persistence structure that was used for storage of lexical items. Results & discussion are presented in Section 5. Lastly, Section 6 presents conclusion and future work.

2 Related Work

2.1 Computational Lexica

Machine Readable Dictionaries (MRDs) and computational lexicons for well-resourced languages such as those reported by Sanfilippo (1994), and ACQUILEX projects I and II[1] were created from existing conventional dictionaries. The aim in those studies was to explore lexical language analysis use cases such as building lexical knowledge-bases. The task of creating MRDs was made easier because the dictionaries used had machine-readable versions that were made available i.e. without copyright restrictions.

In the case of Ry/Rk, such an approach is difficult largely because Ry/Rk dictionaries do not include rich morphosyntax (mainly due to the complex morphology). Additionally, most of the dictionaries are protected by copyright. The lexical semantic relation information (hypernymy and meronymy) provided in the Runyankore and Rukiga thesaurus (Museveni et al., 2012) would be a good starting point but it is also copyrighted.

In addition to having MRDs, well-resourced languages possess the following: large amounts of language data available on the web; prepared corpora of good quality; treebanks (Xiao, 2008; Taylor et al., 2003; Böhmová et al., 2003); and lexical databases such as the original English WordNet (Miller, 1995) and subsequent additions (Christiane and Miller, 1998). Petrolito and Bond (2014) provide a comprehensive survey of different existing language-specific WordNet-based lexical databases and Navigli and Ponzetto (2010) describe a wide-coverage multilingual semantic network derived from combining WordNet and Wikipedia. These resources make the creation of computational lexical resources easier for these languages. It is important to note that the same resources were developed by well-funded research groups.

Among the Bantu languages, computational lexicons have been developed for some languages such as Swahili (Hurskainen, 2004) in East Africa, and isiZulu and isiXhosa (Bosch et al., 2006) in South Africa using XML and related technologies for

[1] see: https://www.cl.cam.ac.uk/research/nl/acquilex/

modelling and annotation. The computational lexicon for Swahili — developed as part of the Swahili Language Manager (SALAMA) — and other South African languages are perhaps the most comprehensive in terms of: (1) the number of lexical items covered and (2) addressing lexical semantic relation issues such as synonymy. The lexical resource for South Africa has been expanded (both by size and number of languages) and converted into the African WordNet (AfWN) to include other southern Africa Bantu languages namely; Setswana, Sesotho, isiNdebele, Xitsonga and Siswati (Griesel and Bosch, 2014, 2020). However, there has been no attempt to create an enriched computational lexical resource for Ry/Rk.

2.2 Computational Lexicon Modelling

With regard to modelling of lexicons for Bantu languages, a Bantu Language Model (BantuLM) was put forward by Bosch et al. (2006, 2018) after eliciting the inadequacies of Lexical Markup Framework (LMF) (Francopoulo et al., 2006) arising from a failure to take such morphologies into account when designing the framework. It was also posited that using BantuLM to prepare lexical resources would encourage cross-language use cases. Bosch et al. (2006) implemented BantuLM using XML and related technologies, while Bosch et al. (2018) switched to an ontology-based approach for describing lexicographic data that combined the best of the Lexicon Model for Ontologies and the Multilingual Morpheme Core Ontology (MMoOnCore) to realise the features envisaged in the BantuLM. Although ontology-based methods encourage the cross-linking of multilingual data, they require a knowledge-base of lexical semantic relations. With the exception of synonym information available in some dictionaries (Taylor and Mapirwe, 2009; Mpairwe and Kahangi, 2013a; Museveni et al., 2009) and basic semantic relations found in a thesaurus for Ry/Rk (Museveni et al., 2012), there are no other sources for such data. Use of ontology-based (semantic networks) for lexical language resources necessitates the formalising the meaning of lexical items beyond word definitions (also called glosses) which current sources do not provide. Going beyond definitions or glosses requires a separate study with huge human and capital resources to turn these resources into lexical semantic networks such as WordNet. YAML[2] was chosen for the preparation, storage and sharing of the Ry/Rk lexicon because for our current purposes we do not require the complex modelling provided for by BantuLM.

3 Data Sources, Curation & Processing

3.1 Existing Data Sources

In total, fourteen linguistic data sources summarised in Table 1 were identified (by web-search, visiting bookshops and publishing houses in Uganda) as the existing data sources that could be used for the development of electronic corpora and or lexica for Ry/Rk. Due to copyright restrictions, we used five of the fourteen sources in whole for lexical resource creation. These five sources (identified as; RRDict1959, RRBibleNew1964, RRSCAWL2004, RRUDofHR and RREthics) are marked using * in that table. However, as explained later in detail in section 3.2.4, we used RRNews2013-2014 (marked with † in the same Table 1) in whole but have made deliberate effort to make sure that only small random fragments of the corpus can be released for demonstration purposes in an academic setting. Other sources marked with ‡ were used solely for reference in case of lack of knowledge.

3.2 Data Curation & Processing

Having obtained sources of data that could be used, the language data contained in those sources had to be extracted and pre-processed in order to obtain individual word tokens. Because the methods used were slightly different for each data source, we explain the process used for each in Sections; 3.2.3, 3.2.1, 3.2.2 and 3.2.4. The procedures used for RRUDofHR and RREthics are identical to those described in section 3.2.2 and 3.2.4 respectively because the former was also scraped from the web while the later required scanning of a hard copy.

3.2.1 RRDict1959

To the best of our knowledge, there is only one MRD for Ry/Rk identified as RRDict1959 in Table 1. It was extracted from the dictionary by Taylor (1959). The MRD is freely available for use as long as one abides by a Bantuist Manifesto.[3] On close inspection of the entries, a number of anomalies were found: (1) singular and plural forms of nouns are entered as separate entries, (2) some entries do

[2] A markup language available at: https://yaml.org

[3] The manifesto can be read at http://www.cbold.ish-lyon.cnrs.fr/Docs/manifesto.html

not qualify as lemmata because they possess additional and unnecessary derivational and inflectional morphemes, (3) lack of conjugation information for verbs, (4) lack of new lemmata that have been introduced to Ry/Rk since 1959, and (5) entries lack synonym information. The first three anomalies were corrected manually by eliminating non-lemma entries, stripping off the unnecessary affixes and providing verbal morpheme endings that guide verb conjugation. For example, we did not agree with the use of the /*ku*/ morpheme as a prefix before a verb because it is unnecessary. Placing /*ku*/ before the verb is akin to placing the word /*to*/ before every verb in English and yet /*to*/ is rarely entered in dictionaries. It is also an unnecessary repetition. The same was done during lemmatisation of verbs from other sources.

3.2.2 RRBibleNew1964

Since a digital version of the New Testament Bible in Runyankore-Rukiga (identified as RRBibleNew1964 in Table 1) is available, it was scrapped from the web after which text pre-processing was done. This pre-processing included text cleaning (removal of HTML markup text, chapter and verse identifiers), text tokenisation, lemmatisation, POS tagging and annotation of each lexical item with simple inflectional morphology i.e. conjugation for verbs, noun class information for nouns, definition glosses for English and synonyms. Lemmatisation and POS tagging were done manually by 4 research assistants. For lemmatisation of verbs, we chose to use the radical concatenated with a final morpheme which most of the time is simply a vowel, called the Final Vowel (FV). This final morpheme is the verbal ending used for the experiential present tense. The open-source machine readable dictionary (RRDict1959) was used to validate our lemmatisation, POS tagging and noun-class identification process for words that existed in the dictionary.

3.2.3 RRSCAWL2004

RRSCAWL2004 is an English–French bilingual list of 1,700 words that was compiled and suggested by Snider and Roberts. (2004) as a useful seed-list for any researcher doing comparative linguistic studies on African languages. Because this list was prepared for Africa, it is highly likely to capture the common concepts used by the ordinary African, such as Ry/Rk speakers. The words in the list are organised semantically under twelve main headings with further subdivisions. The words cover concepts ranging from human to non-human and from concrete to abstract. Since the data is presented within tables of a file in PDF, we used Tabula,[4] a piece of free software to quickly extract these tables locked up in PDF. Tabula is able to export that data into comma separated values (CSV) or Microsoft Office Excel file formats. We hired a professional translator to translate the English glosses to Runyankore and Rukiga. The resulting list was further annotated and fed into Ry/Rk-Lex.

3.2.4 RRNews2013-2014

From scanned images of Orumuri Newspaper, we used the Optical Character Recognition (OCR) feature for English found in Adobe Acrobat Pro DC[5] to extract text from the images. This text was copied and pasted in xml documents that served partially to preserve the structure and content of the newspaper and its articles. Due to the lack of existing OCR software trained specifically on Ry/Rk, errors were encountered and these were corrected manually. Sometimes, it required copying sentence by sentence or paragraph by paragraph. There were two major types of errors: simple spelling mistakes and unrecognisable characters spanning one or several lines of an article. The line errors were mainly associated with Ry/Rk words that contained /ii/ or /aa/ and we are still investigating the reason(s) for this behaviour. Other problems emanated from lists illustrated using bullet points. We used xml to divide the structure of the newspaper into several sections: (1) Amakuru, (2) Amabaruha, (3) Amagara, (4) Shwenkazi, (5) Regional News (Kigezi, Bushenyi, Mabara), (6) Omwekambi and (7) Emizaano. Although the news corpus collected is of poor quality in terms of grammar (Katushemereirwe, personal communication), it is lexically rich and contains words that have been introduced in the languages due to interaction with other languages and globalisation. It therefore contributes significantly to the number of words used currently in contemporary Ry/Rk that are not contained in RRDict1959, RRBibleNew1964, RRVoc2004 and RRSCAWL2004. RRNews2013-2014 was cleaned, tokenised and lemmatised in the same way as RRBibleNew1964 as described in 3.2.2 above.

[4] See: https://tabula.technology/

[5] Version: 221.001.20145 for Mac OS X

3.3 Summing It Up

After pre-processing RRDict1959 to remove the first three anomalies mentioned previously in section 3.2.1, the data obtained was used to validate our lemmatisation, POS tagging and noun-class identification process for lemmata that exist in both RRDict1959 and those that were manually extracted from the completed parts of RRBibleNew1964, RRUDofHR, RREthics, RRSCAWL2004 and RRNews2013–2014. Since text from RRDict1959 and RRBibleNew1964 is dated, the lemmata obtained from the manually created corpus from Orumuri,[6] a weekly Runyankore-Rukiga newspaper, RRUDofHR, RREthics, and lemmata obtained from RRSCAWL2004 and RRVoc2004 (Kaji, 2004) were used to update the RyRk-Lex with words currently used in RyRk. It should be noted that the creation of the RRCorpus and its processing for lexicon extraction is still ongoing.

4 Findings: Ry/Rk-Lex Description

The properties or features for each lemma depend on a number of factors but the major determinants are: the part of speech (POS); the language to which the lemma belongs; and availability of synonyms and definition glosses in English. While the language property is mandatory for all lemma entries, verbs present a problem because the lemma is usually identical for both languages but its method of conjugation differs for each language. We kept the field mandatory for the simple reason that the lemma belongs to both languages although conjugated differently by each language as explained with an example in Subsection 4.2. Otherwise, the properties peculiar to each part of speech are discussed in the subsections below. These properties are illustrated in Table 2 which summarises the structure of Ry/Rk-Lex as specified in a schema[7] we developed whose structure is further described in Section 4.1.

4.1 Ry/Rk-Lex Persistence Structure

For purposes of preparing a shareable resource, we described and stored each entry using YAML. Entries are entered according to a YAML Schema that we designed. Ry/Rk-Lex is shareable because of the schema which communicates the structure of the lexicon. The schema was also utilised for validation of Ry/Rk-Lex in order to identify and correct errors. Manually identified synonyms have been entered for some lemma entries in Ry/Rk-Lex but have not yet been cross-linked.

4.2 Verbs

We have obtained, prepared and stored about 3500 verbs. The verbal features covered include the lemma which is the radical[8] and its final vowel for the experiential present tense (Muzale, 1998; Bamutura et al., 2020). The entry is complemented by a conjugation field that demonstrates how the verb can be conjugated to any of the tenses in Ry/Rk i.e. far past, near past, experiential present, memorial present, near future and far future. Interestingly, the key to performing that conjugation correctly depends on knowing the morpheme for the perfective aspect for the post radical position of the verb. This morpheme is allomorphic and therefor realised differently. The allomorph chosen for a particular verb depends on the following four properties of the verb in experiential present: (1) the syllable structure (2) the penultimate vowel, (3) length of the penultimate vowel and (4) terminal syllable of the verb (Mpairwe and Kahangi, 2013b). Mpairwe and Kahangi (2013b) further attempt at describing these rules for determining the allormorphs as a rule-based procedure or "pseudo" algorithm. Although these rules are natural to a native speaker of the languages, attempts at implementing them as a computer program produced sub-optimal results. .

The verb type field specifies the valency of the verb ignoring any valency increasing derivational suffixes i.e extensions for applicative and causative constructions. Since this lexicon covers two closely related languages, each lemma belonging to the verb POS is annotated with a property for specifying the language. As already mentioned previously, the value for the language field does not depend only on the radical or stem but also the way the verb is conjugated. For instance the verb /reeta/ meaning /bring/ would be conjugated to /reet + sire/ and /ree + sire/ resulting in the surface forms /reetsire/ and /reesire/ in the perfective aspect for Runyankore and Rukiga respectively. Therefore the conjugation field for verbs could be put at top level node but to be more specific it should appear under the conjugation node. We decided to do it at

[6] The publisher, Vision Group terminated the publication of the newspaper in 2020

[7] See appendix I for the full structure

[8] A radical is a sub unit of a stem taken from the base, for details, see Meeussen (1967)

Source	ID	type/Genre	mode	copyright
Taylor (1959)	RRDict1959*	Dictionary	MRD	Free
New Testament Ry/Rk Bible	RRBibleNew1964*	Religion	electronic	Free
Snider and Roberts. (2004)	RRSCAWL2004*	Word List	PDF	Free
Taylor and Mapirwe (2009)	RRDict2009	Dictionary	hard copy	restricted
Kaji (2004)	RRVoc2004‡	Vocabulary List	hard copy	restricted
Orumuri	RRNews2013-2014†	Newspaper	hard copy	restricted
Morris and Kirwan (1972)	RRGrammar1972‡	Grammar book	hard copy	restricted
Mpairwe and Kahangi (2013b)	RRGrammar2013‡	Grammar book	hard copy	restricted
Mpairwe and Kahangi (2013a)	RRDict2013	Dictionary	hard copy	restricted
Museveni et al. (2009)	RRDict2009	Dictionary	hard copy	restricted
Museveni et al. (2012)	RRThes2012	Thesaurus	hard copy	restricted
Karwemera (1994)	RRCgg1994	Book	hard copy	restricted
Universal Declaration of Human Rights	RRUDofHR*	Law	electronic	free
Government communication	RREthics*	Simplified law	hardcopy	free

Table 1: Summary of data sources for corpora and lexical resources. Note: Items marked with * were used without special consideration of copyright. Those with † were used in whole but the resulting corpus will unfortunately not be freely available. Those with ‡ were used solely for reference i.e. lookup of particular information such as synonyms and lemmas for closed categories.

property	type	Optionality	Description
lemma	string	Mandatory	The conventional citation form of a lexical item
lemma_id	integer	Mandatory	The numerical identifier of the lemma
pos	map	Mandatory	The part of speech defined at two levels of granularity.
eng_defn	string	Mandatory	A definition of the lemma in English
synonyms	sequence	Mandatory	A list of synonyms for the lemma
lang	sequence	Mandatory	A list of language identifiers for the lemma
conjugations	sequence of maps	Optional	Non-perfective and perfective Verbal-endings
noun_class	sequence of strings	Optional	Noun class information for nouns

Table 2: Top-level properties for each lemma entry in Ry/Rk-Lex. Each property in column one has a type provided in column two. Column three indicates whether the property is mandatory or optional for each lemma entry while the last column provides a description of the property.

	NC	NCP	Individual Particles		Example		Gloss
ID	Numbers	Particles	Singular	Plural	Singular	Plural	Singular(Plural)
1	β	ZERO_N	n/a	N	n/a	embabazi	n/a (mercy / mercies)
2	σ	N_ZERO	N	n/a	enzigu	n/a	vengeance (n/a)
3	γ	RU_ZERO	RU	n/a	0-ru-me	n/a	dew (n/a)

Table 3: Examples of Ry/Rk nouns without noun classes (NC). Their associated noun class particle (NCP) pairs are shown but the equivalent numeric identifiers as used by the Bleek-Meinhoff system of numbering could not be identified. We therefore used greek letters to represent the unknown.

Part-of-Speech	# of lemmata
Verbs	3532
Common Nouns	4789
Proper Nouns	523
Determiners	124
Pronominal Expressions	85
Adverbs	140
Prepositions	43
Adjectives	148
Conjunctions & Subjunctions	45
Total	9429

Table 4: Number of entries made in Ry/Rk-Lex for each part of speech.

both levels, in order to recognise that the lemma is for both Rukiga and Runyankore but demand any developed parser to further crosscheck for the language property under conjugation.

4.3 Common Nouns and Proper Nouns

In addition to all properties considered mandatory, noun class information was added as an additional field. Both numerical noun classes and textual noun class particles are provided. During lexical collection and processing, three additional categories of nouns that do not fit in the conventional noun class system for Ry/Rk used by Katushemererwe and Hanneforth (2010); Turyamwomwe (2011); Byamugisha et al. (2016) were encountered. An example from each category is illustrated in Table 3.

4.4 Nominal Qualificatives

Nominal qualificatives are expressions that usually qualify nouns, pronouns and noun phrases, and in Ry/Rk include (1) adjectives, (2) adjectival stems and phrases, (3) nouns that qualify other nouns (4) enumeratives (both inclusive and exclusive), (4) relative subject clauses and (5) relative object clauses (Mpairwe and Kahangi, 2013b). Only the nominal qualificatives (1)–(3) were included. Qualificatives (4) and (5) were excluded because they are clauses. Mpairwe and Kahangi (2013b) mention in their grammar book that the notion of adjectives as understood in English results in limited number of adjectives when applied to Ry/Rk. The adjectives are not more than twenty in number. There are however other ways of expressing qualification of nominal expressions in Ry/Rk. We therefore found it difficult to identify and classify all forms of this part-of-speech. In addition to the mandatory properties, four additional properties were required to have adjectives and other nominal qualificatives adequately described. The properties included: position (whether the adjective is located before or after the noun), doesAgree (which indicates whether the adjective changes with respect to the noun class of the nominal being modified), and isProper (a boolean field that captures whether the adjective is a stand-alone or one that requires modification by a suffix). Some adjectival expressions are multi-word expressions (portmateau) such as clauses. These clauses are usually derivational and therefore have been left out of the lexicon.

4.5 Adverbs and Adverbial expressions

Both Schachter and Shopen (2007) and Cheng and Downing (2014) define the adverb as that part-of-speech that modifies all other parts-of-speech apart from the noun. The Universal Dependencies (UD)[9] provides a more concrete definition i.e. "adverbs are words that typically modify verbs for categories such as time, place, direction or manner and they may also modify adjectives and other adverbs". The single exclusion of nouns by all definitions implies that this part of speech is an amalgamation of different words, phrases and clauses as long as they do not modify nouns or noun phrases. For Ry/Rk, Mpairwe and Kahangi (2013b) define it as a word, phrase or clause that answers questions based on the question-words: *where* (for adverbs of place), *when* (for adverbs of time, frequency and condition), *how* (for adverbs of manner and comparison), and lastly *why* (for adverbs of reason or purpose and concession). Most adverbials in Ry/Rk are a single word consisting of two or more words when translated to English. In other words you have a single-word consisting of two or more morphemes belonging to multiple parts of speech. A good example is the word /*kisyo*/ which means /*like that*/ in English and belongs to singular forms of nouns from noun classes 7_8. The associated word /*bisyo*/ for the plural form implies that the stem is /*syo*/. In describing or extracting lemmata for adverbs, we concentrated on adverbial expressions that were easily discernible from a single word. We advise that further work be done for adverbials especially those that span multiple words by obtaining them from professionally annotated corpora alongside detailed annotation guidelines. For instance the multi-morpheme words could obtained from a Ry/Rk corpus that has been anno-

[9] See: https://universaldependencies.org/u/pos/ADV.html

tated using annotation guidelines that are based on a more linguistically sound theory for word class division for Ry/Rk.

4.6 Closed Categories

POS that belong to the closed category are generally few but occur frequently in a corpus. Whereas conjunctions (including subjunctions), prepositions, determiners and quantifiers are actually few in number for Ry/Rk, pronouns constitute a large number. Notably, most POS from the closed category can be adquately covered by working through grammar books such as; (Morris and Kirwan, 1972), (Taylor and Mapirwe, 2009), (Mpairwe and Kahangi, 2013b) and (Ndoleriire, 2020).

4.6.1 Pronouns

Generally, pronouns are words that substitute for nouns or noun phrases and whose meaning is recoverable through anaphora resolution sometimes requiring investigation of linguistic context beyond the sentence. In Ry/Rk, pronominal expressions are either single-word expressions (called pronouns) or pronominal affixes (morphemes) (Mpairwe and Kahangi, 2013b; Katushemererwe et al., 2020). Manually identifying and annotating a single-word pronoun from a tokenised corpus whose sorting is based on most frequent word is much easier than doing the same for pronominal affixes because you lose contextual information that would help with identification. We therefore decided to concentrate on discrete pronouns.

Otherwise, in order to describe and use self-standing or independent pronouns, terms used by (Mpairwe and Kahangi, 2013b,a) and (Katushemererwe et al., 2020) respectively to refer to those pronouns that do not require to be affixed to another POS, the parameters: grammatical gender (noun class), number, person and type of pronoun are required and were captured for this particular POS. Those that have not been covered are affix-based pronouns.

5 Reflections and Discussion

At the time of writing, Ry/Rk-Lex currently consists of 9,429 lemmata of various parts-of-speech summarised in Table 4. From the breakdown we note that verbs and nouns make up the largest share of the total number of lemmata. For the case of verbs, the large number is attributed to the fact that new verbs can be formed via derivation processes such as reduplication, reciprocation and in some cases through the use of applicative and causative constructions common among Bantu languages. Nouns are inherently numerous since they name things. Deverbatives have been excluded so far from Ry/Rk-Lex because they are easy to add once all verbs are known. Despite the low number of proper nouns in Ry/Rk-Lex, this category of nouns is huge and we plan to add more from the Ry/Rk Thesaurus (RRThes2012) after obtaining copyright permission. In Ry/Rk, adverbs are a complicated part of speech. They mostly exist as adverbial expressions constructed from locative noun class particles: */mu/*, */ku/* and */ha/*. As a result, only a few have been considered as lemmata so far but more will be included in future. Parts of speech that belong to closed categories are few and consist of the most frequently used words. For each lemma, we tried our best to enter as much synonym information as we could. However, cross-linking of synonyms has not yet been done due to time constraints but we plan to do it in future. We manually fixed and updated each entry with more information specifically conjugation for verbs and correct noun classes for nouns.

While processing nouns, nouns that did not fall under the accepted noun class numerical system were encountered. In Table 3, examples of such nouns are provided. We suggest that the noun classes used in the numeral system be expanded as some nominal lexical items cannot be brought under the pre-existing numerical system used in literature for Runyankore-Rukiga. Since the notion of adjectives and or nominal qualifiers in Ry/Rk is very limited as mentioned before in subsection 4.4, we found it difficult to identify and classify all forms of this part of speech.

For each lemma entered in the lexicon, a language field is provided to indicate the language the lemma belongs to. A lemma that is used by both languages is annotated with *'all'* while ISO 693-3 three-letter codes *'nyn'* and *'cgg'* are utilised to annotate lemmata that are exclusively used by either Runyankore or Rukiga respectively. It is therefore possible to to automatically extract particular parts of the lexicon for each language. Ry/Rk-Lex attempts to provide a definition in the English language for each lemma despite the fact that this approach to lexical semantics suffers from a number of problems, one of which is circular definitions.

Any current work on lexical resources would expect the inclusion of lexical semantic relations

(synonymy, hypernymy and meronymy) within the resource. Though we have provided some synonym information in Ry/Rk-Lex, we have not yet cross-linked the synonyms. Since YAML provides anchors and references as features, they can be exploited to link synonyms together. Hypernymy and meronymy relations can also be included using a similar method provided knowledge and monetary resources are made available. Since building and maintaining a lexicon is a never-ending process, we are continuously updating it with lemmata as we find more texts written in the language or using free word lists such as: The SPECIALIST LEXICON[10] (Browne et al., 2018); and or the lexicon embedded in the SimpleNLG API and the English Open Word List (EOWL)[11] prepared by Loge (2015). It contains 128,985 words and was extracted from the UK Advanced Cryptics Dictionary (UKACD) Version 1.6.

6 Conclusion and Future Work

In this paper, we have described the creation of Ry/Rk-Lex, a computational lexicon for Ry/Rk. It currently consists of 9,429 lemma entries. Since the languages are under-resourced, we found only fourteen data sources that could be used for its creation. Of the fourteen, only five were utilised as a whole without special consideration of violation of copyright because they are free from copyright. In order to store and make the resource shareable, we designed a schema for structuring the lexicon and used it to organise and annotate all lemmata that have been extracted from the data sources by both manual and automatic methods.

As future work, we plan to build and evaluate conjugation, lemmatisation, morphological analyser and generator, POS tagging software for Ry/Rk that can be used to speed up the process of language resource creation. With these software tools in place, Ry/Rk-Lex can also be used for developing systems for cross-lingual information retrieval (CLIR) especially for people with moderate to poor competence in English but competent in writing Ry/Rk.

For a broader audience, the CLIR system could be augmented with an automatic speech recognition (ASR) module for Ry/Rk targeted towards specific domains. Although Ry/Rk-Lex does not contain all lexical semantic knowledge, our resource can still be used as a starting point for the computational formalisation of the lexical semantics of Ry/Rk and for developing an Ry/Rk WordNet. In its current form, Ry/Rk-Lex has been used to dramatically improve (from 167 to 9,429 lemmata) the lexical coverage of the computational resource grammars of Ry/Rk.

Lastly, there is also need to do more research on establishing a linguistically motivated and sound theory or criteria for word class division and / or drawing the thin line between morphology and lexicon for Ry/Rk as a Bantu language. Using such a criteria would result into lexica that does not appear to be modelled on English and or Latin-based languages. For Ry/Rk-Lex, the word class division was inspired by Indo-European languages and used by GF. However, establishment of a common ground amongst languages in the tradition of the Universal POS tags[12] and the general guidelines put forward by UD version 2 project on the handling of morphology[13] is currently the main focus and future direction this research study.

Acknowledgments

The author would like to acknowledge and thank his academic advisers: Peter Ljunglöf and Peter Nabende for their insightful comments and reviews, research assistants; Ms. Fulgencia Kabagyenzi, Ms. Juliet Nsimire, Ms. Doreck Nduhukire, Mr. Anyine Cranmar for their tireless work in transcribing some of the data sources, and Mr. Jackson Ndyanimanya for translating the RRSAWL2004.

Special thanks to the two reviewers who gave very detailed and helpful comments that have to a great extent improved the quality manuscript.

This work was supported by the Sida / BRIGHT Project 317 under the Makerere-Sweden Bilateral Research Programme 2015–2020 and 2020–2021.

References

David Bamutura, Peter Ljunglöf, and Peter Nebende. 2020. Towards Computational Resource Grammars for Runyankore and Rukiga. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2846–2854, Marseille, France. European Language Resources Association.

[10] Available at https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/web/release/2020.html

[11] see: https://diginoodles.com/projects/eowl

[12] See:https://universaldependencies.org/v2/postags.html

[13] See:https://universaldependencies.org/u/overview/morphology.html

Emily M. Bender. 2013. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague Dependency Treebank. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 103–127. Springer Netherlands, Dordrecht.

Sonja Bosch, Thomas Eckart, Bettina Klimek, Dirk Goldhahn, and Uwe Quasthoff. 2018. Preparation and usage of Xhosa lexicographical data for a multilingual, federated environment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Sonja E. Bosch, Laurette Pretorius, and Jackie Jones. 2006. Towards machine-readable lexicons for south African Bantu languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Allen C. Browne, Alexa T. McCray, and Suresh Srinivasan. 2018. The SPECIALIST LEXICON. Technical report, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Beshesda, Maryland.

Joan Byamugisha. 2019. *Computational Morphology and Bantu Language Learning: An Implementation for Runyakitara*. PhD Dissertation, University of Cape Town, Computer Science Department.

Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2016. Bootstrapping a Runyankore CNL from an isiZulu CNL. In *Controlled Natural Language*, pages 25–36. Springer International Publishing.

Lisa Cheng and Laura Downing. 2014. *The problems of adverbs in Zulu*, pages 42–59. John Benjamins Publishing Company.

Fellbaum Christiane and George A. Miller, editors. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical markup framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Marissa Griesel and Sonja Bosch. 2014. Taking stock of the African Wordnet project: 5 years of development. In *Proceedings of the Seventh Global Wordnet Conference*, pages 148–153, Tartu, Estonia. University of Tartu Press.

Marissa Griesel and Sonja Bosch. 2020. Navigating challenges of multilingual resource development for under-resourced languages: The case of the African Wordnet project. In *Proceedings of the first workshop on Resources for African Indigenous Languages*, pages 45–50, Marseille, France. European Language Resources Association (ELRA).

Arvi Hurskainen. 2004. Swahili language manager: A storehouse for developing multiple computational applications. *Nordic Journal of African Studies*, 13(3):363 – 397.

Shigeki Kaji. 2004. *A Runyankore Vocabulary*. Research Institute for Languages and Cultures of Asia and Africa (ILCAA), Tokyo University of Foreign Studies in English.

Festo Karwemera. 1994. *Emicwe n'Emigyenzo y'Abakiga*. Fountain Publishers, Kampala, Uganda.

Fridah Katushemererwe. 2013. *Computational morphology and Bantu language learning: an implementation for Runyakitara*. Ph.D. thesis, University of Groningen.

Fridah Katushemererwe and Thomas Hanneforth. 2010. Fsm2 and the morphological analysis of Bantu nouns – first experiences from Runyakitara. *International Journal of Computing and ICT research*, 4(1):58–69.

Fridah Katushemererwe, Oswald K. Ndoleriire, and Shirley Byakutaaga. 2020. Morphology: General description and nominal morphology in runyakitara. In Oswald K. Ndoleriire, editor, *Runyakitara Language Studies: A Guide for Advanced Learners and Teachers of Runyakitara*, pages 33–74. Makerere University Press, Kampala, Uganda.

Ken Loge. 2015. English open word list (EOWL). `https://diginoodles.com/projects/eowl`. Accessed: 2021-02-27.

Jouni Filip Maho. 2009. NUGL Online: The online version of the New Updated Guthrie List, a referential classification of Bantu languages. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.603.6490`.

Achille Emile Meeussen. 1967. Bantu grammatical reconstructions. *Africana Linguistica*, 3(1):79–121.

George A. Miller. 1995. Wordnet: A lexical database for English. *Commun. ACM*, 38(11):39–41.

H. F. Morris and Brian Edmond Renshaw Kirwan. 1972. *A Runyankore grammar, by H. F. Morris and B. E. R. Kirwan*, revised edition. East African Literature Bureau Nairobi.

Yusuf. Mpairwe and G.K. Kahangi. 2013a. *Runyankore-Rukiga Dictionary*. Fountain Publishers, Kampala.

Yusuf. Mpairwe and G.K. Kahangi. 2013b. *Runyankore-Rukiga Grammar*. Fountain Publishers, Kampla.

Yoweri Museveni, Manuel J.K Muranga, Alice Muhoozi, Aaron Mushengyezi, and Gilbert Gomushabe. 2009. *kavunuuzi y'orunyankore/Rukiga omu Rugyeresa : Runyankore/Rukiga-English Dictionary*. Institute of Languages, Makerere University, Kampala, Uganda.

Yoweri Kaguta Museveni, Manuel Muranga, Gilbert Gumoshabe, and Alice N. K. Muhoozi. 2012. *Katondoozi y'Orunyankore-Rukiga Thesaurus of Runyankore-Rukiga*. Fountain Publishers, Kampala, Uganda.

Henry R T Muzale. 1998. *A Reconstruction of the Proto-Rutara Tense / Aspect System*. Ph.D. thesis, Memorial University of Newfoundland, Canada.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Oswald K. Ndoleriire, editor. 2020. *Runyakitara Language Studies: A Guide for Advanced Learners and Teachers of Runyakitara*. Makerere University Press, Kampala, Uganda.

Tommaso Petrolito and Francis Bond. 2014. A survey of WordNet annotated corpora. In *Proceedings of the Seventh Global Wordnet Conference*, pages 236–245, Tartu, Estonia. University of Tartu Press.

Aarne Ranta. 2009a. GF: A multilingual grammar formalism. *Linguistics and Language Compass*, 3(5):1242–1265.

Aarne Ranta. 2009b. The GF Resource Grammar Library. *Linguistic Issues in Language Technology*, 2(1).

Antonio Sanfilippo. 1994. *LKB Encoding of Lexical Knowledge*, page 190–222. Cambridge University Press, USA.

Paul Schachter and Timothy Shopen. 2007. Parts-of-speech systems. In TimothyEditor Shopen, editor, *Language Typology and Syntactic Description*, 2 edition, volume 1, page 1–60. Cambridge University Press.

Gary F. Simons and D. Fennig, Charles. 2018. *Ethnologue: Languages of the world*, Twenty-first edition. SIL International, Dallas, Texas. Online version: http://www.ethnologue.com.

Keith Snider and James Roberts. 2004. SIL Comparative African word list (SILCAWL). *The Journal of West African Languages*, 31(2):73–122.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn treebank: An overview. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 5–22. Springer Netherlands, Dordrecht.

Charles Taylor and Yusuf Mapirwe. 2009. *A simplified Runyankore-Rukiga-English English Dictionary*. Fountain Publishers, Kampala, Uganda.

Charles V. Taylor. 1959. *A simplified Runyankore-Rukiga-English and English-Runyankore-Rukiga dictionary : in the 1955 revised orthography with tone-markings and full entries under prefixes*. Kampala : Eagle Press.

Justus Turyamwomwe. 2011. Tense and aspect in Runyankore-Rukiga, linguistic resources and analysis. Master's thesis, NTNU – Norwegian University of Science and Technology.

Richard Z. Xiao. 2008. Well-known and influential corpora. In Anke Ludeling and Merja Kyto, editors, *Corpus Linguistics: An International Handbook*, volume 1 of *Handbooks of Linguistics and Communication Science*. Mouton de Gruyter. This manuscript is not "beautified" so as to fit the publisher's stylesheet. A PDF offprint will be provided when available.

A Appendix

```yaml
% YAML 1.2
---
$schema: "http://json-schema.org/draft-07/schema#"
name: YAML Schema for Ry/Rk-Lex
type : seq
sequence:
  - type: map
    mapping:
      lemma:
        type: str
        required: true
      lemma_id:
        type: int
        required: true
      eng_defn:
        type: seq
        sequence:
          - type: str
        required: true
      pos:
        type: map
        mapping:
          first_level:
            type: str
            required: true
            enum:
              - verb
              - noun
              - adjective
              - adverb
              - preposition
              - pronoun
          second_level:
            type: str
            required: true
        required: true
      synonyms:
        type: seq
        required: false
        sequence:
          - type: str
      lang:
        type: str
        required: true
        enum:
          - all
          - nyn
          - cgg
      conjugations:
        type: seq
        sequence:
          - type: map
            mapping:
              nyn:
                type: str
                required: false
              cgg:
                type: str
                required: false
              all:
                type: str
                required: false
        required: false
      noun_classes:
        type: seq
        sequence:
          - type: str
        required: false
```