Thesis for The Degree of Licentiate of Engineering

Lexical and Grammar Resource Engineering for Runyankore & Rukiga

A Symbolic Approach

David Sabiiti Bamutura

Division of Functional Programming
Department of Computer Science & Engineering
Chalmers University of Technology and Gothenburg University
Gothenburg, Sweden, 2021

Lexical and Grammar Resource Engineering for Runyankore & Rukiga
A Symbolic Approach

David Sabiiti Bamutura

Department of Computer Science & Engineering
Division of Functional Programming
Chalmers University of Technology and Gothenburg University
Gothenburg, Sweden

Cover image: Designed by Andrew Odoch Umahtete of Drew Tete Limited according to the author's ideas. The copyright was transferred to the author.

This thesis has been prepared using LaTeX.
Printed by Chalmers Reproservice,
Gothenburg, Sweden 2021.

“Human knowledge is expressed in language. So computational linguistics is very important.”
- Mark Steedman, ACL Presidential Address (2007)

“If you talk to a man in a language he understands, that goes to his head. If you talk to him in his language, that goes to his heart”
- Nelson Mandela

“Without language one cannot talk to people and understand them; one cannot share their hopes and aspirations, grasp their history, apreciate their poetry, or savour their songs”
- Nelson Mandela

Abstract (English)

Current research in computational linguistics and natural language processing (NLP) requires the existence of language resources. Whereas these resources are available for 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*). 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.

The central focus of this thesis is about enabling the computational analysis and generation of utterances in Ry/Rk. Ry/Rk are two closely related languages spoken by about 3.4 and 2.4 million people respectively. They belong to the Nyoro-Ganda (JE10) language zone of the Great Lakes, Narrow Bantu of the Niger-Congo language family.

The computational processing of these languages is achieved by formalising the grammars of these two languages using Grammatical Framework (GF) and its Resource Grammar Library (RGL). In addition to the grammar, a general-purpose computational lexicon for the two languages is developed. Although we utilise the lexicon to tremendously increase the lexical coverage of the grammars, the lexicon can be used for other NLP tasks.

In this thesis a symbolic / rule-based approach is taken because the lack of adequate languages resources makes the use of data-driven NLP approaches unsuitable for these languages.

Keywords: Language Resources, Bantu Languages, Runyankore, Rukiga, Runyakitara, Grammatical Framework, Resource Grammar Library, Computational lexicon, Computational Grammar, Lexical Resource, Grammar Resource, Grammar Engineering

Abstract (Swedish)

Forskning och utveckling inom datalingvistik och naturlig språkbehandling (NLP) behöver språkresurser. Några språk är resursstarka, och har många olika sorters resurser, men det stora flertalet språk är försummade. De senaste åren har forskare och utvecklare inom NLP börjat inse att språkresurser för försummade språk bör prioriteras mer. Varför? En anledning är att de resursstarka språken kommer från några få språkfamlijer och därför inte kan representera den strukturella mångfalden hos all världens språk.

Denna avhandlings fokus är att möjliggöra automatisk analys och generering av yttranden i Runyankore och Rukiga. Runyankore och Rukiga är två resurssvaga närbesläktade språk som har ca 3,4 respektive 2,4 miljoner talare. Språken tillhör språkzonen Nyoro-Ganda (JE10), och är en del av Great Lakes Bantuspråken, som i sin tur tillhör språkfamiljen Niger-Kongo.

Dessa två språk har implementerats som datorresurser med hjälp av grammatikverktyget Grammatical Framework (GF), och dess resursgrammatikbibliotek (RGL). Förutom grammatiken utvecklar vi också ett datorbaserat lexikon, som vi framför allt använder för att utöka grammatikens lexikaliska täckning, men det kan också användas för andra NLP-uppgifter.

Eftersom språken saknar tillräckliga språkresurser, använder avhandlingen ett symboliskt och regelbaserat tillvägagångssätt. Bristen på språkresurser gör att statistiska och datadrivna NLP-metoder blir oanvändbara.

Nyckelord: språkresurser, bantuspråk, Runyankore, Rukiga, Runyakitara, Grammatical Framework, resursgrammatik, datorbaserat lexikon, datorbaserad grammatik, lexikala resurser, grammatiska resurser

Translated by Assoc. Prof. Peter Ljunglöf

Abstract (Runyankore-Rukiga)

Okucondooza ebikwatiraine n'okweyambisa zaakarimagyezi (computers) omu kushoboorora, okuhandiika n'okugamba endimi (/*computational linguitics*/ nari shi /*Natural Language Processing–NLP*/) nitwetaga ebikwato bihikire. Ebyokweyambisa n'obu birabe biriho aha bw'endimi ezimwe nkyeho, ezirikukira obwingi tizibiine. Abahangu aba NLP nibagira ngu buzima ebyokweyambisa omu kucondooza endimi ezo ezaasigirwe enyima bishemereirwe kutiibwamu amaani. Ahabwenki? Enshonga emwe n'ahabwokuba okurugirira aha biine akakwate n'okukyenga oku orurimi rukushwana (/*language typology*/ omu Rugyereza), endimi ezaakozirweho gye tizirikubaasa kweyambisibwa nk'omusingye gw'okushoboorora ezaasigirwe enyima.

Ekigyendererwa kikuru ky'okucondooza kwangye n'okugyezaho kutaho oburyo bwa zaakarimagyezi kubasa kwega kandi zikabasa kushoma, kukyega na emigambire y'Orunyankore n'Orukiga, ororikweyambisibwa abantu barikuhika obukeikuru bushatu n'emitwaaro makumi ana (3.4 miryoni), n'Orukiga ororikweyambisibwa abantu barikuhika obukeikuru bubiri n'emitwaaro makumi ana (2.4 miryoni). Endimi ezi zombiri eziri omuri ezo ezaasigirwe enyima ziri omu ruganda rw'orurimi ororikumanywa nka Nyoro-Ganda (JE10), ororikushangwa omu Kyanga ky'Enyanja Empango (/*Great Lakes Region*/). Oruganda nirukomooka aha kika ky'endimi ekikumanywa nka Narrow Bantu ekya Niger-Congo.

Okweyambisa zaakarimagyezi omu ndimi ezi zombiri nikihikirizibwa omu kubaga orukanga rw'endimi oku zeemi (Grammatical Framework) n'okutebekanisa n'okwetegyereza gye ei twakubaasa kwiha okututurakyenge gye oku zeemi (Resource Grammar Library). Okwongyerera ahari ebi, hashemereire kubaho enshoboorora ya kaarimagyezi y'endimi ezi. N'obuturaabe nitweyambisa enshoboorora kukanyisa okwetegyereza oku endimi ziri, enshoboorora egi neebaasa kweyambisibwa omu mirimo endijo ya NLP.

Omu kucondooza oku, tweyambisize enkora y'obumanyiso n'ebiragiro (/*symbolic or rule-based approach*/). Ahabw'okushanga hatariho eby'okweyambisa birikumara omu ndimi ezi titurikubaasa kweyambisa enkora ensya za NLP ezi ba keta /*data-driven approaches*/ omu Rugyereza.

Translated by Mr. Tom Namara
with minor edits by
Prof. Peter Kanyandago and David Sabiiti Bamutura

Acknowledgment

In a very special way, I would like to extend my sincere gratitude to my main supervisor: Assoc. Prof. Peter Ljunglöf (Gothenburg University and Chalmers University of Technology) for his guidance and support both academically and emotionally.

Special thanks also go to my co-supervisor: Dr. Peter Nabende (Makerere University); and Examiner, Prof. Aarne Ranta (Gothenburg University and Chalmers University of Technology) who tirelessly addressed my fears and doubts about the eventual impact of this research study. Furthermore, I want to thank Dr. Ng'ang'a Wanjiku for her willingness to become a discussion reader and "opponent" for my Licentiate seminar.

To the principal investigators of the SIDA / BRIGHT Project 317; Prof. Michel Chaudron and Asoc. Prof. Engineer Bainomugisha, I thank them for their financial and moral support.

I want to thank Prof. Fr. Peter Kanyandago who planted the seed in me to work on a topic that applied computer science to the indigenous languages of Uganda during our candid discussion about Africa and Pan-Africanism back in 2007 at Uganda Martyrs University.

In addition, I express my gratitude to Prof. Richard Jones, my former lecturer and master's thesis advisor at University of Kent (UoK) at Canterbury, United Kingdom. He not only introduced me to programming language research but also gave me his blessing to switch to computational linguistics. While at UoK, little did I know that the experience of learning Occam-pi — an "unconventional" and research programming language used for teaching Concurrency Design and Practice by Prof. Peter Welch — would be helpful in getting me acclimatised to the functional programming paradigm that is popular at the Functional Programming Division of the Department of Computer Science and Engineering at Chalmers.

I am also very grateful to my office mates at Chalmers: Inari Listenmaa, Prasanth Kolachina and Herbert Lange for the interesting and ingenious discussions about Grammatical Framework, computational linguistics and linguistics in general. Having had no formal linguistics background, it would have been "mission impossible" to set my foot into the field without them. Special thanks to my office mates and friends back home at Mbarara University of Science & Technology. Dr. Evarist Nabaasa, Dr. Simon Kawuma, Dr. Fred Kaggwa, Ms. Josephine Ayebare, Madam Kate Imanirampa and Ms. Florence Mbabazi for their continuous encouragement.

In a very special way, I would like to thank my family who without their patience, “guidance” and emotional support, this work would not have been possible. I wish to thank my dear wife, Doreck Nduhukire-Bamutura who stood by me during the lonely and tough times, and my sons Philip and Philemon Bamutura who gave me the reason to continue “grinding” during those challenging times. Many thanks to my sister, Dr. Diana Sabiiti Busingye and my dear parents: Hon. Eng. Denis Sabiiti and Mrs. Tophers Tugumisirize-Sabiiti for their relentless words of encouragement and support.

Lastly, I would like to thank my colleagues on the SIDA-BRIGHT project specifically; Swaib Dragule, Michael Kizito, Adones Rukundo (for the beer, wine and whisky parties), Grace Kamulegeya, Rashida Kasauli, Grace Kobusingye and Hawa Nyende. I cannot list everybody by name but to everybody that helped me, I shall forever be grateful.

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

List of Publications

Included publications

This thesis is based on the following publications:

[A] D. Bamutura, P. Ljunglöf, P. Nabende, 2020 "Towards computational resource grammars for Runyankore and Rukiga." *In Proceedings of The* 12^{th} *Language Resources and Evaluation Conference, pages 2846–2854, Marseille, France.* European Language Resources Association.

[B] D. S. Bamutura, 2021 "Ry/Rk-Lex: A computational lexicon for Runyankore and Rukiga languages." *Accepted to the Northern European Association for Language Technology post-proceeding series of the Swedish Language Technology Conference (SLTC 2020)*

Other publications

The following publication was published during my PhD studies. However, it is not appended to this thesis, due to contents overlapping with those of Paper A.

[C] D. Bamutura, P. Ljunglöf, 2019. "Towards a resource grammar for Runyankore and Rukiga." *In WiNLP 2019, the 3rd Workshop on Widening NLP, Florence, Italy, 28th July 2019.*

Statement of Contributions

In Paper A, the search for information and knowledge about the descriptive grammar (morphology and syntax) of the object languages (Runyankore and Rukiga) was done solely by the author. However, the modelling and formalisation of a minuscule grammar of the two languages using the Grammatical Framework (GF) was done by the author in consultation with others. The rest of the standard GF Resource Grammar Library for the two languages — which contributes the largest part to the manuscript — was modelled, formalised and implemented by the author. Though the final manuscript was jointly written, the author's contribution was 75%.

In Paper B, the author's contribution was to search for all possible language data sources for the semi-automatic creation of computational lexica for the object languages — Runyankore and Rukiga. From the fourteen sources found, the author used six of them by performing text extraction, tokenisation, lemmatisation, part of speech (POS) tagging and further annotation of each lemma with additional information. Research Assistants were used later in the project to speed up the tedious and time-consuming aspects of the work i.e. copy-typing hard copy versions of texts. The design of the persistence structure for the lexica and the writing of the manuscript were solely done by the author with the exception of edits and recommendations from the supervisors.

Thesis organisation

This thesis is structured into three parts:

Part I: Introduction and Overview contains two chapters; introduction and background. In the introduction: the research area; problem statement; motivation of the study; research objectives, and the associated research questions are presented; and a brief statement of results is made. In the background we provide a summary of the literature required to understand and explain the ideas in the papers on which this thesis is based. Therefore, the background covers the: genealogy, morphology and syntax of the object languages — Runyankore and Rukiga (Ry/Rk); grammar formalisms and Grammatical Framework (GF) in particular; and related workk on language resources for carrying out computational linguistics and / or natural language processing (NLP) for under-resourced languages.

Part II: Publications contains two chapters; 3 and 4 that are reproductions of two papers: *"Towards computational resource grammars for Runyankore and Rukiga languages"*; and *"Ry/Rk-Lex: A Computation lexicon for Runyankore and Rukiga languages"*. These methodological research papers describe how computational resource grammars and lexical resources for the object languages were developed.

Part III: Discussion, Conclusion and Future Work contains chapters 5 and 6. Chapter 5 provides brief summaries of the research carried out in both papers plus additional research work that was done after their publication. Chapter six concludes the thesis with a general discussion, possible future research directions and a final conclusion.

Contents

List of Figures

List of Tables

Part I

Introduction and Overview

Chapter 1

Introduction

> *"... And since language is our most natural and most versatile means of communication, linguistically competent computers would greatly facilitate our interaction with machines and software of all sorts, and put at our fingertips, in ways that truly meet our needs, the vast textual and other resources of the internet."*
>
> *- Lenhart Schubert (2020)*

Languages maybe classified as natural or artificial. The term natural language usually refers to those languages that come into existence organically. In contrast, an artificial language is a result of purposeful creation by beings. Interestingly, the distinction between natural and artificial language lies on a "continuum" and as we move along that continuum, languages become increasingly restrictive — getting more artificial and formal. For example, spoken language is considered by "some linguists" as the only natural language. Although Text and Braille are instances of human languages, they can be understood as an encoding of speech with additional restrictions. In that context, text and braile are more artificial and formal than speech.

At the extreme "end of the spectrum" we find purely artificial languages such as programming languages (in Computer Science) and Mathematics in and of itself. Formal languages are rigorous mathematical and or computational models created by humans for the sole purpose of modelling theories about other languages in order to test, verify and prove properties about them. The formal languages that are used to account for the theories of natural languages are called linguistic formalisms. However, in this thesis, unless otherwise stated natural language — or simply language — shall refer to the written form of communication used by human-beings commonly referred to as text.

Language and its communicative goal, is undoubtedly indispensable for the survival of all living things especially human beings. Even when the core human senses of sight, speech, smell, touch, hearing and feeling get impaired, human beings have always invented other modes of communication such as written or sign language. Languages provide humans with the ability to express themselves and understand each other thus fulfilling the communicative goal. The existence of many languages is testimony to the creative abilities of the human mind. There are about 7000 distinct natural languages in the world and each contributes to the rich diversity of features in languages.

Linguistics, being the systematic study and description of natural language, has both contributed and benefited from other fields of Psychology, Mathematics, Computer Science to mention but a few. Computational linguistics studies the structure of natural language from a formal, mathematical and computational perspective while covering all subfields of traditional linguistic research. Natural language processing (NLP) is a sub-field of Artificial Intelligence that studies human languages with the objective of simulating processes related to the human linguistic faculty. These two research fields use a plethora of methods / approaches; symbolic / rule-based, data-driven (statistical), machine and deep learning at the extreme. Deep learning aims to build end-to-end NLP systems that largely do not require any form of linguistic intuition through annotation in order to perform NLP tasks. As a result, current research in these fields requires the existence of language resources (text or speech data). Whereas these resources are available for a few "politically advantaged" and well-resourced languages, 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. 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).

The focus of this thesis is an attempt at the formalisation of the grammar and lexicon of Runyankore and Rukiga (henceforce referred to as *Ry/Rk*[1]). Specifically we aim at enabling the computational processing for these languages particularly at syntax level but delving into morphology and semantics when it is unavoidable. The two languages are under-resourced Bantu languages spoken in south-western Uganda. We use Grammatical Framework(GF) (Ranta, 2004, 2009a, 2011a), a symbolic approach, as a means to achieving this task for several reasons: (1) the languages are under-resourced so data-driven approaches are ineffective, (2) being multilingual, GF can be used to develop a number of end-user applications and (3) by leveraging on work done previously by; Kolachina and Ranta (2016), Ranta and Kolachina (2017), Ranta et al. (2017), Kolachina and Ranta (2019), and Ranta et al. (2020), it can be used to bootstrap the development of large enough language data that is amenable to data-driven approaches.

Although our original motivation was the development of a Computer Assisted Language Learning (CALL) application for Ry/Rk, we have chosen the path to continue the development of linguistic resources that can be used for not only CALL but also for empirical evaluation and enable the use of data-driven methods such as development of neural parsers for the languages.

The rest of this chapter is structured as follows: Section 1.1 describes the problem; Section 1.2 provides a motivation for the study: Section 1.3 and its subsections provide the objectives of the research and the associated research questions. The chapter ends with a statement of results in Section 1.4.

[1] The acronym is borrowed from (Byakutaaga et al., 2020) where it is convincingly argued that these two languages are treated as dialects along with Runyoro and Rutoro (Rn/Rt) to form a new language: Runyakitara

1.1 Problem Statement

As already mentioned previously, current research in computational linguistics and NLP requires the existence of language resources. Whereas these resources are available for a few languages, there are many languages that have been neglected. Among the neglected languages and / or under-resourced languages are Ry/Rk notwithstanding the fact that they are spoken by a sizeable population of 3.4 and 2.4 million people (Simons and Fennig, 2018) respectively. Despite the initial exposure to learning Ry/Rk in the first three years of primary school, English becomes the official language of instruction and examination from the fourth year on, severely limiting the continued study of Ry/Rk to higher levels of proficiency. It is also worth to note that although dictionaries, grammar books and an orthography for Ry/Rk exist, Ry/Rk just like other native languages in Uganda largely remain spoken as opposed to written even among those literate in English. Only a dismal few study the language to a level sufficient to achieve proficiency in writing. This results in lack of continuity in learning the grammar of the language. It also explains the Ry/Rk's nearly zero presence on the web hence the lack of any computational language resources for the languages. Because Ry/Rk are highly under-resourced, it is important to take steps in building language resources, encouraging writing in these languages and their continued preservation.

1.2 Research Motivation

In the current era of machine and deep learning, the importance of language resources – both labelled and unlabelled data sets (corpora, treebanks, lexical knowledge-bases) – for all languages cannot be understated. Because Ry/Rk are under-resourced, our motivation for this study is two-fold. In the short term, we seek to enable the computational processing of Ry/Rk using a symbolic approach for the simple reason of lack of language resources. Achieving this comes enables the development of domain-limited applications such as multilingual document authoring (Dymetman et al., 2000), low-coverage multilingual translation (Ranta et al., 2010), domain-specific dialogue systems such as music players (Perera and Ranta, 2007) and Computer-Assisted Language Learning (CALL) (Lange, 2018; Lange and Ljunglöf, 2018b). Another use case is localisation through multilingual dissemination of information especially in multilingual societies Our second motivation for this study is to lay the foundation for making it possible to utilise state of the art statistical learning methods for performing CL and NLP tasks at scale and the development of broad coverage end user applications. Although the former approach yields domain-limited applications, the time to deliver and deploy a working, reliable software product in the market is significantly shorter than the latter approach. Nonetheless, advancing research in NLP using both approaches is worthwhile.

1.3 Research Objectives

The focus of this study is to design and implement computational grammar resources for Ry/Rk by formalising their descriptive grammars as Resource

Grammar Libraries (RGLs) within the Grammatical Framework (GF). We employ GF because it is a rule-based grammar formalism suitable for under-resourced languages.

1.3.1 Specific Research Objectives

S.1 To computationally model and implement the descriptive grammars of Ry/Rk as Grammatical Framework Resource Grammar Libraries (GF-RGL).

S.2 To build general-purpose computational lexical resources for Ry/Rk.

1.3.2 Research Questions

RQ.1 How can we build a computational grammar from dictionaries, grammar books and implicit knowledge of language speakers?

RQ.2 How can we create general-purpose computational lexica for Ry/Rk?

(a) What are the existing linguistic data sources that can be used for the development of computational lexica for Ry/Rk?

(b) Out of the sources identified in *RQ.2* (a), which sources are suitable for creating computational lexica for Ry/Rk?

(c) How can computational lexica for Ry/Rk be modelled or structured in a simple, flexible and extensible manner?

1.4 Results

Our attempt at addressing *RQ1* is detailed in Paper A that we reproduce in Chapter 3. In that paper, we chose GF, a multilingual grammatical formalism out of many and used its domain-specific programming language features; parameters, records, tables and pattern matching to model and formalise significant parts of the morphology and syntax of Ry/Rk. Because the grammatical tense and aspect system of Ry/Rk is very different from that of English and many Indo-European languages, we established a mapping between the tense and aspect system used by Standard GF-RGL in order to maintain the multilingual capabilities of GF. The complex nominal and verbal morphology for the two languages was sucessfully modelled despite the complexity introduced by the large noun class system in the languages, its impact on concordial agrrement with other POS such as verbs, adverbial expressions, nominal qualificatives, determiners and numerals. Before modelling, the author used his intuitive knowledge of the spoken language, consulted grammar books, dictionaries and also asked experts on the languages for help when stuck. However, after paper A, the GF-RGL for the two languages has been extended to cover all the six tenses and seven aspects as extensions to the standard GF-RGL.

For *RQ2*, after carrying out a manual search both on the web and visiting bookshops and libraries for possible linguistic data sources that could be used for computational lexicon construction, we found fouteen data sources (see Chapter 4). Out of those, we used five fully without any restrictions. Another data source, Orumuri newspaper, was also fully utilised despite restricted by copyright but we decided that the corpus so obtained shall never be released for commercial gain. However random sentences can be released and used for non-commercial educational and research purposes.

Text extraction (using both copy-typing for hard-copy sources and web-scraping for online digital text), text cleaning, tokenisation, lemmatisation, and anootation tasks asuch as pos tagging, attaching definition glosses for English and synonyms were done. All other sources were used as references since they are restricted by copyright. We used YAML to store the lexicon according to a schema we designed to preserve its structure and allow easy sharing of data whose structure and content can be validated before use by machines and programs.

Currently we have 12,500 lexical items. (Note that paper B reported 9,400 but we continued our lexical extraction even after submitting it for review). We have used the general lexicon developed to tremedously increase the lexical coverage (from 167 lexical items to 12, 500) of the resource grammar developed under *RQ1*.

Chapter 2

Background

"Language is a system of signs that express ideas, and is therefore comparable to a system of writing, the alphabet of deaf-mutes, symbolic rites, polite formulas, military signals, etc. But it is the most important of all these systems"
- Ferdinard de Saussure (1916)

2.1 Bantu Languages

Since Ry/Rk are Bantu Languages, it is prudent to give a brief overview of languages with respect to genealogy, typology and the socio-political issues afecting their continued development. The Bantu languages belong to the Benue-Congo branch of the Niger-Congo Language family (Simons and Fennig, 2018).This family spans the area from Dakar, Senegal, eastwards along a line through Western, Central, Eastern and Southern Africa. In the Benue-Congo branch, they are placed under the Bantoid group which is divided into a northern and a southern subgroup. Out of 11 further divisions among the southern subgroup, the Bantu is the largest division consisting of about 500 languages (Hinnebusch et al., 1981).

Bantu Languages have been studied since the 19th Century by several linguists such as; Bleek's treatment of the phonology and nominal morphology of South African languages (Bleek, 1862, 1869); Koelle's lexicon-based comparative studies on Niger-Congo languages (Koelle, 1854); and Meinhof et al.'s work on characterising the noun class system of Bantu languages (Meinhof et al., 1915). Joseph H. Greenberg and Diedrich Herman Westernam refined Meinhof et al.'s comparative classification scheme in addition to extending his work. Guthrie (1948) is credited for his geographically motivated classification of Bantu languages by subdividing them into several zones and his attempt at a comparative study of Bantu languages (Malcom, 1967). Currently, Maho's geographical classification is the most recent and widely accepted. Malcom worked alongside Meeussen (1967) though the latter specialised on the languages of Belgian Congo and Rwanda, and Uganda. The two worked on the reconstruction of a Proto-Bantu language (common ancestor of Bantu languages) using both lexical (Bostoen and Bastin, 2016; Meeussen, 1980) and grammar (Meeussen, 1967) approaches. Other more recent and notable Bantu language scholars include Hinnebusch, Nurse, and Mould who covered Bantu language classification in East Africa (Hinnebusch et al., 1981).

Typologically, the Bantu languages are agglutinating in nature, with a tonal system that varies from mild to high. Tone may be marked or unmarked in written text depending on the orthography adopted for the language. Each noun in these languages inherently belongs to a particular noun class and the number of possible noun classes in a language can be as large as 20. The charcateristic noun class system was probably first identified by Meinhof et al. (1915) and refined by others such as Meeussen (1967). It notably dictates a system of cordial agreement acting both within and across various phrasal categories. The languages have largely Subject-Verb-Object word order with a Consonant-Vowel (CV) structure in their word morphologies. Orthographically, the morphology within the verbal unit may be: conjunctive e.g. Ry/Rk, isiZulu (Taljard and Bosch, 2006); disjunctive e.g. Northern Sotho (Taljard and Bosch, 2006); or sometimes a hybrid of the two e.g. Setswana (Pretorius et al., 2009) is used. Otherwise the verbal template remains more or less the same.

We now turn to the problem of lack of a critical number of people reading and writing in their native languages. Among the Bantu, language use is highly skewed towards oral or verbal communication at the expense of the written word. In heavily multilingual nations especially East and Central Africa, mother-tongue literacy is to a great extent about speech with only listening and oratory skills — mainly acquired from homes and social communities — at the expense of writing, reading and comprehension. The lack of native language writing, reading and comprehension skills severely affects the ability of the Bantu people to engage in higher-order tasks such as acquiring new knowledge through reading; and expressing knowledge through the written word.

This sad state of affairs is not only exclusive to Bantu speaking Africa but also appears in other areas of the world. Due to the delay in the development of indigenous languages (i.e. documentation of orthographies; descriptive grammars; and the writing of dictionaries) and the elevation of colonial languages as vehicles of learning against the native languages , there has been little effort by native speakers that are ”orally proficient” in their native languages to develop written resources in these languages. This explains the lack of computational resources and a low prescence on the web for the largest number of larguages across the world.

Another factor that exacerbates the situation is the fact that African countries were created by European colonialists who divided them without consideration of the many languages spoken by different communities. It is therefore not uncommon to find an African nation comprised of 2-40 languages. It therefore becomes difficult to choose one language over another as the official language hence the need of an external language. Countries south of Tanzania and Congo that have considerably fewer languages have not escaped this problem too. South Africa has taken a different stance to constitutionally recognise all 11 languages as national languages and encourages their education. The government of South Africa has also invested a lot money into general linguistic and computational linguitic research of thier languages. Kenya and Uganda have policies recognising the importance of mother-tongues in their education systems for the lower primary while Tanzania has silenced all mother-tongue languages in place of Swahili as a unifying language and English a global language (Lisanza, 2015).

The observations made about the usage patterns of Bantu languages discussed above negatively affects attempts by researchers in language technology to advance research for these languages simply because basic text language resources are negligibly small. Speech data is also difficult to obtain without violation of the now ubiquitous laws on copyrights and right to privacy. There are other disadvantages associated with this trend but since the subject of this thesis is focused on solving the computational aspects of such languages we do not delve into such matters.

2.2 Runyankore and Rukiga (Ry/Rk)

Runyankore and Rukiga are languages spoken in South-western Uganda by about $3,420,000$ and $2,390,000$ people (Simons and Fennig, 2018) respectively. Their ISO 6390-3 codes are *nyn* for Runyankore and *cgg* for Rukiga. Their genealogical trees are shown in Figures 2.2 and 2.3. They belong to the *JE10* zone (Maho, 2009) of the Great Lakes, Narrow Bantu of Niger-Congo language family. The two peoples hail from and / or live in the regions of Ankole and Kigezi — both located in south western Uganda (See Figure 2.1), East Africa. Just like any other Bantu language from the JE10 Nyoro-Ganda group — consisting of; Runyankore, Rukiga, Runyoro, Rutoro, Luganda, Lusonga, Lugwere, Runyala among others — of the Great Lakes Bantu, Ry/Rk are *mildly tonal* (Muzale, 1998), *highly agglutinating* (see Examlpe (2.1) below) with a *large Noun Class System* of 17-20 classes (Byamugisha et al., 2016; Katushemererwe and Hanneforth, 2010b). They exhibit high incidencies of *phonological conditioning* (Katushemererwe et al., 2020). These characteristics make the computational analysis and generation of these languages more complex to deal with.

Despite the Ethnologue classifying these languages as distinct, Byakutaaga et al. (2020) consinder them as dialetcs. The fact that they share of the same dictionaries (Mpairwe and Kahangi, 2013a; Taylor and Yusuf, 2009), grammar books (Morris and Kirwan, 1972; Mpairwe and Kahangi, 2013b; Taylor, 1985), orthographies (Karwemera, 1995; Taylor, 2008) and when we also take into account the high level of lexical similarity suggests that the claim by Byakutaaga et al.'s about the languages being dialects of each other is a strong one.

Historically, they have always been considered dialects. Before the 1950s, the two languages were considered as part of one bigger language called Runyoro that also included Runyoro and Rutoro as the other two dialects and had one common Bible (Turyamwomwe, 2011). At a conference called in 1946 with representatives from the four dialects (i.e. Runyankore, Rukiga, Rutooro and Runyoro) to agree on a single orthography for them, the representatives for the Banyankore and Bakiga communities respectfully rejected the idea of using 'Runyoro' and its orthography for their languages. One of the main reasons for the rejection was the desire to preserve their language and cultural heritage. Later in 1954 a standard orthography for Runyankore-Rukiga (Taylor, 2008) was adopted at a separate conference in Mbarara.

It should also be noted that the separation of Runyankore and Rukiga from

Runyoro and Rutoro can be attributed to the high lexical similarity between Runyankore and Rukiga i.e. 84%–94% as compared to 78%–93% of Runyoro and Rutoro (Lewis et al., 2018; Turyamwomwe, 2011). Currently, the four languages are collectively referred to as the Runyakitara language.

(2.1) Runyankore

ti-n-ka-mu-reeb-a-ho-ga

not-PNEG-I-1SG.SUBJ.CL1-had-PASTRM.PERF-him/her-3SG.OBJ.CL2 see-RAD-fvinf-never-LOC-ever-EMPHATIC

not-I-had-him/her-see-FV-never-ever

‘I had never ever seen him / her.’

2.3 Morphology and Syntax of Runyankore and Rukiga

2.3.1 Nominal Morphology

The morphological structure of nouns in Ry/Rk depicted in Figure 2.4 at the most basic level consists of two parts, a *class prefix* and a *noun stem.* The class prefix is further divided into an *Initial Vowel* (IV) and a *noun class particle* (NCP) (Mpairwe and Kahangi, 2013b) also known as a *Class Prefix* (CP). The initial vowels can be any of */a/*, */e/* and */i/* or none which we label as "∅" in our glosses. The NCPs / CPs give an indication or clue as to which noun class the noun belongs as well as its grammatical number. The *noun stem* usually bears the bulk of the semantic meaning of the noun. The number of noun classes varies from author to author but twenty noun classes for Runyankitara (an amalgamation of Runyankore, Rukiga, Rutoro and Runyoro) are suggested in (Katushemererwe and Hanneforth, 2010b) and they use a numbered system of classification originally devised in the 19^{th} Century (called the Bleek-Meinhoff system). The justification for the numbered system as suggested by Maho (2009) was to easily map noun classes across different Bantu languages based on their etymology but considering that different languages have different number of noun classes, the argument falls short. It is perhaps only useful for comparative linguistics.

For Ry/Rk, Mpairwe and Kahangi (2013a,b) make use of NCPs in place of noun classes. The NCPs are:
/-ba-/, */-bi-/*, */-bu-/*, */-ga-/*, */-gu-/*, */-ha-/*, */-i-/*, */-ka-/*, */-ki-/*, */-ku-/*, */-ma-/*, */-mi-/*, */-mu-/*, */-n-/*, */-ri-/*, */-ru-/* and */-tu-/*,
to which we add */baa-/* as an extra used when referring to a group of people with a familial relationship (Katushemererwe and Hanneforth, 2010b) on page 38. Apart from the locative particles */-ha-/*, */-mu-/* and */-ku-/* , all other particles can be arranged in singular-plural pairs for nouns with singular and plural forms. We generalise such a pairing using the notation [Ψ_Ω] where Ψ and Ω are noun class particles chosen from the sets:
$S = \{BU, GU, HA, I, KA, KI, KU, MU, N, RI, RU\}$ of singular and
$P = \{MA, GA, MA, BU, TU, BI, BA, MI, N, BU, BA, BAA\}$ of plural noun class particles respectively. We use the upper case for the NCPs to fulfil the syntactic requirements for parameters in GF as mentioned in (Ranta, 2011b) and discussed briefly in Section 2.5.

We borrow the use of the number ZERO (0) from Mpairwe and Kahangi (2013a) in their Runyankore-Rukiga dictionary to denote absence of either singularity or plurality in order to maintain the pairing for such nouns. Hence the pairs [Ψ _ ZERO], [ZERO _ Ω] and [ZERO_ ZERO] which represent nouns that are always singular, plural and those that collectively neither have an initial vowel nor noun class particle respectively as depicted in table 2.1. We chose to use the noun class particles (class prefixes) over noun classes because they provide a more fine-grained classification of nouns according to both gender and agreement concords that should be used with other parts of speech of Ry/Rk. These are conveniently and explicitly mentioned for each lexical entry for nouns and other "special" parts of speech. By special parts of speech we mean those that have no direct equivalent to those used for Indo-European

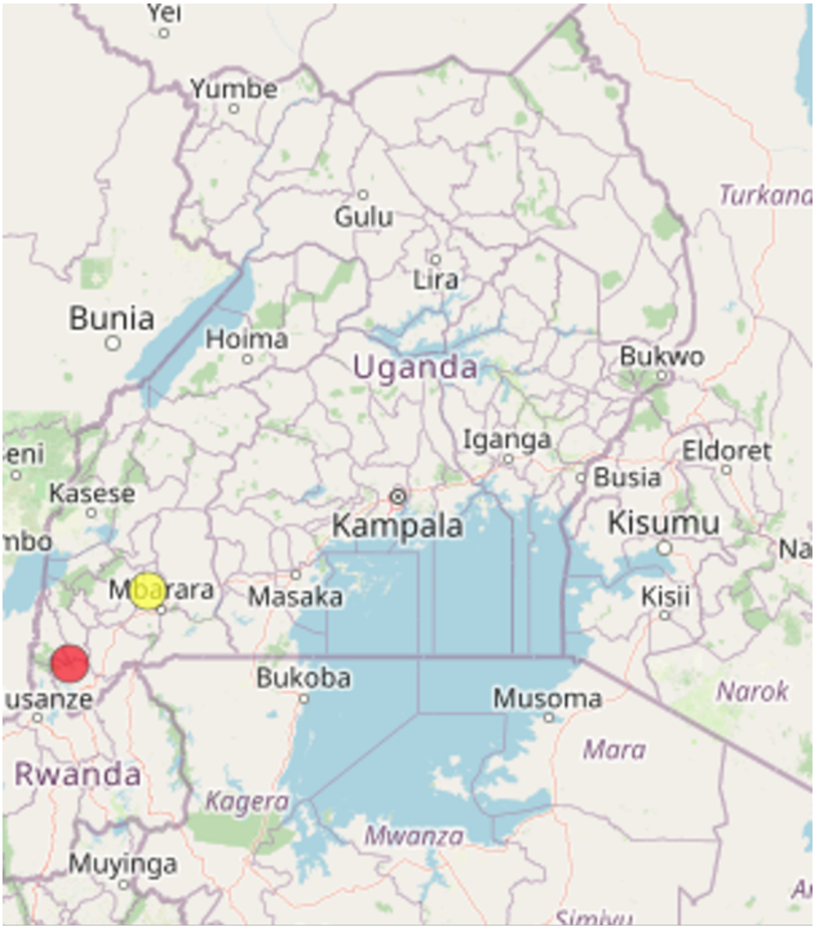
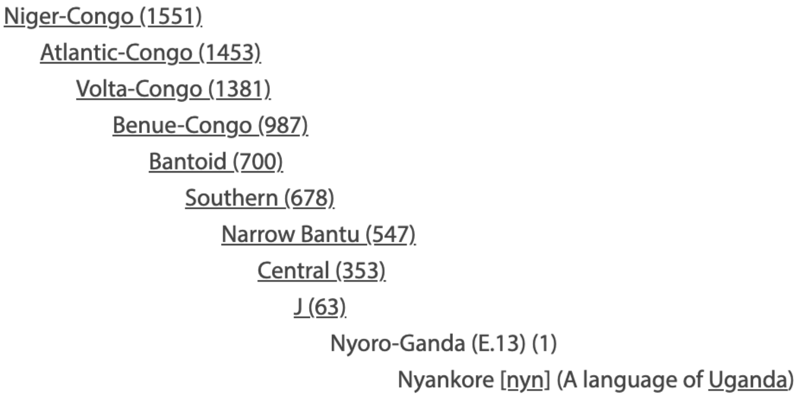
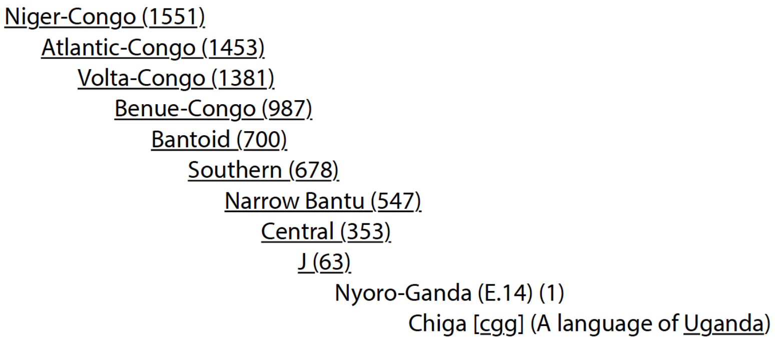


Figure 2.1: Places where Runyankore (yellow) and Rukiga (red) are predominatly used on the map of Uganda. The map was obtained from Glottolog at: https://glottolog.org/resource/languoid/id/nkor1241.bigmap.html#6/1.077/31.146



- Niger-Congo (1551)
  - Atlantic-Congo (1453)
    - Volta-Congo (1381)
      - Benue-Congo (987)
        - Bantoid (700)
          - Southern (678)
            - Narrow Bantu (547)
              - Central (353)
                - J (63)
                  - Nyoro-Ganda (E.13) (1)
                    - Nyankore [nyn] (A language of Uganda)

Figure 2.2: Collapsed genealogical tree for Runyankore obtained from Ethnologue at `https://www.ethnologue.com/subgroups/nyoro-ganda-e13`.



- Niger-Congo (1551)
  - Atlantic-Congo (1453)
    - Volta-Congo (1381)
      - Benue-Congo (987)
        - Bantoid (700)
          - Southern (678)
            - Narrow Bantu (547)
              - Central (353)
                - J (63)
                  - Nyoro-Ganda (E.14) (1)
                    - Chiga [cgg] (A language of Uganda)

Figure 2.3: Collapsed genealogical tree for Rukiga obtained from Ethnologue at `https://www.ethnologue.com/subgroups/nyoro-ganda-e14`

languages and are used in the dictionary by Mpairwe and Kahangi (2013a). The dictionary further provides a comprehensive table of Concords for the affixes required for denclension of various parts of speech that depend on the NCP. For a computational linguist implementing a computational grammar, such information is important and simplifies their work.

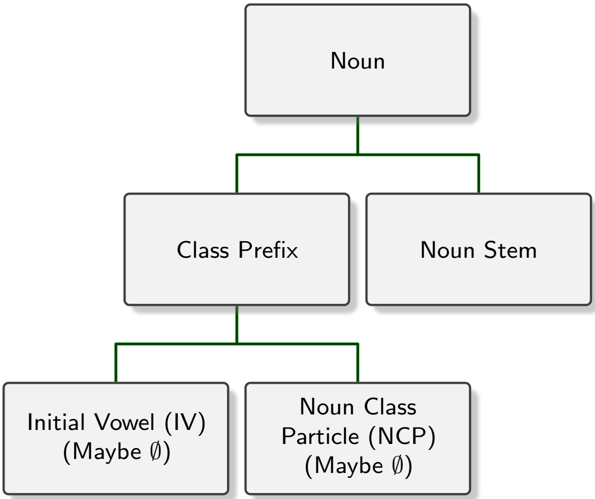

Noun
Class Prefix
Noun Stem
Initial Vowel (IV) (Maybe ∅)
Noun Class Particle (NCP) (Maybe ∅)

Figure 2.4: Structure of a Noun in Ry/Rk

2.3.2 Verbal Morphology

In Meeussen's original construction, the Bantu verbal unit consists of a *pre-stem* and *stem* as depicted in Figure 2.5 below. The stem is further divided into a *base* and *final vowel (FV)* as shown in Figure 2.7. The base is also divided into a *radical (Rad)* and *extensions* (see Figure 2.6). Further subdivisions in each of these parts results into 11 slots (Katushemererwe and Hanneforth, 2010a; Turyamwomwe, 2011), each with a set of morphemes that may appear in a particular slot for a particular purpose such as primary or secondary negative polarity (Pneg / Sneg), subject (S), object (O), tense, aspect and other markers. Figure 2.8 is an attempt as depicting the full verbal unit in one diagram and all the slots within the template of the verb.

	NC	NCP	Individual Particles		Example		Gloss
ID	Numbers	Particles	Singular	Plural	Singular	Plural	Singular(Plural)
1	1_2	MU_BA	MU	BA	o-mu-shaija	a-ba-shaija	man (men)
2	1a	MU_ZERO	MU	n/a	o-mu-hangi	n/a	creator (n/a)
3	1b/2b	ZERO_BAA	n/a	BAA	swhento	baa-shwento	Uncle(s)
4	2a	ZERO_BA	n/a	n/a	n/a	ba-ryakamwe	n/a (inner circle / group)
5	3_4	MU_MI	MU	MI	o-mu-ti	e-mi-ti	tree(s)
6	3a	MU_ZERO	MU	n/a	o-mwisyo	n/a	breath (n/a)
7	4a	ZERO_MI	n/a	MI	n/a	e-mi-gyendere	n/a (way of walking)
8	5_6	RI_MA	RI	MA	e-ri-sho	a-ma-isho	eye(s)
9	5a	I_MA	I	MA	e-i-teeka	a-ma-teeka	law(s)
10	5b	I_ZERO	I	n/a	e-i-tétsi	n/a	pampering(n/a)
11	6a	ZERO_MA	n/a	MA	n/a	a-ma-te	milk (milk)
12	7_8	KI_BI	KI	BI	e-ki-ti	e-bi-ti	stick (stick)
13	7	KI_ZERO	KI	n/a	e-ki-niga	n/a	anger (n/a)
14	8	ZERO_BI	n/a	BI	n/a	e-bi-bembe	(n/a) leprosy
15	9_10	N_N	N	N	e-n-te	e-n-te	cow(s)
16	9	N_N	n/a	n/a	e-bahaasa	e-bahaasa	envelope(s)
17	10	ZERO_ZERO	n/a	n/a	bwîno	bwîno	ink (ink)
18	11_10	RU_N	RU	N	O-ru-shózi	e-n-shózi	mountain(s)
19	12_14	KA_BU	KA	BU	a-ká-bunza	o-bu-bunza	question mark(s)
20	12	KA_ZERO	KA	n/a	a-ka-bi	n/a	danger (n/a)
21	14	ZERO_BU	n/a	BU	n/a	o-bu-cécezi	n/a(being humble)
22	13	ZERO_TU	n/a	TU	n/a	o-tu-ro	n/a (sleep)
23	15_6	KU_MA	KU	MA	o-ku-guru	a-ma-guru	leg(s)
24	16	HA_ZERO	HA	n/a	a-ha-kaanyima(*)	n/a	behind the house (n/a)
25	17	KU_ZERO	KU	n/a	o-ku-z/'imu	n/a	Underground (n/a)
26	18	MU_ZERO	MU	n/a	o-mu-nda	n/a	in the stomach (n/a)
27	20_21	GU_GA	GU	GA	o-gu-kazi	a-ga-kazi	*bad* woman (women)
28	11_14	RU_BU	RU	BU	o-ruro	o-bu-ro	one millet grain (many)
29	14_6	BU_MA	BU	MA	o-bu-ta	a-ma-ta	bow(s)
30	β	ZERO_N	n/a	N	n/a	embabazi	mercy (mercies)
31	σ	N_ZERO	N	n/a	enzingu	n/a	vengeance (n/a)
32	γ	RU_ZERO	RU	n/a	o-ru-me	n/a	dew (n/a)
33	δ	RI_ZERO	RI	n/a	e-ri-ana (eryana)	n/a	childishness (n/a)

Table 2.1: Table showing the Runyankore and Rukiga noun class (NC) system and noun class particles (NCP) derived from several sources (Katushemererwe and Hanneforth, 2010b) and (Mpairwe and Kahangi, 2013a,b)). Examples of lexical items in both singular and plural are provided. However, the labels used from ID 30 to 33 under Numbers are greek-letters because we failed to place them under the existing system.

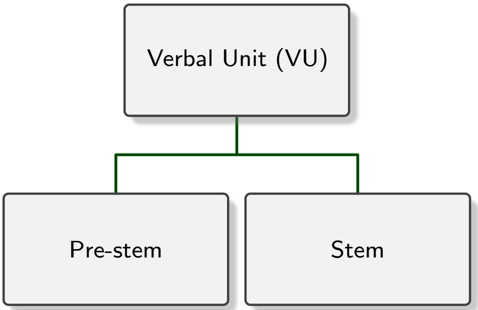


Figure 2.5: The structure of Bantu verbal unit at depth 1

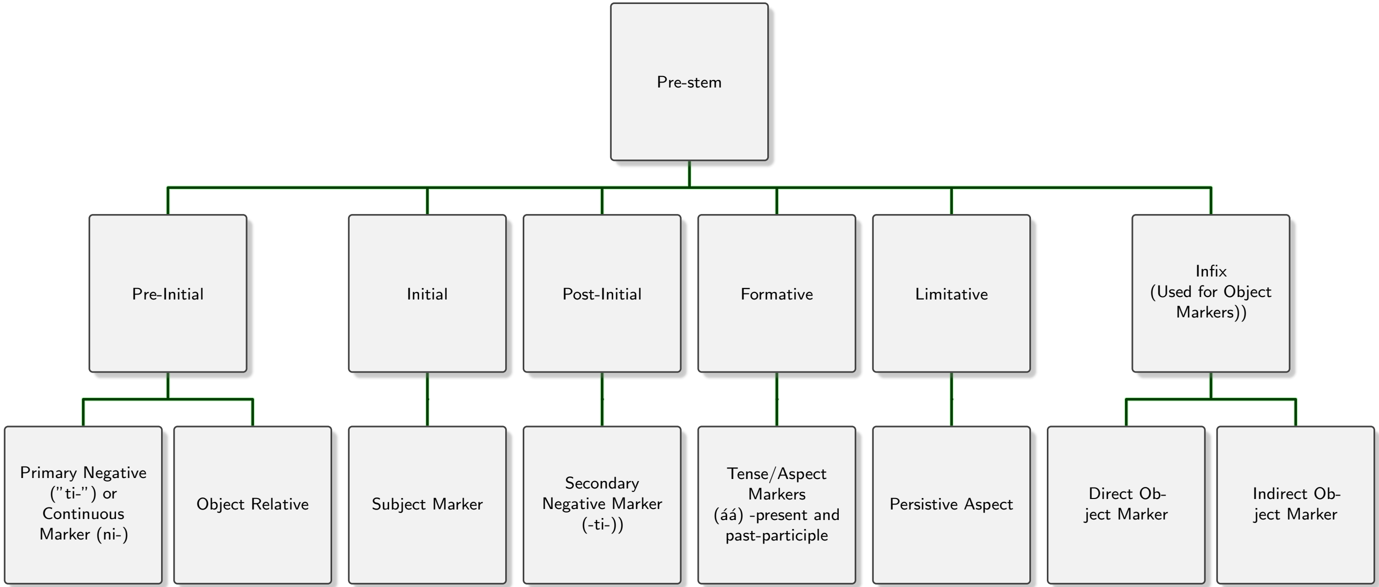

Pre-stem
Pre-Initial
Initial
Post-Initial
Formative
Limitative
Infix (Used for Object Markers))
Primary Negative ("ti-") or Continuous Marker (ni-)
Object Relative
Subject Marker
Secondary Negative Marker (-ti-))
Tense/Aspect Markers (áá) -present and past-participle
Persistive Aspect
Direct Object Marker
Indirect Object Marker

Figure 2.6: Structure of Pre-stem component of Bantu verbal unit .

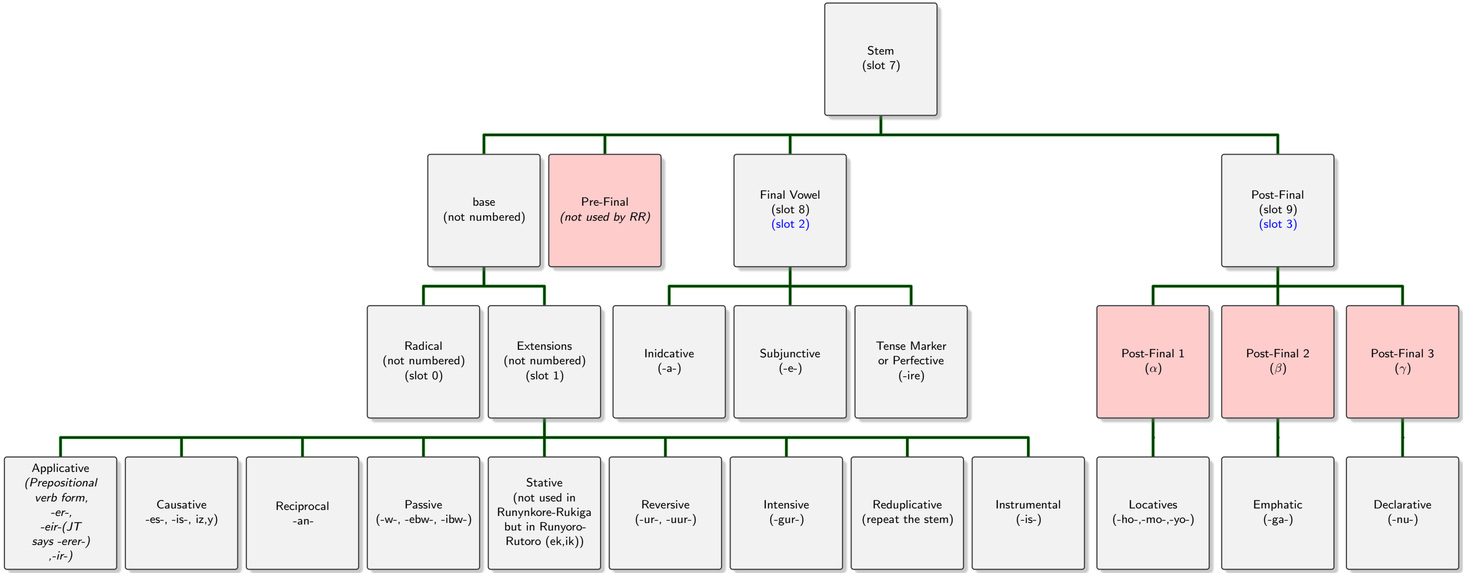


Figure 2.7: Structure of Stem component of the Bantu verbal unit .

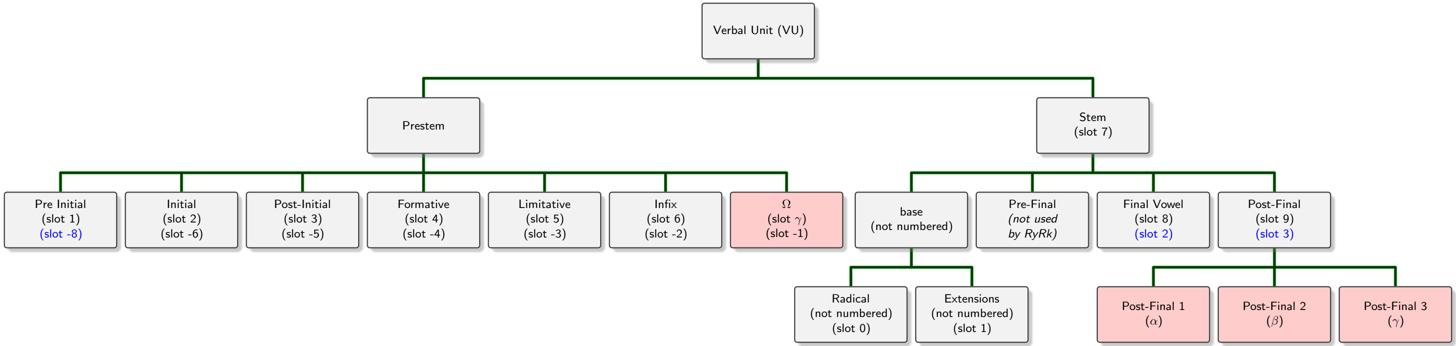

Verbal Unit (VU)
Prestem
Stem (slot 7)
Pre Initial (slot 1) (slot -8)
Initial (slot 2) (slot -6)
Post-Initial (slot 3) (slot -5)
Formative (slot 4) (slot -4)
Limitative (slot 5) (slot -3)
Infix (slot 6) (slot -2)
Ω (slot γ) (slot -1)
base (not numbered)
Pre-Final (not used by RyRk)
Final Vowel (slot 8) (slot 2)
Post-Final (slot 9) (slot 3)
Radical (not numbered) (slot 0)
Extensions (not numbered) (slot 1)
Post-Final 1 (α)
Post-Final 2 (β)
Post-Final 3 (γ)

Figure 2.8: Slots in black font colour are obtained from Derek Nurse (2003)'s template for Bantu while those in blue were improvements by Katushemererwe and Hanneforth (2010a) for Runyakitara. The Postfinal 3 was introduced by Turyamwomwe (2011) for the declarative.

Mpairwe and Kahangi (2013b) opine that regular verbs in Ry/Rk appear in four major verb forms, though they prefer to call them "functional categories". The forms are imperatives, subjunctives, perfectives and infinitives. Each of these verb forms can be further subdivided into the simple, prepositional (which we interpret as the applicative) and causative. Each sub-division can be rendered in active and passive voice.

2.3.3 Grammatical Tense and Aspect in Ry/Rk

The subject of grammatical tense and apsect among linguists has been studied extensively for Indo-European languages with Hewson and Bubeník (1997)'s work as an example. The T/A system for Ry/Rk has also been studied by a number of scholars: Muzale (1998), Katushemererwe and Hanneforth (2010a), Turyamwomwe (2011) and Ndoleriire (2020) each with varying level of coverage, agreements and disagreements which are mainly limited to the names they give the tenses. While (Muzale, 1998) shows how different T/A markers have developed through time (diachronicaly) up to their current forms (as of 1998) among the Rutara group of languages, Katushemererwe and Hanneforth (2010a) and Ndoleriire (2020) confine their work to Runyakitara but Turyamwomwe (2011)'s work is restricted to Runyankore.

Traditionally, tense is divided into past and non-past. Non-past is further divided into present and future. However, in Ry/Rk the past is split into the Remote Past, Near Past and Immediate Past (Turyamwomwe, 2011). Muzale (1998) calls the immediate past – which refers to an event that took place recently like earlier today – the memorial present. We also found that the Memorial Present identified in (Muzale, 1998) and Immediate Past (Katushemererwe and Hanneforth, 2010a; Turyamwomwe, 2011) are one and the same i.e. they mean the same and use identical tense and polarity agreement markers. The Universal Tense is identical to Muzale (1998)'s Experiential Present. The Future is divided into the Near and Far or Remote Future. As an example, Table 2.2 shows how different morphemes are combined to form a verb for the seven tenses while omitting markers for direct and indirect objects. The present tense is divided into universal tense (referred to as simple present tense in English) and the Continuous / Progressive which are similar to Muzale's Experiential and Memorial Present. The Future is divided into the Near and Far / Remote Future. In the verbal unit of Ry/Rk, tense and aspect are marked using particular morphemes which may be simple (a single morpheme) or compound (multiple morphemes). The tense markers for all these tenses are summarised in Table 2.2. As an example Table 2.2 shows how different morphemes are combined to form a verb for the seven tenses. Note that this is simply a general template that applies to verbs whose extensions slot is empty and hence Final vowel is /*a*/ in imperative. This final vowel /*a*/ would be replaced by /*ire*/ in the anterior (Perfective) in the simplest case but there are thirty-eight rules for converting an imperative to a perfective. The rules depend on: the number of syllables in the verb (monosyllabic, disyllabic, trisyllabic etc.); the length of the penultimate vowel and the letters composing or modifying the terminal syllable such as; /*-sa*/,/*-sh-*/,/*-za*/,/*-zya*/ or the semi-vowels /*-w*/ or /*-y*/.

For example, the verb entered in the dictionary as /*gyenda*/ is annotated with /{*da-zire*}/ to mean that in order to convert the imperative into perfective, replace the */da/* in */gyenda/* with */zire/* to form */gyenzire/* in the perfective. Muzale (1998) and Turyamwomwe (2011) both have different aspects for Runyankore and Rukiga. The difference could be attributed to Turyamwomwe's emphasis on perfective versus imperfective aspects i.e. perfective, progressive, persistive and habitual ignoring the full spectrum of aspects possible. However Muzale (1998) covers Retrospective, Resultative, Persistive and Remote Retrospective in addition to that covered by Turyamwomwe (2011).

Traditional Tense System	**Tense in Ry/Rk**	**Pol**	***/To see/***	**Generalization**
Past	Remote Past	Pos	S-ka-reeb-a	S-ka-Rad-FV
		Neg	ti-S-rá-reeba -ir-e	Pneg-S-TM-Rad-TM-FV
	Near Past	Pos Neg	S-∅-reeb-ir-e ti-S-∅-reeb-ir-e	S-∅-Rad-TM-FV Pneg-S-∅-Rad-TM-FV
	Immediate Past	Pos Neg	S-áá-reeb-a ti-S-áá-reeb-a	S-TM-Rad-∅-e Pneg-S-TM-Rad-∅-FV
Present	Memorial Present	Pos	S-áá-reeb-a	S-TM-Rad-FV
		Neg	ti-S-áá-reeb-a	Pneg-S-TM-Rad-FV
	Experiential Present	Pos	S-∅-reeb-a	S-∅-Rad-FV
		Neg	ti-S-∅-reeb-a	Pneg-S-∅-Rad-Fv
Future	Near Future	Pos	ni-S-ija/za ku-reeb-a	CM-S-ija /za ku-Rad-FV
		Neg	ti-S-ku-ija/ku-za ku-reeb-a *or* ti-tu-ra-reeb-FV	Pneg-S-ku-ija /za ku-Rad-FV *or* Pneg-tu-ra-Rad-FV
	Remote Future	Pos	S-riá-reeba-a	S-TM-Rad-FV
		Neg	ti-S-riá-reeba-a	Pneg-S-TM-Rad-FV

Table 2.2: How different morphemes are combined to form a verb. CM = Continuous Tense Marker, Pneg = Primary Negative marker, Sneg= Secondary Negative marker, S = Subject Marker, followed by a Tense Marker (TM), ∅ = absence of TM, Rad = Radical and FV = Final Vowel. Note: Pos = Positive and Neg = Negative. The Immediate Past and memorial present are one and the same referring to an event the occurred a moment earlier.

2.3.4 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), (5) relative subject clauses and (6) relative object clauses (Mpairwe and Kahangi, 2013b). 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 achieving qualification of nominal expressions in Ry/Rk. Some adjectival expressions are multi-word expressions (portmateau) such as clauses. Because such clauses are usually derivational they cannot be considered lexical items. As a resul it is therefore difficult to identify and classify all forms of this part of speech without a sound theory for word class division and possibly morphemic tags.

Among the adjectival stems and phrases, they are further divided into three types, adjectival stems whose concord is conjunctive with the stem and two others where the concord is disjunctive, but taken from two different classes of concords i.e adjectival clitics and genitive clitics. Some adjectival stems exist in the language but others can be derived from verbs that bear the same or similar semantic meaning of the adjective in mind. This derivation is achieved by affixing the conjugated copulative verb */ri/* i.e. (Subject Prefix + */ri/*) as a prefix to the the verb. An example is */-ri-kutagáta/* comes from the verb */kutagáta/* meaning */to be warm/*. Lastly, depending on the nominal expression, it can either occur before or after the nominal (noun, noun phrase or pronoun). We note that Katushemererwe et al. (2020) i.e. (see Byakutaaga et al., 2020, chap. 2, pgs. 67-73) provide the most recent treatment of the morphology of adjectives.

2.3.5 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)[1] (Nivre et al., 2016) 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/*.

2.3.6 Numerals

Since numbers can be nouns, quantifiers, determiners, adjectives or adverbs, modelling them becomes difficult because we have to track agreement concords attributed to gender. Numerals are inherently nouns since they give names to entities used for counting (Ordinals) and order (cardinals). However, Numerals are also quantifiers of nouns i.e. they give an indication of how much or big other nouns are. Being a noun, each numeral belongs to a noun class and therefore has an initial vowel and a noun class particle. When used in quantification of other nouns, the numeral drops the initial vowel for all numbers with a few exceptsions (see Mpairwe and Kahangi, 2013b, chap. 26, pg. 274) and acquires the prefix of the noun or noun phrase it quantifies. The agreement marker (Noun Prefix) acts as a prefix to the last word of the number. For instance, take the example */two hundred and forty people/*. The number */two hundred and forty/* in Ry/Rk is *magana abiri na ana* while the noun phrase

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

/two hundred and forty people/ is: */abantu magana abiri na ba-a-na/* whose actual surface form is: */abantu magana abiri na bana/*. The initial vowel */a/* of */ana/* i.e. */one/* is dropped. Some numerals can be pluralised while others cannot for example you can have */one 6/* (*/o-mu-kanga gumwe/*) and */two groups of 6/* (*/emikanga ebiri/*). The counting system is awash with synonyms attributed to the evolution of the language over time and the influence of English. The surface form of numerals depends on whether the numeral is Cardinal or Ordinal. When numerals are used in noun phrases the surface form of the number (signified) depends on the actual number(signifier) and noun class of the head noun in the noun phrase.

2.3.7 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) (Katushemererwe et al., 2020; Mpairwe and Kahangi, 2013b). 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.

For Ry/Rk, pronouns can exist as either discrete words or affixes. Apart from the noun class MU_BA, the rest of noun classes use only the third person because it is only humans that can use all the three persons. A fair explanation of pronouns can be found in (see Mpairwe and Kahangi, 2013b, chap. 20) but Katushemererwe et al. (2020) provide a thorough and recent explanation of the morphology of pronouns in (see Katushemererwe et al., 2020, pg. 60-66)

2.4 Grammar Formalisms, Frameworks and Resource Grammars

Grammars have been studied since 6^{th} century BC, first by Yaska and later Panini. A grammar is a collection of rules that describe both the structure of a language and a method of establishing whether an utterance in the language is well-formed. This definition appeals to both traditional grammar (descriptive and prescriptive) and formal grammar. Description grammars provide only a narrative description of natural languages. During the process of designing computational grammars of such languages, software developers require both such descriptions and a design or specification language for translating these narratives into pseudo-code before actual coding.

2.4.1 Grammar Formalisms and Frameworks

In computational linguistics, the emphasis has been put on the use of grammar formalisms as rigorous formal and mathematical (theoretical) devices for studying and characterising languages. A framework can be defined as a common set of assumptions and tools that is used when grammatical theories of a natural language are formulated (Stefan, 2016), or as a set of guiding principles for syntactic inquiry (Bender, 2008). These frameworks are usually based on particular linguistic theory that is used to explain various phenomena at different levels of language analysis, such as morphology, phonology, syntax and semantics. This led to advancement of various 'theories /approaches of grammar' such as unification grammars which include: Phrase Structure Grammars (PSG), different extensions to the basic PSG grammars such as Generalized Phrase Structure Grammar (GPSG) (Gadzar et al., 1985), Lexical-Functional Grammar (LFG) (Joan Bresnan, 1982), Categorial Grammar (CG) (Adjukiewicz, 1935; Bar-Hillel et al., 1960) and its variants, Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994), Tree Adjoining Grammar (TAG) as well as various forms of dependency grammars (Stefan, 2016). There are numerous formalisms for expressing both formal and natural languages and their expressive power can be summarized by the augmented Chomsky Hierarchy found in (Jäger and Rogers, 2012; Jurafsky and Martin, 2009) i.e. regular languages, context-free Languages (CFGs), mildly-context-sensitive languages, context-sensitive languages and recursively enumerable languages listed in order of increasing generative / expressive power. With CFGs, rules get cumbersome once we try to deal with: (1) permutation (changing order of constituents); (2) suppression (the omission of certain constituents eg. dropping of subject, direct and indirect objects in Ry/Rk and other pro-drop languages); (3) reduplication, (4) agreement; and (5) specification of additional context (ie. on a CFG production rule) under a multilingual setting (Ranta, 2011b). Generally, there is a need for more user-friendly formalisms for natural languages. Note that CFGs can theoretically handle only the first four. The last one is usually handled by context-sensitive grammars.

However, most if not all natural language do not usually require the highly expressive power of context-sensitive grammars or formal languages whose parsing complexity is non-polynomial as shown by Joshi (1985). Joshi (1985) suggested that mildly context-sensitive grammars (MCSG) are the grammars with sufficient properties (expressive power) for modelling and formalising the features of all natural languages. Parallel Multiple Context-free Grammar (PCMFG) that lies between mildly context-sensitive grammar and context-sensitive grammar can formally describe more complex languages languages than mildly context-sensitive grammars. However, formalisms based on context-sensitive languages such as Head Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) and Lexical Functional Grammar (LFG) (Kaplan, 1997) definitely describe more complex languages than PMCFG. Examples of mildly context-sensitive grammar formalisms whose equivalence was identified by (Vijay-Shanker and Weir, 1994) are: Tree Adjoining Grammar(TAG) (Joshi et al., 1975), Head Grammar, Linear indexed Grammar and Combinatory Categorical Grammar. Grammatical Framework (GF) discussed in Section 2.5 below can express any language as long as it is PCMFG.

2.4.2 Resource Grammars

Resource Grammars can be defined as broad coverage machine-readable implementations of traditional grammars of a particular natural language using a grammar formalism augmented with a development, or programming environment. There are numerous resource grammars that have been developed such as the English Resource Grammar based on HPSG and developed using the Lexical Knowledge Builder (LKB) grammar engineering environment (Copestake and Flickinger, 2000). Other resource grammars of substantial size developed using the LKB environment include; Japanese, German, Spanish, Portuguese, Korean, Modern Greek and Norwegian. Medium-sized grammars of; French, Mandarin Chinese, Bulgarian, Wambaya, Hausa, Russian, Dutch, Hebrew and Indonesian, and some experimental grammars of other languages have also been developed. These grammars have been used to develop applications in semantic analysis, semantic parsing, summarisation, textual entailment, POS tagging, Ontology acquisition, Machine Translation, Grammar Tutoring etc.

Grammatical Framework (GF)(Ranta, 2004) and its Resource Grammar Library discussed in detail in section 2.5 is an alternative environment to LKB with over 30 languages supported as resource grammars of substantial size. Whereas resource grammars implemented within the LKB framework usually describe aspects ranging from phonology to syntax and semantics, GF Resource grammars are multilingual broad coverage syntactic grammars augmented with simple inflectional functional morphology. They are implemented in the form of software libraries exposed by a common Application Programming Interface (API) (Cooper and Ranta, 2008) which can be utilised by domain-specific grammars. (referred to as Application Grammars). Development of broad coverage grammars usually requires two kinds of experts; linguists who understand the inner workings of the grammar of a natural language, as well as programmers that can best model the domain-specific knowledge required by the NLP application they wish to implement. GF provides a separation of concerns between designers of resource grammars and application grammars to enhance productivity by letting each of them concentrate on their area of expertise while contributing to the overall development of an NLP application. This separation of concerns and the emphasis on domain-specific applications allows the realisation of useful NLP applications using a subset of the grammatical functions. Hence it is quicker to obtain benefits from resource grammars using a minimal lexicon as compared to large lexical resources without a grammar.

2.5 Grammatical Framework (GF)

Note: This section was written together with Peter Ljunglöf

Grammatical Framework (GF) is a grammar formalism based on type theory and a special purpose functional programming language for defining grammars of both formal and natural languages (Ranta, 2009a, 2011b). Its main feature is the separation of abstract and concrete syntax, which makes it very suitable for writing multilingual grammars. GF is modular and highly expressive (Ljunglöf, 2004) making it suitable for engineering libraries, and expressing long distance dependencies among natural languages. It is suitable for under-resourced languages since it does not need any additional linguistic resources, and being multilingual, it can be used to develop resources for under-resourced languages using existing resources of other languages already covered in its Resource Grammar Library (Kolachina and Ranta, 2016; Ranta, 2009b).

The main idea of GF is the separation of abstract and concrete syntax. The abstract syntax defines a set of abstract syntactic structures, called abstract terms or trees, which can be used to define a language-independent or semantic meaning representation. The concrete syntax defines a relation between the abstract structures and their language-specific constructions. This makes it possible to define several concrete syntaxes for one single abstract syntax, which then can act as an interlingua between different languages. GF also has a rich module system which facilitates grammar writing as an engineering task, by reusing common grammars.

2.5.1 GF Abstract Syntax

The abstract theory of GF is a version of Martin-Löf's dependent type theory (Nordström et al., 1990). In this research study, as with most GF grammars, we only use non-dependent categories, which makes the abstract syntax equivalent to a context-free grammar. An abstract GF grammar consists of category declarations introduced using the keyword *cat* and function declarations or type signatures (as they are called languages descending from imperative languages) introduced using the keyword *fun*. They are declared using the functional programming paradigm. The categories and functions are enclosed in a special module called abstract module. Category declarations simply provide an typed "abstract concept" whose typed "concrete realisation" is determined in another module (typically concrete modules see 2.5.2). The data structure attached to the category is "unrestricted" in abstract modules. Unrestricted here means it can be nay of the allowed datatypes (strings, parameters, records and tables) Given the following simple context-free grammar:

```
S  → NP VP
VP →V2 NP
NP →Det N
```

The abstract syntax would be represented by first declaring the categories using *cat* and then declaring the functions using *fun* as follows:

```
cat S ;  VP ; NP ; V2; Det ; N;
```

and the functions as follows:

```
fun
  PredVP : NP →VP →S;
  ComplV2 : NP →V2 →VP;
  DetN    :  N → Det → NP;
```

Compared to a standard context-free grammar, the order of the arguments are switched. This is a reflection of GF's background from type theory and functional programming. The meaning of these declarations is that PredVP takes an NP and a VP as arguments and returns an S, and that DetN takes a Det and aa N and returns an NP. Thus, the functions are equivalent to the context-free rules. GF does not distinguish between phrasal and lexical rules – the lexicon is simply a number of GF rules that take no arguments called constant functions.

2.5.2 GF Concrete Syntax

The concrete syntax of an abstract grammar is a compositional transformation of the abstract syntax trees into concrete representations. This means that every abstract category C has a corresponding *linearisation type* C°, and that every abstract function $f : C_1 \to \cdots \to C_n \to C$ has a corresponding *linearisation* $f^\circ : C_1^\circ \to \cdots \to C_n^\circ \to C^\circ$. The concrete syntax defines a relation between the abstract structures and their language-specific constructions. This makes it possible to define several sets of concrete “syntaxes” for one single abstract syntax. The single abstract syntax then acts as an interlingua between different languages. The concept of a shared abstract syntax is the reason for the multilingual capabilities of GF. Linearisation types are declared by the keyword *lincat*, where the following says that the linearisation type C° is T:

$$\textbf{lincat } C = T;$$

Correspondingly, linearisations are declared by the keyword *lin*:

$$\textbf{lin } f\ x_1 \ldots x_n = t[x_1, \ldots, x_n]$$

Here, each variable x_i is bound to a linearisation term with type C_i°, and t is then a term with type C°. To ensure decidability and efficiency of parsing, the linearisation type C° is restricted to any combination of the data types and structures: strings, finite parameters, tables and records.

$$C^\circ := \mathrm{Str} \mid P \mid P \Rightarrow C'^\circ \mid \{r_1 : C_1^\circ; \ldots; r_n : C_n^\circ\}$$

The *parameter types* are defined using datatype declarations, similar to how to define types in functional programming languages:

$$\textbf{param } P = p_1\ \alpha_1 \mid \ldots \mid p_k\ \alpha_k$$

For a thorough treatment of GF as a domain-specific programming language the interested reader is referred to (Ranta, 2011b) and the GF webpage[2] that contains numerous resources and tutorials on GF.

2.5.3 Using GF Grammars as libraries

The module system of GF makes it possible to use one grammar as a library when defining a new grammar, and this new grammar can in turn be used as a library when defining yet another grammar. This makes it possible to create *resource grammars* that can be used when writing grammars for domain-specific applications. It is also possible to use several grammar libraries at once, giving a hierarchy of grammars, much like when a programmer uses different libraries when solving a problem.

2.5.4 The GF Resource Grammar Library (GF-RGL)

The idea behind resource grammars are explained in section 2.4.2, and for GF there exists a multilingual Resource Grammar Library (GF-RGL), which consist of several natural language grammars built with a common abstract syntax (Ranta, 2009b). Currently the GF-RGL contains more than 30 languages from several different language families. Most of the GF-RGL language

[2] See: `https://www.grammaticalframework.org/`

implementations are of European languages, as well as some Middle-East and Asian languages. However, there is a lack of languages from other regions such as Africa, and in particular the GF-RGL does not cover many Bantu languages.

The GF-RGL does not attempt to cover all grammatical and morphological structures in all languages, but instead it has a focus on constructions that are common between many languages. It implements more that 50 grammatical categories and almost 200 construction functions. Because of the expressive module system of GF, it is possible to extend the common GF-RGL with language-specific constructions. One example of a language-specific extension in Ry/Rk is the additional verb tenses that are explained in section 2.3.3.

2.6 Related Work

2.6.1 NLP Resources for Under-resourced Languages

The term Language Resource refers to any speech or text data processed and formatted for use in building or improving NLP systems and applications. The definition of under-resourced languages by Besacier et al. suggests four characteristics such a language should possess: 1) lack of a unique writing system or stable orthography, 2) limited presence on the web, 3) lack of linguistic expertise and 4) lack of electronic resources for speech and language processing, such as monolingual corpora, bilingual electronic dictionaries and / or computational lexicons, transcribed speech data, pronunciation dictionaries, vocabulary lists etc. Runyankore and Rukiga (Ry/Rk) exhibit two of those characteristics i.e. 2) and 4) of which research study seeks to address.

The lack of raw text whether digital or non-digital implies a lack of corpora, because without the former you cannot have the latter. This challenge makes the development of NLP systems and applications with meaningful performance using the latest data-driven approaches (statistical and machine learning) very challenging. However, rule-based approaches, though old, can be used to formalise the grammars of these languages thus producing resources that can in turn can be leveraged to develop localised applications, tools and other language resources.

The absence of computational grammars and lexicons for most Bantu languages and the perceived challenges of developing them from traditional grammar books and dictionaries (Keet and Khumalo, 2014) has prompted computational linguistics researchers for Bantu languages to concentrate on natural language generation (NLG) whose input is an ontology and controlled natural language (CNL) — which is an engineered and highly specific subset of the target natural language with a specific narrow domain — so as to verbalise axioms of those ontologies. However, a more ambitious and difficult approach that of developing a comprehensive grammar as a software library which can be used by software developers to develop applications for any domain. Deeveloping a resource grammar for a natural language using GF takes such an approach.

2.6.2 Computational Lexical Resources

Machine Readable Dictionaries (MRDs) and computational lexicons for well-resourced languages such as those reported by Sanfilippo (1994), and ACQUILEX projects I and II[3] 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 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 have: large amounts of language data available on the web, prepared corpora of good quality, treebanks (Böhmová et al., 2003; Taylor et al., 2003; Xiao, 2008) 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 about different language-specific WordNet-based lexical databases that have been created while 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 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.

2.6.3 Modelling Computational Lexical Resources

With regard to modelling of lexicons for Bantu languages, a Bantu Language Model (BantuLM) was put forward by Bosch et al. (2018, 2006) after eliciting the inadequacies of Lexical Markup Framework (Francopoulo et al., 2006) arising from a failure to take such morphologies into account when designing the

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

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 (Mpairwe and Kahangi, 2013a; Museveni et al., 2009; Taylor and Mapirwe, 2009) and basic semantic relations found in thesaurus (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. We chose to use YAML[4] 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.

2.6.4 NLP Resource Engineering Using GF

There has been an increased interest in work on low-resource languages using GF. Any work on any NLP application of any language must begin with an implementation of a GF-RGL. The GF-RGL may be: miniature in the sense that it is very small i.e. it covers a small but interesting fragment of the language; or the more ambitious goal of covering as much of the language as possible. Examples of miniature implementations include: Swahili(Ngángá, 2012) and Tswana (Pretorius et al., 2017) from South Africa. However, the Swahili implementation has been greatly improved by Kituku (2019) who has also worked on more comprehensive GF-RGLs for Kikamba (Kituku et al., 2019) and Engekusi (Kituku et al., 2021).

Outside GF, previous work on Ry/Rk include: morphological analyzers by Katushemererwe and Hanneforth (Katushemererwe and Hanneforth, 2010a,b), a Controlled Natural Language for Runyankore (Byamugisha et al., 2016) and a Noun pluraliser (Byamugisha et al., 2018). However, this work has been limited to small fragments of the languages.

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

Part II

Publications

Chapter 3

Paper A: Computational Resource Grammars for Ry/Rk

This chapter is a reproduction of the following paper. The only thing that has changed is the formatting, no changes were made to the content.

D. Bamutura, P. Ljunglöf, P. Nabende, 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.

Errata

[A] The citation for the RGL in Section 3.1.1 should have been (Kolachina and Ranta, 2016; Ranta, 2009b) instead of (Kolachina and Ranta, 2016; Ranta, 2009a)

[B] In Table 3.2 the generalisation for Immediate Past should have been:

(a) S-TM-Rad-∅-e for positive polarity

(b) Pneg-S-TM-Rad-∅-FV for negative polarity

[C] words and sentences in the object languages Ry/Rk are not typeset according to the generic styles for linguistics for example:

(a) *tinkamureebagaho* should be /*tinkamureebagaho*/

(b) ti-n-ka-mu-reeb-a-ga-ho should be /*ti-n-ka-mu-reeb-a-ga-ho*/

(c) the English literal and idiomatic translations in running text ought to be treated similarly.

Note: In this publication, we used the acronym R&R for Runyankore-Rukiga because we had not come across the current de facto acronym Ry/Rk. Therefore R&R should be interpreted as Ry/Rk.

Abstract

In this paper, we present computational resource grammars of Runyankore and Rukiga (R&R) languages. Runyankore and Rukiga are two under-resourced Bantu Languages spoken by about 6 million people indigenous to South Western Uganda, East Africa. We used Grammatical Framework (GF), a multilingual grammar formalism and a special-purpose functional programming language to formalise the descriptive grammar of these languages. To the best of our knowledge, these computational resource grammars are the first attempt to the creation of language resources for R&R. In Future Work, we plan to use these grammars to bootstrap the generation of other linguistic resources such as multilingual corpora that make use of data-driven approaches to natural language processing feasible. In the meantime, they can be used to build Computer-Assisted Language Learning (CALL) applications for these languages among others.

3.1 Introduction

Runyankore & Rukiga (hereafter R&R) are two heavily under-resourced Bantu languages. Their limited presence on the web makes it difficult to develop substantial computational linguistic resources for these languages. Consequently, the lack of such resources makes the use of data-driven Natural Language Processing (NLP) approaches unsuitable for these languages. However, rule-based approaches such as grammars, can be used to bootstrap the creation of such resources. In this paper we present computational resource grammars of these two languages developed using Grammatical Framework (GF).

3.1.1 Grammatical Framework (GF)

GF is a multilingual grammar formalism , a logical framework and a special-purpose functional programming language for defining grammars of both formal and natural languages (Ranta, 2009a, 2011b). We chose GF because it does not need any additional linguistic resources, and being multilingual, it can be used to develop resources for under-resourced languages by using existing linguistic resources of well-resourced languages already covered in its Resource Grammar Library (RGL) (Kolachina and Ranta, 2016; Ranta, 2009a).

3.1.2 Abstract and Concrete Syntax

Each grammar in GF consists of an *abstract and concrete syntax*. The abstract syntax defines a set of abstract syntactic structures, called abstract terms or trees, which are used to define a language-independent or semantic meaning representation. The concrete syntax defines a relation between the abstract structures and their language-specific constructions. This makes it possible to define several sets of concrete "syntaxes" for one single abstract syntax. The single abstract syntax then acts as an interlingua between different languages. The concept of a shared abstract syntax is the reason for the multilingual capabilities of GF.

3.1.3 Resource & Application Grammars

Grammars designed in GF are of two types: *resource* and *application grammars*. Resource grammars are broad-coverage grammars developed from scratch for the purpose of formally describing the morphology and syntax of natural languages while application grammars model semantic information about a specific application domain. Using GF's modular system, Resource Grammars are packaged together and exposed by both a common API (that is based on the common abstract syntax) and language specific APIs into what is called the GF Resource Grammar Library (GF-RGL) Ranta (2009b). Application grammars make use of general linguistic functions implemented in resource grammars by accessing them through the GF-RGL. Resource grammars have been used successfully in domain-limited application areas such as Multilingual Document Authoring (Dymetman et al., 2000), low-coverage multilingual translation (Ranta et al., 2010), domain specific dialogue systems such as music players (Perera and Ranta, 2007) and Computer-Assisted Language Learning (CALL) (Lange, 2018; Lange and Ljunglöf, 2018b).

Another important use case in the area of localisation is the multilingual dissemination of weather information especially in multilingual societies. Our immediate motivation is therefore to utilise the GF-RGL for R&R to leverage the work done by Lange (2018) on CALL for the Latin language in order to build, localise and improve tools that can be used to create automatic exercises for learning R&R grammar to higher levels of proficiency accessible to all.

3.2 Related Work

Previous work on the computational modelling of the grammar of R&R include: noun and verb morphological analysers by Katushemererwe and Hanneforth (2010a,b), a Controlled Natural Language for Runyankore (Byamugisha et al., 2016) and a Noun pluralizer (Byamugisha et al., 2018). However, this work has been limited to small fragments of the languages. Within the GF community, there has been work on computational modelling of Bantu languages: Kikamba (Kituku et al., 2019), Tswana (Pretorius et al., 2017), and Swahili (Ngángá, 2012). While we consulted the Swahili implementation during initial development, we found that Swahili is morphologically and syntactically less complex than R&R. Additionally, its coverage of the GF-RGL functions was very small. Little insight was generated from that grammar. Likewise the Tswana GF-RGL was limited to modelling the proper verb for declarative sentences which is small in scope. Twsana's use of both a disjunctive and conjunctive orthography as compared to R&R's conjunctive morphology also provided limited insights into how to implement the grammars of R&R. Work on Kikamba and R&R was done during the same time-frame and hence both of us benefited from the sharing of ideas.

3.3 Runyankore & Rukiga (R&R)

R&R are languages spoken in South-Western Uganda by about 6 million people (Simons and Fennig, 2018). They belong to the **JE10** zone (Maho, 2009) of the Niger-Congo Bantu language family. Just like any other Bantu languages, morphologically, R&R are **highly agglutinating** (e.g., the single word *tinkamureebagaho* (ti-n-ka-mu-reeb-a-ga-ho) is a sentence meaning "I have never seen him/her"), exhibit high instances of **phonological conditioning** and a **large Noun Class System** of 17 noun classes (Byamugisha et al., 2016; Katushemererwe and Hanneforth, 2010b). This noun class system dictates a complex concordial system of agreement among phrasal categories. These properties make the morphology of the languages more complex to computationally model as compared to analytic languages such as English. Since both languages share the same dictionaries (Mpairwe and Kahangi, 2013a; Taylor and Yusuf, 2009) and grammar books (Morris and Kirwan, 1972; Mpairwe and Kahangi, 2013b) their grammar is largely identical while the lexicon differs by 6%–16% (Simons and Fennig, 2018; Turyamwomwe, 2011).

3.3.1 Nominal Morphology

The morphological structure of nouns in R&R consists of two parts, a **class prefix** and a **noun stem**. The class prefix is further divided into an **Initial Vowel (IV)** and a **Noun Class particle (NCP)** (Mpairwe and Kahangi, 2013b). The **noun stem** usually bears the bulk of the semantic meaning of the noun. Each Noun in R&R, belongs to a particular **Noun Class (NC)**. The group of possible noun classes is given in Table 3.1 adapted from (Katushemererwe and Hanneforth, 2010b) with modifications. The predominant naming scheme of noun classes in Bantu languages (called the Bleek-Meinhoff system) makes use of a combination of a numeral and optionally letters (see column labelled Numbers in Table 3.1). However, we discovered an alternative scheme that uses NCP (refer to "Particles" column in the same table) utilised by Mpairwe and Kahangi (2013a,b) in their dictionary and grammar books. Since we make heavy use of these books, we have found it convenient to use the latter scheme in order to avoid an additional step of mapping between the two systems during our implementation of the grammar as explained in section 3.4 Apart from locative particles -ha-, -mu- and -ku- , most of the other particles can be arranged in singular-plural pairs for common nouns. We generalise such a pairing using the notation $[\alpha - \beta]$ where α & β are noun class particles chosen from the sets of singular & plural particles respectively. We borrow the use of the number ZERO (0) from (Mpairwe and Kahangi, 2013a) in their Runyankore-Rukiga dictionary to denote absence of either singularity or plurality in order to maintain the pairing for such nouns. Hence the pairs $[\alpha - 0]$, $[0 - \beta]$ and $[0 - 0]$ which represent nouns that are always singular, plural and those that collectively neither have an IV nor noun class particle respectively. It is important to note that classes 9_10 and 9 in the table are both assigned N_N because the set of agreement concords for the two classes are the same. More noun classes are used in our implementation to cater for Numerals which are a special set of nouns for naming entities used to count (**ordinals**), or encode order (**Ordinals**).

3.3.2 Verbal Morphology

In Meeussen's 1967 original construction, the Bantu verbal unit consists of a **pre-stem** and **stem**. The stem is further divided into a **base** and **final vowel (FV)**. The base is also divided into a **radical (Rad)** and **extensions**. Further subdivisions in each of these parts results into 11 slots (Katushemererwe and Hanneforth, 2010a; Turyamwomwe, 2011), with each slot taking a set of morphemes for a particular purpose such as Primary/Secondary negative (*Pneg* / *Sneg*), subject (*S*), object, tense, aspect and other markers. Regular verbs can be classified into four base-forms: Imperatives, Subjunctives, Perfectives and Infinitives. They can be rendered in active or passive voice and within each voice, the verb can take the form of Simple, Prepositional and Causative.

In the verbal unit of R&R, Tense and Aspect (T/A) are marked using morphemes which may be simple or compound. However, in our attempt to model the grammar of R&R, we have combined the constructions suggested by Muzale (1998), Katushemererwe and Hanneforth (2010a) and Turyamwomwe (2011), based on omissions and coverage made by each. While Muzale (1998) shows how different T/A markers have developed through time (diachronicaly)

	Class		Individual Particles		Example		Gloss
ID	Numbers	Particles	Singular	Plural	Singular	Plural	Singular(Plural)
1	1_2	MU_BA	MU	BA	o-mu-shaija	a-ba-shaija	man (men)
2	1a	MU_ZERO	MU	n/a	o-mu-hangi	n/a	creator (n/a)
3	1b/2b	ZERO_BAA	n/a	BAA	swhento	baa-shwento	Uncle(s)
4	3_4	MU_MI	MU	MI	o-mu-ti	e-mi-ti	tree(s)
5	3a	MU_ZERO	MU	n/a	o-mwisyo	n/a	breath (n/a)
6	4a	ZERO_MI	n/a	MI	n/a	e-mi-gyendere	n/a (way of walking)
7	5_6	RI_MA	RI	MA	e-ri-sho	a-ma-isho	eye(s)
8	5a	I_MA	I	MA	e-i-teeka	a-ma-teeka	law(s)
9	5b	I_ZERO	I	n/a	e-i-tétsi	n/a	pampering(n/a)
10	6a	ZERO_MA	n/a	MA	n/a	a-ma-te	milk (milk)
11	7_8	KI_BI	KI	BI	e-ki-ti	e-bi-ti	stick (stick)
12	7	KI_ZERO	KI	n/a	e-ki-niga	n/a	anger (n/a)
13	8	ZERO_BI	n/a	BI	n/a	e-bi-bembe	(n/a) leprosy
14	9_10	N_N	N	N	e-n-te	e-n-te	cow(s)
15	9	N_N	n/a	n/a	e-bahaasa	e-bahaasa	envelope(s)
16	10	ZERO_ZERO	n/a	n/a	bwîno	bwîno	ink (ink)
17	11_10	RU_N	RU	N	O-ru-shózi	e-n-shózi	mountain(s)
18	12_14	KA_BU	KA	BU	a-ká-bunza	o-bu-bunza	question mark(s)
19	12	KA_ZERO	KA	n/a	a-ka-bi	n/a	danger (n/a)
20	14	ZERO_BU	n/a	BU	n/a	o-bu-cécezi	n/a(being humble)
21	13	ZERO_TU	n/a	TU	n/a	o-tu-ro	n/a (sleep)
22	15_6	KU_MA	KU	MA	o-ku-guru	a-ma-guru	leg(s)
23	16	HA_ZERO	HA	n/a	a-ha-kaanyima(*)	n/a	behind the house (n/a)
24	17	KU_ZERO	KU	n/a	o-ku-z/'imu	n/a	Underground (n/a)
25	18	MU_ZERO	MU	n/a	o-mu-nda	n/a	in the stomach (n/a)
26	20_21	GU_GA	GU	GA	o-gu-kazi	a-ga-kazi	bad woman (women)
27	11_14	RU_BU	RU	BU	o-rur-o	o-bu-ro	one millet grain (many)
28	14_6	BU_MA	BU	MA	o-bu-ta	a-ma-ta	bow(s)
29	γ	RU_ZERO	RU	n/a	0-ru-me	n/a	dew (n/a)

Table 3.1: The Runyankore and Rukiga noun class system (both the numerical system and that based on Individual particles) and examples of both singular and plural. Adapted from (Katushemererwe and Hanneforth, 2010b) and updated using the dictionary by Mpairwe and Kahangi (2013a)

up to their current forms (as of 1998) among Rutara, Katushemererwe and Hanneforth (2010a) confine their work to Runyakitara and Turyamwomwe (2011) restricts himself to T/A in R&R. Therefore our design was based first on (Muzale, 1998) followed by (Katushemererwe and Hanneforth, 2010a) and lastly (Turyamwomwe, 2011) for verbs. Traditionally, tense is divided into Past and Present and Future. However, in R&R the past is split into the Remote Past, Near Past and Immediate Past (Turyamwomwe, 2011). We found that the Memorial Present identified in (Muzale, 1998) and Immediate Past (Katushemererwe and Hanneforth, 2010a; Turyamwomwe, 2011) are one and the same i.e. they mean the same and use identical tense and polarity agreement markers. The tense markers for all these tenses are summarised in Table 3.2.

Universal Tense	**Tense in Ry/Rk**	**Pol**	**“To see”**	**Generalization**
Past	Remote Past	Pos	S-ka-reeb-a	S-ka-Rad-FV
		Neg	ti-S-rá-reeba -ir-e	Pneg-S-TM-Rad-TM-FV
	Near Past	Pos	S-∅-reeb-ir-e	S-∅-Rad-TM-FV
		Neg	ti-S-∅-reeb-ir-e	Pneg-S-∅-Rad-TM-FV
	Immediate Past	Pos	S-áá-reeb-a	S-TM-Rad-TM-e
		Neg	ti-S-áá-reeb-a	Pneg-S-TM-Rad-TM-FV
Present	Memorial Present	Pos	S-áá-reeb-a	S-TM-Rad-FV
		Neg	ti-S-áá-reeb-a	Pneg-S-TM-Rad-FV
	Experiential Present	Pos	S-∅-reeb-a	S-∅-Rad-FV
		Neg	ti-S-∅-reeb-a	Pneg-S-∅-Rad-Fv
Future	Near Future	Pos	ni-S-ija/za ku-reeb-a	CM-S-ija /za ku-Rad-FV
		Neg	ti-S-ku-ija/ku-za ku-reeb-a **or** ti-tu-ra-reeb-FV	Pneg-S-ku-ija /za ku-Rad-FV **or** Pneg-tu-ra-Rad-FV
	Remote Future	Pos	S-riá-reeba-a	S-TM-Rad-FV
		Neg	ti-S-riá-reeba-a	Pneg-S-TM-Rad-FV

Table 3.2: How different morphemes are combined to form a verb. CM = Continuous Tense Marker, Pneg = Primary Negative marker, Sneg= Secondary Negative marker, S = Subject Marker, followed by a Tense Marker (TM), ∅ = absence of TM, Rad = Radical and FV = Final Vowel. Note: Pos = Positive and Neg = Negative. The Immediate Past and memorial present are one and the same referring to an event the occurred a moment earlier.

The Universal Tense is identical to Muzale (1998)'s Experiential Present. The Future is divided into the Near and Far / Remote Future. As an example, Table 3.2 shows how different morphemes are combined to form a verb for the seven tenses while omitting markers for direct and indirect objects. With regard to Aspect, Muzale (1998) identifies Retrospective, Resultative, Persistive and Remote Retrospective in addition to Perfective, Progressive, Persistive and Habitual identifed by Turyamwomwe (2011).

3.3.3 Reason for lack of resources

Despite the initial exposure to learning R&R in the first three years of primary school, English becomes the official language of instruction and examination from the fourth year on, severely limiting the continued study of R&R to higher levels of proficiency. It is also worthy to note that although dictionaries, grammar books and an orthography for R&R exist, R&R just like any other native languages in Uganda largely remain oral as opposed to written even among those literate in English. Only a dismal few study the language to a level sufficient to achieve proficiency in writing which implies lack of continuity in learning the grammar of the language. This explains the nearly zero presence on the web hence the lack of any computational language resources. As a result, the languages are highly under-resourced. It is therefore important to take steps in building language resources, encouraging writing in these languages and their preservation.

3.4 GF-RGL Implementation of R&R

In this section, we explain how the grammars for R&R were implemented using GF. The GF-RGL does not attempt to cover all grammatical and morphological structures in all languages, but instead focus is put on constructions that are common amongst the many languages of the world. It implements more than 50 grammatical categories and almost 200 construction functions. Because of the expressive module system of GF, it is possible to extend the common GF-RGL with language-specific constructions. The task is to write concrete modules for each abstract module.

3.4.1 Lexicon

When building an RGL for any language, the first thing to tackle is the lexicon. For each lexical item defined in the abstract module of the GF-RGL lexicon, a concrete mapping must be implemented for the language under investigation. This concrete mapping involves the enumeration of all possible morphological inflectional forms of the lemma provided. It is impossible to have a strict one-to-one mapping due to the existence of synonyms and lexical gaps. Synonyms are treated as separate GF lexical categories, so we selected a single word from the set of synonyms and left other synonyms to be catered for by an Extension module for the Lexicon. For lexical gaps in R&R which are a result of cultural differences, modernisation and lack of universality in language, we employed loan words (influenced by English) and adapted them according to the orthography of R&R. For the problem of a lack of a rich notion of adjectives particularly with respect to **degree**, we used circumscription. Just like GF-RGLs for other languages, we minimised the requirement of explicitly enumerating all the inflectional forms of a lexical item from a given category through the use of morphological paradigms. If a lexical entry ω of a given lexical type C has surface forms $\langle \omega_1, \omega_2 \ldots \omega_n \rangle$, then these paradigms are special functions that take between one surface form (base form) and at most $n-1$ surface forms and other information to produce the full set of inflected word-forms of that lexical entry. Paradigms that take one surface form, called smart paradigms (Détrez and Ranta, 2012), are restricted to lexemes whose inflection is regular.

3.4.1.1 Common Nouns and Proper Nouns

In R&R, common nouns inherently belong to a noun class. It is possible to use these nouns in either their **Complete** or **Incomplete** forms and each of these is inflected for number (refer to Table 3.3 for an example). We therefore declared parameters for NounState, Number and Gender (lines 2–4), a linearisation type for Nouns (line 18) in code listing 3.1 on the next page. We also declared paradigms for computing inflection tables for nouns. We used a composite parametric data type similar to algebraic data types from functional programming to encode agreement with respect to noun class, Person and Number in lines 4-12 of listing 3.1. Under normal circumstances Proper Nouns do not inflect with number. They are all in the third person but belong to different noun classes based on the common noun they give a name to. It was therefore necessary to keep track of information about Agreement and whether the noun refers to a location or place (refer to line 19 in listing 3.1). The smart-paradigm we implemented for nouns (smartNoun) is a very accurate "pluraliser" which handles most of the cases using pattern-matching. Incomplete nouns are used to compose noun phrases from determiners and nouns for example ("every person" is realised as "buri muntu" with the initial vowel of "person" removed).

3.4.2 Verbs

In R&R verbal inflection depends on tense,[1] Anteriority[2] (2), Polarity (2), Noun Class of Subject, Direct Object and Indirect Object markers (33 * 6 (Person and Number) each) bringing the total possible number of combinations to 124,198,272 inflections which are impractical to enumerate and cannot be handled by the GF compiler at the moment. Apart from the Subject marker (S), Object and Indirect Markers are optional because their use eliminates the need to mention the direct and indirect object(s) in declarative sentences of R&R. We therefore decided to cater for only Subject markers bringing the number down to 3,168 inflections. We found that this number was still prohibitive to successful compilation of the grammar. In light of the above, it was impossible to design a smart-paradigm for verbs. Our solution to the problem involved building the verb at sentence level by designing smaller tables

	Singular	Plural
Complete	omuntu	abantu
Incomplete	muntu	bantu

Table 3.3: The possible inflectional forms for the noun "omuntu" meaning person.

[1] We implemented using GF-specific language-independent tense system consisting of Past, Present, Future and Conditional)

[2] Anteriority is a phenomenon used to model grammatical aspect in a manner universal to all languages. It divides each tense into those in which the action is completed (Anterior) versus lack of completeness (Simultaneous)

```
param
  Number = Sg | Pl;
  NounState = Complete | Incomplete ;
  Gender = MU_BA | MU_ZERO ... RU_ZERO;
  Case = Acc | Nom | Gen;
  ConjArg = Nn_Nn | Nps_Nps | Pns_Pns | RelSubjCls | Other;
  AgrConj = AConj ConjArg;
  Agreement = AgP3 Number Gender | AgMUBAP1 Number
      | AgMUBAP2 Number | NONE;
  AgrExist = AgrNo | AgrYes Agreement;
  Position = Post | Pre;
  RCase = RSubj | RObj;
  RForm = RF RCase | Such_That;
  -- Possible Complement types held by a ClSlash
  ComplType = Nn | Ap | Adverbial | AdverbialVerb | Empty;
  VVFForm = VVImp | VVPerf | VVBoth;
oper
  Noun : Type = {s : Number ⇒NounState ⇒Str; gender : Gender} ;
  ProperNoun : Type = {s : Str; a : Agreement; isPlace : Bool};
  mkXClitic : Agreement →Str = \a →case a of {
      AgMUBAP1 n ⇒mkClitics "n" "tu" n;
      -- about 20—30 more table rows
      ...
      };
  mkXCliticTable : Agreement ⇒Str = table {
      AgMUBAP1 n ⇒mkClitics "n" "tu" n;
      -- about 20—30 more table rows
      ...
      };
  Adjective : Type = {s : Str; position : Position; isProper : Bool; isPrep : Bool};
  mkAdjective: Str → Position → Bool → Bool → Adjective =
      \a, pos, isProper, isPrep → {
          s = a; position = pos; isPre = True; isProper = isProper; isPrep = isPrep
      };
  Adverb : Type = {s : Str; agr : AgrExist};
  mkAdv : Str →AgrExist →Adverb = \str, agr → {s = str; agr = agr};
  NounPhrase : Type = {s : Case ⇒Str; agr : Agreement};
  VerbPhrase : Type = {
      s : Str; pres : Str; perf : Str; isPresBlank : Bool; isPerfBlank : Bool;
      isRegular : Bool; comp : Str ; comp2 : Str; ap : Str; isCompApStem : Bool;
      agr : AgrExist; adv : Str; containsAdv : Bool; adV : Str; containsAdV : Bool
  };
  Clause : Type = {
      s : Str ; subjAgr : Agreement; root : Str; pres: Str; perf: Str;
      isPresBlank : Bool; isPerfBlank : Bool; compl : Str
  };
```

Listing 3.1: Pseudo Code for Parameter, Record & Table Types & Operations in Resource Module

```
--Determiners can be lexical types or Phrasal Type
--especially through DetQuant,DetQuantOrd
Determiner : Type = {
    s : Str; s2 : Agreement ⇒Str; ntype : NounState;
    num : Number; pos : Position; doesAgree : Bool;
    firstFieldisEmpty : Bool; isQuant : Bool
};
-- prepositions sometimes have two kinds, near or far
-- i.e omu or omuri
-- Can be genetive
Preposition = {s : Str; other : Str; isGenPrep : Bool};
```

Listing 3.2: Category linearisation Types in StructuralCgg.gf

and morpheme-generating operations in Resource modules of both languages. These operations are simply used when necessary as we dynamically built the verb from the radical up to its full form. The operations are of the form “mkXClitics” and “mkXCliticTable” depicted in lines 20–29 of listing 3.1. The X stands for agreement concords obtained from (Mpairwe and Kahangi, 2013a). It should be noted that the verbal template example provided in Table 3.2 is very trivial because conjugation of the R&R verb *reeba* i.e. “to see” from Universal to Perfective form is easy. You simply replace the final vowel *“a’* with the morpheme *“ire”*. In actual sense there exists thirty-eight rules for converting a verb in the imperative mood to the perfective mood. The rules depend on the number of syllables in the verb (mono-, di- and tri- syllabic among others), the length of the penultimate vowel and the letters composing or modifying the terminal syllable such as ***-sa,-sh-,-za,-zya*** or the semi-vowels ***-w or -y***. This can be encoded as a smart paradigm for verb conjugation but the dictionary already gives the set of terminal letters of the verb that must be replaced with the right perfective ending. For example, the entry for the verb *gyenda* in the R&R dictionary by Mpairwe and Kahangi (2013a) is marked by *da-zire* to mean that in order to convert the imperative into perfective, replace the *“da”* in *gyenda* with *“zire”* to form *gyenzire*. We did not cover the full spectrum of grammatical aspects possible apart from those required for the language-independent implementation using the concept of Anteriority in GF-RGL. We aim to provide these aspects in a separate Tense / Aspect system within the GF-RGL as extensions in the future.

3.4.3 Determiners

In R&R it is impossible to express the definite and indefinite articles as distinct words. However Asiimwe (2007) suggests that definiteness can be expressed morpho-syntactically using the Initial Vowel on the noun and other constituents in the noun phrase. Demonstrative determiners are peculiar in that the word used depends on its position on the spatial dexis (Proximal, Medial and Distal), resulting in three words for each noun class. We chose to implement the former two as standard but leave the third form for implementation as an extension. The determiner agrees with number and noun class of the noun. Determiners can be derived from composition of other lexical types (such as Quantifiers and Numerals) via abstract functions: "DetQuant" and "DetQuantOrd" in the abstract syntax. This implies that these non-constant functions add complexity to the modelling of the determiner. Different determiners may appear either before or after the noun hence the need to have a field to track the position they take in Noun Phrases constructed for example by "DetCN" and "DetNP". For the linearisation category type we used a record within another record (refer to lines 3–7 in listing 3.2) . The string field in the outer record is for determiners that appear before a noun and do not inflect with the Noun while the table of Agreement to Strings inside the inner record is for demonstrative determiners which agree with the noun. The words "every"meaning *buri* and "much" meaning *-ingi* are examples of determiners that take "Pre" and "Post" positions of a noun. Additionally, we have to track whether the determiner 1) composes with either a Complete or Incomplete noun in the "nounCat" field, and 2) is obtained from one of the composing functions. This example for determiners demonstrates the kind of thinking process involved. This process necessitates redesigning types as one encounters new knowledge about the behaviour of Syntactic categories.

3.4.4 Adjectives

The two languages have two major kinds of adjectives; those that stand alone as their Indo-European counter parts and adjectival stems that require adjectival prefixes derived from the noun class particle of the noun they qualify (Mpairwe and Kahangi, 2013b). Stand–alone adjectives are of two types, those that require the use of possessive pronouns such as *ya* ("of" in noun class *MU_BA*). Some adjectival stems already exist but a large number can be derived from verbs that bear the same or similar semantic meaning of the adjective in mind. Derivation is done by affixing the conjugated copulative verb "ri" i.e. (Subject Prefix + ri) as a prefix to the verb. An example is "-ri-kutagáta" which is derived from the verb kutagáta (to be warm). Lastly, depending on the adjective, it can either occur before or after the nominal (noun/noun phrase). A summary of this information is given in Table 3.4 and the linearisation category type for Adjective is given on line 30 in listing 3.1.

Adjective Type	**Example**
Self-standing	Kaganga (Very Large)
Self-standing (Genitive Prepositional)	kijubwe (green) emotoka ya kijumbwe
Adjectival Stem	-rungi (nice) -kwostya (others)

Table 3.4: The various forms of the adjectives possible.

3.4.5 Numerals

We implemented Numerals for R&R by following the abstract syntax designed by Hammarström and Ranta (2004). This abstract syntax attempts to give a general yet prototypical representation of numbers of several languages taken from different parts of the world. Since numbers can be nouns, quantifiers, determiners, adjectives or adverbs, modelling them becomes difficult because we have to track agreement concords attributed to gender. Numerals are inherently nouns since they give names to entities used for counting (Ordinals) and order (cardinals). However, Numerals are also quantifiers of nouns i.e. they give an indication of how much or big other nouns are. Being a noun, each numeral belongs to a noun class and therefore has an initial vowel and a noun class particle. When used in quantification of other nouns, the numeral drops the initial vowel and acquires the prefix of the noun or noun phrase it quantifies. The agreement marker (Noun Prefix) acts as a prefix to the last word of the number. For instance, take the example "two hundred and forty people". The number "two hundred and forty" in R&R is *magana abiri na ana* while the noun phrase "two hundred and forty people" is *abantu magana abiri na ba-ana*. Some numerals can be pluralised while others cannot for example you can have "one 6" (*o-mu-kanga gumwe*) and "two groups of 6" (*emikanga ebiri*). The counting system is awash with synonyms attributed to the evolution of the language over time and the influence of English. The surface form of numerals depends on whether the numeral is Cardinal or Ordinal. When numerals are used in noun phrases the surface form of the number depends on the number and noun class of the head noun in the noun phrase. Therefore we modelled the numeral using tables to store the numeral with its various inflectional forms while keeping the gender and number information as record fields.

3.4.6 Phrasal Categories

Phrasal categories are derived from the combination of one or more lexical items. The rules for creating phrasal categories are declared in the abstract syntax as functions that take lexical categories as arguments. In GF-RGL abstract syntax, common nouns, proper nouns and pronouns by themselves can be noun phrases. They can also be formed from the combination of a determiner with a noun. The linearisation category type of the noun phrase (refer to line 37 in listing 3.1) stores all forms of the surface string dependent on case. A record field is used to store the agreement information for the noun contained in the noun phrase. Verb phrases are formed from verbs and their complements. Complements maybe noun phrases, adverbial phrases and adjectival phrases. The number of complements the verb may take are one, two or none. All this complement information is stored using fields in the record for verb phrase. In GF-RGL, the clause type is used as a phrasal category to store information for various components of a sentence i.e. Subject (usually a noun phrase) and Verb Phrase. We modelled the clause using a record structure that stores: the Subject as a string and agreement information to be used at the sentence level for determining the Subject marker located in the verb. At the sentence level, clauses are converted to strings according to tense, polarity and Simultaneity (GF-RGL way of covering aspect in language neutral way) to form actual strings for the sentence. Since we could not carry around big tables from Verb-level to Sentence level, we kept the different agreement concords in table structures that can be called upon when needed. The formation of sentence is perhaps the most complicated because morphemes for tense, aspect, polarity and subject markers within the verb must be determined and placed in their various positions according to the verbal template given in Table 3.2.

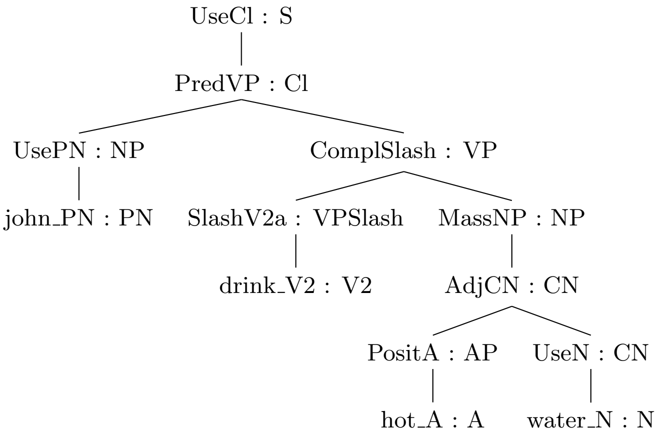


Figure 3.1: A GF abstract syntax tree generated from parsing “John drunk hot water”

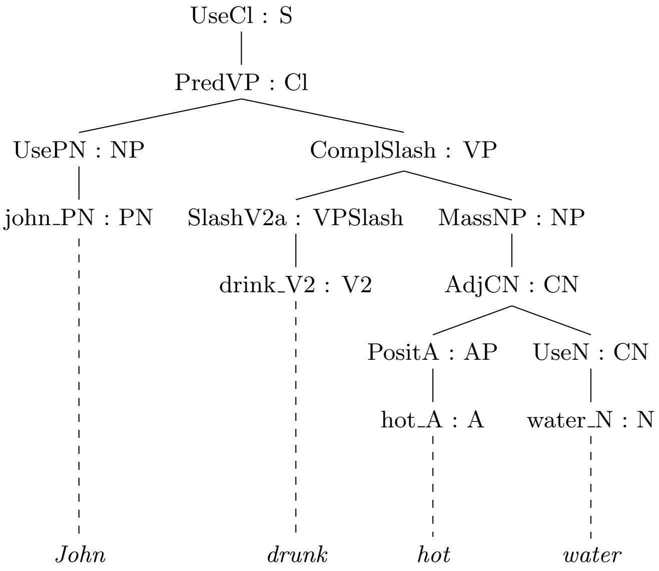


Figure 3.2: A GF concrete syntax tree generated from linearising the parse tree of 3.1 into English

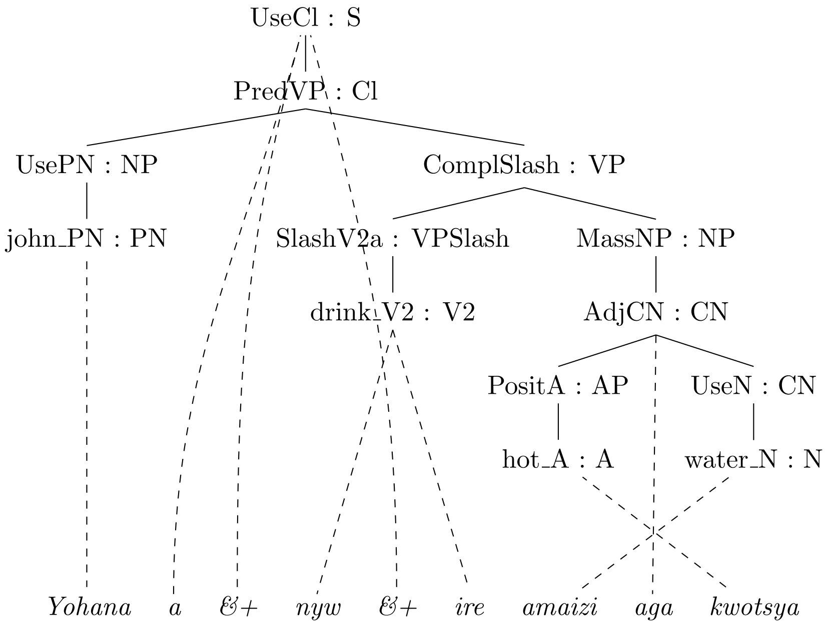


Figure 3.3: A GF concrete syntax tree generated from linearising the parse tree of 3.1 into Runyankore

3.5 An Example and Observations

In this section, we explain how an example GF abstract syntax tree depicted in figure 3.1 linearises (linearisation is the process of generating strings in a particular language from a parse tree) to Runyankore and Rukiga. The example was generated from parsing the English sentence "John drunk hot water" using GF for the purpose of generating a parse tree. Actually GF generated three parse trees but we chose just one of them for which we had all syntax functions implemented for both Runyankore and Rukiga in the RGL. GF generates *Yohana anywire amáàíîzi aga kwotsya* and *Yohana azáànywire amáàíîzi aga kwosya* as Runyankore and Rukiga linearisations for the abstract tree in figure 3.1. The nodes of the parse tree are GF-RGL syntax functions and their return types (the linearisation categories). When we linearised this abstract tree to English, Runyankore and Rukiga, we obtained concrete syntax trees for the languages in figures 3.2 and 3.3 for English and Runyankore respectively. We have left out the tree for Rukiga because it it is similar to that of Runyankore. The only difference is the spelling of "hot" being *kwosya* for Rukiga as opposed to *kwotsya* for Runyankore. The English concrete syntax tree is straight forward with each word from the sentence linearised from the leaves of the abstract syntax tree in figure 3.1. For the two R&R, a special bind symbol "&+" is used for concatenation i.e combining morphemes without spaces. Translation via GF is direct translation so the translations obtained may not be what a native speaker would use. However, they are grammatical i.e. they follow the syntax rules of the language. We made two observations about Runyankore & Rukiga from the parse trees: 1) concrete syntax trees are similar for the two languages and 2) the parse trees of Runyankore and Rukiga are more complicated in relation to English. The explanation for the first observation is that the grammar of the two languages are nearly identical with the exception of a few grammar rules and lexical items. The second observation stems from the fact that the languages are agglutinating resulting in several morphemes within a given word that are connected with grammatical features such as tense, aspect, mood, grammatical number, Person and noun classes. The function "play_V" responsible for linearisation of the verb play cannot have all its forms conjugated in a paradigm because of the millions of possibilities as discussed already in section 3.4.2, hence we decided to handle it at sentence level. While implementing the grammar of these languages, we also observed that Runyankore has more resources in terms of grammar books and dictionaries with most books concentrating on Runyankore as opposed to Rukiga.

3.6 Discussion

During the implementation of GF-RGL for Runyankore and Rukiga we observed that the difference between these languages lies only in a few lexical items. We therefore implemented Rukiga and reused its grammar for the implementation of Runyankore. The only changes we had to make were lexical items specific to Runyankore i.e those not shared by the two languages and a few rules for tenses. In total, we have implemented 290 abstract functions of which, 167 are lexical rules while 123 are phrasal rules. The missing rules consist of 400 lexical and 280 phrasal rules. We computed the 50 most used functions on wordnet and found that we implemented 43 of those functions which is not bad coverage. We plan to perform a proper evaluation in the future after compiling huge lexica and building application grammars for language-learning applications based on this GF-RGL. We simplified the verbal template by ignoring the use of the direct and indirect Object-markers because use of such markers would require anaphoric resolution, which occurs at the discourse rather than the syntactic level. GF-RGL's ability to do multilingual translation based on its universal abstract syntax prevented us from implementing all forms of lexical and syntactic categories because it would break multilingual translation. However, GF-RGL is flexible enough to allow the grammarian to implement language specific features as extensions, which we have done for structural words and intend to do for other syntactic categories. During the development of the grammar, we used regression tests by repeated linearisation of GF abstract syntax trees to English, Runyankore and Rukiga to check for grammatical correctness and ensure our changes did not break existing functions. Phonological conditioning is a particular problem for R&R which we have managed to solve only in our smart noun paradigm. A global solution would require development of morphological analyser and generator for the two languages.

3.7 Conclusion and Future Work

In this paper, we have described our work on the development and implementation of computational resource grammars for Runyankore & Rukiga Languages. We have succeeded in the modelling and implementation of the morphology and syntax of the languages using GF. The result has been a resource grammar for each language that together have been made freely made available under an open-source licence on GF's Github. In the near future we plan to: complete the Resource Grammar Libraries for the two languages by including language-specific tense and aspectual forms for verbs packaged as additional modules and development of morphological analysers and generators as efficient tools for handling phonological conditioning. We would also like to collect a corpus on which we shall perform an evaluation of the performance of the resource grammars developed. We are currently compiling a large computational lexicon for the two languages which shall increase the coverage of our lexicon. The increase in lexical coverage improves the quality of end user applications developed using resource grammars. Lastly, we will build application grammars in the domain of Computer-assisted language Learning for teaching learners of the two languages about the mechanics of the grammars of these languages.

3.8 Acknowledgements

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

Chapter 4

Paper B: A Computational Lexicon for Ry/Rk

This chapter is a reproduction of the following paper. The only thing that has changed is the formatting, no changes were made to the content.

D. S. Bamutura, 2021 “Ry/Rk-Lex: A computational lexicon for Runyankore and Rukiga languages.” *Accepted to the Northern European Association for Language Technology post-proceeding series of the Swedish Language Technology Conference (SLTC 2020)*

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 tasks such as: morphological analysis; part of speech tagging (POS); named entity recognition (NER); 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.

4.1 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 (hereafter 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.

4.1.1 Runyankore and Rukiga Languages

Runyankore and Rukiga (*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* (Byamugisha et al., 2016; Katushemererwe and Hanneforth, 2010b). They exhibit high incidences of *phonological conditioning* Katushemererwe et al. (2020) that makes them complex to deal with computationally. It is therefore more difficult to develop a computational grammar for these languages using symbolic approach. For details about the nominal and verbal morphology of these languages from the perspective of computational linguistics, the reader should see (Bamutura et al., 2020; Byamugisha, 2019; Katushemererwe et al., 2020; Katushemererwe and Nerbonne, 2013).

4.1.2 Challenges of Creating Computational Lexica for Ry/Rk

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 tasks such as: morphological analysis; part of speech tagging (POS); named entity recognition (NER); applications such as spell and grammar checking ; and cross-lingual information retrieval (CLIR).

4.1.3 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 4.2 previous related work. The data used for the study, its sources, curation and processing is provided in section 4.3. Section 4.4 describes Ry/Rk-Lex in terms how the different parts of speech were handled, the persistence structure used for storage of lexical items. Results & discussion are presented in section 4.5. Lastly, Section 4.6 presents the conclusion and future work.

4.2 Related Work

4.2.1 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

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

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 (Böhmová et al., 2003; Taylor et al., 2003; Xiao, 2008); 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 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.

4.2.2 Computational Lexicon Modelling

With regard to modelling of lexicons for Bantu languages, a Bantu Language Model (BantuLM) was put forward by Bosch et al. (2018, 2006) after eliciting the inadequacies of Lexical Markup Framework (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 (Mpairwe and Kahangi, 2013a; Museveni et al., 2009; Taylor and Mapirwe, 2009) and basic semantic relations found in thesaurus (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. We chose to use YAML[2] 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.

4.3 Data Sources, Curation & Processing

4.3.1 Existing Data Sources

In total, fourteen linguistic data sources summarised in table 4.1 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 are marked using * in that table. However, as explained later in detail in section 4.3.2.4, we used RRNews2013-2014 (marked with †in the same table 4.1) 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.

4.3.2 Data Curation & Processing

labeldata-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 4.3.2.3, 4.3.2.1, 4.3.2.2 and 4.3.2.4. The process for RRUDofHR and RREthics are identical to those described in section 4.3.2.2 and 4.3.2.4 respectively because the former was also scraped from the web while the later required scanning of a hard copy.

4.3.2.1 RRDict1959

To the best of our knowledge, there is only one MRD for Ry/Rk identified as RRDict1959 in table 4.1. 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, 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) 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

[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

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.

4.3.2.2 RRBibleNew1964

Since a digital version of the New Testament Bible in Runyankore-Rukiga (RRBibleNew1964) 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, part of speech (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 part of speech 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.

4.3.2.3 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 a Ry/Rk speaker. 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.

4.3.2.4 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

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

[5] Version: 221.001.20145 for Mac OS X

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 4.3.2.2 above.

4.3.3 Summing It Up

After pre-processing RRDict1959 to remove the first three anomalies mentioned previously in section 4.3.2.1, 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 New1964, 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.

4.4 Findings: Ry/Rk-Lex Description

The properties or features for each lemma depend on a number of factors but the major determinant is the part of speech (POS), the language to which the lemma belongs, 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 4.4.2. Otherwise, the properties peculiar to each part of speech are discussed in the following subsections. These properties are illustrated in table 4.2 which summarises the structure of Ry/Rk-Lex as specified in the schema[7] we developed whose structure is further described in section 4.4.1.

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

[7] See appendix I for the full structure

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 4.1: 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 4.2: 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.

	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 4.3: Examples of Runyankore and Rukiga nouns and their associated noun class particle pairs whose 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 4.4: Number of entries made per part of speech.

4.4.1 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.

4.4.2 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 Bamutura et al. (2020); Muzale (1998). 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 but implementing them as a rule-based computer program produced sub-optimal results although these rules are natural to a native speaker of the languages.

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

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

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

4.4.3 Common Nouns and Proper Nouns

In addition to all properties considered mandatory, we added noun class information as an additional field. We provide both the numerical noun classes and the textual noun class particles. We note that during our lexical collection work, we encountered three additional categories of nouns whose examples are illustrated in table 4.3 that do not fit in the conventional noun class system for Ry/Rk used by Byamugisha et al. (2016); Katushemererwe and Hanneforth (2010b); Turyamwomwe (2011).

4.4.4 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). We included nominal qualificatives (1)–(3) but excluded (4) and (5) 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.

4.4.5 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

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

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

4.4.6 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 (Byakutaaga et al., 2020; Morris and Kirwan, 1972; Mpairwe and Kahangi, 2013b; Taylor and Mapirwe, 2009).

4.4.6.1 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) (Katushemererwe et al., 2020; Mpairwe and Kahangi, 2013b). 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, 2013a,b) 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.

4.5 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 4.4. 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 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 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, we encountered nouns that did not fall under the accepted noun class numerical system. In table 4.3, 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 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 4.4.4, 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.

4.6 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 manual methods.

As future work, we plan to build and evaluate conjugation, lemmatisation, morphological analyser and generator, POS tagging software for Runyankore and Rukiga 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, we have used it to improve the lexical coverage of the computational resource grammars of Ry/Rk. There is also need to do more research on establishing a linguistically motivated and sound theory or criteria for word class division and 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, we are more focused on establishing 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]

[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

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 RR-SAWL2004.

Special thanks to the two reviewers who gave very detailed and helpful comments that have to a great extent been addressed.

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

Part III

Discussion, Conclusion and Future Work

Chapter 5

Results

This chapter describes my research contributions with respect to each paper by distinguishing the contributions made in the paper and those made after the paper was published.

5.1 Paper A

Paper A is a methodological type of paper in which we described our work on modelling, formalising and implementation of computational resource grammars for Ry/Rk. To a great extent, we succeeded in implementing the most important functions for the standard GF Resource Grammar Libraries (GF-RGLs) for these languages. The result has been a resource grammar for each language that together have been made freely made available under an open-source licence on GF's Github. This paper made the following contributions: a description of a methodological approach to modelling, formalisation and implementation of computational grammars for two Bantu languages in the JE10 zone; introduced a resource that can enable computers analyse and generate Ry/Rk, and do multilingual literal translation. Although the translation requires post-editing to achieve idiomatic translation for broad coverage translation usecases, idomatic translation can be achieved if the computational resource gramars are utilised as libraries by application grammars. The accuracy and precision improve because the application use-cases are domain-specific and care is taken by the designer to choose the correct lexical items, their inflections in relation to agrement with noun classes, grammatical number, tense and aspect etc.

5.1.1 Additional work after publication

Since computational grammar engineering and language resource creation is always a never-ending task, we extended the standard RGL by accounting for all the six tenses of Ry/Rk i.e. remote past, near past, memorial present, experiential present, near future and remote future. We reiterate that the immediate past mentioned by Turyamwomwe (2011) is equivalent to the memorial present mentioned in (Muzale, 1998) as we stated in Chapter 2, Section 2.3.3. In addition to work on tenses, the grammatical aspects namely; performative, perfect, resultative, Retrospective, habitual, progressive, and Persistive were also modelled, formalised and implemented for each tense where they applied. The tables of grammatical tenses against their aspects with example verbs and their glosses for Runyankore and Rukiga respectively provided by Muzale (1998, appendices. I-B3 and I-B4) were very instructive for this exercise.

5.2 Paper B

Paper B is also a methodological type of paper that described the semi-automatic creation of Ry/Rk-Lex, a computational lexicon for Ry/Rk. Since the task is also a never-ending one, we have now increased the number of lexical items from 9,429 to about 12,500 using only six out the fourteen sources we identified. Out of these sources, only five were utilised fully i.e. without special regard to any possibility of copyright violation because they are free from copyright. The text from the newspapers have been converted into a corpus but we have decided to take precautionary steps to only release random subsets of sentences extracted from the corpus. In order to store and make the lexical resource shareable, we designed a YAML schema for structuring the lexicon and used it to organise and annotate all lemmata that have been extracted from the data sources.

Chapter 6

Discussion, Conclusion and Future Work

6.1 Discussion

In this research study, we set out to model, formalise and implement the lexica and grammars of two under-resourced languages; Runyankore and Rukiga (Ry/Rk). The motivation was to: (1) to enable computers process (analyse) understand and generate utterrances in Ry/Rk; and (2) develop general-purpose computational lexical resources for Ry/Rk. We used Grammatical Framework (GF) and its Resource Grammar Library (GF-RGL) for the modelling and formalisation of the grammars of these languages. We simplified the verbal template by ignoring the use of the direct and indirect Object-markers because use of such markers would require anaphoric resolution, which occurs at the discourse rather than the syntactic level. GF-RGL's ability to do multilingual translation based on its universal abstract syntax prevented us from implementing all forms of lexical and syntactic categories because it would break multilingual translation. However, GF-RGL is flexible enough to allow the grammarian to implement language specific features as extensions, which we have done for structural words. We have also implemented the full tense and aspect system of the two languages. During the development of the grammar, we used regression tests by repeated linearisation of GF abstract syntax trees to English, Runyankore and Rukiga to check for grammatical correctness and ensure our changes did not break existing functions. Phonological conditioning is a particular problem for Ry/Rk which we have managed to solve in part within in our smart noun paradigm. A global solution would require development of morphological analyser and generator for the two languages.

Where as narrow-coverage lexicon of 167 lexical items was sufficient for grammar development we found that 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 constructed *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, at the time of writing this thesis, Ry/Rk-Lex currently consists of about 12,500 lemmata of various parts-of-speech. Ry/Rk-Lex 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 mentioned in Section 6.1.1.

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 (Museveni et al., 2012) 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 will be expanded in future after taking into account recent work done by Katushemererwe et al. (2020).

Parts of speech that belong to closed categories are few and consist of the most frequently used words. 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 that did not fall under the accepted noun class numerical system. In Table 4.3, 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 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 4.4.4, we found it difficult to identify and classify all forms of this part of speech.

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.

6.1.1 Use cases

Resource grammars are useful in domain-limited application areas such as Multilingual Document Authoring (Dymetman et al., 2000), low-coverage multilingual translation (Ranta et al., 2010), domain specific dialogue systems such as music players (Perera and Ranta, 2007), Computer-Assisted Language Learning (CALL) (Lange, 2018; Lange and Ljunglöf, 2018a,b) etc. Given the lack of reading comprhension and writing skills amongst Ry/Rk speakers, the resource grammar for Ry/Rk in the GF-RGL could be utilised to teach the grammar under the umbrella of Computer Assisted Language Learning (CALL) is another good use case.

Dialogue systems restricted to the domain of Patient-Doctor communication in a health facility can help medical personnel communicate effectively with patients especially in settings where there exists a language barrier between them. Such systems can help improve health outcomes through 1) doctor obtaining an accurate history for diagnosis, 2) patient gets a satisfactory explanation about the importance of adhering to advice and prescription. Another important use case in the area of localisation is the multilingual dissemination of weather information especially in multilingual societies.

Developing linguistic resources: By leveraging on public and freely accessible resources of well resourced languages supported by GF-RGL and using bootstrapping techniques and algorithms in (Kolachina and Ranta, 2016, 2019; Ranta et al., 2020; Ranta and Kolachina, 2017; Ranta et al., 2017).

6.2 Future Work

For future work, we plan:

S.1 To complete the RGL for the two languages and cater for the similarities, and collaborate with other researchers working on Bantu languages in GF.

S.2 To build application grammars to demonstrate the usefulness of the GF-RGLs developed and other linguistic resources for the two languages.

S.3 To build a small labelled / annotated parallel English-Runyankore-Rukiga bilingual corpus obtained from Bible text and Universal Declaration of Human Rights.

S.4 To develop a Runyankore-Rukiga UD treebank by leveraging the English, Runyankore and Rukiga GF-RGLs, UD to GF and GF to UD GF conversion tools developed in: (Kolachina and Ranta, 2016), (Ranta and Kolachina, 2017), (Ranta et al., 2017), (Kolachina and Ranta, 2019) and (Ranta et al., 2020).

S.5 To design and evaluate a machine-learned parser for Runyankore-Rukiga using the treebank obtained in S.4 above.

6.3 Final Conclusion

In this study, set out to carry out computational lexical and grammar engineering for Runyankore and Rukiga using a GF, a symbolic approach. This was justified by the lack of computational language resources that make computational linguistics and NLP research for these languages using data-dirven techniques possible with good results. We have suceeded at developing both lexical and grammar resources. We have therefore made contributions to the field in two ways: provision of previously non-existent resources and providing a methodological process of doing the same for other languages. We hope to extend this work in the next phase of research.

Bibliography

Kazimierz Adjukiewicz. 1935. Die syntaktische Konnexität. *Studia Philosophica*, 1:1–27. English translation "Syntactic Connexion" by H. Weber in McCall, S. (Ed.) *Polish Logic*, pp. 207–231, Oxford University Press, Oxford, 1967.

Allen Asiimwe. 2007. Morpho-syntactic patterns in Runyankore-Rukiga. Master's thesis, NTNU – Norwegian University of Science and Technology.

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.

Y. Bar-Hillel, C. Caifman, and E. Shamir. 1960. *On Categorial and Phrase-structure Grammars*. Weizmann Science Press.

Emily M. Bender. 2008. Grammar engineering for linguistic hypothesis testing. In *Proceedings of the Texas Linguistics Society X Conference: Computational linguistics for less-studied languages,*, pages 16–36, Stanford, CA. CSCLI Publications ONLINE.

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.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85 – 100.

Wilhelm Heinrich Immanuel Bleek. 1862. *A comparative Grammar of South African Languages*, volume 1. Gregg International Publishers Ltd., Farnborough, Hants, Eng.

Wilhelm Heinrich Immanuel Bleek. 1869. *A comparative Grammar of South African Languages*, volume 2. Trübner & Co., London.

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 Language 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).

Koen Bostoen and Yvonne Bastin. 2016. Bantu lexical reconstruction. In *Oxford Handbooks Online*, pages 1–31. Oxford University Press.

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.

Shirley Byakutaaga, Gilbert Gomushabe, Fridah Katushemererwe, Levis Mugumya, Edith Togboa Natukunda, Oswald K. Ndoleriire, and Celestino Oriikiriza. 2020. *Runyakitara Language Studies: A Guide for Advanced Learners and Teachers of Runyakitara*. Makerere University Press, Kampala, Uganda.

Joan Byamugisha. 2019. *Ontology Verbalization in Agglutinating Bantu Languages: A Study of Runynakore and Its Geralizability*. 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.

Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2018. Pluralizing nouns across agglutinating Bantu languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2633–2643. Association for Computational Linguistics.

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.

Robin Cooper and Aarne Ranta. 2008. Natural languages as collections of resources. In Robin Cooper and Ruth Kempson, editors, *Language in Flux: Relating Dialogue Coordination to Language Variation, Change and Evolution*. College Publications, London.

Ann Copestake and Dan Flickinger. 2000. An open source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Gérard Philippson Derek Nurse, editor. 2003. *The Bantu Languages*, chapter seven. Routledge.

Grégoire Détrez and Aarne Ranta. 2012. Smart paradigms and the predictability and complexity of inflectional morphology. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–653, Avignon, France. Association for Computational Linguistics.

Marc Dymetman, Veronika Lux, and Aarne Ranta. 2000. Xml and multilingual document authoring: Convergent trends. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING 00, pages 243–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

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).

Gerald Gadzar, Ewan Klein, Geoffrey k. Pullum, and Ivan A. Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge, MA.

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).

Malcom Guthrie. 1948. *The Classification of Bantu Languages bound with Bantu Word Division*, first edition. 9781315105536. Routledge, London.

Harald Hammarström and Aarne Ranta. 2004. Cardinal numerals revisited in gf. In *Workshop On Numerals In The World's Languages*, Leipzig, Germany.

John Hewson and Vít Bubeník. 1997. *Tense and aspect in Indo-European languages. [electronic resource] : theory, typology, diachrony.* Amsterdam studies in the theory and history of linguistic science: Series IV Current issues in linguistic theory v. 145. Amsterdam ; Philadelphia : J. Benjamins, c1997.

Thomas H. Hinnebusch, Derek Nurse, and Martin Mould. 1981. *Studies in the Classification of Eastern Bantu Languages*, volume 3 of *Sprache und Geschichte in Afrika: Beiheft*. Helmut Buske, Hamburg.

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

Gerhard Jäger and James Rogers. 2012. Formal language theory: refining the chomsky hierarchy. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 367(1598):1956–1970.

Ronald M. Kaplan Joan Bresnan. 1982. Introduction: Grammars as mental representations of language. In Joan Bresnan, editor, *The mental representation of grammatical relations*, MIT Press Series on Cognitive Theory and Mental Representation. MIT Press, Cambridge, MA/London.

Aravind K Joshi. 1985. How much context sensitivity is necessary for characterizing structural descriptions: Tree adjoining grammars. *Natural language parsing: Psychological, computational and theoretical perspectives*, pages 206–250.

Aravind K. Joshi, Leon S. Levy, and Masako Takahashi. 1975. Tree adjunct grammars. *Journal of Computer and System Sciences*, 10(1):136 – 163.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

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

Ronald M. Kaplan. 1997. Lexical resource reconciliation in the xerox linguistic environment. In *Computational Environments for Grammar Development and Linguistic Engineering*.

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

Festo Karwemera. 1995. *Empandiika ya Runyankore-Rukiga: egufuhaziibwe*. Fountain Publishers, Kampala, Uganda.

Fridah Katushemererwe and Thomas Hanneforth. 2010a. Finite state methods in morphological analysis of Runyakitara verbs. *Nordic Journal of African Studies*, 19(1):1–22.

Fridah Katushemererwe and Thomas Hanneforth. 2010b. 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.

Fridah Katushemererwe and John Nerbonne. 2013. Computer-assisted language learning (call) in support of (re)-learning native languages: the case of runyakitara. *Computer Assisted Language Learning*, 28:1–18.

C. Maria Keet and Langa Khumalo. 2014. Toward verbalizing ontologies in isizulu. In *Controlled Natural Language*, pages 78–89, Cham. Springer International Publishing.

Benson Kituku. 2019. Grammar Engineering for Swahili Language. *International Journal of Computer and Information Technology*, "08"("06").

Benson Kituku, Wangiku Nganga, and Lawrence Muchemi. 2019. Towards kikamba computational grammar. *Journal of Data Analysis and Information Processing*, 07(04):26.

Benson Kituku, Wanjiku Nganga, and Lawrence Muchemi. 2021. Grammar engineering for the Ekegusii language in Grammatical Framework. In *European Journal of Engineering and Technology*, volume 6, pages 20–29. European Open Science Publishing.

Sigismund W. Koelle. 1854. *Polyglotta Africana or Comparative Vocabulary of Nearly Three Hundred Words and Phrases in more than One Hundred Distinct African Languages.* Church Missionary House, London.

Prasanth Kolachina and Aarne Ranta. 2016. From Abstract Syntax to Universal Dependencies. *Linguistic Issues in Language Technology*, 13(3):1–57.

Prasanth Kolachina and Aarne Ranta. 2019. Bootstrapping UD treebanks for delexicalized parsing. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 15–24, Turku, Finland. Linköping University Electronic Press.

Herbert Lange. 2018. *Computer-Assisted Language Learning with Grammars. A Case Study on Latin Learning*. Licenciate thesis, University of Gothenburg, Sweden.

Herbert Lange and Peter Ljunglöf. 2018a. MULLE: A Grammar-based Latin Language Learning Tool to Supplement the Classroom Setting. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA '18)*, pages 108–112, Melbourn. Australia. Association for Computational Linguistics.

Herbert Lange and Peter Ljunglöf. 2018b. Putting control into language learning. In *CNL 2018: Sixth International Workshop on Controlled Natural Language*, volume 304 of *Frontiers in Artificial Intelligence and Applications*, pages 61–70, Maynooth, Ireland. IOS Press.

M. Paul Lewis, F. Simons Gary, and Charles D. Fennig. 2018. Uganda - Languages — Ethnologue.

Esther M. Lisanza. 2015. Language policies in east africa. In Eunice N. Sahle, editor, *Globalization and Socio-Cultural Processes in Contemporary Africa*, pages 121–145. Palgrave Macmillan US, New York.

Peter Ljunglöf. 2004. *Expressivity and complexity of the grammatical framework.* Ph.D. thesis, School of Computer Science and Engineering, Chalmers University of Technology.

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`.

Guthrie Malcom. 1967. *Comparative Bantu: An introduction to the comparative linguistics and prehistory of the Bantu languages.* Farnborough, Gregg.

A. E. Meeussen. 1980. Bantu lexical reconstructions. Reprint by Tervuren: Koninklijk Museum voor Midden-Afrika.

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

Carl Meinhof, Alice Werner, and Bernhard Struck. 1915. *An Introduction To The Study of African Languagess*. 9781110858910 1110858914. Kessinger Publishing, LLC. Translated from German to English.

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*. East African Literature Bureau Nairobi.

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

Y. 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*, chapter Verbal Morphology: Tense and Aspect in Runyakitara. Makerere University Press, Kampala, Uganda.

Wanjiku Ngángá. 2012. Building swahili resource grammars for the grammatical framework. In Diana Santos, Krister Lindén, and Wanjiku Ngángá, editors, *Shall We Play the Festschrift Game? Essays on the Occasion of Lauri Carlson's 60th Birthday*, pages 215–226. Springer Berlin Heidelberg, Berlin, Heidelberg.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Bengt Nordström, Kent Petersson, and Jan Smith. 1990. *Programming in Martin-Löf's Type Theory*. Oxford University Press.

Nadine Perera and Aarne Ranta. 2007. Dialogue system localization with the gf resource grammar library. In *Proceedings of the Workshop on Grammar-Based Approaches to Spoken Language Processing*, SLP '07, pages 17–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

Laurette Pretorius, Laurette Marais, and Ansu Berg. 2017. A GF miniature resource grammar for Tswana: modelling the proper verb. *Language Resources and Evaluation*, 51(1):159–189.

Rigardt Pretorius, Ansu Berg, Laurette Pretorius, and Biffie Viljoen. 2009. Setswana tokenisation and computational verb morphology: Facing the challenge of a disjunctive orthography. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 66–73, Athens, Greece. Association for Computational Linguistics.

Aarne Ranta. 2004. Grammatical framework. *Journal of Functional Programming*, 14(2):145–189.

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).

Aarne Ranta. 2011a. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

Aarne Ranta. 2011b. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford.

Aarne Ranta, Krasimir Angelov, Normunds Gruzitis, and Prasanth Kolachina. 2020. Abstract syntax as interlingua: Scaling up the grammatical framework from controlled languages to robust pipelines. *Computational Linguistics*, 46(2):425–486.

Aarne Ranta, Krasimir Angelov, and Thomas Hallgren. 2010. Tools for multilingual grammar-based translation on the web. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 66–71, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aarne Ranta and Prasanth Kolachina. 2017. From Universal Dependencies to abstract syntax. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 107–116, Gothenburg, Sweden. Association for Computational Linguistics.

Aarne Ranta, Prasanth Kolachina, and Thomas Hallgren. 2017. Cross-lingual syntax: Relating grammatical framework with Universal Dependencies. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 322–325, Gothenburg, Sweden. Association for Computational Linguistics.

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.

Lenhart Schubert. 2020. Computational Linguistics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, spring 2020 edition. Metaphysics Research Lab, Stanford University.

Gary F. Simons and D. Fennig, Charles. 2018. *Ethnologue: Languages of the world*, 21 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.

Müller Stefan. 2016. *Grammatical theory: From transformational grammar to constraint-based approaches*. Language Science Press, PB - Language Science Press.

Elsabé Taljard and E. Sonja Bosch. 2006. A comparison of approaches to word class tagging: Disjunctively vs. conjunctively written Bantu languages. In *Nordic Journal of African Studies*, volume 15.4. Nordic Association of African studies.

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. 1985. *Nkore-Kiga (Croom Helm Descriptive Grammars)*. Routledge Kegan & Paul.

Charles Taylor. 2008. *A handbook of Runyakore-Rukiga Orthography*. Fountain Publishers, Kampala.

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.

Charles V Taylor and Mpairwe Yusuf. 2009. *A simplified Runynakore-Rukiga-English DICTIONARY*. Fountain Publishers, Kampala.

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

K. Vijay-Shanker and D. J. Weir. 1994. The equivalence of four extensions of context-free grammars. *Math. Syst. Theory*, 27(6):511–546.

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.

Appendix A

YAML Schema for Ry/Rk-Lex

```
% YAML 1.2
---
$schema: "http://json-schema.org/draft-07/schema#"
name: YAML Schema for Ry/Rk—Lex
desc: |
    A schema describing the structure of Ry/Rk—Lex 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.
# listing continued next page
# listing continued here
        pos:
            type: map
            name: A mapping of pos tags at various levels
            mapping:
                 first_level :
                    type: str
                    required: true
                    enum:
                      - verb
                      - noun
                      - adjective
                      - adverb
                      - preposition
                      - pronoun
                the second_level:
                    type: str
                    required: true
            required: true
        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
```

Appendix B

Part of Speech and Morphological Tags

Part of Speech Tag (POS)	**POS description**
ADJ	Adjective
ADJC	Adjective Comparative
ADJS	Adjective Superlative
ADV	Adverb
ADVplc	Place Adverb
ADVtemp	Temporal Adverb
APPL	Applicative
AUX	Auxiliary
BEN	Benefactive
CARD	cardinal numeral
CIRCP	Circumposition
CN	Common Noun
COMP	Complementizer
COND	Conditional
CONJ	Conjunction
CONJC	Coordinating Conjunction
CONJS	Subordinating Conjunction (e.g when, although)
COP	Copula
DECL	declarative
DEF	Definite
DEM	Demonstrative
DET	Determiner
DIST	DISTAL
GLOSSES	**POS description**

Table B.1: Glossary of Part of Speech Tags and their description

GLOSSES	**POS description**
1	first person
2	second person
3	third person
ACTV	Active Voice
ADJ>N	Noun derived from Adjective
ASP	Aspect (underspecified)
AUX	Auxiliary (morpheme)
CL	noun class marker
MU	noun class MU. Equivalent to CL1??
BA	noun class BA
CL1	noun class 1
CL2	noun class 2
CL3	noun class 3
CL4	noun class 4
CL5	noun class 5
CL6	noun class 6
CL7	noun class 7
CL8	noun class 8
CL9	noun class 9
CL10	noun class 10
CL11	noun class 11
CL12	noun class 12
CL13	noun class 13
CL14	noun class 14
CL15	noun class 15
CL16	noun class 16
CL17	noun class 17
CL18	noun class 18
CL20	noun class 20
CL21	noun class 21
CL22	noun class 22
CL23	noun class 23
CLITadv	Cliticised Adverb
CLITdet	Cliticised determiner
ClITn	Cliticised Noun
CLITp	cliticised preposition
CLITpron	cliticised pronoun
CLITv	cliticised verb
CONSEC	Consecutive
DIM	Diminutive
DIR-SP	Direct Speech
DIST	Distal ’remote’
DIST2	Far Distal
DO	Direct Object
EMPH	Emphatic
EXPLET	expletive

Table B.2: Part 1: Interlinear glosses.

GLOSSES	**POS description**
FUT	Future
FUTclose	Close Future
FUTImmed	Immediate Future (Same as ???)
FUTnear	Near future
FUTrel	Relative future
FUTrm	remote future
FV	Final Vowel
GEN	Genitive
HAB	Habitual
IMP	Imperative
IMPF	Imperfective
IND	Indicative
IND-SP	Indirect Speech
INF	Infinitive
ITR	Intransitive
LOC	Locative
MAVM	Main Clause Affirmative
MEDIAL	Medial
N>A	Noun-to-Adjective
N>ADJ	derives an adjective from a noun
N>N	noun derivation
NOM	Nominative
OBJ	object
OBJ2	Second object
OBJcogn	Cognate Object
OBJind	Indirect object
OM	Object Marker
PASS	passive
PASThst	Hesternal past: yesterday or earlier but not remote
PASTim	Very recent, in the last minute or so
PASTpast	Past in the past
PASTrel	Relative past
PASTrm	Remote past
PFV	Perfective
PL	Plural
PNCT	Punctual ???
PRES	Present
PRF	Perfect
PROG	Progressive
PROX	Proximal
PRSTV	Persistive Aspect
PSSEE	Possessee
PSSOR	P

Table B.3: Part 2: Continuation of Interlinear glosses.