RRLex: A Computational Lexicon for Runyankore and Rukiga Languages

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

Abstract

In this paper, we report on RRLex, a moderately large computational lexicon for Runyankore and Rukiga (R&R) that we constructed from various existing data sources. R&R are under-resourced languages with virtually no computational resources. About 4,700 lemmata have been entered so far. RRlex has been enriched with syntactic and lexical semantic features, with the intent of becoming a reference computational lexicon for R&R. We have used RRlex to increase the lexical coverage of previously developed computational resource grammars for R&R.

1 Introduction

Almost all NLP research areas require the use of computational language resources. However, such resources are available for a few well-resourced languages. As a result, the greater set of other languages remain neglected. Recently, the NLP community has started to acknowledge that resources for under-resourced languages should also be given priority (Bender, 2013). This study is done as a follow-up to a previous, but related study on the engineering of computational resource grammars for R&R (Bamutura et al., 2020), using the Grammatical Framework (GF) and its Resource Grammar Library (Ranta, 2009a,b). In that study, a narrow-coverage lexicon was sufficient for grammar development. In order to both encourage wide use of the grammar (in real-life NLP applications) and fill the need of computational lexical language resources for R&R it was imperative to develop a general-purpose lexicon. Consequently, we set out to create *RRLex*, a computational lexical resource for R&R.The lack of open source language resources presented a number of challenges that we discuss in subsections 2.1 and 4.1.

2 Background

2.1 Runyankore and Rukiga

Runyankore and Rukiga are two similar Bantu languages spoken by about 6 million people (Simons and Fennig, 2018) in the south-western region of Uganda in East Africa. They are under-resourced languages with an almost zero existence on the web. When it comes to the creation of computational language resources (especially lexical resources) for R&R, four major problems stand out: 1) language resources must be collected manually 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 lexical information, 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.

2.2 Related Work

Machine Readable Dictionaries (MRDs) and computational lexicons for well-resourced languages such as those reported in (Sanfilippo, 1994),and ACQUILEX projects I and II[1] were created from existing conventional dictionaries with the purpose of exploring lexical language analysis use cases such as building lexical knowledge-bases. The dictionaries used not only had human-readable paper versions but also machine-readable versions which made lexicon creation easier. In addition to lemma entries and their Part-of-Speech (PoS) tags, these lexicons contained richer information in terms of subcategorisation features for verbs and nouns. In the case of R&R and other Bantu languages, such an approach is difficult largely because R&R dictionaries do not include rich morphosyntax (mainly due to the complex mor-

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

phology), and lexical semantic relation information (hypernymy and meronymy). 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 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 number of lexical items covered and addressing lexical semantic relation issues such as synonyms. 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 R&R. Therefore in this paper, we present such a resource for R&R using available data resources.

2.3 Existing lexical resources for R&R

We found only one MRD for R&R identified as RRDict1959 in table 1. It was extracted from the dictionary by Taylor (1959). The MRD is freely available for use subject to abiding by a Bantuist Manifesto[2]. On close inspection of the entries, we found a number of anomalies: 1) singular and plural forms of nouns are entered as separate entries, 2) some entries do not qualify as lemmata because they possess additional and unnecessary derivational and inflectional morphemes, 3) the lack of conjugation information for verbs, 4) the lack of new lemmata that have been introduced to R&R 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.

2.4 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 (Francopoulo et al., 2006) arising from the complex morphology of Bantu Languages. 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. We chose to use YAML[3] for the preparation, storage and sharing of the R&R lexicon because for our current purposes we do not require the complex modelling provided for by BantuLM.

3 Data Sources

In total, eleven data sources summarised in table 1 were identified (by web-search and 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 R&R. Due to copyright restrictions, we used only five of the eleven sources for lexical resource creation. These five sources are marked with * in table 1. The RRVoc2004 data source required copy typing entries from its vocabulary to enrich the lexical coverage of the computational lexicon under development. Likewise hard-copies of RRNews, and RRUDofHR were copy-typed and combined with RRBIble text to form an electronic text corpus[4] (hereafter referred to as RRCorpus) which was later processed as discussed in subsection 4.1.

4 RRLex Implementation

4.1 Data Curation and Processing

The RRCorpus was further cleaned, tokenised, lemmatised and annotated manually with PoS tags and definition glosses. For lemmatisation of verbs, we chose to use the radical concatenated with a final morpheme as used in the Memorial and Experiential Present Tense. In most of the cases this morpheme is simply a vowel and is called the Final Vowel (FV). After pre-processing RRDict1959 to remove the first three anomalies mentioned previously in section 2.3, the data obtained was used to validate our lemmatisation, part-of-speech tagging and noun-class identification process for lemmas that exist in RRVoc2004 and those that were manually extracted from the completed parts of RRCorpus. The lemma entries found in the corrected MRD but were absent in RRCorpus, were

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

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

[4] The copy-typing is still on-going

Source	ID	type/Genre	mode	copyright
Taylor (1959)	RRDict1959*	Dictionary	MRD	Free
New Testament R&R Bible	RRBible*	Religion	electronic	Free
Taylor and Mapirwe (2009)	RRDict2009	Dictionary	hard copy	restricted
Kaji (2004)	RRVoc2004*	Vocabulary List	hard copy	restricted
Orumuri	RRNews*	Newspaper	hard copy	Free
Mpairwe and Kahangi (2013)	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)	RCgg1994	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 & or lexical resources. Note: Items marked with * were used when creating RRlex. Orumuri is a weekly Runyankore-Rukiga newspaper

property	type	Optionality	Description
lemma	string	Mandatory	The simplest form of a lexical item
lemma_id	integer	Mandatory	The numerical identifier of the lemma
pos	map	Mandatory	The part of speech.
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 the lexicon. 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.

used to update RRLex. It should be noted that the creation of the RRCorpus and its processing for lexicon extraction is still ongoing.

4.2 RRLex Persistence Structure

For purposes of preparing a shareable resource, we described and stored each entry using Yet Another Markup language (YAML). Entries are entered using a YAML Schema that we designed. RRLex is shareable because of the schema which communicates the structure of the lexicon. The schema was also utilised for validation of RRLex in order to identify and correct errors. Table 2 summarises the structure of RRlex as specified in the schema we developed[5]. Manually identified synonyms have been entered for some lemma entries in RRLex but have not yet been cross-linked. The implemented lexicon can be accessed through yaml libraries implemented for the vast majority of programming languages.

5 Results and Discussion

At the time of writing, RRLex currently consists of 4,722 lemmata of various parts-of-speech summarised in table 4. From the breakdown we note that verbs and nouns take the largest share of the total number of lemmas. 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 by nature numerous because they name things. Deverbatives have been excluded so far from RRLex because they are easy to add once all verbs are known. Despite the low number of proper nouns in RRlex, this category of nouns is huge and we plan to add more from the R&R Thesaurus (RRThes2012) after obtaining copyright permission. In R&R, adverbs are a complicated PoS. 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 will be expanded 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 hope 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, we encountered nouns

[5] See appendix I for the full structure

	Class		Individual Particles		Example		Gloss
ID	Numbers	Particles	Singular	Plural	Singular	Plural	Singular(Plural)
1	α	KA_TU	KA	TU	a-ka-syo	o-tu-syo	kife (small knives)
2	β	ZERO_N	n/a	N	n/a	embabazi	n/a (mercy)
3	π	RI_BA	RI	BA	o-ri-kwera	a-ba-rikwera	Caucasian (Caucasians)
4	δ	ZERO_BA	n/a	BA	n/a	a-ba-tuuraine	n/a (neighbours)
5	σ	N_ZERO	N	n/a	n/a	kahembe ka mushwekye	evil spirit (n/a)
6	γ	RU_ZERO	RU	n/a	0-ru-me	n/a	dew (n/a)

Table 3: The Runyankore and Rukiga noun class pairs that were obtained during annotation. They lacked equivalent numeric identifiers as used by the bleek-meinhoff system of numbering

Part-of-Speech	**# of lemmata**
Verbs	2488
Common Nouns	1797
Proper Nouns	30
Determiners	97
Pronouns	59
Adverbs	104
Prepositions	29
Adjectives	80
Conjunctions & Subjunctions	38
Total	4722

Table 4: Table showing the number of entries made per part of speech.

that did not fall under the accepted noun class numbers. In table 3, we give examples of such nouns. We suggest that the noun classes used in the numeral system be expanded as some nominal lexical items cannot fall under the pre-existing numerical system used in literature for Runyankore-Rukiga. Since the notion of adjectives in R&R is very limited as mentioned in (Bamutura et al., 2020), we found it difficult to identify and classify all forms of this PoS. 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 use a computer program to extract particular parts of the lexicon for each language. RRlex 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 meornymy) within the resource. Though we have provided some synonym information in RRLex, 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.

6 Conclusion and Future Work

In this paper, we have created RRLex, a computational lexicon for R&R. It currently consists of 4,722 lemma entries. Since the languages are under-resourced, we found only eleven data sources that could be used for its creation. Of the eleven, only five were actually utilised because of restrictive copyrights attached to the other six. 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 manual methods. Having obtained this resource, we plan to build and evaluate conjugation, lemmatisation, morphological analyser and generator, part-of-speech tagging software for Runyankore and Rukiga that can be used to speed up the process of language resource creation as future work. With these in place, RRLex 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 R&R. For a broader audience, the CLIR system could be augmented with an Automatic Speech Recognition (ASR) module for R&R targeted towards specific domains. Although RRLex 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 R&R and developing an R&R WordNet. In its current form, we have already used it to improve the lexical coverage of the computational resource grammars of R&R.

Acknowledgments

This work was supported by the Sida / BRIGHT Project 317 under the Makerere–Sweden Bilateral Research Programme 2015-2020 and 2020-2022.

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.

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.

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).

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.

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

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.

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.

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.

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.

A Appendices

A.1 Appendix I

```
% YAML 1.2
---
$schema: "http://json-schema.org/draft-07/schema#"
$id: "http://nlp.gemcs.biz/rrlexicon/draft-01/RRLex.yaml"
name: YAML Schema for RRLex
desc: |
  A schema describing the structure of RRLex and
  constraints to typing data.
type : seq
sequence:
  - type: map
    mapping:
      lemma:
        type: str
        required: true
        name: The lemma of a lexical item
        desc: The form of a word after lemmatization
      lemma_id:
        type: int
        required: true
        name: lemma entry identifier
        desc: a uinque identifier for the lemma item
      eng_defn:
        type: seq
        sequence:
          - type: str
        required: true
        name: A definition of the lemma in English
        desc: |
          The main semantic information available in the lexicon.
          The other being the synonyms field.
      pos:
        type: map
        name: A mapping of pos tags at various levels
        mapping:
          first_level:
            type: str
            required: true
            enum:
              - verb
              - noun
              - adjective
              - adverb
          second_level:
            type: str
            required: true
        required: true
# listing continued next page
# listing continued here
      synonyms:
        type: seq
        required: false
        desc: |
          should be optional if the word has no known synonyms
        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
```