RIJKSUNIVERSITEIT GRONINGEN

Computational Morphology and Bantu Language Learning: An Implementation for Runyakitara

Proefschrift

ter verkrijging van het doctoraat in de Letteren
aan de Rijksuniversiteit Groningen op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op dinsdag 25 juni 2013
om 9.00 uur

door

Fridah Katushemererwe
geboren op 7 juli 1972
te Rubaga, Uganda

Promotor	:	Prof. dr. ir. J. Nerbonne
Copromotores	:	Prof. dr. A. Hurskainen Dr. R. Baguma
Beoordelingscommissie	:	Prof. dr. C.L.J. de Bot Prof. dr. M. Mous Prof. dr. G. van Noord

ISBN: 978-90-367-6346-2

Dedication

To God be the Glory, this work is dedicated to my family: Robert, Naomi, Jeremiah, Jerome, Jethro, Jenninah & Jotham.

Acknowledgements

This space cannot be enough to thank everyone who helped me during a journey of my PhD studies. I thank you all and I pray that God gives you the grace. Allow me; however, mention those I cannot miss to point out.

First of all, I thank God, the source of wisdom, who led me into all this and has given me the strength to keep 'right on up to the end of the road'. To Him be the Glory.

In a special way, I extend my infinite appreciation to my supervisors: Prof. John Nerbonne whose contribution to this work is invaluable; Prof. Arvi Hurskainen who laid a firm foundation to this work and continued polishing it up to the end; Dr. Rehema Baguma whose insights have greatly shaped this thesis; Dr. Thomas Hanneforth who enabled me to work on the basic but core part of this thesis; and Prof. Irina Zlotnikova whose knowledge and advice have contributed greatly to this work. I thank you all and may the Almighty reward you abundantly.

I also extend my gratitude to Prof. Venansious Baryamureeba and Dr. Jude Lubega for encouraging me to do a PhD and for working tirelessly to secure NUFFIC funds from which I have benefited to complete this dissertation. I thank Dr. Josephine Nabukenya, the Dean of School of Computing and Informatics Technology, and all the academic staff in the School for the conducive academic atmosphere and their availability for consultation.

My sincere thanks go to all NUFFIC administrators in Uganda and in the Netherlands for their endless support in ensuring that my PhD journey is pushed ahead. Special thanks go to Erik Haarbrink, Gonny Lekerveld, Peace Tumuheki, and Margreet van der Giezen for their tireless effort in making sure that my stay on a PhD program is comfortable.
I would like to thank DAAD (German Academic Exchange Service) and its administrators both in Uganda and in Germany for a six month research grant that enabled me do great work on this PhD program. Special thanks go to Dr. Gerald Heusing and Jennifer Schenk (DAAD 2009) who facilitated my stay in Germany up to the end of the grant period.

I wish to thank Dr. Sake Jager and André Rosendaal for their great academic contribution and for allowing me to use Hologram to test Runyakitara ideas. In the same vain, I thank Dr. Geoffrey Ondogah and Mr. Milton Kaye for developing a software interface for one of the prototypes in this dissertation.
I extend my sincere gratitude to Marjoleine Sloos who agreed unconditionally to be my paranymph during the defense ceremony at the University of Groningen. Marjoleine, thanks for your kindness.

I sincerely thank my respondents both in Kampala and other areas. Special thanks go to the staff and students of Runyakitara at Makerere University. Mention goes to Dr.

Celestino Oriikiriza, Ms. Allen Asiimwe, Mr. Gilbert Gumoshabe and Mr. Innocent Mugabe. Your insights contributed greatly to this dissertation.
I also extend my sincere thanks to my Colleagues at the former Institute of Languages, now School of Languages, Literature and Communication for their support during this research journey. Mention goes to Prof. Oswald Ndoleriire, Dr. Edith Natukunda, Mr. Ahmed Kaggwa, Dr. Susan Kiguli, Dr. Saudah Namyalo and Ms. Jane Alowo. Your encouragement kept me on the road.

My colleagues on a PhD program at Makerere University, University of Groningen, and University of Potsdam, I thank you. You were a wonderful source of encouragement. The PhD jokes you cracked, the different PhD songs and poems you shared were a major source of inspiration up to the end of the journey. Just to mention but a few, I sincerely thank Prossy Olango, Irene Nakiyimba, Gideon Kotze, Peter Nabende, and Florian Kuhn for your support and encouragement throughout the PhD journey.
Last but not least, my family members who missed my presence while I was away for my PhD studies. I am greatly indebted to my husband, Mr. Robert Tweheyo, my children: Musiimenta Naome, Nabimanya Jeremiah, Nabaasa Jerome, Naahurira Jethro, Naayebare Jenninah and Nareeba Jotham for being patient on this long journey. You were a great source of encouragement to continue and complete the PhD.

Table of Contents

Dedication ..... *iii*

Acknowledgements ..... *v*

Table of Contents ..... *vii*

Chapter 1 ..... *1*

General Introduction *1*

1.1 Introduction **1**

1.2 Background to the study **3**

1.2.1 Conceptual background 3

1.2.2. Theoretical background 5

1.2.3 Context of this study 13

1.3 Reason for the study 18

1.4 General and specific objectives 18

1.5 Structure of the dissertation 19

Chapter 2 Finite State Methods in Morphological Analysis of Runyakitara Verbs *21*

2.1. Introduction **21**

2.2. Runyakitara Verb Morphology and the Computational Challenge **22**

2.2.1 Number of morphemes involved 23

2.2.2 Excursus on morphological complexity 25

2.2.3 Morpheme combination 26

2.2.4 Morpheme order 27

2.2.5 Allomorphy 28

2.2.6 Vowel harmony 28

2.3. Formalization and Implementation **28**

2.3.1 The structure of RUNYAGRAM 30

2.3.2 Symbol signature 30

2.3.3 Word grammar 31

2.3.4 Context-dependent rewriting rules: morpho-phonological and orthographical rules 34

2.3.5 Sample output 35

2.4. Testing **37**

2.5. Conclusion and future research **38**

2.6 Future research **39**

Appendix A – Detailed description of Runyakitara Morphology **41**

Appendix B – *fsm2* Script for Creating the Verbal Analyzer **42**

Chapter 3 Fsm2 and the morphological analysis of Bantu nouns: initial experiences with regard to Runyakitara *43*

3.1 Introduction **43**

3.2 Previous work on the morphological analysis of the Bantu languages **44**

3.3 Methodology **45**

3.4 Highlights of Runyakitara noun morphology **46**
3.4.1 Runyakitara noun classification system 46
3.5 Formalization **49**
3.6. Implementation **51**
3.7 Application to Runyakitara nouns **52**
3.7.1 A symbol specification module 52
3.7.2 Noun grammar module 52
3.7.3 Morphotactics 54
3.7.4 Replacement rules 55
3.8. Grammatical output **55**
3.9. Testing **56**
3.10. Applications of the Runyakitara noun system analyzer **57**
3.11. Conclusion and future research **57**
3.12. Future research **57**
Chapter 4 RUMORPH: A morphological analyzer of Runyakitara - approach, results and issues ***59***
4.1. Introduction **59**
4.2 Runyakitara: a four-in-one cluster of languages **60**
4.3 Previous work on the morphological analysis of the Bantu languages **62**
4.4 Highlighted features of Runyakitara morphology and considerations for computation **63**
4.4.1 Verbs and their affixes 63
4.4.2 Nouns 67
4.4.3 Adjectives 68
4.4.4 Pronouns 69
4.4.5 Other word categories 69
4.5 Coverage/Scope **70**
4.6. Approach used in RUMORPH **71**
4.7 The Architectural structure of RUMORPH **71**
4.7.1. Symbol Specification/Signature Module 80
4.7.2. Grammar Module 80
4.7.3. Context-Dependent Rewriting Rules 76
4.8 Results and discussion **78**
4.8.1 Testing and evaluation 79
4.8.2 Error analysis 80
4.8.3 General issues 81
4.9. Conclusion and future work **82**
4.10. Future work **83**
Chapter 5 Language Teaching and Learning in Uganda: situation analysis and the need for Computer Assisted Language Learning (CALL) ***85***
5.1. Introduction **85**

5.2. Languages, language teaching and learning in Uganda 85
5.2.1 Foreign language teaching and learning in Uganda: Policy and practice 87
5.2.2 Primary, secondary and tertiary levels (excluding university) 88
5.2.3 University level 88
5.2.4 Local language teaching and learning in Uganda 90
5.3. Methods used in language teaching and learning in Uganda 92
5.4. Computer Assisted Language Learning (CALL) 93
5.5. Applications of CALL 94
5.6. CALL in Uganda 94
5.7. Untapped opportunities for CALL in Uganda 95
5.8. Needs assessment for Runyakitara CALL 96
5.8.1 Objectives 96
5.8.2 Study design 96
5.8.3 Study area and participants/subjects 96
5.8.4 Data collection and analysis 97
5.8.5 Presentation and discussion of results 97
5.9. Conclusion 101
5.9.1 Focused summary 101
Chapter 6 Computer-Assisted Language Learning of Runyakitara: A Pilot Study 103
6.1. Introduction 103
6.2. Runyakitara morphological structure and how it challenges language learners 105
6.3. Morphological analyzers as aids for the learning of morphology 107
6.4 Runyakitara morphology instruction in Hologram 107
6.5 Development 108
6.6. User study 110
6.7. Results and discussion 113
6.8. Conclusion 114
Chapter 7 Computer Assisted Language Learning (CALL) in support of (re)-learning native languages: the case of Runyakitara 117
7.1. Introduction 117
7.2. Motivation 118
7.3. Related research 119
7.4. Highlights of Runyakitara noun morphology and consideration for RU_CALL 121
7.5. RU_CALL: design and implementation 122
7.5.1 RU_MORPH (The Morphological Analyzer of Runyakitara) 123
7.5.2 RU_CALL tutoring module 124
7.5.3 Theory 125
7.5.4 Learner Performance Monitoring 125
7.5.5 Feedback 125

7.6. The RU_CALL system ... 125
7.6.1 User's view of the system ... 125
7.6.2 Learner ... 125
7.6.3 RU_CALL tutoring module ... 126
7.6.4 User Performance ... 127
7.6.5 Interface to morphological analyzer ... 128

7.7. Evaluation of RU_CALL ... 128
7.7.1 Study design ... 129

7.8. Results and Discussion ... 130
7.8.1 Results from experts ... 130
7.8.2 Results from learners ... 131
7.8.3 Learners' evaluation of RU_CALL ... 133

7.9. Conclusion and pointers to future research ... 133

Chapter 8 Toward a CALL system for Runyakitaran syntax ... 135

8.1. Introduction ... 135

8.2 Highlights of Runyakitara grammar focused on here ... 137

8.3. Design of the learning system ... 140
8.3.1 Morphological Analyzer ... 141
8.3.2 Dealing with ambiguity ... 142
8.3.3 Detection of spelling errors ... 144
8.3.4 Correction of word order ... 145
8.3.5 Correction of concord (Concord Module) ... 147

8.4. Learning applications ... 149
8.4.1 Interactive dialogues ... 149
8.4.2 Guided tours ... 150
8.4.3 Learning by means of interactive grammatical dialogue ... 151
8.4.4 Learning from guided tours ... 153

8.5. Conclusion ... 153

8.6. Future work ... 154

Chapter 9 Summary, conclusion and directions for future research ... 155

9.1. Conspectus ... 155

9.2. Contributions ... 157
9.2.1 Contribution to theory ... 157
9.2.1 Contribution to practice ... 158

9.3. Limitations of the study and future research ... 159

References ... 161

Samenvatting ... 175

Curriculum Vitae ... 177

Chapter 1

General Introduction

1.1 Introduction

This dissertation focuses on computational morphology applied to language learning, particularly with regard to the Runyakitara group of Bantu languages. Runyakitara shares with most Bantu languages a notoriously complex morphology which is a major challenge to language learners. Traditional language learning, with highly trained teachers, is infeasible due to the lack of trained teachers and the expense their work represents. However, before exploring the importance and applicability of this discipline to the learning of these languages, the importance of language in general should first be highlighted.

Today, where development depends on knowledge and where knowledge is delivered through language, language learning at all levels of human development is crucial. For people to share knowledge, they must share a language. If they do not share a language by virtue of birth or upbringing, then at least some of them need to learn the others' language. Computational morphology is a core component in natural language processing and has previously been applied in language learning.

Worldwide, language is important in facilitating human-life enjoyment. It is a social instrument which facilitates and enriches communication amongst all human beings (Zahram, 2001). Some education experts have argued that "Language is not 'Everything' in Education, but without language, 'Everything' is nothing in Education" (Alidou et al., 2006). This underscores the importance of language skills in education.

Language is also important, if any meaningful development is to be realized. According to Wolff (2005), there is a strong relationship between language and development, which is either ignored or not understood by many policy makers and leaders. Stressed by Crystal (1997), language is an important object of study because of its unique role in capturing the breadth of human thought and endeavour. Crystal argues that human beings are able to see back as well as plan ahead through language. Briefly, without language, there is little intelligent activity in human life.

Language learning is important to meaningful learning at all levels of schooling and in all subject areas. It is also important in shaping the learner's cognitive, emotional and social development. It has been observed that incompetence in basic language skills leads to difficulty in learning at later stages or levels such as upper primary and secondary as well as post-secondary levels (Kingston, 2003). Language learning not only supports educational tasks but also facilitates communicating in other languages in addition to a

first language. This creates a foundation for intellectual growth in other languages and cultures.

While the importance of languages is well appreciated, most African languages are less studied/documented and regarded as inferior – even by native speakers themselves. For example, some have argued that African languages cannot handle scientific terminology (Asiimwe, 2008). Due to this background, some of these languages are even verging on imminent extinction, possibly resulting in the collapse of their community's social and economic system (Crawhall, 1998). Zahram (2001) elaborates that, in most African countries, the study of English literature and language is, for example, preferred to the study of Kiswahili or any other African language because English is regarded as a key to social success in society both domestically and abroad. Such attitudes have contributed to the slow progress in the linguistic description and technological implementation of African languages. In addition, most countries do not use African languages as languages of instruction in education. As a consequence, some parents do not consider them as important languages for the success of their children (Alidou et al., 2006).

Although there has been a gradual change of attitude towards African languages, as well as progress in their description and technical processing, there are still many neglected languages and cultures in Africa today, and a lot needs to be done. At the same time, it has been recognized that modern development relies on scientific and technical knowledge, which comes to Africa through foreign languages (Zahram, 2001). It is also acknowledged that, for any development to take root, the majority of the population must be involved, and the majority of Africans do not speak foreign languages. Basic literacy skills in native languages are surely an auspicious basis for literacy in a foreign language. In this situation, literacy programs for indigenous languages should involve large numbers in national development programs. Therefore, efforts to develop indigenous languages should be supported.

As the African continent proceeds towards meeting the millennium development goal of "education for all", most countries in Africa, including Uganda, are using local languages for instruction at the lower levels of primary education. The objective of such practices is to provide knowledge to learners in a language and culture with which they are familiar. In such a strategy, materials for language learning become an important requirement for instruction in these languages. Investing in language learning resources for local native languages, specifically African languages will support the success of such a strategy. This research is, in fact, largely motivated on this basis.

For a long time, instructional books and teachers have occupied the central position in methods of language learning, and many think that they are adequate in this regard. This dissertation will not advance arguments that books and teachers need to be replaced by software, and conservative education administers should be wary of any such radical suggestions. But textbooks and teachers are in short supply, and the demand is greater than may be appreciated. In the current situation, there are no adequate descriptive textbooks for most African languages, and there are seldom trained language teachers for any but the largest languages. We turn to software development as a step toward

satisfying the demand for language-learning materials, and we aim for this to be useful not only in situations where there is no alternative, but also in those situations where there is a sufficient supply of textbooks. In those situations, too, educators often feel a need to prepare supplementary materials based on their own insights and experience in order to deal with unpredictable events that arise in the process of language learning (Mugane, 1997; Lai & Kritsonis, 2006).

We hinted at a hidden demand for language-learning materials in the last paragraph. We believe that such a demand exists from experience in offering language courses in Kampala, where we have noted that many students are the children of parents who migrated to the capital city to seek work and were thereby cut off from learning their first languages completely. We return to this group in Chap. 7. These learners represent a sizable group who have attracted the attention of scholars of language endangerment, but less the attention of second-language learning specialists.

Providing and improving computational language learning resources may have an impact on the existing language situation in Africa, facilitating literacy in native languages and thereby allowing citizens to access to government services using languages they understand. Improved native language proficiency also improves the communication among the citizens of a country and better enables them to exploit its (normally) rich cultural heritage (Prah, 2008).

1.2 Background to the study

The previous section has provided a motivation for studying languages, while focusing on the role of African languages in the global arena. This section introduces the reader to the concepts, theories and context of the study.

1.2.1 Conceptual background

Three major concepts shall be introduced here: language learning, morphological analysis, and computational morphology.

Language learning involves two concepts: learning and language. We assume that the notion 'language' is familiar enough to require no special discussion. Learning as a "process of acquiring modifications in existing knowledge, skills, habits, or tendencies through experience, practice, or exercise",[1] and we focus in this dissertation on what is involved in learning a language when the language is not (thoroughly) learned in childhood. We have foremost in mind people who learn second and further languages as adults, and we will pay a bit of attention to learners whose learning of a first language was interrupted.

[1]The Columbia Electronic Encyclopedia® Copyright © 2007, Columbia University Press.

According to Krashen (1981), adults have two distinctive ways of developing linguistic competence: learning and acquisition. For Krashen, language learning is a conscious process which results in conscious knowledge about the language (e.g. knowledge of grammatical rules), whereas language acquisition is a sub-conscious process resulting in sub-conscious knowledge of the language. This latter process is similar to how children acquire their first language. Although Krashen's theory has enjoyed a lot of popularity in second language teaching, it has also attracted a lot of criticism based on empirical research and peoples' experience, as practical data indicates that accurate language competence requires conscious knowledge of grammatical rules combined with practice (Gregg, 1984). We also adopt the view that choosing to focus on learning or acquisition as defined by Krashen (1981) should depend on the targeted knowledge or skill in a language. Language by nature is a complex multi-faceted system. It consists of a wide array of skills ranging from listening, reading, writing and speaking as well as the knowledge of vocabulary, grammar, etc.

We borrow from Krashen's distinction the insight that discursive knowledge about language, such as that found in a dictionary or reference grammar, is insufficient to support the active use of the language for speaking, listening, reading and writing. Language learners must actively practice language skills if they are to be automated to the point needed for straightforward communication. This means that extensive practice material is indispensible, and we shall aim to provide that through computer-supported exercises.

Language learning is a multi-level task that integrates elements such as words, syntax, pronunciation, interaction and culture (Heilman & Eskenazi, 2006). Some parts of language learning, for example, learning how to read, are very different from more structured domains, since tasks such as reading may involve tens of thousands of knowledge components – words, constructional patterns within words (morphology), and constructional patterns among words (phrases) – rather than a few hundred (Heilman & Eskenazi, 2006). Research has further established that the set of grammatical or lexical items that a language student must know is very large. It is also often difficult to accurately assess the importance of knowing any single item because of the various contexts in which words may occur.

We are particularly interested in reading and writing. For anyone to be considered literate in a language, he/she must have the ability to combine individual words in the ways required to make phrases, to combine phrases into sentences and to arrange sentences into paragraphs. He/she must be able to communicate not only directly and face to face but also over long distances using the more permanent medium of writing. To gain such ability, learners must have knowledge of a language system and be able to use it to communicate effectively.

However, such ability cannot be gained overnight. Learning a language is a complicated process, a point previously stressed by many (Heilman & Eskenazi 2006). It has been established that learners of normal intelligence with a strong foundation in their first language require continuous language exposure over five to seven years to gain peer-

level language proficiency (Gülitz, 1996). Therefore, no single investigation can claim to cover the task of learning an entire language system; instead, research in language learning adopts different foci, each of which contributes important knowledge about the overall process.

Morphological analysis is the process of breaking down words into their constituent meaningful parts called morphemes (Bellomo, 2009). Consider the English word "reader", which is comprised of two meaningful units – the base *read* (the act of instruction) and *er* which conveys the meaning of an agent (a person or object) who performs the action in the base. Thus, the reader is the one that reads. Studies have shown that understanding morphemes can significantly enhance vocabulary, reading and grammatical accuracy (Casalis & Louis-Alexandre, 2000; Bellomo, 2009).

Computational morphology, the field associated with the focus of this study, is the branch of Computational Linguistics concerned with automatic word analysis and generation (Gasser 2009). In morphological analysis, a word form is analyzed into a lexical representation, consisting of the word's component morphemes; "going" can, for example, be analyzed as "going: go[V-Root] ing[V-suf-prog]." This indicates that "go" is a root and "ing" is a "suffix", which therefore follows the root and that grammatically marks its "progressive" aspect. In morphological generation, a lexical representation is converted to a surface word form (e.g. "go[V-root] ing [V-suf-prog] => going"). Morphological analysis identifies the meaningful components of a word. An automatic morphological analyzer is a software component/system that takes a word as its input, breaks it into its smaller meaningful components and outputs it with its linguistic tags (Beesley & Karttunnen, 2003). This piece of software can be employed in a number of other important linguist applications, including language learning.

1.2.2. Theoretical background

The study of language learning has been of interest to many researchers from a wide range of disciplines, such as psychology, sociology, culture and linguistics, just to mention a few. For example, Phil's[2] "English for Foreign Learners" support site lists 50 theories of language learning. This abundance indicates the importance and interest associated with the subject. The cross-disciplinary nature of our topic leads us to draw on theories and models from morphology, computational morphology and language learning. It should be noted that the wide range of theories advanced in the disciplines reflect the fact that researchers have failed to agree on several important points.

The objective of this section is not to suggest new theories or hypotheses, plenty of which are already available, but rather to provide a basic understanding of the theories whose principles we have adopted to underpin our study. We discuss these points because a sound theoretical underpinning is likely to improve the quality of a Computer Assisted Language Learning (CALL) program (Ma & Kelly, 2006; Jager, 2009). At the same time,

[2]www.philself/support.com/learning.htm

the underlying theoretical principles constitute a very important component of the methodology.

We discuss morphological theories, computational morphology and language learning theory in succession in the remainder of this subsection.

a) Morphological theories

There are quite a number of morphological theories, but the relevant theories used in this study are the Item-and-Arrangement and Item-and-Process theories of Hockett (1954, 1958), bearing in mind that there is no single theory that can entirely explain the morphological system of a Bantu language.

The Item-and-Arrangement (IA) theory

In Item-and-Arrangement (IA) theory, words are formed from unambiguously delimited items called morphemes and certain arrangements for the ordering of these items/elements called rules. In this model, each piece of morpho-syntactic material is paired with some morphological information. The IA view is supported and enhanced by Distributed Morphological Theory (Halle & Marantz, 1993). To elaborate with an example, implementing IA theory enables us to divide the word *cats* into two components: a root and a plural marker:

i) root [+plural]

/cat/ /s/

This indicates a one-to-one mapping of morphemes, minimal meaning-bearing items and morphs or forms that are pronounced. The major shortcoming of the theory is that the mapping of morpho-syntactic information and phonological information is not always a one-to-one relationship. This is exemplified in most languages of the world. For example, the noun *omuntu* (person) in the Runyakitara languages contains the morpheme *mu*, which provides two sorts of morpho-syntactic information: class prefix and number. Thus,

ii) /o-mu-/ /ntu/

class1 singular root

Hockett himself was quick to realize that morphemes do not always occur in a one-to-one relation with forms and proposed another theory called Item-And-Process (IP).

Item-and-Process (IP) Theory

Instead of using morpheme combination as the sole basis of word formation processes, Hocket (1958) also proposes that words are derived from the operations of abstract rules,

commonly called word formation rules. Thus, a root can be paired with a set of morpho-syntactic features as illustrated below:

[+N]
[+Pl]
/cat/ → /cats/ - (*cats* is a plural form of *cat.)*

It is important to note here that from the IP perspective the resultant /cats/ is a single piece, not a composite of two morphs. IP has been defended and enhanced in the works of Anderson (1992) and Aronoff (1994).

In our research, we have specifically integrated ideas from the two theories (IA & IP) in the Runyakitara morphological analyzer. The morphological system of the Runyakitara group of languages is predominantly a morpheme-based system, however, so that IA is the more relevant perspective. Phenomena which are not compatible with IA are then handled by employing IP principles. The integration of the two theories is the basis of the Runyakitara morphological analyzer discussed in later chapters of this dissertation.

b) Finite State Morphology

Commonly referred to as automata theory, finite state theory deals with the mathematical modelling of abstract machines. A finite state machine (network) is represented as an abstract machine that accepts input symbols, generates output symbols and changes its inner state in accordance with some predefined plan (Beesley & Karttunen, 2003). As defined by Hopcroft et al. (2001), finite state machines share the following characteristics:

a) a finite set of defined states, one of which is defined as the initial state of the machine, and a subset of which are final (or accepting) states;
b) a set (alphabet) of defined inputs;
c) a set (alphabet) of defined outputs;
d) a set of transitions between selected states, each responsible for reading, writing or transducing a fixed amount of input and/or output; and
e) a shared single state at any instant of time.

Our study draws on certain notions from finite state linguistic theory (Johnson, 1972; Kaplan & Kay, 1994; Koskenniemi, 1983). Johnson (1972) was the first to realize that finite state machines could be used to model phonological phenomena. Kaplan and Kay independently re-discovered much later that finite state theory can simplify the modelling of phonology and morphology. Koskenniemi's influential work on two-level morphology (Koskenniemi, 1983) was popularized by Karttunen (1983) and led to the wide use of finite state machines in the development of morphological analyzers.

For finite state linguistic researchers Beesley & Karttunen (2003), such analysis would do the following:

1. Represent, a "language" as a set of strings in a simple finite state network consisting of states and arcs that are labelled by atomic symbols. This can be illustrated below with a Runyakitara string: tu-shom-e (let us read).

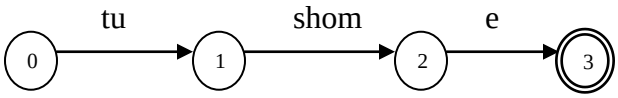


Fig.1-A simple finite state network accepting one string (word)

Where 1,2,3, are states, ⟶ are transitions, 0 is the initial state and ◎ is an accepting state. Note that the machine was in the state 0 before beginning to accept any string. When a string ***tu*** was input, the machine transitioned to 1, then on to 2 as another string was input, and still further until it reached an accepting state 3.

2. Represent a "relation" as a set of ordered pairs of strings in the form of a finite state transducer. The arcs of the transducer are labelled by pairs of symbols. Each path of the transducer represents a pair of strings in the relation. The first element of each ordered pair belongs to the input (upper) language, while the second belongs to the output (lower) language of the relation.

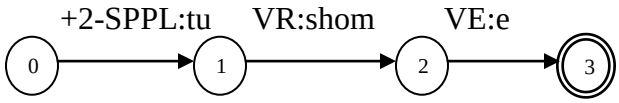


Fig. 2 - A simple finite state transducer for *tushome*

The arcs are now represented as pairs of elements, the first of which is abstract and denotes grammatical information, while the second is the form, also used in Fig. 1

3. Construct new languages and relations using set operations such as union, concatenation and composition. These operations can also be defined for finite state networks to create a network that encodes the resulting language or relation.

Simplified by Beesley & Karttunen (2003), a lexical network and transducer rules can be combined into a single network (lexical transducer) containing all the morphological information about a language (morphemes, derivation, inflection, compounding, etc.).

The three steps above form the basis of our Runyakitara morphological system. We emphasize that the task of the morphological system is to provide well-formed Runyakitara words, given abstract specification of their grammatical properties (1^{st} pers. Singular, preterite form with pronominal object of the 12^{th} class), or conversely to provide the grammatical properties given the word. The morphological system is not a model for Runyakitara speakers except in that it encodes this knowledge, nor does it in any way seek to model Runyakitara learners and the stages of learning they progress through.

c) Language Learning Theories

In the context of Computer Assisted Language Learning (CALL) design, Ma & Kelly (2006) view language learning theory as a general term referring to the program designer's assumptions about the nature of language, language learning and the process of learning. In addition, some theoretical consideration has to be given to the type of learner involved, as that may also greatly influence CALL program design.

Language learning theories are numerous, and we accept the argument by Ma & Kelly (2006) that any choice of specific language learning theory as the appropriate background from which CALL software should be designed an implemented should depend on the particular elements of language knowledge or skills on which the CALL program would like to focus. For example, CALL programs for vocabulary learning should primarily be based on learning theories or research findings specific to vocabulary learning. In line with this view, the following questions shall guide the discussion of the theories relevant to our study:

- Who is the learner in this research?
- What is a learner supposed to learn? (nature of language and language elements to consider)
- How does a learner learn a language? (process of language learning)

Since we should prefer that our software development work be as useful as possible, we also prefer that it rely as little as possible on differentiating aspects of language learning theory. In this way we hope that the practical work will be widely useful even if one or another aspect of the learning theory that informs it turns out to be disfavoured – or outright wrong. We nonetheless review the usual questions asked to be as forthcoming as possible about the development phase and the assumptions that informed it.

Who is the learner?

In trying to answer the above question, theory acknowledges a range of learners with a variety of characteristics. In terms of age, learners can be children or adults. Adult learners can be students or professionals. Learners can also be people with special needs, such as people with disabilities. In this study, we concentrate on adult learners; therefore, theories of adult learning apply.

Cross (1981) presents the Characteristics of Adults as Learners (CAL) model in the context of her analysis of lifelong learning programs. The CAL model consists of two classes of variables: **personal characteristics** and **situational characteristics**. Personal characteristics include: age, life phase and developmental stage. These three dimensions have different characteristics as far as lifelong learning is concerned. Advanced age, for example, results in the deterioration of certain sensory-motor abilities (e.g. eyesight, hearing, reaction time) while intelligence abilities (e.g. decision-making skills, reasoning, vocabulary) tend to improve throughout adulthood (before old age). Life phases and

developmental stages (e.g. marriage, job changes, retirement) involve a series of plateaus and transitions which may or may not be directly related to age. Situational characteristics consist of part-time versus full-time learning, and voluntary versus compulsory learning. The administration of learning (i.e. schedules, locations, procedures) is strongly affected by the first variable, while the second pertains to the self-directed, problem-centred nature of most adult learning.

Given the current state of curriculum development, it is difficult to specify exactly the learner of Runyakitara who we aim to assist. Our assumption is that digital instruction in the Runyakitara languages can serve native speakers, such as those wishing to improve their literacy (i.e. proficiency in reading and writing) and those otherwise interested language learners. It should be clearly noted that children are not a focus of this study. Therefore, the learner is assumed to be an adult, whether or not a student. In chapter 7 below we shall pay special attention to an unusual group of learners we have encountered, who may constitute a new field of application for CALL. These are adults whose native language learning was interrupted when their parents migrated to a large urban center in search of work. Our reading of the literature on language endangerment suggests that this may be a substantial number of people. As learners they are unusual in that they have some ability in a "native" language, but much less than "native speakers" are normally assumed to have.

What is a learner supposed to learn?

In its broadest sense, the learner is learning a language. This involves learning a wide array of knowledge (vocabulary, grammar and discourse) and skills (listening, reading, writing, speaking and translation). The training needed is broad and includes correct pronunciation, rules of grammar and vocabulary. In short, language learning involves learning the sound system (phonology), word formation system (morphology) sentence formation system (syntax), semantics (meanings) and use (pragmatics).

It is important to focus on one or a few aspects of language learning because language learning is complex. Our focus is on learning selected, particularly complex elements of Bantu grammar relevant to the Runyakitara language group. Bantu languages are difficult to learn, particularly when it comes to the concord system and noun classes. Research has established that, to effectively learn a Bantu language, it is important to learn the structure, which is unusual for most learners (Hurskainen, 2009; Taylor, 1985). This is also true for other languages with complex morphology, such as Turkish. For example, Kuruoğlu et al. (2000) argue that providing a solid structural basis greatly benefits the learners. Kuruoğlu et al. (2000) note that more communicative techniques can only be employed at advanced levels when learners get more comfortable with the structure of the language. This is re-emphasized by Amaral (2006) who argues that it is difficult to target communicative goals until students have mastered the appropriate language forms and rules.

Jager (2009) also notes that many language instructors who explicitly advocate for communicatively oriented instruction nonetheless turn to form-based exercises as

supplementary activities. This dissertation follows Jager (2009) in assuming that there is a role for exercise drills in language learning even if the ultimate goal is the development of communicative skills, and even if the dominant form of *classroom* exercise (as opposed to supplementary activity) is practice in situationally specific communication.

How does a learner learn a language?

Given the nature of our study, we, as language instructors, integrate principles from different theories to underpin our research in the language learning process. We employ conversational theory from cognitive theorists, guided discovery learning from constructivist theorists and some principles of Behaviourist theory. In other words, this research borrows principles from the mentioned theories to design the language learning model for selected elements of Bantu grammar, adopting the Runyakitara languages as a specific composite case.

Conversational Theory (CT)

The original idea of conversational theory was developed by Pask, a cybernetics specialist, between 1966 and 1996. The theory was later enhanced and supported by many researchers based on empirical studies (Ford, 2001; Pangaro, 2001; Pask. 1975). The main concepts of the theory are "conversation" and "understanding". The theory conceives understanding as that which results from a conversation between different conceptual participants (p-individuals), which may or may not correspond to mechanical participants (m-individuals, including people and machines) (Ford, 2004). Ford (2004) summarizes the major points of conversational theory that stimulate this study:

> A conversation consists of interactions between p-individuals in which both agree on the nature and derivation (the "why" and the "how") of one or more concepts. Where agreement is reached, the concept can be shared by both in further intellectual activity. Differences may result in new concepts being available to both. This is the intellectual activity that results in changes in the individual's knowledge structure that we refer to as "learning" and generates what we refer to as "information needs".

Pask's CT has informed many educational methods and technologies across many disciplines, such as Information Systems, Education and Psychology. In our interpretation, CT emphasizes the fact that learning can occur through conversation. We focus therefore on a teacher (system) and a learner (human) engaging in conversation.

Guided Discovery Approach

Guided discovery approach has its roots in the discovery learning theory of Bruner (1966). Ormrod (1995) describes discovery learning as "an approach to instruction through which students interact with their environment by exploring and manipulating objects, wrestling with questions and controversies, or performing experiments". It is a form of inquiry-based, constructivist learning theory that takes place in problem-solving

situations where the learner draws on past experience and existing knowledge to discover facts, relationships and new truths. Bruner's theory of instruction is based on the following principles:

- Experiences should be designed to help students become willing and able to learn.

- Knowledge should be appropriately structured, by which Bruner means that educators should determine how a body of knowledge should be structured in order to facilitate understanding by learners.

- Any domain of knowledge or problem or concept within that domain can be represented in three ways or modes: a set of actions, a set of images or graphics that stand for the concept and a set of symbolic or logical statements.

- The nature and pacing of rewards and punishments should be specified. Bruner suggests that movement from extrinsic rewards, such as teacher's praise, toward intrinsic rewards inherent in solving problems or understanding the concepts is desirable. Feedback to the learner is critical in the development of knowledge.

Implications of theory to this research

i) Integration of multiple theories and approaches

Various researchers tend to view theoretical issues through a mono-disciplinary lens, which means that morphological theories have been treated differently from language learning theories. Given the interdisciplinary nature of the topic and the benefits of providing an analytical framework for the issues at hand, it is important to integrate relevant theories and associated findings from various empirical studies. This requires the adoption of an interdisciplinary approach to the study of computational morphology and Bantu language learning.

ii) Development of a Runyakitara learning model

The relationship between morphological analysis and language learning has been extensively explored in both theory and practice. Outside the computational domain, morphological analysis has been applied to the teaching of reading skills (Nerbonne & Smit, 1996; Keiffer & Lesaux, 2007) and vocabulary (Osburne & Mulling, 2001; O'Sullivan & Ebel, 2004; Bellomo, 2009). All the research mentioned has been centrally concerned with the manner in which knowledge of word formation (understanding the combination of roots and affixes) enhances vocabulary acquisition and reading comprehension.

In computational work, research has confirmed that automatic morphological analysis enhances intelligent feedback on word forms (Amaral & Meurers, 2006), aids in vocabulary acquisition and reading (Nerbonne et al., 1998), and facilitates the learning of grammar (Hurskainen, 2009).

Adding to what theoretically and empirically exists in the literature, it is desirable to develop an integrated model that incorporates computational morphology and relevant theories of language learning. Several studies have contributed to the issue without, to the best of our knowledge, any of them developing an integrated model of morphological analysis and Bantu language learning. The specific addition we have in mind here is the inclusion of grammatical exercises produced automatically using software for morphology.

1.2.3 Context of this study

In traditional face-to-face or teacher-centred learning environments, textbooks are regarded as important for the provision and support of language instruction. This, however, limits the content that learners encounter, the space and time of language learning and the opportunities for interaction with peers or with automated tools. One educator has remarked that "textbooks are neither descriptively adequate tools nor accurate models of what takes place in the process of learning a language" (Mugane, 1997). We understand Mugane to suggest that textbooks require additional supplementary tools for language learning, which are, unfortunately, largely unavailable in countries like Uganda. Mugane (1997) further points out that, textbooks may be one of the weakest pedagogical tools in language learning.

The advent of computer technology has made it possible to improve on the use of textbooks as the sole learning aid both inside and outside the classroom. Computers, which are now fast, easy to use, convenient and cheap, offer great opportunities for developing powerful language learning systems, while providing global access to less documented and studied languages. Educators now recognize that utilizing computer technology and language learning software can be conducive to the creation of effective independent and collaborative learning environments in which students can be provided with new language experiences (Kung, 2002; Jager, 2009).

Shalaan (2005) observes that the overwhelming majority of language learning systems have been developed for English, followed by Japanese, French and German. Other languages, including most African languages, are not part of the technological development in language learning. This means that the majority of the world's languages are not benefiting from the developments of computer technology with its related advantages. Shalaan calls for more research on techniques that combine natural language processing with language learning systems. Research on a language learning system for Bantu languages, specifically the Runyakitara group, partly responds to his appeal. We turn now to a brief description of Runyakitara and why our focus on this language also represents an innovative aspect of the current study.

Runyakitara

Runyakitara is a name for four closely related Bantu languages spoken in western Uganda. Bernsten (1998) refers to Runyakitara as a name for four major dialects of western Uganda: Runyankore, Rukiga, Runyoro and Rutooro. The documented linguistic data about the Runyakitara languages and their implications for language learning include the following.

In terms of mutual intelligibility, the four languages are mutually intelligible at a level greater than 70%, with the following table detailing their lexical similarity:

Languages	**Lexical similarity**
Nyankore and Chiga	84% - 94%
Nyoro and Rutooro	78% - 93%
Nyoro and Nyankore	77% - 96%
Nyoro and Chiga	67%

Table 1: Lexical similarity in Runyakitara (adapted from**:** Lewis, 2009)

Geographically, the languages are spoken by approximately six million (6,000,000) people (native speakers) in nineteen districts of Western Uganda (Uganda Bureau of Statistics 2002). There are other speakers in some parts of Tanzania (Haya) and Democratic Republic of Congo (Songora). Some speakers of Runyakitara languages also live in cities and towns such as Kampala, and their children have lost regular touch with the native languages.

Socially, the languages of Runyakitara are used in the media, taught in schools and used in day-to-day business transactions. In a recent development of instructing in local languages in lower levels of primary education in Uganda (Bukenya 2008), the languages are now used as a medium of instruction from Primary 1 to 3 in western Uganda.

The status of the Runyakitara language is important to this study in several ways. First, although the language has six million speakers, that is in part a trick of regarding the closely related languages as the same. Separately they would each count for fewer, and some would be candidates for endangerment. Our work may help preserve the languages for those who wish to speak them in the future. On the other hand, six million speakers is an excellent number if the effort in language politics succeeds, and the varieties indeed come to function as one. Second, literacy rates are low in all the Runyakitaran varieties, and language learning software (and, indeed, other applications of the computational morphology) may serve to improve literacy. Third, because Runyakitara is an aggregation of closely related varieties, real native speaker abilities in the combined language are rare, and many "native speakers" may benefit from additional training in the language, particularly with respect to writing. Fourth, as we discovered in the course of the research for this dissertation, there is also a Runyakitara diaspora, and there is interest in learning the language among the children of the emigrants. We discuss this group in Chap. 7 (below).

Fifth and finally, there has been relatively little work on computer-assisted language learning for Bantu languages and none at all for Runyakitara. This study therefore enlarges the empirical base of study on which general ideas and theories for CALL may draw.

Runyakitara morphology

There is no recent systematic and comprehensive publication on Runyakitara morphology. The information in this study is derived from various sources, such as grammar books (Taylor, 1985), manuscripts (Ndoleriire & Oriikiriza, 1995) and also other Bantu language studies (Nurse & Phillipson, 2003). Morphology is chosen as a focus in this study because of the complexity of Runyakitara morphology, which is a stumbling block for learners. We elaborate on this complexity in Chap. 2, especially in Sec. 2.1, where we also attempt to quantify this and compare it to the complexity of European languages that are regarded as morphologically difficult such as Latin, Russian and Finnish.

Following the linguistic typology in Comrie (1989), the Runyakitara group falls under the group of synthetic, agglutinative languages. It has been observed that most languages cannot be categorized exclusively with respect to a single typology, and Runyakitara also exhibits some features of fusion, as its verbs are highly inflected. Inflection, derivation, compounding and reduplication are productive features in Runyakitara languages, and there is a noun classification structure marked by morphology which is unique to Bantu languages. This makes the morphology of Runyakitara a complex topic, the learning of which may especially benefit from computer support. Chapters 2 and 5 include further discussion of this issue. Below is a selection of topics that contribute to the Runyakitara morphological complexity that have to be taken into account. We intend this introductory presentation to provide a flavour of the complexity.

a) Agglutination: some words in Runyakitara are formed through a process where morphemes are added together, each contributing a meaning to the whole. For example, in a verb **"titukamureebaga"** (we have never seen him/her) morphemes are added as in the example below:

I)

ti –	tu	-ka-	mu	-reeb-	a-	ga
↓	↓	↓	↓	↓	↓	↓
NG-	2PPL-	FP-	OM-	V-	VE-	ADV

The example in I above shows that the following morphemes have been added to the root: ***ti*** is a negation marker, ***tu*** is a 2nd person plural subject marker, ***ka*** is a tense marking for "far past", **mu** is an object marker for persons, ***reeb*** is a verbal root, ***a*** is a verb ending (indicative) and ***ga*** is an adverbial marker for 'ever'.

b) Reduplication: words are formed through copying/doubling a part or a whole word. The following illustrates some of the reduplication that may occur in Runyakitara:

	kukwata 'to catch'	*kukwata-kwata*	'to touch "in a funny way"'
II)	*baareeba* 'they have seen'	*bareeba-reeba*	'they…seen "suspiciously"'
	ibiri 'two'	*ibiri-ibiri*	'two and two'
	babiri 'two people'	*babiri-babiri*	'two by two'
	omuntu 'a person'	*omuntu-ntu*	'a "stupid" person'
	ogu 'this'	*ogu-ogu*	'this one'

Note that, for verbs and nouns, reduplication only affects the root while the entire word is repeated in the case of pronouns and numbers.

c) Inflection: there is an extensive system of inflection for Runyakitara verbs, but, as we have noted, there currently is no consistent and accurate documentation of this subject in the Runyakitara languages. Comparing Runyakitara with other Bantu languages (Katamba, 2003), we noted that Runyakitara verbs can be inflected for negation, subject, tense, aspect, object, mood and adverbial markers. The examples below illustrate the different forms of inflection of a Runyakitara verb, the inflections being indicated in bold-facetype:

i) Mood – *shom-**a*** 'read'
*mu-shom-***e** 'you should read'

ii) Tense – *mu-shom-**ire*** 'you read' – there are 7 tenses in Runyakitara, each with a different tense morph.

iii) Aspect – *n-***aa**-*shom-ire* 'I have read'

iv) Negation – **ti**-*naa-shom-a* 'I have not read'
*tu-***ta**-*shom-a* 'we shouldn't read'

v) Adverbial marker(s) *mu-shom-***er**-*e* 'read for him/her, e.g. read on his behalf'. These are commonly called verb extensions. We identified 7 verb extension markers in Runyakitara (see Chap. 2).

d) Allomorphy: Runyakitara has various allomorphs, that is, a single morpheme can be realized in two or more different ways. A case in point here is a causative morpheme which has six different realizations [*es/is/iz/ez/sy/y*]. The following illustrates these:

*reeb**es**a* 'cause … to see'
*kwat**is**a* 'cause … to touch'
*guru**sy**a* 'cause … to jump'
*r**iz**a* 'cause … to cry'
*te**ez**a* 'cause … to beat'
*ham**y**a* 'make … firm'

Applicative, passive, reversive and intensive morphemes also behave in a more or less similar manner. The same point is elaborated in Chap. 2, 4 and 5

e) Noun classification: Similar to all other Bantu languages, nouns in Runyakitara are categorized into noun classes. A detailed description of the noun structure and

the ways in which it is morphologically computed is found in Katushemererwe & Hanneforth (2010) (Chap. 2 below). A noun class in Runyakitara serves two roles in morphology but has other functions in syntax. In morphology, it marks the class and number of the noun (i.e. whether a noun is singular or plural). In syntax, a noun class is part of a larger concord system, which we address detail in Chap. 8 of this dissertation.

All the above morphological phenomena need to be considered in providing comprehensive and relevant language resources in the Runyakitara languages. We will focus on developing the morphological analyzer as a building block for language-learning applications in Runyakitara.

Having said that, there are no computational resources for classroom learning nor for individual learning needs in most Bantu languages, let alone the Runyakitara group. In order to realize educational software for learning Bantu languages (Hurskainen, 2009), it is important that such a tool be developed in order to provide learning content and support to Bantu language learners, and specifically to Runyakitara learners.

Local Benefits

Considering the language situation in Uganda, there are several benefits to be derived from learning Runyakitara in the contemporary context.

In the 21st century, there are opportunities to improve Uganda's economy but only if Ugandans keep up with the changing world. If Uganda wishes to attract more investors, then, it must participate on the global stage. Uganda has to make sure that its own people understand its languages and cultures in order to provide them with a firm foundation on which to operate in the world. A further important aspect involved here relates to cultural acceptance: it is important that Uganda exposes its languages and cultures to foreigners to enable intercultural comparison and, consequently, to promote tolerance and acceptance.

Many Ugandans believe that it is not necessary to learn their indigenous/native languages because English, an official language of Uganda, is sufficient given its international status. They argue that a person is better off knowing English than knowing the local languages (Asiimwe, 2006). This is surely a false dichotomy given that English is not the sole language of everyday communication in Uganda. There are many instances when local languages have to be used. Therefore, a person in Uganda benefits from learning the local languages *and* English, rather than English alone.

In addition, many people around the world learn additional languages for personal enjoyment and enrichment. These people are life-long learners who are always seeking to enrich their personal lives by accessing various arts, entertainment and information sources available to speakers of other languages.They also seek out and take advantage of travel opportunities. Providing Runyakitara content in digital form can cater to the wishes of such a category of people.

1.3 Reason for the study

This research is motivated by the fact that (i) most African countries need to use local languages at the lower levels of education (Bukenya, 2008); (ii) universities need instructional materials (to educate the teachers at these levels); and (iii) Africans and other people interested in African languages require digital content.

In natural language processing (NLP) applications, computational morphology is a basic layer over which other layers such as syntactic and semantic analysis are built (Jurafsky & Martin, 2008). Morphological analysis has not been fully exploited for the learning of grammar, especially with regard to the Bantu languages. This situation confirms the more general statement made by Nerbonne et al. (1998) that, although automatic morphological analysis has long been well established, it has not been fully utilized in language learning. Morphological analysis has certainly been utilized in various methods of vocabulary learning/extension used in European and Asian languages (Bellomo 2009; Kieffer et al, 2007; Nerbonne 1998), but there is limited literature documenting the ways in which morphological analysis can aid the learning of Bantu languages with their complex morphology.

Considerable research has been done on NLP systems for Bantu languages in general including work on computational morphology (Hurskainen, 1992; Muhirwe, 2007; Elwell, 2006); Pretorius & Bosch, 2003; Okemwa & Ng'ang'a, 2008; Karttunen, 2003), speech recognition systems (Badenhorst & Van Heerden, 2009; Gumede & Plauché, 2009), a Swahili Language Manager (Hurskainen, 2004), and a parallel corpus for English and Swahili, (De Pauw & Wagacha, 2009). However, utilizing morphological analysis for language learning in Bantu languages has not been a focus of research. This dissertation is the first to deploy morphological analysis to support language learning for the Runyakitara group of Bantu languages.

Resource poor languages need comprehensive tools that can improve the documentation and accessibility of language resources. The lack of language learning software in Runyakitara prompted the need to research the ways in which morphological analysis might be utilized for Bantu language learning, specifically basing our study on grammar instruction/learning in Runyakitara.

1.4 General and specific objectives

The general theoretical objective of this research is to contribute to understanding the extent to which a morphological analyzer can be utilized to support Bantu language learning. To achieve this general aim, the following specific objectives are pursued:

i) To critically review and identify the different application areas of computational morphology in language learning;

ii) To develop and evaluate a morphological analyzer for Runyakitara, accounting for many of the word forms required for learning;

iii) To design and implement a language learning system for Runyakitara that does not restrict the learner to limited vocabulary;

iv) To evaluate empirically the effectiveness of the system in learning Runyakitara.

The first objective relates to the different areas where a morphological analyzer has been used for language learning, with specific attention to the manners in which it can be applied to Bantu language learning. It is concerned with the language knowledge and skills that language learners may be able to gain as a result of applying a morphological analyzer.

The second objective concerns the development of a morphological analyzer of Runyakitara. This objective has been adopted because Runyakitara had no morphological analyzer on which to base our study. It is concerned with the language formalization, implementation and testing of the Runyakitara morphological system.

The third objective deals with the design considerations and implementation issues of a language learning model for Runyakitara based on a morphological analyzer.

The model is empirically evaluated as part of achieving the fourth objective.

Our practical objective in undertaking this research is to begin with language technology for the Runyakitara group of languages and in particular with software to improve learners' chances in dealing with its complex morphology.

1.5 Structure of the dissertation

The rest of the dissertation is structured as follows:
In Chapter 2, we present Runyakitara verb morphology, highlighting complex elements that represent a challenge to computation and the solutions provided. We develop the argument here that a processing system is necessary if one is to provide a wide range of material to learners. While one might at first blush consider a database of forms, filling that database manually would be error prone and ultimately infeasible, meaning that an automated system is necessary in this case as well. This chapter details the development of a computational model for the grammar of Runyakitara verbs. It therefore provides a building block to be employed in a comprehensive morphological analyzer.
Chapter 3 describes the Runyakitara noun morphology and its implementation using the finite-state approach. It presents the noun classification system of Runyakitara and the manner in which it is accounted for computationally. The result here also creates another building block for a comprehensive morphological analyzer.
The description of the comprehensive morphological analyzer of Runyakitara is provided in chapter 4. The results, language related issues and its ability to analyze the four languages of Runyakitara are discussed. We also deepen the argument in this chapter that

Runyakitara's morphological complexity requires a rule-based treatment and that it is likely to be challenging for language learners (Sec. 4.4).

In chapter 5**,** we analyze the situation of language teaching and learning in Uganda, highlighting the major issues to focus on in the design of computer-assisted language learning applications for Ugandan languages. We attempt to assess the perceived need for CALL in support of learning local languages.

Chapter 6 presents the results of a pilot study which was carried out to further understand the issues that emerged in the previous study (reported in Chapter 5).

Chapter 7 presents the morphology learning system and an experiment in using it in a university-level language course. It turned out that some of the course participants were the children of emigrants from the Runyaktiara area. They had learned the language somewhat as children, but were no longer competent to speak and understand it, and were effectively illiterate. Design and evaluation results are reported.

An idea (and a simple prototype) for a more ambitious Intelligent Computer Assisted Language Learning (ICALL) system for Runyakitara is discussed in Chapter 8. Selected topics in syntax that can be supported by the morphological analyzer are the focus of the exercise presented there. The design and implementation are described. Further areas of application are also highlighted.

In chapter 9, we summarize our key results, highlighting our contributions to theory and practice and indicating directions for future research.

Chapter 2

Finite State Methods in Morphological Analysis of Runyakitara Verbs

(*An earlier version of this chapter was published in Nordic Journal of African Studies, Vol. 19(1) 1-22, 2010 as: Fridah Katushemererwe & Thomas Hanneforth, Finite State Methods in Morphological Analysis of Runyakitara Verbs.)*

Abstract

To partly address the lack of an automatic analyzer and generator for the word forms of Runyakitara, this chapter presents a computational model for grammatical Runyakitara verbs. This model, which we are labelling RUNYAGRAM, is based on freely-available, open-source, finite-state methods and, in particular, the fsm2 interpreter. To capture morphotactic structures, it uses non-recursive context-free grammars supported by fsm2 and morpho-phonological alternations with a finite composition of commonly used context-dependent string rewriting rules. Their combination results in a finite state transducer that can be exported and used in numerous software-developing efforts. The obtained transducer is an important building block that can be employed in comprehensive morphological analyzers, syntactic parsers, spell-checkers, text-to-speech synthesizers, and machine translation systems. Currently, 86% of the verb forms are recognized. It is possible to increase the scope, or alternatively, to adapt the approach of the RUNYAGRAM system to suit specific languages.

Keywords

Morphological analysis, Finite State Methods, Runyakitara Verb

2.1. Introduction

One of the key enabling technologies required in natural language processing applications is a morphological analyzer. It is an established fact in computational linguistics that a morphological analyzer provides the basis for many natural language processing applications (Pretorius & Bosch 2003; Yona & Wintner 2005).

Computational morphology deals with automatic word-form recognition and generation. The general challenges posed by a computational morphological analyzer, as described by Pretorious and Bosch (2003), are twofold:

- The morphemes that make up words do not combine at random; their combinations and orders are selective. A morphological analyzer needs to know which combinations of morphemes (morphotactics) are valid.

- Morphemes may be realized in different ways, depending on their context. A morphological analyzer needs to recognize the morpho-phonological changes between lexical and surface forms (morpho-phonological alternation).

Automatic morphological analyzers and generators must take the above two issues into consideration.

Comprehensive morphological analyzers are available for well documented languages such as English, Swedish, German, Arabic, and Finnish (Karttunen & Beesley 2005).

Considerable progress has also been achieved in applying finite state methods to Bantu language analysis, as exemplified by the Kiswahili morphological analyzer (Hurskainen 1992; 1996; 2004), the Zulu analyzer prototype (Pretorius & Bosch 2003), Lingala verb morphology (Karttunen 2003), Ekegusii verb morphology (Elwell 2005), Kinyarwanda (Muhirwe & Trosterud 2008), and Setswana verb morphology (Pretorius, Berg, & Pretorius 2009).

However, the fact that there are over five hundred Bantu languages means that almost all of them have not been subject to any such analysis. Although the Bantu languages are classified as largely agglutinative and exhibit significant inherent structural similarity, they differ so extensively in terms of their phonological features that each Bantu language is likely to require an independent morphological analyzer.

Runyakitara is a language group belonging to the under-resourced Bantu languages with no computational morphology. Bernsten (1998) splits Runyakitara into four major dialects: Runyankore, Runkiga, Runyoro, and Rutooro. Guthrie (1967) groups these four dialects into two languages belonging to Narrow Bantu branch of the Niger-Congo family, Nyankore-Kiga (E.13) and Nyoro-Ganda (E.11). There is no recent survey which can guide us in regard to Runyakitara typology. For purposes of this paper, Runyakitara will be taken to mean two major language clusters mentioned above: Runyoro-Rutooro and Runyankore-Rukiga, denoted by R-R in the following.

Runyakitara is spoken by approximately six and half million (6,500,000) people in nineteen districts of Western Uganda. As a significant language group in Uganda, some parts of Tanzania and the Democratic Republic of Congo, it is important that R-R is given computational attention, especially since it has a large number of speakers, is used by the media in western Uganda (two regular newspapers – one online) and possesses a rich history and culture that should be preserved. Furthermore, Runyakitara languages are used as languages of instruction in the lower levels of primary education in Western Uganda, and we shall later consider how computational efforts may add value to their educational status. As emphasized by other Bantu researchers, (Hurskainen 1992; Elwell 2005), the morphology of a verb in R–R represents one of the more complex morphological systems known, which means that it requires special attention on that score alone.

2.2. Runyakitara Verb Morphology and the Computational Challenge

A verb in a typical Bantu language may acquire many prefixes and suffixes. The Runyakitara verb morphology poses the following challenges to computational modelling because of the following features: a) number of morphemes, b) morpheme order, c) morpheme combination, d) allomorphs, and e) vowel harmony. Each of these will be discussed in the sub-sections below.

2.2.1 Number of morphemes involved

The Bantu verb template described in many studies (Maho 2007, Nurse & Philippson 2003) suggest that there are about 8 to 15 morpheme slots, which may be represented as in Table 1:

Slot	**1**	**2**	**3**	**4**	**6**	**7**		**8**	**9**
Meaning	Pre-initial	Initial	Post-initial	Tense marker	OM	Verbal base		Final	Post-final
Morpheme	NEG	SM	NEG	Tense	Object marker	Root	Verb ext.	Mood, aspect, NEG	

Table 1: Bantu Verb Template (Nurse & Philippson 2003)

Notes: NEG – negative, SM – subject marker, OM – object marker, Verb ext – verb extension

The above generic template raises many questions, particularly regarding the definition of a morpheme when applied to R-R morphology. What exactly is considered a morpheme in terms of this template? If verb extension (in Slot 7) is a morpheme, does that mean that extensions such as causative, applicative or passive markers are allomorphs of the same morpheme? This and many other questions prompted us to devise a R-R verb template to cater more specifically to the number of morphemes present in this language group.
Since the morphemes involved in the formation of R-R verbs are more numerous, it is important to expand the template. R-R verbal morphemes can be broadly classified as prefixes, (morphemes to the left of Slot 0) root (Slot 0) and suffixes (morphemes to the right of Slot 0). The following template provides a more accurate indication of the morphemes involved in the formation of Runyakitara verbs:

-7	**-6**	**-5**	**-4**	**-3**						**-2**		**-1**	**0**	**1**							**2**			**3**	
Ng1	Asp	Sp	Ng 2	Tense/aspect markers						Object pronouns		Asp	R	Verb extension morphemes (VEXT)							Verb end (VE)			Pf1	Pf2
				Inf	Hab		Pf	ff	Rp	Op 1	Op 2	ref		Ca	Apl	Rec	Pas	Int	Stat	Rev	Ind	subj	past		
ti	ni	18	ta	ku	Ø		aa	ria	ka	18	18	e		es is iz y sy	er ir	an	w ebw ibw	erer irir	ek ik	uk ur uur	a	e	ire	ho mu yo	ga

Table 2: Runyakitara Verb Template

In the above table, slot 0 represents the root; to the right of 0 are suffixes to the root. Slot 1 is for verb extensions such as: Ca – causative, Apl – applicative, Rec – reciprocal, Pas – passive, Int – intensive, Stat – stative, Rev – reversive. Slot 2 represents Verb end: Ind – indicative, subj – subjunctive, past – past tense. Slot 3 indicates post final morphemes: pf1 – post-final 1; pf2 – post-final2. To the left of slot 0, -1 Asp – aspect, -2 – object pronouns, -3 Tense/aspect markers,[inf: infinitive, Hab: habitual, pf: perfective, ff: far future, Rp: remote past] -4 – Ng2 – Negative 2, -5 Sp – subject prefix; -6 Asp – aspect; -7 Ng1 – Negative 1. For a more detailed description and examples, see Appendix A.

Runyakitara has the typical characteristics of the template morphology outlined by Spencer (1991). As noted by Spencer (1991), template morphology poses a computational challenge because it represents a morphological system in which a verb stem or root consists of one or more obligatory affix(es) as well as a set of optional affix(es). Such combinations of morphemes make automatic analysis difficult because it is first necessary to identify the affixes attached to the root required to compose specific verb forms.

Adding to the number of morphemes involved, subject and object pronominal markers display agreement with the classes of the nouns to which they refer. If the subject is not otherwise indicated, they serve as subject and object pronouns. These markers appear on the verb root as prefixes to the root. R-R has eighteen (18) noun classes, there are therefore as many as 18 subject and object pronoun markers in each case. In addition, R-R are type 3 languages according to the classification provided by Maho (2007), which means that they may have two or more objects in a given construction. Specific evidence from Runyakitara confirms that these languages can have a double object construction, which means that a verb can have a marker for both direct and indirect objects in the same construction. An example in this case is *mu-mu-n-kwat-ire* (you grab/hold him for me), where *mu-n* indicates both direct and indirect objects representing **him** and **me**. This will add to the number of morphemes, increasing the challenge that the morphological multitude poses.

2.2.2 Excursus on morphological complexity

Given the goals of this thesis, namely to provide software support for learners of Runyakitara, we wish to note that the morphological complexity of the language requires that a rule-based system be implemented if the software is to support a wide range of language use. The rule-based implementation is capable of dealing with the enormous number of word forms each Runyakitara verb form is capable of forming. A list, or database, of forms would be impractical (unless it were created by a rule-based system).

One way to quantify the complexity of an inflectional system is to count how many inflected forms there are per lemma. English is regarded as relatively simple because it normally has one or two forms per tense-aspect combination, but it has a perfect participle, and a progressive participle and three single forms for the infinitive, the imperative and the moribund subjunctive mood. Ignoring very irregular verbs such as *be* and *have*, this adds up to fewer than ten inflected forms per verb. Latin's verbal morphology is regarded as relatively complex among European languages because it combines six person/number forms, with six tenses (present, imperfect, future, perfect, pluperfect and future perfect), two modi (active and passive), and two moods (indicative and subjunctive) yielding 144 forms per lemma (plus an infinitive, a supine, and a couple of imperative forms, bringing the total to nearly 150), but this count ignores the absence of the future and future perfect forms in the subjunctive mood, which would reduce the paradigm by two (tenses) times six (person/number forms), bringing the total down to about 125 forms per lemma. Ostler (2007) estimates the number more carefully at 106 inflected forms per lemma. In fact this is a common means of assaying morphological complexity. Hajič and Hladká (1998) count roughly 90 forms per lemma in Czech (p.485) as evidence of its complexity, and El Kholy & Habash (2012) reason that Arabic

is morphologically complex due to its "thousands of inflected forms per lemma" (p.91). Let us apply this yardstick to Runyakitara.

The template in Appendix A has eleven slots in the Runyakitara verb paradigm, where the root is obligatory, but where there are ten possible verb extensions in slot 1, three verb ends in slot 2, three post-final affixes in slot 3, two affixes (reflexive or zero) in slot 4, five tense/aspect markers in slot 6, two markers (negation or zero) in slot 7, two aspect markers in slot 9 (progressive or zero), and two polarity markers in slot 10, yielding $10 \times 3 \times 3 \times 2 \times 5 \times 2 \times 2 \times 2 = 7.200$ combinations, ignoring subject, object and indirect object affixes. And the complement affixes are distinguished by classifier, of which there are eighteen. Since up to three noun phrase complements may occur with each verb, this yields an additional $18 \times 18 \times 18 = 5.832$ combinations, each of which may combine with any of the other forms, yielding an impressive forty-two million forms! According to this criterion then, English is simplest, Latin and Czech are an order of magnitude more complex, Arabic a second order of magnitude more complex (thousands of forms) and Runyakitara morphology six orders of magnitude more complex than English (tens of millions of forms). In fact the complexity of Runyakitara is aggravated by its morpheme combination rules, allomorphy and vowel harmony, each of which will be discussed a bit below. We argued here for Runyakitara morphological complexity on the basis of the size of its paradigms in order to keep the argument simple.

The calculation made above should not be taken as the last word on morphological complexity. One might wish to correct for syncretism, i.e. the phenomenon that two positions in a paradigm might systematically always have the same value. In fact theorists are in general agreement that regularity must be factored in systematically, and Bane (2008) and Martens (2011) have independently proposed that (morphological) complexity ultimately be measured in terms of the length of the minimal description required to describe a phenomenon completely, a formulation they derive from information theory. It would go beyond the scope of this dissertation to apply their ideas to Runyakitara inflectional paradigms, but we can add that, other things being equally, large paradigms will count as more complex than small ones in their construal as well.

From this discussion we wish to draw two conclusions relevant to the argument of this dissertation. First, Runyakitara morphology is quite complex, which means that it is difficult to learn, and therefore worthy of special effort in developing supporting materials for instruction. This means that the development of software to support learning Runyakitara morphology is worthwhile. Second, the morphology is so large in scope that it would infeasible to list the verb forms associated with each lemma. Aside from the sheer amount of time needed, a hand-crafted list would be susceptible to error and difficult to maintain and expand.

2.2.3 Morpheme combination

Despite the studies that have been carried out on morpheme combinations in the Bantu languages, (Hyman 2007), there is limited available research on Runyakitara morpheme combinations. This lack of research particularly pertains to verb extensions. As earlier noted by Hyman, (2007), verb extensions are difficult to analyze mainly because of their quantity, functional diversity and frequent occurrence in long successions. Runyakitara

has seven (7) verbal extensions which can be added to the root individually or in combination. For example, a verb may have verb extensions such as the following:

*reeb-**a** (see)*
*reeb-**es-a*** (see with),
*reeb-**an-a*** (see each other),
*reeb-**w-a*** (be seen),
*reeb-**es-an-a*** (make each other to see),
*reeb-**an-is-a** (*make to see each other*)*,
*reeb-**es-an-is-ibw-a*** (be made to make them see each other).

In the last example, [***es, an, w, is, ibw***] are all verb extensions that have different functions. The position of the causative morphs ***es*** and ***is*** in the above example is also different, but there is no study available that establishes if the combination of verbal extensions and their sequence is significant in Runyakitara.

2.2.4 Morpheme order

Although the Bantu verb template is presumed to present a fixed order of morphemes and provides Slot 4 in table 1, for example, as a slot for tense aspect markers, some morphemes in Runyakitara violate the order. Specific cases are: progressive ***ni,*** reflexive ***e*** and past ***ire,*** which have positions that differ from the order of the Bantu template. As indicated by the Runyakitara template represented in table 2, ***ni*** comes before the subject marker in the construction while other tense/aspect markers follow the subject marker:

***ni**-ba-mu-reeb-a* (they are seeing him)
*ba-**ka**-mu-reeb-a* (they saw him [last year or some months back]).
*Ba-mu-reeb-**ire***(they saw him [yesterday])

In the above verb constructions, ***ni, ka***, and ***ire*** are tense/aspect markers but appear in different positions with respect to the root.

Also, the order of verb extensions in the template does not necessarily mean fix the affix order in an actual utterance. On the contrary, the position of verbal extensions frequently depends on the argument structure. This suggests that there is no fixed order in which they are supposed to appear in the construction of the verb. For example, a verb root may have the following combinations of extensions:

reeb-a (see)
reeb-es-a (see with)
reeb-an-a (see each other)
reeb-es-an-a (make each other to see)
reeb-an-is-a (make … to see each other)
reeb-an-is-ibw-a (be made to make … see each other).
reeb-er-a (see for)
reeb-er-an-a (see for each other)

Since, in the above example, *is* and *es* are both causatives, there appears to be some flexibility in the position of causatives in the verb structure and, consequently, in the manner in which morphemes may precede and follow one another.

2.2.5 Allomorphy

Runyakitara has various allomorphs (i.e. different realizations of the same morphemes). A case in point here is the causative morpheme which has four different realizations [**es/is/iz/s/y**]. Applicative, passive, stative and reversive morphemes are similarly no exception. All such allomorphs pose a challenge to computational modelling.

2.2.6 Vowel harmony

Katamba (1984) analyzes the vowel harmony of the verb extensions in Luganda, a language closely related to the Runyakitara group. His analysis, which divides harmonizing vowels into mid and non-mid groups, contributes to the understanding of the vowel harmony in the language. It is not, however, very helpful when determining morphemes for computational purposes, as the position of mid and non-mid vowels in the string is difficult to specify (with any degree of predictive confidence). The suggestion provided by Morris and Kirwan (1972) with regard to penultimate syllables may, however, be useful here. A penultimate syllable, a syllable preceding the final one (penultimate meaning 'before last'), may help to position morphemes when vowel harmony is involved. For example, in the word *bo-ro-go-ta*, (flow of water) the penultimate syllable is '**go**' and the one preceding it also has /o/ as its vowel. By analogy, we can conclude with respect to the causative that, when a penultimate syllable is /**e/** or /**o/**, the causative extension will be ***es***. On the other hand, when the penultimate syllable is /***a/, /i/*** or /***u/***, the causative extension will be ***is*** or ***iz***. Similar variations apply to applicative, intensive and stative morphemes.

2.3. Formalization and Implementation

Given the nature of Runyakitara morphology, the choice of an appropriate approach was a significant issue. The hierarchical nature of Runyakitara morphology might have been represented using Phrase Structure Grammar (PSG) as proposed by Selkirk (Spencer 1991), who formulates phrase-structure-like rules written as W+A for suffixing and A+W for prefixing. Once it became clear that Selkirk's rules overlook important local morpho-phonological and orthographical processes, we began using replacement rules for that purpose. Since no recursion was needed, we turned to the framework of finite-state acceptors (FSA)/transducers (FST) to describe both the concatenative rules and the phonological processes involved in Runyakitara verb formation (but see Section 3.3 below as well). Our approach relies heavily on the closure properties of these automata in relation to intersection, composition and substitution (see Hopcroft & Ullman 1979; Kaplan & Kay 1994).

Specific application to Runyakitara involved the use of *fsm2* (Hanneforth 2009), a scripting language within the framework of finite state technology. Finite-state technology is considered the preferred model for representing the phonology and morphology of natural languages (Wintner 2007). The model has been used to computationally analyze natural languages such as English, German, French, Finnish, Swahili, to mention just a few (Beesley and Karttunen 2003). One main advantage of this model is that it is bidirectional – it applies to both analysis and generation. This bidirectionality was the principal reason that the technology was selected for use in the morphological grammatical analysis of R-R.

Further reasons that *Fsm2* was chosen as a resource tool to be used in a morphological grammar of R-R include the following:

- It supports a full-set of algebraic operations defined on both un-weighted and weighted finite state automata and weighted finite state transducers (Hanneforth 2009). Algebraic operations are useful in designing complex morphological analyzers in a modular way.

- *fsm2* supports a number of equivalence transformations, which change or optimize the topology of a weighted automation without changing its weighted language or relation. This means that automata may be minimized, determined, optimized etc;

- *fsm2* uses symbol signatures to map symbols to numbers that are internally recognized by the automata. Symbol signatures are useful in language modelling, since every word in a language is given an alphabetic symbol, and one of the developer's tasks is to define symbols that represent morphemes and their categories.

- *fsm2* provides an efficient way of compiling morphological grammars that easily account for the co-occurrence of roots and inflectional affixes, a common feature of Runyakitara.

- *fsm2* is open-source software. The source code can be downloaded from *www.fsmlib.org*.

- *fsm2* is able to load lexicons, grammars and replacement rules defined by the morphology developer and to automatically transform various rule formats into transducers.

2.3.1 The structure of RUNYAGRAM

RUNYAGRAM has a modular structure comprising a special symbol module/file, a grammar module and a replacement rule module. The three are combined to produce a single finite state transducer.
The following diagram illustrates the overall architecture of RUNYAGRAM:

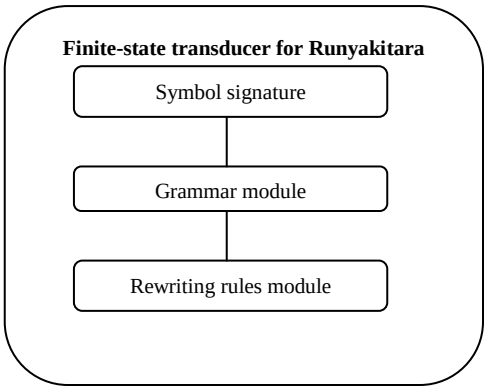


Figure 1: Sketch of the architecture of RUNYAGRAM

The output that RUNYAGRAM generates can be used as input for other applications such as:

- a spell checker for Runyankore-Rukiga
- a dictionary, since RUNYAGRAM outputs lemmas
- a syntax analyzer for Runyakitara
- a language learning system for vocabulary and grammar, which must be further developed.

The remaining sub-sections illustrate the construction of the sub-analyzer for verbs in Runyakitara.

2.3.2 Symbol signature

Like AT&T Lextools (see Roark and Sproat 2007), *fsm2* uses a *symbol signature* to define the basic entities of the grammatical description. Fig. 2 shows some sample entries.

```
Letter      a b c d e f g h i j k l m n o p q r s t u v w x y z
Category:   VERB_ROOT_SIMPLE1 Simple1
Category:   VERB_PREF_TENSE Tense
```

Figure 2: Sample entries of the RUNYAGRAM symbol signature

The entries are of two types:

1. **Supertype – subtype** definitions
2. **Category** definitions, a category consisting of a name and a (perhaps empty) list of features.

The first line in Fig. 2 defines `Letter` as the supertype of the subtypes `a`, `b`, `c`, etc. The following lines define two categories `VERB_ROOT_SIMPLE1` and `VERB_PREF_TENSE`, with features `Simple1` and `Tense` defined elsewhere in the signature. Features themselves are again treated as supertypes, having their subtypes as their values. Each symbol in the signature – whether type or category name – is mapped by *fsm2* onto a unique integer used internally in the compiled automata.

2.3.3 Word grammar

To specify morpheme order, we do not use the "classical" *continuation-class mechanism* of Koskenniemi (1984) but instead employ a *context-free word grammar* for that purpose.[3] In our view, this sort grammar provides a much more natural way of defining orders and groupings of elements compared to the continuation-class method, which basically amounts to the hand-coding of a finite state automaton within the lexicon. Since the generative capacity of context-free grammars exceeds the capacity of finite-state automata, we restrict ourselves to a subset of context-free grammars along the lines of the quasi-context free grammars by Mohri & Sproat (1996). This subset may include left- or right-side recursive rules, but excludes all forms of centre-embedding.

In the *fsm2* framework, grammar rules have the form $A \rightarrow \beta$, where *A* is a designated non-terminal symbol and β is an arbitrary regular expression (which may even use intersection or negation).
The compilation approach is based on the ordering of the non-terminals in the grammar, creating finite-state automata (FSA) for each grammar symbol and substituting the FSA for the individual grammar symbols into the rules for the right side of the previously computed order. In the right sides of grammar rules, the morphemes of Runyakitara alternate with grammatical categories bearing grammatical information for the morphemes preceding them.
The grammar module consists of a set of quasi context-free rules accounting for the concatenative nature of Runyakitara morphology. The grammar contains a large number of rules, of which we will present just a sample, exemplifying the principles underlying the overall grammatical organization. We will follow the approach of elaborating the minimum form of a verb until the maximum number of morphemes is reached, thus accounting for every form of the verb. Fig. 3 provides some (simplified) sample rules of the verb sub-grammar.

[3] A *context-free grammar* (see Aho & Ullman, 1979) is 4-tuple $\langle \Sigma, N, S, P \rangle$ where Σ is a finite set of alphabetic symbols, *N* is a finite set of non-terminal symbols (phrase symbols), $S \in N$ is the start (sentence) symbol of the grammar and P is a set of rules $A \rightarrow \beta$, where $A \in N$ and $\beta \in (N \cup \Sigma)^*$. This means that the left side of a grammar rule is restricted to a single phrasal symbol, whereas the right side can contain an arbitrary combination of alphabetic and phrasal symbols.

```
# Verb structure rules

#Minimum number of morphemes a verb takes
[VERB]              →     [VROOT] [VEND]

# Maximum number of morphemes a verb takes
[VERB]              →     [VPREFNEG] [VPREFSP] [VPREFTM] [VPREFOP][VPOP2] \
                          [VROOT] [VEXT1] [VEXT2] [VEXT3] [VEND] [POSTV]

# Morpheme insertion rules (morphemes are in bold-face)
[VROOT]             →     (reeb|teer|kwat|shom)\
                                [VERB_ROOT_SIMPLE Simple=simpleverb]
[VEND]              →     a     [VERB_END_IND Ind=mood]
[VPREFNEG]          →     ti    [VERB_PREF_NEG Neg=polarity]
[VPREFSP]           →     a     [VERB_PREF_SPM3S Spm3s=agrmt3]
[VPREFTM]           →     aa    [VERB_PREF_PERF Perf=perfective]
[VPREFOP]           →     bu    [VERB_PREF_OPM13 Opm13=objectprefix13]
[VPREFOP]           →     bu    [VERB_PREF_OPM13 Opm13=objectprefix13]
[VEXT]              →     es    [VERB_PREF_CAUS Caus=causative1]
[POSTV]             →     mu    [VERB_SUFF_POST Post=postverbal]
```

Figure 3: Sample rules of the verb grammar (Non-terminals are enclosed in square brackets:

[VPREFNEG] = *verb prefix negative;* [VPREFSP] = *verb subject prefix;* [VPREFTM] = *verb prefix tense marker;* [VPREFOP] = *verb prefix object marker;* [VROOT] = *verb root;* [VEXT] = *verb extension;* [VEND] = *verb end;* [POSTV] = *Verb suffix post verbal. Symbols after morphemes in bold-face indicate categorical information. | means disjunction.)*

To compile a grammar like the one in Fig. 3 into an unweighted finite-state acceptor, the grammar rules are converted into a directed graph according to the following principle: for all non-terminals *A* and *B*, if there exists a rule $A \rightarrow \ldots B \ldots$, then the graph contains an edge $A \rightarrow B$. After this pre processing step, a *topological order* (cf. Cormen et al. 2001) of the resulting graph is computed. If the graph is cyclic (which means that the underlying grammar is recursive), the (acyclic) *component graph* of all *strongly connected components* is used instead.[4] All the right-side grammar rules that share the same left side are disjunctively combined and, for every non-terminal *A*, a finite-state acceptor *FSA*(*A*) representing all the right sides for *A* is computed. In a final step, each non-terminal A is substituted by its corresponding FSA in *reverse topological order*, beginning with the FSAs for the grammar rules which do not have further non-terminals in their right sides. Note that the grammar need not be in a special format (right-linear etc.) to apply this procedure.

To illustrate these steps, Fig. 4a shows the FSA for non-terminal VERB, while Fig. 4b shows the FSA for VROOT according to our grammar fragment. The FSA for VROOT of Fig. 4b is substituted into the one in Fig. 4a, replacing the two occurrences of VROOT

[4]The regularity check also takes place at this stage: all non-terminals in a strongly connected component (there may be more than one in case of mutual recursion) must occur in *either a right- or left-linear* form in the sub-grammar restricted to these non-terminals. This for example excludes rules like $S \rightarrow a\ S\ b \mid c$ which generates a non-regular language.

(transitions 0 → 1 and 5 → 6). All other symbols in Fig. 4a are replaced in a similar way by their corresponding automata, yielding a finite-state acceptor representing the whole grammar fragment.

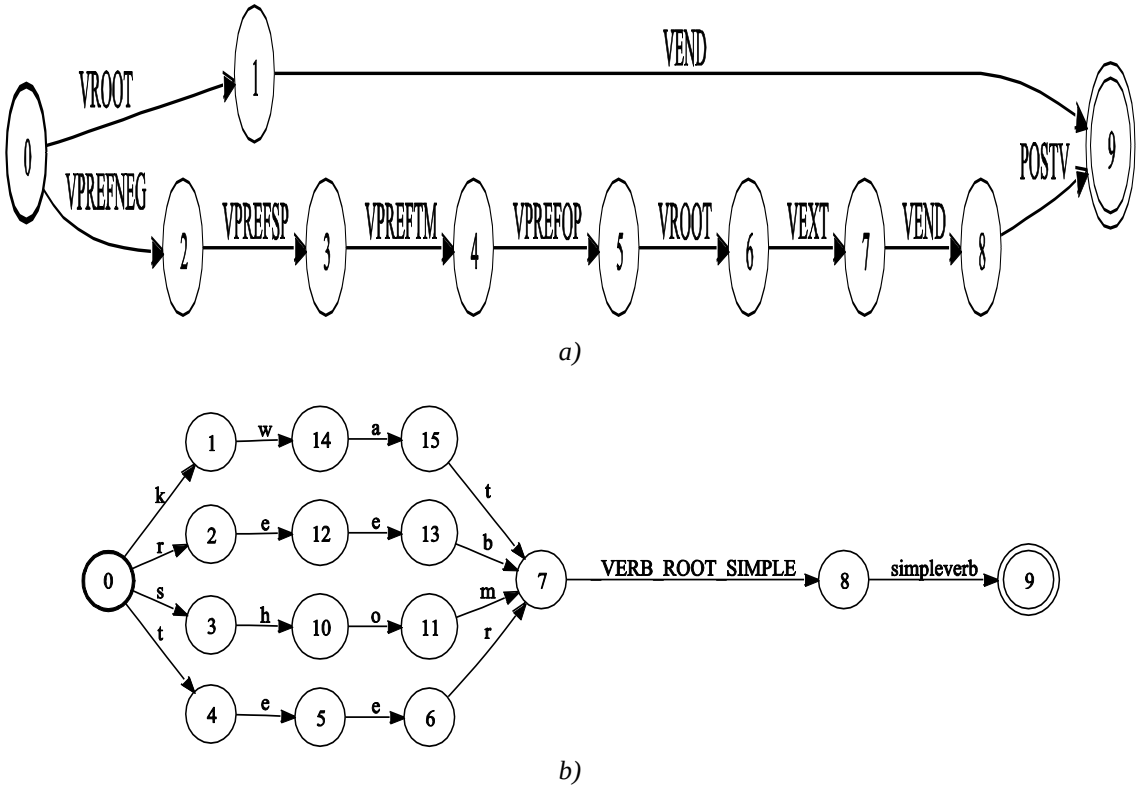


Figure 4: FSAs corresponding to grammar rules of Fig. 3. a) FSA for VERB*, b) FSA for* VROOT.

The grammar fragment in Fig. 3 accounts for verb forms like *teera* 'beat', *reeba* 'see', *kwata* 'catch' and *shoma* 'read'. However, we need also to make provision for *shutama* 'sit', *gyenda* 'go', etc, which are not represented by the fragment. The grammar fragment is simplified, since it would be computationally too expensive to include the complete set of Runyakitara verb stems, resulting in grammars with tens of thousands of rules. We therefore partitioned the set of verb stems into eight equivalence classes, each class containing all the verb stems that participate in the same word-grammatical constructions and represented by a unique symbol in the grammar. After compiling the word grammar into a finite-state acceptor $\boldsymbol{A_G}$, a final processing step then substitutes each equivalence-class-denoting symbol by the set of its corresponding verb roots. This also simplifies the addition of new verb roots, since the grammar automaton remains unchanged and only the final substitution has to be recomputed. Nevertheless, the construction of a grammar with approx. 330 rules and allowance for subsequent substitution takes less than a quarter of a second on a modern CPU, resulting in a finite-state acceptor with ≈ 800 states and ≈ 1,200 transitions.

The language (in the technical sense) generated by the grammar is still just a set of morpheme concatenations forming strings, some of which are nothing more than abstract concatenations (morphotactics) without proper phonological and orthographical representation. Fig. 5 shows some of the strings described by the grammar.

```
a      [VP_SPM3S Spm3s=agrmt3s]
aa     [PERF Perf=perfective]
bu     [VP_OPM13 Opm13=agrt13]
reeb   [VERB_ROOT_SIMPLE]
a      [VERB_END End=indicative]

a      [VP_SPM3SSpm3s=agrmt3s]
aa     [PERF Perf=perfective]
bu     [VP_OPM13 Opm13=agrt13]
reeb   [VERB_ROOT_SIMPLE]
a      [VERB_END End=indicative]
mu     [POST Post=postverbial]
```

Figure 5: Some elements of the language generated by the verb grammar (morphemes are in bold face, strings like `End=indicative` *indicate feature-value pairs).*

Both **a-aa-bu-reeb-a** and **aa-bu-reeb-a-mu** are valid underlying forms in Runyakitara, representing correct grammatical information, but are not correctly spelt and well pronounced words. The correct forms are *yaabureeba* and *yaabureebamu*, which require a change of the first **a** to **y**.

To deal with this kind of allomorphic variation, we switch from the *Item-and-Arrangement* model inherent in the above grammatical approach to a more process-oriented *Item-and-Process* model (see Hockett 1954 for a description of these models).

2.3.4 Context-dependent rewriting rules: morpho-phonological and orthographical rules

Rewriting rules cover morpho-phonological and orthographical issues and are of the abstract form:

$$\alpha \rightarrow \beta / \gamma\_\delta$$

This means that an instance denoted by α is replaced by an instance β, if α is preceded by a γ and followed by a δ. It is well-known (Johnson 1972, Kaplan & Kay 1994) that rules of this kind stay within the realm of regular devices if certain conditions apply: (i) α, β, γ and δ must denote regular language features and (ii) rules are not allowed to apply to their own output.

For example, the replacement rule

```
a→ y/_[VP_SPM3S Spm3s=agrmt3s] aa [PERF Perf=perfective]
```

states that **a** is replaced by **y**, whenever **a** (a verb prefix marker for third person singular) occurs before **aa** (verb prefix marker for perfective). This kind of rule will change **a-aa-reeb-a** to *y-aa-reeb-a* (he has seen), a well formed R-R word.

We developed a set of 34 context-dependent replacement rules for R-R verbs. The rules in this category are able to delete, substitute, and insert symbols in the string as long as the context is clearly defined. Each replacement rule $\boldsymbol{RR_i}$ – which corresponds to an infinite *regular relation* (see Kaplan & Kay 1994) – is compiled into a finite-state transducer, and all resulting rule transducers are in turn composed, resulting in one big transducer representing all the rules simultaneously (∘ denotes composition):

$$\boldsymbol{RR} =_{def} \boldsymbol{RR_1} \circ \boldsymbol{RR_2} \ldots \circ \ldots \boldsymbol{RR_k}$$

In terms of computational complexity, compiling these kinds of rules is the most expensive step of the whole construction.[5] Compilation needed approx. 1.5 seconds, yielding a finite transducer $\boldsymbol{RR}$ with ≈ 170 states and ≈ 93,000 transitions. To apply the replacement rules to the strings generated by the grammar, both finite-state machines are composed:

$$\boldsymbol{A_G} \circ \boldsymbol{RR}$$

All the allomorphic changes performed by the combined rule transducer $\boldsymbol{RR}$ manifest themselves on the output tape of $\boldsymbol{A_G} \circ \boldsymbol{RR}$. But these changes have to occur at the surface, input level. We achieve the desired effect by *inverting* the transducer, which is accomplished by switching the input and output tape. But before doing so, we have to get rid of the categorical information (introduced in the stem and affix lexicons) still present on both tapes of the transducer. For that purpose, we define a simple unconditional rewriting rule which replaces each category by ε, the empty string, effectively deleting all categories:

$$[\texttt{<Category>}] \rightarrow \varepsilon$$

Here `<Category>` is a special meta-symbol, denoting all the grammatical categories defined in the symbol signature. The transducer for the Runyakitara verb morphology is then defined as follows ($^{-1}$ denotes inversion):

$$(\boldsymbol{A_G} \circ \boldsymbol{RR} \circ ([\texttt{<Category>}] \rightarrow \varepsilon))^{-1}$$

This transducer maps Runyakitara verb forms (incorporating all the allomorphic changes) as sequences of underlying forms alternating with categorical information about these morphemes (see the next section for sample output).

2.3.5 Sample output

The output of the system includes morphemes, their categories and features. Fig. 6 presents some sample output.

[5] This is due to the various complementary operations for restricting the replacements to the correct contexts (*P-iff-S*-operator, see Kaplan & Kay, 1994).

```
mukakubaasa:        mu       [VERB_PREF_SPM2P Spm2p=agrmt2p]
                    ka       [VERB_PREF_FPAST Fpast=remotepast]
                    ku       [VERB_PREF_OPM15 Opm15=agrt15]
                    baas     [VERB_ROOT_SIMPLE Simple=simpleverb]
                    a        [VERB_END_IND Ind=mood]

mukakubaaga:        mu       [VERB_PREF_SPM2P Spm2p=agrmt2p]
                    ka       [VERB_PREF_FPAST Fpast=remotepast]
                    ku       [VERB_PREF_OPM15 Opm15=agrt15]
                    baag     [VERB_ROOT_SIMPLE Simple=simpleverb]
                    a        [VERB_END_IND Ind=mood]

zizigyegyenesa:     zi       [VERB_PREF_SPM10 Spm10=agrmt10]
                             [VERB_PREF_PRESENT Present=habitual]
                    zi       [VERB_PREF_OPM10 Spm10=agrt10]
                    gyegyen  [VERB_ROOT_SIMPLE1 Simple1=simpleverb1]
                    es       [VERB_EXT_CAUS Caus=true]
                    a        [VERB_END_IND Ind=mood]

zizigyegyenera:     zi       [VERB_PREF_SPM10 Spm10=agrmt10]
                             [VERB_PREF_PRESENT present=habitual]
                    zi       [VERB_PREF_OPM10 Spm10=agrt10]
                    gyegyen  [VERB_ROOT_SIMPLE1 Simple1=simpleverb1]
                    er       [VERB_EXT_LOC Loc=prep]
                    a        [VERB_END_IND Ind=mood]

zizigyegyenera:     zi       [VERB_PREF_SPM10 spm10=agrmt10]
                             [VERB_PREF_PRESENT Present=habitual]
                    zi       [VERB_PREF_OPM10 Opm10=agrt10]
                    gyegyen  [VERB_ROOT_SIMPLE Simple1=simpleverb1]
                    er       [VERB_EXT_APPL Appl=prep]
                    a        [VERB_END_IND Ind=mood]

zizigyegyenerera:   zi       [VERB_PREF_SPM10 Spm10=agrmt10]
                             [VERB_PREF_PRESENT Present=habitual]
                    zi       [VERB_PREF_OPM10 Opm10=agrt10]
                    gyegyen  [VERB_ROOT_SIMPLE Simple1=simpleverb1]
                    erer     [VERB_EXT_INT Int=degree]
                    a        [VERB_END_IND Ind=mood]
```

Figure 6: Sample output of RUNYAGRAM

Taking the first word of the above output as an example, *mu-ka-ku-baas-a* ‘you then managed it (this tense starts from last month onwards) has morphemes ***mu-*** serving as a subject prefix marker for class two and a plural marker having an agreement function; ***ka-*** is a tense marker indicating remote past***, ku-*** is an object prefix marker for class 15 that also conveys agreement, ***baas-*** is a verb root for simple verbs, and ***-a*** is a verb end for the indicative mood.

2.4. Testing

Testing is one of the more complex tasks in morphological analyzer development (Beesley and Karttunen 2003) and therefore needs a lot of care and patience. One of the important aspects of *fsm2* is its testing functionality, which helps developers test and debug morphological analyzers. The *fsm2* testing functionality can be represented as:

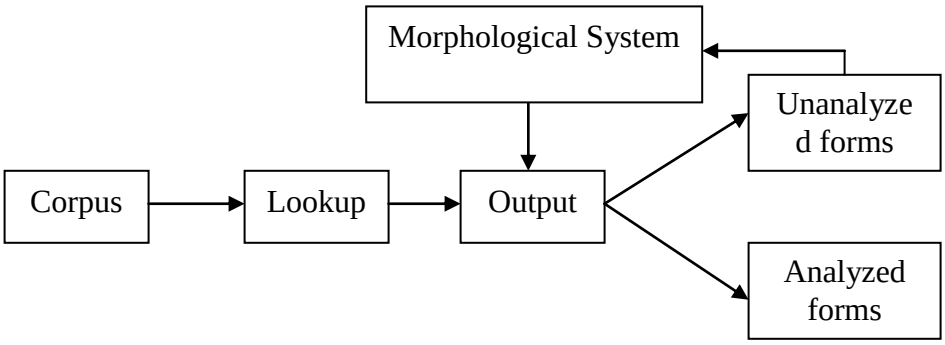


Figure 6: Testing process in fsm2

To test applicability to Runyakitara, a list of 3971 Runyankore-Rukiga verbs was extracted from the dictionary and a Runyakore orthography reference book (Taylor 1957). This list constituted the raw material for testing. Using the lookup operation provided by *fsm2*, the words were looked up in the analyzer and the results were stored in two files: one with the analyzed forms and another containing the unanalyzed forms. The unanalyzed forms were re-examined for possible subsequent inclusion in the morphological system.

The following table presents the results for RUNYAGRAM:

Corpus	3971 tokens	Percentage
Analyzed forms	4604	86%
Unanalyzed forms	559	14%
Precision (correctly analyzed)	3820	82%

Table 3: Testing Results.

The above results indicate that the R-R verb system analyzer, at its current stage of development has been successful in analyzing 86% of running text. The precision for the system is at 82%.

There is no general concensus over what level of accuracy is required in a morphological analyzer tested/subjected to real world text (Hurskainen 2004). In well-documented

languages, the required level of precision and recall depends greatly on the application targeted, for instance, one needs a high precision for spell-checking. Such a target may be realised because the corpus is in most cases accurate. In less documented languages, it may be difficult to achieve a high percentage because of the lack of accurate corpus. In other words, even when the morphological analyzer is 100% accurate, it may not achieve 100% recall because corpora in Runyakitara for example contain words that are not part of the morphological analyzer such as proper names, contractions, abbreviations, foreign words and spelling errors.

We consider our 82% precision as a positive outcome at the current level of development with great hope that it will improve. Even when the required level of accuracy was 100%, 82%, reflects positively on the ability of *fsm2* to analyze the verb morphology of R-R at its current development. But since we focus on language learning system, we will be able to avoid using material that is not properly analyzed by RUNYAGRAM.

2.5. Conclusion and future research

This study demonstrates the successful application of a finite state approach to the analysis of Runyakitara verb morphology. Although the finite-state approach is already considered a standard model in the morphological analysis of languages, its specification via a context free grammar to analyze a Bantu language had not been explored. Language-specific knowledge and insight have been applied to classify and describe the morphological structure of the language, and quasi context-free rules and rewriting rules have been formulated to analyze and generate the grammatical verbs of Runyakitara.

The above-described results represent a preliminary effort at building a morphological analyzer for Runyakitara, a group of closely-related Bantu languages. RUNYAGRAM, which is based on a combination of the *Item-and-Arrangement* and *Item-and-Process* models proposed by Hockett (1954; 1958), shows how these models may be applied to Runyakitara morphology.

Specifically, this study has provided:

1. The first computational description of the orthography of the Runyakitara verbs
2. A proof that the *fsm2*-based approach (context-free grammar + rewriting rules) is applicable to a morphologically complex set of Bantu languages such as the Runyakitara languages.
3. An enrichment of the common Bantu template to account for the more specific situation in R-R. The elaborated template improves our understanding of Bantu morphology, implying that Bantu languages may differ in certain morphological aspects.

2.6 Future research

The overall plan for this research is to develop a means of accounting for all the Runyakitara word categories to be analyzed by *fsm2*. This will result in a comprehensive morphological analyzer for Runyakitara, which will provide input for many other planned applications, such as learning systems and machine translation tools.

Acknowledgments
The authors would like to acknowledge the continued support and expert advice of Prof. Arvi Hurskainen (Helsinki, Finland) and Prof. John Nerbonne (University of Groningen, The Netherlands). We are especially grateful to an anonymous reviewer for his elaborate and astute comments on an earlier version of this paper. Finally, we would like to thank the editor, Axel Fleisch, for his work.

This chapter and the earlier version of it published in Nordic Journal of African Studies 19(1) is a result of a six month visiting research fellowship in Germany funded by DAAD (The German Academic Exchange Service). We are grateful for this support.

Appendix A – Detailed description of Runyakitara Morphology

Slot	Meaning	morpheme	Word formed	Gloss
0	Verb root	Vroot	*gyend*-a	go
1	Verb extensions (VEXT)	Ca – causative (es) Apl – applicative (er) Rec – reciprocal (an) Pas – passive (w) Int – intensive (erer) Stat – stative (ek) Rev - reversive	gyend-*es*-a gyend-*er*-a gyend-*an*-a reeb-*w*-a gyend-*erer*-a gyend-*ek*-a teek-*uur*-a *Also possible:* gyend-es-ebw-a gyend-an-is-a gyend-an-is-ibw-a	make to go go for go with be seen go specifically for - remove (on stack)
2	Verb end (VE)	Ind – indicative (a) Subj – subjunctive (e) Past – past tense (ire)	y-aa-gyend-*a* n-gyend-*e* n-gyenz-*ire*	he has gone may I go I went
3	Post final	Pf1 – adverbial (ho, yo, mu) Pf2 – mitigator (ga)	gyend-a-*yo* ti-n-ka-gyend-a-*ga*	go there I have never gone
4	Aspect marker	Asp – reflexive (e)	ku-*e*-reeb-a	to see oneself
5	Object pronouns	Op1 – object pronouns (18) Op2 – object pronouns (18)	*ba*-gyend-e mu-*mu*-*n*-reeb-er-e	Let them go You see him for me
6	Tense/aspect markers	Inf – infinitive (ku) Hab – habitual (ø) Pf – perfective (aa) Ff – far future (ria/rya) Rp – remote past (ka)	*ku*-gyend-a n-gyend-a n-*aa*-gyend-a n-*dya*-gyend-a n-*ka*-gyend-a	to go I go (everyday) I have gone I will go (far future) I went (last year)
7	Negation marker	Neg2 – negative (ta)	ku-*ta*-gyend-a	not to go
8	Subject pronouns	Sp – subject pronouns (18)	*n*-aa-gyenda *tw*-a-gyend-a	I have gone we have gone
9	Aspect marker	Asp – progressive (ni)	*ni*-ba-gyenda	they are going (now)
10	Negation marker	Neg1 – negative 1 (ti)	*ti*-baa-gyend-a	they have not gone

Appendix B – *fsm2* Script for Creating the Verbal Analyzer

```
# Define a macro mapping verb equivalence class symbols (%SYMBOL%)
# to sublexicons stored in a file called %SYMBOL%.lex.
macroverb_substitution(%SYMBOL%)
       loadlexicon %SYMBOL%.lex
       optimize
       map %SYMBOL%
endmacro

# Load symbol signature
load symspec ../../symbols/rr.sym

# Load all verb roots stored in a number of files and associate them
# with a symbol denoting the verb's equivalence class (VERBFORMx).
# This creates a substitution map associating each verb class with the
# verb roots in this class

callverb_substitution(VERBFORM1)
callverb_substitution(VERBFORM2)
callverb_substitution(VERBFORM3)
callverb_substitution(VERBFORM4)
callverb_substitution(VERBFORM5)
callverb_substitution(VERBFORM6)
callverb_substitution(VERBFORM7)
callverb_substitution(VERBFORM8)

# Compile the verb subgrammar and optimize it
load grammar verbs
optimize

# Perform the verb root substitution and optimize the result
substitute
optimize

# Compile the rewriting rules
# and compose them with the result of the step before
load contextrules verbs.rules
compose

# Delete all the category information from the lower tape
regex "[<category>] --> []"
compose
# Finally, swap input and output tape of the transducer
invert
optimize
```

Chapter 3

Fsm2 and the morphological analysis of Bantu nouns: initial experiences with regard to Runyakitara

(This was published in International Journal of Computing and ICT Research, Special Issue, Vol. 4(1) 58-69, October 2010 as: Fridah Katushemererwe & Thomas Hanneforth, Fsm2 and the Morpho- logical Analysis of Bantu Nouns – First Experiences from Runyakitara.

Abstract

This paper describes the application of finite state methods, fsm2 in particular, to the automatic analysis of Bantu nouns, in particular, Runyakitara. This study represents an initial effort in developing a computational analysis of Runyakitara. It provides a detailed description of Runyakitara noun classes and the manner in which they were analyzed using fsm2. At the current stage of system development, 80% of Runyakitara nouns are correctly analyzed, and no forms were incorrectly analyzed. This is a positive outcome, providing further corroboration that fsm2 can be successfully used to analyze the morphology of the Bantu languages.

Key words: **Finite-State methods, fsm2, morphological analysis, Bantu languages, Analysis of Runyakitara**

3.1 Introduction

Although computational morphology is an essential input for other text analysis applications, the literature on its development with regard to most of the Bantu languages is still sparse. Morphological analysis of natural languages is a well-studied field, and the effectiveness of finite state methods in analyzing the morphology of natural languages has been well demonstrated (Karttunen 2003). Finite-state technology is considered the preferred model for representing the phonology and morphology of natural languages (Wintner 2007), and the model has been used to computationally analyze natural languages such as English, German, French, Finnish, and Swahili, to mention just a few (Beesley and Karttunen 2003). Most implementations for Bantu languages (Pretorious & Bosch 2003, Hurskainen 1992, Muhirwe 2007, Elwell 2005) have used *lexc* and *xfst*, of Beesley and Karttunen (2003). We chose fsm2 as opposed to lexc/xfst because:

i) fsm2 has a notion of grammar as opposed to lexc and xfst. All one does in xfst is to encode a finite state automata representing grammar in lexc's class mechanism. Given the nature of Runyakitara as described in the introduction and in Chap. 2 (above), we preferred a tool that codifies grammar as the best solution for the behaviour of Runyakitara morphology.

ii) Lexc/xfst is proprietary as opposed to fsm2, which is open source software. Although there is an open source xfst clone, FOMA, developed by Hulden (2009), the choice had already been made to develop Runyakitara morphological analyzer with fsm2.

Finite-state methods (*fsm2*) were hence used to compile a comprehensive system containing all the significant lexemes of Runyakitara nouns. To date, the Runyakitara noun morphological analyzer is a combination of a symbol specification, a noun grammar module and a replacement rule module. The purpose of developing the tool is to provide sharable morphological grammar rules for Runyakitara nouns in an organized framework so that they can be used in other applications. Currently, there are no such rules for Runyakitara. If, however, important language applications like spell-checkers are to be developed for the language, a word grammar checker (a morphology) is required.

Although the Bantu languages are classified as largely agglutinative and exhibit significant inherent structural similarity, they differ substantially in terms of their phonological features to such an extent that each Bantu language likely requires an independent morphological analyzer.

This chapter focuses on the treatment of the nouns of Runyakitara, a closely-related group of Bantu languages, in a finite-state programming environment. The decision to focus on nouns was taken because nouns constitute a major word category in Runyakitara and play a major role in syntactic analysis. Secondly, the noun classification system in Runyakitara is computationally interesting because of the number of noun classes involved and the derivational, compounding and reduplication phenomena Runyakitara nouns are comprised of. Thirdly, we wish to focus on nouns in the application of computational morphology to computer-assisted language learning (later in this dissertation).

3.2 Previous work on the morphological analysis of the Bantu languages

A considerable amount of work has been performed on the application of finite state methods to the analysis of the Bantu languages. Using Xerox Finite State tools, Karttunen (2003) implements a realizational framework to model Lingala verb morphology. This approach focuses on the use of replacement rules to gradually construct the verb from the root, piece by piece.

The Xerox Finite State technology has also been utilized in the development of a prototype analyzer for Zulu (Pretorius & Bosch 2003). This analysis uses lexc, xfst and replacement rules to account for the morphotactics, morpho-phonology and orthographical issues in the Zulu language. To account for the long-distance dependencies found in Zulu morphology, Pretorius and Bosch use flag diacritics as described in Beesley & Karttunen (2003).

Work has also been carried out on Swahili using a language-specific morphological parser (Hurskainen 1992) known as SWATWOL. This parser is a two-level analyzer that similarly accounts for the morpho-syntax and morpho-phonology of Swahili.

Related to the above is the Swahili language manager SALAMA (Hurskainen 2004), which represents a primary process in the development of multiple computational applications. SALAMA is a computational environment that manages written Swahili and provides linguistic processing with a view to supporting various kinds of language applications. It comprises the standard Swahili lexicon, a full morphological and morpho-phonological description of Swahili, a rule-based system for solving word-level ambiguities, a rule-based system for tagging text syntactically, a rule-based system for

handling idiomatic expressions, proverbs and other non-standard clusters of words and semantic tagging and disambiguation system for defining correct semantic equivalents in English. SALAMA and SWATWOL are language specific and do not account for all the types of problems encountered in Runyakitara.

Muhirwe (2007) describes the computational analysis of Kinyarwanda morphology. He applies the Xerox Finite State compiler to model Kinyarwanda phonological alternations, concentrating on orthographical rules. It is important to note that such rules are language dependent. Therefore, the rules for Kinyarwanda or Kiswahili language are not directly applicable to other Bantu languages, even though they belong to the same group – the Bantu language group.

Finite state methods have also been applied to the analysis of Seswana verb morphology (Pretorius 2008); tonal marked Kinyarwanda (Muhirwe 2010; Hurskainen 2009) and solutions for reduplication in Kinyarwanda (Muhirwe & Trosterud 2008).

Considerable work has therefore been done on specific languages, mainly applying Xerox finite state methods of morphological analysis, and a number of implementations have been successful. However, the fact that there are over five hundred (500) Bantu languages means that a large number of them have not yet been subject to any such analysis. The lack of literature dealing with the Runyakitara languages suggests that they belong to the latter category. In addition, *fsm2* as a scripting language has not yet been applied or implemented with regard to any of the Bantu languages. This makes a publication on the application of *fsm2* to the automatic analysis of Runyakitara nouns both unique and relevant.

3.3 Methodology

The design of the system was carried out in three phases: formalization, implementation, and testing. Formalization involved most of the linguistic investigation required throughout the course of the design. Nouns were extracted from a dictionary, '*Kashoboorozi y'Orunyankore-Rukiga'* (Oriikiriza 2007). Initially, manual coding was undertaken to identify the sub-classes of the main classes of nouns. Classes without prefixes had to be identified manually. This was complicated by the fact that the entries in the *Kashoboorozi* do not indicate the noun class prefixes that apply to nouns, enormously increasing the amount of manual work required.

The core of the system is a grammar written using *fsm2* formalism (Hanneforth 2009). All the regular aspects of nouns were encoded as regular expressions in compliance with a quasi context-free grammatical framework. The replacement rules were encoded as regular expressions supported by fsm2. The grammar and rules were composed together using a composition operator that *fsm2* also provides.

When the model was completely implemented, it was tested using the lookup tool, also included in fsm2 (Hanneforth 2009). Testing was conducted on a corpus of Runyakitara nouns extracted from a weekly newspaper (Orumuri) and a teachers' handbook of *Runyankore-Rukiga* orthography.

3.4 Highlights of Runyakitara noun morphology

Similar to all Bantu languages, Runyakitara has a noun class system. Demuth (2003) describes how the Bantu noun classification system is conveyed by a set of grammatical morphemes rather than independent lexical items. The classes are morphologically signalled by noun class prefixes and agreement markers. The latter indicate that nouns also function as part of a larger concord system. There is evidence to suggest that the noun class prefix of a noun agrees with all the constituents of a noun phrase, such as adjectives, pronouns and numerals. Researchers of the Bantu languages agree that that noun class features are determined by grammatical number, semantics (i.e. whether they are human/animal/non-living things) and, in some cases, arbitrarily (Aikhenvald 2006; Katamba 2003).

In Runyakitara, a noun can consist of a root and an affix, the affix usually being a prefix. Suffixation is also possible and mainly involves forms derived from verbs, adjectives and adverbs. Affixation occurs by adding an appropriate class prefix in the majority of cases or by replacing the final stem vowel of the derived forms. Such nouns comply with the requirements of their respective classes; for example, a noun *omu-shom-i* (reader) is derived from *ku-shom-a* (to read) but complies with class 1/2 for humans, the derivational process being irrelevant in this case.

Nouns in Runyakitara are also associated with an initial vowel as a pre-prefix to the root. According to Ndoleriire & Oriikiriza (1990), these are **a,** (***aba****ntu*) **e**, (***eki****tookye*) and **o**, (***omu****ntu*). There are rules that govern the occurrence of the initial vowel. If the noun class prefix has the vowel **a** (e.g. **ba**, **ma**), the initial vowel will be **a,** thus ***ama****ta* 'milk' ***aba****kazi* 'women'. When the noun prefix has **i** or -, the initial vowel is **e,** for example ***e****k****i****tookye*, ***e****m****i****ti*, etc. The initial vowel is o when the noun class prefix has **u**, ***o****m****u****ntu* 'person', ***o****m****u****ti* 'tree'. When a noun is preceded by a preposition such as **omu** 'in', **aha** 'at', the initial vowel is dropped e.g. *omu muti* 'in the tree'.

Although Bantu languages have a general noun classification system, each language has its own unique sub-classification system. The noun classification of Runyakitara is regarded as specific to the language group and needs to be dealt with separately.

Whereas nominal morphology is a well-studied element of the Bantu languages, classification systems still lack detailed descriptions, especially to the extent required for computational analysis. In part, it is for this reason that a detailed description and computational analysis of Runyakitara morphological grammar is being undertaken in this study.

3.4.1 Runyakitara noun classification system

The noun class system used in this analysis has borrowed a great deal from Katamba (2003) and Taylor (1985). Katamba (2003) provides a detailed comparative analysis of different classification systems, singling out the Bleek-Meinhof system and its revisions as the benchmark. This study has provided important insights for Runyakitara analysis. To cater to the needs of Runyakitara, Taylor (1985) details a classification system of Runyakitara nouns describing 17 classes, but with few or limited sub-classes. The description of a noun class system of Runyakitara provided in Ndoleriire & Oriikiriza (1990) has twenty (20) noun classes. However, this description falls short of a numbering

system and a detailed description of the sub-classes belonging to either singular or plural. The table below, therefore, redresses these omissions in providing a more comprehensive description of the Runyakitara noun class system.

Class	**Singular**	**Plural**	**Semantics**	***Example***	**Gloss**	**Usage**
1/2	o-mu-	a-ba	Human	*o-mu-kazi* *a-ba-kazi*	Woman Women	Takes on both singular and plural
1a	o-mu-	-	Names referring to deity	*o-mu-hangi*	Creator	Only singular
1b/2b	-	baa-	Human, kinship	*shwento* *baa-shwento*	Uncle Uncles	Takes on both singular and plural, but no prefix for singular
2a	-	a-ba-	Human, group	*a-ba-ryakamwe*	Group name	Only plural forms
3/4	o-mu	e-mi-	Plants, fruits,	*o-mu-ti/e-mi-ti*	Tree(s)	Both singular & plural
3a	o-mu-	-	Uncountable	*o-mu-isyo*	Breath	Singular only
4a	-	e-mi-	Abstract names	*e-mi-gyendere*	Way of walking?	Only plural
5/6	e-ri-	a-ma-	Some parts of the body	*e-ri-isho/a-ma-isho*	Eye(s)	Both singular & plural
5a	ei-	a-ma-	Miscellaneous	*ei-teeka/a-ma-teeka*	Policies	Both singular & plural
5b	ei-	-	Abstract names	*ei-tetsi*	Pampered?	Only singular
6a	-	a-ma-	Mass nouns	*a-ma-te*	Milk	Only plural
7/8	e-ki-	e-bi-	Objects, misc	*e-ki-ti/e-bi-ti*	Tree (s)	Both singular & plural
7	e-ki-	-	Abstract	*e-ki-niga*	Anger	Only singular

Class	Singular	Plural	Semantics	*Example*	Gloss	Usage
8	-	e-bi-	Mass nouns	*e-bi-bembe*	Leprosy	Plural only
9/10	en-	en-	Animals and borrowed words	*e-nte*	Cow(s)	Singular and plural
9	-	-	borrowed words, derived words	*ebahaasa*	Envelope (s)	Singular & plural
10	-	-	borrowed words	*bwino*	Ink	Singular & plural
11/10	o-ru-	en-	Insects, plants miscellaneous	*o-ru-shozi*	Mountain(s)	Singular & plural
12/14	a-ka-	o-bu-	Small items, miscellaneous	*a-ka-buuza*	Question mark(?)	Singular & plural
12	-aka-	-	Abstract nouns	*a-ka-bi*	Danger	Abstract
14	-	o-bu-	abstract nouns	*o-bu-cureezi*	To be humble	Abstract
13	-	o-tu-	Abstract and diminutives	*o-tu-ro*	Sleep	Abstract
15/6	o-ku-	a-ma-	Some body parts	*o-ku-guru/amaguru*	Leg(s)	Singular & plural
16	aha-	-	Location	*aha-kaanyima*	Behind the house	Singular
17	oku-	-	Location	*oku-zimu*	Underground	Singular
18	omu-	-	Location	*omu-nda*	In the stomach	Singular
20/21	o-gu-	a-ga-	derogatory	*o-gu-kazi/a-ga-kazi*	Bad/ugly woman	Singular & plural

Table 1: Noun classification system of Runyakitara

There are generally twenty noun classes in Runyakitara, although only eighteen are in use, as two are derogatory and tend to be ignored especially in written contexts. Most of the classes are paired in singular and plural, but there are exceptional cases where a class is in either singular or plural, as illustrated above. As indicated in the table, the status of either singular or plural may be marked by a null prefix in either case.

It should also be noted that some Runyakitara nouns do not take affixes but still belong to their semantic classes (e.g. *taata* 'Dad' in class one and *ebaafu* 'basin' in class 9 have neither prefixes nor suffixes). The class such nouns belong to is conveyed by the concord markers on nominal constituents such as verbs or adjectives (e.g. ***e**baafu **e**yangye n'**e**yera* 'my basin is clean').

Derivation is productive in Runyakitara, where nouns are derived from verbs, adjectives and adverbs. This process involves the addition of an appropriate class prefix and replacement of the final stem vowel (e.g. *o-mu-egi* 'student', which is derived from *o-ku-ega* 'to study'). Such nouns are treated under their respective classes as marked by prefixes.

Compound nouns are also productive in Runyakitara. Compound nouns result from combining two or more words of different meanings to form one word with a single meaning. The combinations mainly involve a noun and another noun, verb and noun or noun and adjective. The case of such nouns is based on the prefix of the first noun, but most belong to class nine, which is open to new words.
Reduplicated nouns are rare in Runyakitara, although they can occur in abusive speech (e.g. *omuntuntu* 'not-worthy of/as a person'). Runyakitara also allows other reduplicated forms of nouns that belong to the core of the language.

3.5 Formalization

Given the above highlighted features of Runyakitara noun morphology, a quasi context-free grammar, specifically employing the simple substitution approach proposed by Mohri & Sproat (1996), is preferred as the appropriate model for Runyakitara morphotactics because:

- Rules to constrain the order of morphemes are easily written and can output strings
- Noun classes with their semantic roles can easily be accounted for in quasi context-free grammar.

Formally, a context-free grammar is represented as follows:

- **Context-free grammar: G = [T, N, S, R]**
  - **T = a set of terminal symbols**
  - **N= a set of non-terminal symbols**
  - **S = a start symbol**
  - **R = a set of production rules in the form:**
    - **N' → X = replace N' by X, where N' ϵ N and X is a sequence of symbols from T U N**

Modelling Runyakitara nouns using the above approach can occur as follows:
Non-terminal symbols: [N] → [NP] [NR]
Terminal symbols [NP] → (omu|mu)
[NR] → ntu
Where N = noun; NP = noun prefix and NR = noun root.
This then raises the issue concerning the need to write rules for each and every noun root, which would hardly be feasible. A more pragmatic solution is to categorize noun roots according to their classification scheme. As a result, the categorized roots which belong to the known noun classes of Runyakitara were labelled with an abstract class identifier. For example, class 1-2 was labelled PEOPLE, so that all roots that belong to that class are

given that specific root. The PEOPLE symbol class is then substituted into the grammar (as a general term relevant in specific contexts).
Below is a table detailing the symbols given to the different classes:

Class	**Prefix**	**Semantics**	**SYMBOL FOR COMPUTATIONAL PURPOSES**
1/2	omu-aba	People	PEOPLE
1a	omu	Creator	CREATOR
1b/2b	baa	Kinship	KINSHIP
2a	aba	Group	GROUP
3/4	omu-emi	Plants	PLANT
3a	omu	uncountable	UNCOUNTABLE
4a	emi	Abstract	ABSTRACT4
5/6	eri-ama	Miscellaneous	MISC
5a	ei-ama	Some Body parts	BODY
5b	ei	Seasons	SEASONS
6a	ama	Mass	MASS
7/8	eki-ebi	Objects	OBJECTS
7	eki	Abstracts	ABSTRACT7
8	ebi	Mass nouns	MASS8
9/10	en-en	Animals	ANIMALS
9	-	Abstract nouns	ABSTRACT9
10	-	Mass nouns	MASS10
11/10	oru-en	Insects	INSECTS
12/14	aka-obu	Diminutives	ABST12
12	aka	Small and tinny	SMALL
14	obu	Abstract	ABSTRACT
13	otu	Mass nouns	MASS13
15/6	oku	Body parts	BPARTS
16	aha	Locative	LOCATION
17	oku	Locative	LOCA1
18	omu	Locative	LOCA2

Table 2: Classes and symbols representing roots

When the symbols representing roots are incorporated into the grammar, it looks like the extract below:

Non-terminal Symbols
[NOUN] --> [NOUN_PREF1][NOUN_ROOT1]
[NOUN] --> [NOUN_PREF1][NOUN_ROOT1A]
Terminal Symbols
[NOUN_PREF1] --> (omu|mu) [NOUN_PREF_1S 1s=npref1s]
[NOUN_PREF2] --> (aba|ba) [NOUN_PREF_2P 2p=npref2p]
[NOUN_ROOT1] --> [PEOPLE] [NOUN_ROOT_PS Ps=class1]
[NOUN_ROOT1A] --> [CREATOR] [NOUN_ROOT_1SI 1Si=singular1]

Runyakitara noun grammar extract

3.6. Implementation

The grammar is implemented using *fsm2* [Hanneforth 2009), a scripting language within the framework of finite state technology. Finite-state technology is considered the preferred model for representing the phonology and morphology of natural languages (Wintner 2007). The model has been used to computationally analyze natural languages such as English, German, French, Finnish, Swahili, just to mention a few (Beesley and Karttunen 2003). One of its main advantages is that it is bidirectional – it applies to both analysis and generation. This bi-directionality was a principal reason that the technology was selected for application to the morphological grammatical analysis of Runyakitara nouns.

fsm2 was chosen as a resource tool for a morphological grammar of Runyakitara nouns for a number of reasons:

i) It supports a full-set of algebraic operations defined on both un-weighted and weighted finite state automata and weighted finite state transducers (Hanneforth 2009). Algebraic operations are useful in designing complex morphological analyzers in a modular way.

ii) *fsm2* supports a number of equivalence transformations which change or optimize the topology of a weighted automation without changing its weighted language or relation, which means that an automaton can be minimized, determined, optimized etc.

iii) *fsm2* uses symbol signatures which map symbols as numbers that are internally recognized by the automata. Symbol signatures are useful in language modelling, since every word in a language is given an alphabetic symbol, and one of the tasks of a developer is to define symbols that represent morphemes and their categories.

iv) *fsm2* provides an efficient way of compiling morphological grammars that easily account for the co-occurrence of roots and inflectional affixes, a common interdependence in Runyakitara.

v) *fsm2* is able to load lexicons, grammars and replacement rules defined by the morphology developer and to automatically transform rules into transducers.

3.7 Application to Runyakitara nouns

The noun morphological system has a modular structure comprising a special symbol module/file, a noun grammar module and a replacement rule module. The three are combined to produce a single finite state transducer.
The following diagram demonstrates the overall architecture of the noun morphological system:

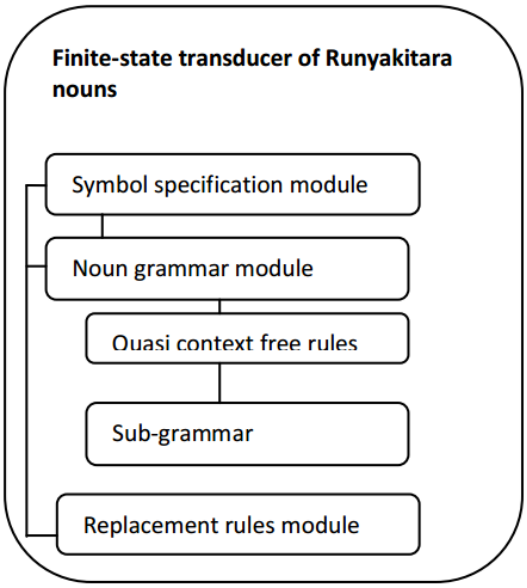


Figure 1: Noun morphological system architecture

3.7.1 A symbol specification module

This module provides a unique mapping of user-defined alphabetic symbols and categories as integers (numbers), which are used internally by the automata and the operations. A symbol signature relates symbols to their internal integer representation on a one-to-one basis in order to allow computation symbols (Hanneforth 2009). A symbol specification module for a noun is first loaded in *fsm2* before any other file is loaded.

3.7.2 Noun grammar module

The grammar module consists of a sub-set of quasi context-free rules accounting for concatenative nature of Runyakitara noun morphology. The grammar contains a large number of rules, but we present a sample, exemplifying the principles underlying the overall organization of the grammar:

Non-terminal Symbols

[NOUN] → [NP1S] [NROOT1] # omu-ntu (person)
[NOUN] → [NP2P] [NROOT1] # aba-ntu (people)
[NOUN] → [NP3S] [NROOT3] # omu-ti (tree)
[NOUN] → [NP4P] [NROOT3] # emi-ti (trees)

Terminal symbols:

[NP1S] → omu
[NP2P] → aba
[NROOT1] → ntu
[NP3S] → omu
[NP4P] → emi
[NROOT3] → ti

An extract of a noun context free grammar

Notes
[NP1S] – Noun prefix class 1 singular
[NP2P] – Noun prefix class 2 plural
[NP3S] – Noun prefix class 3 singular
[NROOT] – Noun root 1
[NROOT3] – Noun root for class 3
[NP4P] – Noun prefix class 4 plural

The above rules provide for prefixes and roots codify the allomorphy of the classifiers, an in particular how the choice of classifier prefix depends on the roots. We note at this point that this allomorphy (involving rough eighteen classes, shown in Table 2) is a further complexity in Runyakitara morphology. This in turn further confirms the need for instructional material to support learning such complex material.

Since it is not feasible to write rules for each root, *fsm2* provides two options for including roots when developing a morphological analyzer:

a) **Include statement**

This '#include' statement allows one to write a sub-grammar containing roots (either verbal or nominal), preferably in a separate file, then includes the roots in the grammar of prefixes and suffixes. For example, (i) is a grammar for class one nouns:

(i) [NOUN] → [NP1S][NROOT1]
[NP1S] → omu
#Include [NROOT1]

[NROOT1] → ntu
[NROOT1] → shaija
[NROOT1] → kazi
[NROOT1] → gyenyi

#Include VROOT1 will include *ntu, shaija, kazi*, and *gyenyi* noun roots in a grammar of 'noun prefix 1 singular and noun root1.' The above grammar applies to the following nouns: *omu-ntu a* 'person', *omu-shaija* 'man', *omu-kazi 'woma'* and *omu-gyenyi* 'visitor'.

This is the approach selected for implementation of Runyakitara noun morphology because it is easy to implement.

b) The second approach is to write lexicons for each class of noun roots, then use a substitution method to include them in the grammar. Roots already categorized into noun classes are arranged in lexicons, one lexicon for each noun class, compiled into finite state machines, and then assigned to non-terminal symbols. The comprehensive finite-state machine representing the entire nominal morphology refers to the non-terminal symbols, but replaces them by component finite state machines in its final realization. The substitution operation supported by *fsm2* effectively substitutes the roots into the grammar. Therefore, NROOTI, as illustrated above, can stand for many roots of the same noun class.

3.7.3 Morphotactics

The output from the context-free nominal grammar is still a set of morpheme concatenations forming strings, but some are still abstract concatenations (morphotactics) without proper phonological and orthographical representation. The following represents a sample of output from a Runyakitara noun grammar using *fsm2*:

```
omuegi : omu[NOUN_PREF_1S 1s=npref1s]egi[NOUN_ROOT_PS Ps=class1]
omuegizo : omu[NOUN_PREF_3S 3s=npref3s]egizo[NOUN_ROOT_3SI 3Si=singular3]
omuegoojooro : omu[NOUN_PREF_3S 3s=npref3s]egoojooro[NOUN_ROOT_3SI 3Si=singular3]
omueguzi : omu[NOUN_PREF_1S 1s=npref1s]eguzi[NOUN_ROOT_PS Ps=class1]
```

Output extract from a noun grammatical system

The above four examples: **omuegi, omuegizo, omuegoojooro** and **omueguzi** are valid morpheme sequences in Runyakitara, representing correct grammatical information, but are not correctly spelt words that reflect pronunciation accurately. The grammatical forms

are *omwegi*, *omwegizo, omwegoojooro* and *omweguzi*. This calls for a change of <u> to <w> in all cases. These and many similar cases of a phonological and orthographical nature are accounted for by replacement rules.

3.7.4 Replacement rules

Rules here cover morpho-phonological and orthographical occurrences. These phenomena are the subject of replacement rules, which are compiled into finite-state automata. *fsm2* provides for conditional and unconditional replacement rules. An expression:

$$\alpha \text{ -> } \beta \text{ / } \gamma\_\delta$$

indicates that alpha is replaced in *fsm2* by beta whenever alpha occurs in the context of gamma on the left and delta on the right. An example of a replacement rule that was included to account for morpho-phonological and orthographical processes is indicated below:

a) **u -> w / m_ (a | o | i)**

The above rule means that **u** is replaced by **w**, whenever **u** occurs between **b** and **a** or **o** or **i**. This kind of rule will change *om**u**-egi* to *om**w**egi*, *om**u**-egizo* to *om**w**egizo, omu-egoojooro* to *om**w**egojooro* and *om**u**-eguzi* to *om**w**eguzi*, resulting in each case in well formed Runyakitara words.

A subset of replacement rules for Runyakitara nouns was developed in accordance with the above framework. The rules in this category are able to delete, substitute and insert symbols in the string as long as the context is clearly defined.

The grammar transducer and the context rule transducer are combined to produce a single transducer whose output comprises grammatically correct Runyakitara nouns.

3.8. Grammatical output

The output of a noun analyzer includes morphemes, their categories and features. The following is sample output of a noun morphological analysis system:

```
abaakiizi : aba[NOUN_PREF_2P 2p=npref2p]akiizi[NOUN_ROOT_PS Ps=class1]
abaambari : aba[NOUN_PREF_2P 2p=npref2p]ambari[NOUN_ROOT_PS Ps=class1]
abaambuzi : aba[NOUN_PREF_2P 2p=npref2p]ambuzi[NOUN_ROOT_PS Ps=class1]
abaami : aba[NOUN_PREF_2P 2p=npref2p]ami[NOUN_ROOT_PS Ps=class1]
byetengo : bi[NOUN_PREF_8P 8s=npref8p]etengo[NOUN_ROOT_8PL 8Pl=plural8]
byevugo : bi[NOUN_PREF_8P 8s=npref8p]evugo[NOUN_ROOT_IT It=class7]
byeyariro : bi[NOUN_PREF_8P 8s=npref8p]eyariro[NOUN_ROOT_8PL 8Pl=plural8]
byeyemekye : bi[NOUN_PREF_8P 8s=npref8p]eyemekye[NOUN_ROOT_IT It=class7]
byeyera : bi[NOUN_PREF_8P 8s=npref8p]eyera[NOUN_ROOT_8PL 8Pl=plural8]
byeyerezo : bi[NOUN_PREF_8P 8s=npref8p]eyerezo[NOUN_ROOT_IT It=class7]
```

The above output accounts for all the information pertinent to nouns. Taking the first noun as an example, *aba* is recognized as a noun prefix for class two and a plural marker, *akiizi* is a noun root for people in singular form.

Interestingly, the nouns *abaami* and *baami* are essentially identical but are used in different circumstances. As already mentioned, nouns which do not have prefixes like *baami* (chiefs/men) are preceded by a preposition.

3.9. Testing

Testing is one of the more complex tasks in morphological analyzer development (Beesley and Karttunen 2003) and therefore needs a lot of care and patience. One of the important aspects of *fsm2* is its testing functionality, which helps developers test and debug morphological analyzers. The *fsm2* testing functionality can be represented as:

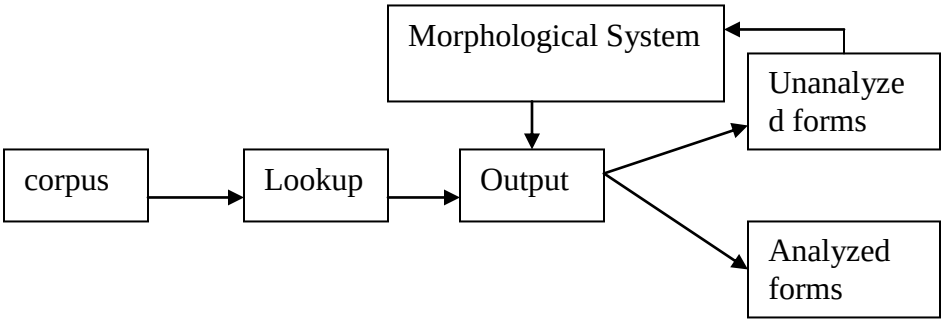


Figure 2: Testing process in fsm2

To test applicability to Runyakitara, Runyankore-Rukiga nouns were extracted from the weekly newspaper (*Orumuri*) and a Runyakore orthography reference book, (Taylor 1957). These nouns constituted the raw material for testing. Using the lookup operation provided by *fsm2*, the words were looked up in the analyzer and the results saved in two files: one for analyzed forms and another for unanalyzed forms. The unanalyzed forms were re-examined for possible subsequent inclusion in the noun morphological system.

The following table provides the results:

Corpus (nouns)	**Analyzed forms**	**Percentage (recall)**	**Unanalyzed forms**	**Percentage**	**Correctly analyzed**	**Precision**
5599	4472	80%	1127	20%	4472	100%

Table 3: results

The above results indicate that, at its current stage of development, the Runyakitara noun system analyzer has been successful in analyzing 80% of running text. The 1127 forms which were not analyzed were not yet included in the system. All of the 4472 analyzed strings were correctly analyzed. So recall is 80%, but precision is 100%. This is a positive

outcome, reflecting well on the ability of fsm2 to analyze the noun morphology of Runyakitara. It means that our morphology is incomplete, but correct in everything it does. During the debugging process (which is our next step), the 1127 unanalyzed forms will be processed and included into the noun grammar.

3.10. Applications of the Runyakitara noun system analyzer

The output which the noun system analyzer of Runyakitara generates can be used as input for other applications such as:

- a spell checker of Runyankore-Rukiga
- a dictionary since the system can output lemmas
- a syntax analyzer of Runyakitara
- a language learning system for vocabulary and grammar, depending on how it is developed

3.11. Conclusion and future research

This study demonstrates the applicability of the finite state approach to the analysis of Runyakitara noun morphology. Language specific knowledge and insight have been used to classify and describe the morphological structure of the language, while quasi context-free grammar and replacement rules have been written to account for the grammatical nouns of Runyakitara.

The above-described results represent a preliminary effort at building a morphological analyzer for Runyakitara, a group of Bantu languages with limited electronic resources. The analyzer, which is based on a combination of the Item-and-arrangement and Item-and-Process models proposed by Hockett, (1954; 1959), shows how the models may be applied to Runyakitara morphology.

Specifically, this study has provided:

a) The first computational description of the orthography of the Runyakitara nouns
b) A proof that the fsm2-inspired approach (context free grammar plus Replacement rules) is applicable to a morphologically complex Bantu language, Runyakitara
c) A computational framework for the noun classification system of Runyakitara, which did not previously exist in any Runyakitara text book but was devised during this research for computational purposes.

3.12. Future research

A future goal of this research is to develop a means to account for all the Runyakitara word categories within the *fsm2* implementation in order to eventually produce a comprehensive morphological analyzer for Runyakitara. The morphological analyzer will

provide input for many other planned applications, such as learning systems and machine translation tools.

---

* The authors would like to acknowledge the continued support and expert advice of Prof. Arvi Hurskainen, of Helsinki, Finland and Prof. John Nerbonne of the University of Groningen, The Netherlands.

This chapter and a previous version published in the International Journal of Computing and ICT research – Special Issue – are the results of a six month visiting research fellowship in Germany funded by DAAD (The German Academic Exchange Service). We are grateful for this support.

Chapter 4

RUMORPH: A morphological analyzer of Runyakitara - approach, results and issues

(An earlier version of this chapter was presented at the 8th International Confererence of Computing and ICT Research, August 5-7, 2012, Kampala, Uganda, as: Fridah Katushemererwe & Rehema Baguma, RUMORPH: A morphological analyzer of Runyakitara - approach, results and issues. It is available at www.cit.mak.ac.ug/iccir/?p=iccir_12.)

Abstract

This paper reports on the performance of a comprehensive morphological analyzer of Runyakitara - RUMORPH. Runyakitara is a name given to four closely-related Bantu languages (Runyankore-Rukiga and Runyoro-Rutooro). As a group of languages, Runyakitara is spoken by about six million people in Uganda. The research was motivated by lack of an automatic analyzer and generator for the word forms of Runyakitara. In addition, few researchers have reported on the performance of an analyzer for the entire morphological system of the Bantu languages. The model, RUMORPH, is based on freely available open-source finite-state methods and, in particular, the *fsm2* interpreter. It is able to account for morphotactic structures using quasi context-free grammars supported by fsm2 and morpho-phonological alternations by means of a finite composition of commonly used context-dependent string rewriting rules. Their combination results in a finite state transducer that can be exported and used in a number of software-developing platforms. The transducer is an essential component of syntactic parsers, spell-checkers, text-to-speech synthesizers, language learning systems and machine translation tools. The RUMORPH system was developed primarily on the basis of Runyankore-Rukiga texts, with the aim of understanding the extent to which it can analyze other languages in Runyakitara. Currently, it recognizes 62% of real-world newspaper corpus written in Runyankore-Rukiga, 75% of literary texts extracted from story books written in the same language set and 51% of literary texts in Runyoro-Rutooro. However, precision is above 90% in all these cases, which reflects positively on its development. There are a number of issues related to the nature of Runyakitara itself, its corpora and the scarcity of literature in the four languages. However, it may be possible to expand the scope of the system. In conclusion, the RUMORPH system is the first of its kind, and we hope that it will be useful to other researchers, based on the fact that its approach can be adapted to related languages.

Categories and Subject Descriptors: J.5 (**Computer Applications**): Arts and humanities -- linguistics; I.2.7 **(Computing Methodologies)** Natural language processing -- Language generation, Language models, Language parsing and understanding, text analysis.
Additional key words: Morphological Analyzer, Runyakitara, Bantu languages, Finite-State Methods, Context-free grammar.

4.1. Introduction

Words are important building blocks in natural language applications because, together, they form phrases, clauses, sentences, etc. Computational morphological analysis deals

with automatic word analysis that results in a word form with its linguistic information. The product of automatic morphological analysis is a morphological analyzer. A morphological analyzer has already been recognized as an important artefact in natural language processing.

Despite the recognized importance of automatic morphological analysis, most Bantu languages, including the Runyakitara, do not have morphological analyzers. There was neither a rule-based nor a data-driven morphological analyzer of Runyakitara. In addition, few reports in literature deal with the performance of morphological analyzers for the Bantu languages (honourable exceptions being Hurskainen 1992 and Pretorious and Bosch 2003). Reporting on the performance of the entire morphological analyzer will provide a comprehensive view as well as insights into the progress of automatic morphological analysis as a whole.

Previously, we reported on RUNYAGRAM, an automatic analyzer and generator of Runyakitara verb forms. RUNYAGRAM is an important building-block for a comprehensive morphological analyzer of Runyakitara (chapter 2 above and its previous version by Katushemererwe and Hanneforth 2010). In another chapter, we reported on the analysis of Runyakitara nouns. This chapter reports on the results of RUMORPH, the comprehensive morphological analyzer of Runyakitara dealing with all its word categories, and reviews the issues/challenges related to its performance. Since it involves the work of the previous two chapters, there is some inevitable overlap with them, as well, which we try to keep to a tolerable level. The following section will provide an overview of the features of Runyakitara relevant to this discussion. The rest of the presentation will include a review of related work as well as discussions of the analyzer's formalization, design and implementation, along with the results of the study, conclusion and suggestions for future work.

4.2 Runyakitara: a four-in-one cluster of languages

There is a heated debate among Ugandan linguists and native speakers of languages in Runyakitara on whether Runyakitara should be considered one language or four. Currently, Runyakitara is used as a name given to the two major language clusters spoken in Western Uganda, namely, Runyankore-Rukiga and Runyoro-Rutooro. Researchers, such as Bernsten (1998), refer to these languages as four major dialects: Runyankore, Runkiga, Runyoro, and Rutooro. Ethnologue describes Runyakitara as a standardized version of four western Uganda languages to aid in teaching purposes, especially at the university level but the language spoken is not Runyakitara (Lewis 2009).

But, what are the linguistic facts about Runyakitara and what do they imply for morphological analysis? In terms of mutual intelligibility, the four languages are similar to an extent that exceeds 70%, with the following table detailing their lexical overlap:

Languages	**Lexical similarity**
Nyankore and Chiga	84% - 94%
Nyoro and Rutooro	78% - 93%

Nyoro and Nyankore	77% - 96%
Nyoro and Chiga	67%

(adapted from: Lewis, M. Paul [ed.] 2009.)

Table 1:Lexical similarity of Runyakitara languages

The lexical differences are actually minor when it comes to the written Runyankore-Rukiga and Runyoro-Rutooro texts. Some cases of larger differences are listed below:

English	**Runyankore-Rukiga**	**Runyoro-Rutooro**
'to see',	kureeba	kurora.
'cultivator'	omuhingi	omulimi
'to sit'	kushutama	kwikarra
'to love'	kukunda	kugonza

Table 2: lexical differences between Runyankore-Rukiga and Runyoro-Rutooro

All linguists and native speakers agree that there are not many cases of this nature. Also, there is mutual understanding in social gatherings and exchanges; when a *Mutooro* speaks of *okurora,* a *Munyankore* understands the meaning, although there is no documented research and evidence on the extent of this mutual understandability.

The major differences are probably due to sound change, which are technically called phonological changes.The following are a few examples:

English	**Runyankore-Rukiga**	**Runyoro-Rutooro**
man	omushaija: /omuʃeiʒa/	omusaija: /omuseidʒa/.
maize cob	ekicoori : /etʃitʃoːri/	ekicooli: /etʃi tʃoːli/
to spread (disease)	okuturira: /okuturira/	okuturra: /okutuːɹa/

Table 3: Phonological differences between Runyoro-Rutooro and Runyankore-Rukiga

It should be noted that the morphological structure of the noun classification system, inflectional nature of verbs and the concord system of all the four languages is similar.

Geographically, the languages are spoken by approximately six and half million (6,500,000) people in nineteen districts of Western Uganda. There are other speakers in some parts of Tanzania (Haya) and Democratic Republic of Congo (Songora). Socially, the languages of Runyakitara are used in the media, taught in schools and used in day-to-day business transactions. In addition, the languages are now used as a medium of instruction in lower levels of primary education in Western Uganda.

Based on the above linguistic facts, we assume that a single morphological system can be developed for the four languages of Runyakitara. The morphological structures of the languages are not very different, except for some morpho-phonological features, which can be described in terms of rules. Geographical and social facts indicate that the Runyakitara group is important and should be given computational attention.

4.3 Previous work on the morphological analysis of the Bantu languages

This section repeats material from Section 3.2 above. It was included in the journal publication which appeared independently of and is therefore included here as well.
A considerable amount of work has been performed on the applicability of finite state methods to Bantu language processing. Using Xerox Finite State tools, Karttunen (2003) realized a system to model Lingala verb morphology. This approach focuses on the use of replacement rules to gradually construct the verb from the root, piece by piece.

The Xerox Finite State technology has also been utilized in the development of a prototype analyzer for Zulu (Pretorius and Bosch 2003), reports of which have been published. This analysis uses *lexc, xfst* and replacement rules to account for the morphotactics, morpho-phonology and orthographical issues in the Zulu language. To provide for the long distance dependencies found in Zulu morphology, Pretorius and Bosch use flag diacritics, as described in Beesley and Karttunen (2003).

Work has also been carried out on Swahili using a language-specific morphological parser (Hurskainen 1992) known as SWATWOL. This parser is a two-level analyzer that similarly accounts for the morpho-syntax and morpho-phonology of Swahili.

Related to the above is the Swahili language manager, SALAMA (Hurskainen 2004). This language manager is a common preprocessor in the development of multiple computational applications for Swahili. It is a computational environment for managing the written Swahili language and for developing various kinds of language applications. It comprises the standard Swahili lexicon, a full morphological and morpho-phonological description of Swahili, a rule-based system for solving word-level ambiguities, a rule-based system for tagging text syntactically, a rule-based system for handling idiomatic expressions, proverbs and other non-standard clusters of words, and semantic tagging and disambiguation system for defining correct semantic equivalents in English.

Muhirwe (2007) describes a computational analysis of Kinyarwanda morphology, specifically looking at morphological alternations. He applies the Xerox Finite State compiler to model Kinyarwanda phonological alternations concentrating on orthographical rules. It is important to note that rules are language dependent. Therefore, the rules for Kinyarwanda or Kiswahili language are not directly applicable to other Bantu languages, although they belong to the same group – the Bantu language family.

Finite state methods have also been applied to the analysis of Seswana verb morphology (Pretorius 2008); tonal marked Kinyarwanda (Muhirwe 2009; Hurskainen 2009) and solutions for reduplication in Kinyarwanda (Muhirwe and Trosterud 2008).

Based on the literature reviewed above, we may conclude that there has been considerable work conducted on specific languages, mainly applying Xerox Finite State

methods to morphological analysis, and a number of these implementations have been successful. However, the fact that there are over five hundred Bantu languages means that a large number of Bantu languages, well over 95%, have not yet been subject to any such analysis. The lack of literature dealing with the Runyakitara languages suggest that they belong to the latter category. In addition, *fsm2* has not yet been implemented, as a scripting language with regard to any of the Bantu languages. This makes a publication on the application of *fsm2* to the automatic analysis of Runyakitara both unique and relevant.

4.4 Highlighted features of Runyakitara morphology and considerations for computation

Just as in the case of many Bantu languages, the morphology of Runyakitara is an extremely complex system involving many morphological processes such as concatenation, morpho-phonology, inflection, derivation, compounding, reduplication and in a few cases, infixation. Important aspects relevant to the implementation of RUMORPH are discussed below:

4.4.1 Verbs and their affixes

An initial description of the computational analysis of Runyakitara verb morphology can be found in a preceding chapter (Chap. 2, Katushemererwe and Hanneforth (2010). This sub-section contains a summary of the major issues that pose a challenge to computing Runyakitara verb morphology:

A verb in Runyakitara possesses a rich morphology in terms of the word forms per lexeme, the average number of morphemes per word, and the number of morphologically expressed grammatical categories and irregularity. As we noted in Chap. 2 (above), verbs appear in a staggeringly large variety of forms. While Chap. 2 (above) focused on the aggregate size of verbal paradigms, this section will detail how these are constituted. We wish to further support the arguments that Runyakitara is complex enough to warrant a focused development effort in computational morphology and that it is complex enough so that learners are likely to benefit from the presence of extensive learning materials such as exercises.

a) Moods: a Runyakitara verb can occur in one of the following moods:

i) Infinitive form: in its basic form, a Runyakitara verb appears in infinitive form with a prefix marker **ku** indicating non-finiteness. An example here is *ku-gyend-a* ‘to go’ where **ku** marks infinity, **gyend** is a root, while **a,** is final vowel. An infinitive can combine with subject prefix markers and negative markers e.g.

Infinitive + object marker:	*ku-ru-gyend-a* ‘to travel it’
Infinitive + negative:	*ku-ta-gyend-a* ‘not to go’
	ku-ta-ru-gyend-a ‘not to travel it’

ii) Imperative mood: A Runyakitara verb can also occur in an imperative mood. Below, we describe the details of an imperative mood, i.e. the different forms an imperative verb takes:

a) Imperative: ***gyend-a*** 'go'
b) Negated imperative: ***o-ta-gyend-a*** 'don't go/you shouldn't go'
you-not-go

Note that, where there is **o** 'you' on b) above, all other noun class markers (for the 18 classes we considered) are also accounted for, as follows:

		(A)	
1.	1ps	-	***n****-ta-gyend-a* 'I shouldn't go'
2.	2ps	–	***o****-ta-gyend-a* 'you shouldn't go'
3.	3ps	-	***a****-ta-gyend-a* 's/he shouldn't go'
4.	1pl	-	***tu****-ta-gyend-a* 'we shouldn't go'
5.	2pl	–	***mu****-ta-gyend-a* 'you shouldn't go'
6.	3pl	-	***ba****-ta-gyend-a* 'they shouldn't go'
7.	C3s	-	***gu****-ta-gyend-a* 'it shouldn't go'
8.	C4p	-	***e****-ta-gyend-a* 'they shouldn't go'
9.	C5s	-	***ri****-ta-gyend-a* 'it shouldn't go'
10.	C6p	-	***ga****-ta-gyend-a* 'they shouldn't go'
11.	C7s	-	***ki****-ta-gyend-a* 'it shouldn't go'
12.	C8p	-	***bi****-ta-gyend-a* 'they shouldn't go'
13.	C9s	-	***e****-ta-gyend-a* 'it shouldn't go'
14.	C10p	-	***zi****-ta-gyend-a* 'they shouldn't go'
15.	C11s	-	***ru****-ta-gyend-a* 'it shouldn't go'
16.	C12p	-	***ga****-ta-gyend-a* 'they shouldn't go'
17.	C13s	-	***ka****-ta-gyend-a* 'it shouldn't go'
18.	C14p	-	***tu****-ta-gyend-a* 'they shouldn't go'
19.	C15s	-	***ku****-ta-gyend-a* 'it shouldn't go'
20.	C16s	-	***ha****-ta-gyend-a* 'it shouldn't go'
21.	C17s	–	***ku****-ta-gyend-a* 'it shouldn't go'
22.	C18s	-	***mu****-ta-gyend-a* 'it shouldn't go'

Notes: C = class, s = singular, p = plural. There are 22 subject prefixes (in bold) representing 18 noun classes. Note that the subject prefixes are not chosen at will but must agree grammatically with the expressed or unexpressed subject.

iii) Subjunctive mood: a Runyakitara verb can also appear in subjunctive mood by adding **e** as **a** final vowel instead of **a**, for example, *n-gyend-**e*** 'may I go'. Note that where there is **n-** on *n-gyend-e*, other classes can be included as in (A) above.

b) Tense and aspect markers

There are a number of tense and aspect markers in Runyakitara just as in other Bantu languages. The most notable ones include:

i) **Present/habitual** (**Ø**): this tense describes the action that takes place regularly e.g. daily, hourly, etc. It is marked by a zero morpheme, e.g. a-**Ø**-gyend-a ·he goes (everyday)'. This construction can be instantiated using the 22 subject prefixes above and 22 different object affixes, i.e. in 22^2= 484 different forms.

ii) **Present progressive (ni)**: this marks both tense and aspect. It indicates an action that is taking place now and is still ongoing. The marker is **ni** before the subject prefix e.g. ***ni****-n-gyend-a* 'I am going'. The verb in progressive can also follow the trend in i) above.

iii) **Past:** the past is divided into three parts: immediate past, recent past and far past as follows:

a. **Immediate past** (**ire**): this describes an action that has just been completed mainly in past hours, e.g. *n-**aa**-gyenz-**ire*** 'I have already gone'.
b. **Recent past (ire**): this describes an action that took place strictly the previous day e.g. *n-gyenz-**ire*** 'I went – (yesterday)'.
c. **Far past** (**ka**): This describes an action or event that took place some time in the past, but beginning with two days back, e.g. *n-**ka**-gyend-a* 'I went – (last year)'.

Note: you can add subject and object prefixes to a, b, and c above, so that the construction follows the same trend as in i) above.

iv) **Future:** this is divided into the near future and the remote future:

a. **Near future (ni)**: describes the action or event that will take place in the near future beginning with the next hour/minutes from now, e.g. ***ni****-tu-gyend-a* 'we will go (in the next three hours)'. Here, **ni** is not a progressive marker but a near future marker. This can also be combined with subject and object markers as noted above to make 484 verb forms.

b. **Remote future (ria)**: This describes the time in the remote future beginning with next month, e.g. *tu-rya-gyend-a* 'we will go (next year or in the years to come)'. The remote future also takes on subject and object markers to make 484 verb constructions.

c. Subject and Object markers: a verb in Runyakitara takes on subject and object pronouns as prefixes to the root. These represent noun classes. As already demonstrated in (A) above, there are 22 subject affixes and 22 object affixes. There are cases where a double object comes in, so that the object markers alone total 484 in combination, altogether raising the number of subject and object affix combinations to 10,648 in one verb construction.

d. Negation (Negative markers ti & ta): Runyakitara has two types of negative markers **ti** and **ta**. ***ti*** always precedes a subject pronoun, while ***ta*** comes after a subject pronoun. The two never occur together in the same verb construction. Examples:

ti*-n-aa-mu-reeb-a* 'I have not seen him/her'

*tu-****ta****-mu-reeb-a* 'we shouldn't see him/her'

e. Verb extensions: Runyakitara has seven 'valency change' markers which Bantu researchers have preferred to call verb extension markers (Lodhi 2002). These are: causative, applicative, stative, intensive, reciprocal, reversive and passive. The complexity of verb extension morpheme in Runyakitara is that each of them has two or more allomorphs. Let us consider a causative with a subjunctive as follows:

- *n-gyend-****es-e*** **'**may I cause to go**'** – causative marker is **es**, subjunctive marker **e**.
- *n-gamb-****is-e*** 'may I cause to talk/speak' - causative marker is **is**, subjunctive marker **e**.
- *n-gum-****y-e*** 'may I make firm' - causative marker is **y**, subjunctive marker **e**.
- *n-d-****iz-e*** 'may I make to cry' - causative marker is **iz**, subjunctive marker **e**.
- *n-du-****sy-e*** **'**may I cause to get tired**'** causative marker is **sy**, subjunctive marker **e**.
- *m-paa-****zy-e*** 'may I make … satisfied' causative marker is **zy**, subjunctive marker **e**.

Another issue related to verb extensions is that they occur in arbitrary combinations where it is difficult to generalize or to specify an order in which they occur. A case in argument is a verb with two causative markers where the second represents a causative relation with respect to the causative relation represented by the first.

ba- ka- ba- reeb-es-an- is- a

↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

2spl- far past-opl- see- caus- reciprocol- caus- indicative

'They were made to make them see each other'.

Note that each and every construction here is multiplied by 484 (or perhaps even 10,648 to account for all subject and object pronoun affix combinations.

As we noted in Chapter 2, Runyakitara verbs occur in a myriad of different inflected forms. A verb in Runyakitara minimally includes two morphemes, but may include up to 10 morphemes, for example: *ti-ba-ka-mu-mu-kwat-kwat-ir-ho-ga* 'they have never touched (with intensity) him there on his behalf'**.** This possibility enriches the morphology because each component adds its specific component of meaning.

But other factors complicate Runyakitara morphology as well. We consider irregularity and reduplication below. As an example of irregularity we consider past tense formation in Runyakitara, which we discussed above under ' Tense and Aspect'.As noted above, **ire** marks the past tense e.g. *ba-reeb-ire* 'they saw'. Here, **ire** is just added after the root **reeb**. However, irregularity arises when a sound segment in the root is affected by an affix, as in the following examples:

Present (habitual)	**past tense**
a) *a-gyend-a* 'he goes'	*a-gyenz-ire* 'he went' – **d** on the root **gyend** becomes **z**
b) *a-kwat-a* 'he holds'	*a-kwas-ire* 'he held' **t** changes to **s** *a-kwais-e* 'he held' **t** changes to **is**
c) *ba-reeb-an-a* 'they see each other'	*ba-reeb-ain-e* 'they saw each other' **i** is inserted in a verb extension **an,** and final **a** changes to **e**.
d) *bon-a* 'find'	*a-boin-e* 'he got' **i** is infixed in the root **bon**
e) *reeb-a* 'see'	*reeb-a-reeb-a* 'see 'strangely''

Reduplication is so productive in Runyakitara verbs that almost every verb is affected. Only the verb stem reduplicates, so that the prefixes are not reduplicated. For example, in *ba-ka-reeb-a-reeb-an-a* 'they saw* each other [strangely]', only **reeb** is reduplicated. Reduplication has already been identified troublesome in natural language processing specifically using finite state technology (Hurskainen, 1992).

This concludes our discussion of the complexities of Runyakitara verb morphology. We established in this section that the enormous combinatorics we noted in Chap. 2 correspond to a variety of meanings that arises often in everyday languages, and therefore may not be avoided either by morphological components in computational systems or by learners who wish to function in Runyakitara. Given the points above, one may conclude that Runyakitara verb morphology poses a computational challenge and a language learning challenge.

4.4.2 Nouns

The noun classification system of Runyakitara categorizes nouns into 20 noun classes, but only 18 are given consideration (Chap. 3, Katushemererwe and Hanneforth 2010). This classification was mainly motivated by the needs of computation. A typical Runyakitara noun consists of a pre-prefix, a prefix and a root. In the example of *o-mu-ti* 'tree', ***o*** is a pre-prefix, *mu* is a prefix and *ti*, a root. For purposes of this study, we combine a pre-prefix and a prefix together into one prefix. A prefix indicates class and number (e.g. *omu:* Class 1 singular), while a root indicates the actual meaning of a noun. There are, however, two other categories of nouns: derived nouns and compound nouns in Runyakitara. Derived nouns are nouns based on other word classes, such as verbs and adjectives. For example, *omushomi* 'a reader' and *omushomesa* 'teacher' are derived from the verb *kushoma* 'to read'. These derivational phenomena are not part of this system. Compound nouns are the result of combining two or more words of different meanings to form one word with a single meaning. A noun such as *endiira-kukinduka* 'the person who eats a lot' is a compound in Runyakitara. Compounds are mainly formed of components consisting of a noun and a noun, a verb and a noun, or a noun and an adjective. Such nouns are treated on the basis of the prefix of the first noun segment, but most are in class nine, which is open to new words.

4.4.3 Adjectives

Literature classifies adjectives in Runyakitara according to manner, time, quality, etc (Ndoleriire and Oriikiriza 1995). This classification is less helpful for our computational purposes, than a classification that would break an adjective down into smaller components - morphemes. For this reason, adjectives were classified in our study according to the affixes that they have, a procedure revealing that, like nouns, adjectives also had 20 classes. As might be expected, the adjective classes corresponded to the twenty noun classes that they qualify. Like nouns, 18 classes were given consideration. The two excluded classes are not known to have independent nouns. The following table illustrates the manner in which adjectives are included in the Runyakitara morphological system:

Noun Class	**Prefix in Singular**	**Prefix in Plural**	**Semantics**	***Example***	**Gloss**	**Usage**
1/2	o-mu-	a-ba	Human	*o-mu-rungi* *a-ba-rungi*	A beautiful one Beautiful ones	Takes on both singular and plural
1a/2a	-	baa-	Human	*kaganga* *baa-kaganga*	Extremely fat	Takes on both singular and plural, but no prefix for singular
3/4	o-mu	e-mi-	Plants, fruits,	*o-mu-rungi* *e-mi-rungi*	Good one(s)	Both singular & plural
5/6	e-ri-	a-ma-	Some parts of the body	*e-ri-rungi/a-ma-rungi*	Good one(s)	Both singular & plural
7/8	e-ki-	e-bi-	Objects, misc	*e-ki-rungi/e-bi-rungi*	Good one(s)	Both singular & plural
9/10	en-	en-	Animals and borrowed words	*En-rungi (enungi)*	Good one(s)	Singular and plural
11/10	o-ru-	en-	Insects, plants miscellaneous	*o-ru-rungi* *en-rungi(enungi)*	Good one(s)	Singular & plural
12/14	a-ka-	o-bu-	Small items, miscellaneous	*a-ka-rungi* *o-bu-rungi*	Good one(s)	Singular & plural
13	-	o-tu-	Abstract and diminutives	*o-tu-rungi*	Good one(s)	Both singular & plural
15/6	o-ku-	a-ma-	Some body parts	*o-kurungi*	Good one(s)	Singular & plural
16	aha-	-	Location	*aha-rungi*	Good one(s)	Singular
17	oku-	-	Location	*aha-rungi*	Good one(s)	Singular
18	omu-	-	Location	*aha-rungi*	Good one(s)	Singular
20/21	o-gu-	a-ga-	derogatory	*o-gu-rungi* *a-ga-rungi*	Good one(s) Good one(s)	Singular & plural

Note that the adjective rungi (good) may take the prefixes of every noun class, as do other adjectives of quality.

4.4.4 Pronouns

Pronouns in Runyakitara are categorized into free and bound pronouns (Taylor 1985). We call free pronouns independent while bound pronouns are dependent. By independent pronouns, we mean pronouns that stand alone without affixes. Dependent pronouns are bound morphemes that are affixed to the roots of verbs. Like nouns and adjectives, pronouns, whether dependent or independent, assume the noun prefixes of the nouns that they represent. They are, therefore, classified according to the noun classes. The following table gives an example of how demonstrative and possessive pronouns behave with different noun classes:

Noun class	**Noun Prefix**	**Demonstrative pronoun**	***Possessive pronoun***	**Free forms**
1	**o-mu-**	*ogu*	*o-w-*	*nyowe* 'me' *iwe* 'you' *we* 'him/her'
2	**a-ba-**	*aba*	*a-ba-*	*imwe* 'you' *itwe* 'we' *bo* 'they'
3	**o-mu-**	*ogu*	*o-gw-*	*gwo* 'they'
4	**e-mi-**	*egi*	*e-y-*	*yo*'they'
5	**e-ri-**	*eri*	*e-ri-*	*ryo* 'they'
6	**a-ma-**	*aga*	*a-ga-*	*go* 'they'
7	**e-ki-**	*eki*	*e-ki-*	*kyo* 'they'
8	**e-bi-**	*ebi*	*e-bi-*	*byo* 'they'
9	**en-**	*egi*	*e-ya-*	*yo* 'they'
10	**en-**	*ezi*	*e-za-*	*zo* 'they'
12	**a-ka-**	*aka*	*a-ka-*	*ko* 'they'
14	**o-bu-**	*obu*	*o-bu-*	*bwo* 'they'
13	**o-tu-**	*otu*	*o-tu-*	*two* 'they'
15/6	**o-ku-**	*oku*	*o-ku-*	*kwo* 'they'
16	**aha-**	*aha*	*a-ha-*	*ho* 'there'
17	**oku-**	*oku*	*a-ha-*	*ho* 'there'
18	**omu-**	*omu*	*a-ha-*	*mwo* 'there'
20/21	**o-gu-**	*ogu*	*o-gu-*	*gwo* 'they'

Note that the free forms and demonstrative pronouns are independent. In other words, they can occur as independent words, while possessive pronouns cannot. We note further in passing that the table above also convincing suggests that the free forms and affixes, even though very close semantically, are different enough to require significant additional learning.

4.4.5 Other word categories

Word categories such as conjunctions, prepositions, interjections and selected names do not inflect; and, therefore, do not pose any computational challenge. Although some adverbs inflect, all adverbs were treated as adverbs without sub-categorization. These

word categories were given parts of speech tags and considered as free morphemes in the lexicon. For example, the conjunction *na* 'and' was considered as *na*[CONJ], *ni*[PROP], *munonga*[ADVERB]. Names were included after the first test of RUMORPH on a newspaper corpus. After the test, it was clear that names of places and administrative units such as sub-counties needed to be included to prevent a high rate of non-recognition in real life text.

4.5 Coverage/Scope

Most of the words used in the development of RUMORPH system were drawn from a Runyankore-Rukiga dictionary (Oriikiriza 2007) that claims in its introduction to include all the lexemes from other dictionaries of Runyankore-Rukiga. The scope covered in this system encompasses word categories along with their morphological structure and lexical coverage. In the presentation of this section, we adopt the format of Pretorius and Bosch (2003) to illustrate the scope of the Runyakitara morphological analyzer:

Word category	**Sub-category**	**Morphemes**	**Entries**
Nouns	Proper nouns	-	
	Other nouns	Prefixes	18
		Roots	4274
Pronouns	Possessive	Prefixes	17
		Roots	20
	Demonstrative	Roots	74
Adjectives		Prefixes	18
		Roots	608
Verbs		Roots	2931
		Negative markers	02
		Subject prefixes	18
		Tense/aspect markers	07
		Object markers	19
		Verb extensions	13
		Verb final markers	02
		Post verb markers	02
Adverbs			213
Prepositions			11
Conjunctions			26
Interjections			69

Table 5: scope of RUMORPH

4.6. Approach used in RUMORPH

Given the nature of Runyakitara morphology, it was important to carefully select an appropriate approach. The concatenative tendency of Runyakitara can be represented using a Phrase Structure Grammar [PSG] along the lines of the one developed by Selkirk (Spencer 1991), who proposes phrase-structure-like rules written as W+A for suffixing and A+W for prefixing. However, it was clear that the rules proposed by Selkirk only account for the concatenative feature of morphology. It was important therefore to also think of a way to handle morpho-phonological and orthographical features. Since recursion is not required, both the concatenative rules and phonological processes could be described within the framework of finite-state acceptors (FSA)/transducers (FST). Our approach relies heavily on the closure properties of these automata in relation to intersection, composition, and substitution (see Hopcroft and Ullman 1979, Kaplan and Kay 1994).

The implementation was carried out using *fsm2* (Hanneforth 2009), a scripting language within the framework of finite state technology. Finite-state technology is considered the preferred model for representing the phonology and morphology of natural languages (Wintner 2008). The model has been used to computationally analyze natural languages such as English, German, French, Finnish, Swahili, to mention just a few (Beesley and Karttunen 2003). Its main advantage is that it is bidirectional – it works for both analysis and generation. This bidirectionality was the principal reason that the technology was selected to be applied on the morphological grammatical analysis of Runyakitara. *fsm2* was specifically identified as a resource tool for a morphological grammar of Runyakitara for the reasons noted in chapters 2 and 3, to which we refer the reader.

4.7 The Architectural structure of RUMORPH

The analyzer has a modular structure comprising a special symbol module/file, and a combination of modules for the word classes of the entire language. The general architecture of the entire system is illustrated below:

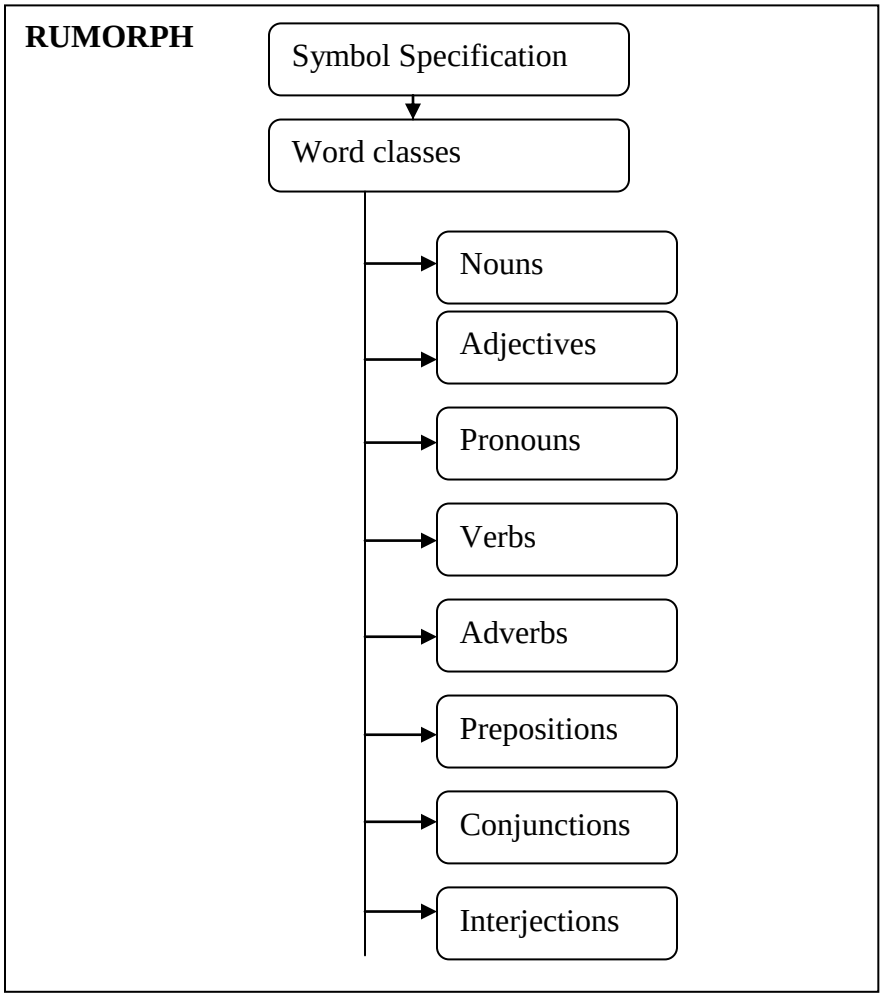


Fig. 1 General RUMORPH architecture

Each word class/category consists of the following modules: a symbol specification module, a grammar module and a replacement rules module. The grammar and rules modules are combined to form a single finite state transducer for each word category. The following diagram demonstrates the architecture of each word category:

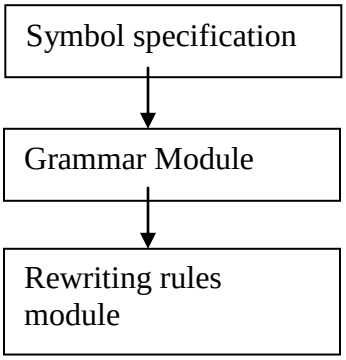


Fig. 2 Sketch of the architecture of an individual word category.

4.7.1 Symbol Specification/Signature Module

fsm2, like AT&T Lextools (see Roark and Sproat 2007), uses a *symbol signature* to define the basic entities of the grammatical description. In the case of nouns, figure three below shows some sample entries:

Letter a b c d e f g h i j k l m n o p q r s t u v w x y z
Letter A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Category: NPREF_1S 1s
Category: NROOT_PS Ps

Fig 3. Sample entries of RUMORPH symbol signature.

The entries are of two types:
1. **Supertype** – **subtype** definitions
2. **Category** definitions, a category consisting of a category name and (perhaps) an empty list of features.

The first two lines in Fig. 3 define Letter as the Supertype of the subtypes a, b, c, A, B, C, etc. The following lines define two categories NPREF_1S 1s and NROOT_PS Ps, with features 1s (1 Singular) and Ps (People singular) defined elsewhere in the Symbol Signature file. Each symbol in the signature – whether type or category name – is mapped by *fsm2* as a unique integer used internally in the compiled automata.

4.7.2 Grammar Module

To specify the morpheme order, we do not use the "classical" *continuation class mechanism* of Koskenniemi (1984) but a *context-free word grammar*. In our view, a grammar is a much more natural way of determining orders and groupings of elements than the continuation-class method, which basically amounts to hand-coding a finite state automaton within the lexicon. Since the generative capacity of context-free grammars is beyond that of finite-state automata, we restrict ourselves to a subset of context-free grammars along the lines of the quasi-context-free grammars formulated by Mohri and

Pereira (1996). This subset may include left- or right recursive rules, but rules out all forms of centre-embedding.

In the *fsm2* framework, grammar rules have the form $A \rightarrow \beta$, where A is a designated non-terminal symbol and β is an arbitrary regular expression (which may even use intersection or negation). The compilation approach is based on the order of the non-terminals of the grammar, creating finite-state automata (FSA) for each grammar symbol and substituting the FSA for the individual grammar symbols into the right side of the rules in the previously computed order. On the right sides of the grammar rules, morphemes of Runyakitara alternate with grammatical categories bearing grammatical information for the morphemes preceding them.

The grammar module consists of a set of quasi context-free rules accounting for the concatenative nature of Runyakitara morphology. The grammar of each word category contains a large number of rules, but we present just a sample (and only from verb grammar), exemplifying the principles underlying the overall organization of the grammar. We devised our own work method of elaborating a verb from its minimum number to its maximum number of morphemes. This was done to account for every form of the verb form. The following provides some sample rules for the verb sub-grammar:

Verb structure rules
Minimum number of morphemes that a verb takes – the result is an imperative verb e.g. *shom-a* 'read'
[VERB] → [VROOT] [VEND]

Maximum number of morphemes that a verb takes e.g. *ti-n-ka-mu-shom-er-a-ho-ga* 'I have never read for him/her'
[VERB] → [VPREF_NEG] [VPREFSP] [V_PREFTM] [V_PREFOP] [VROOT] [VEXT] [VEND] [POSTV1] [POSTV2]

 # Morpheme insertion rules (morphs are in bold-face)
[VROOT] → [**gyend**|**zin**|**gamb**|**shom]** VROOT_SIMPLE Simple=simpleverb]
[VEND] → **a** [V_END_IND Ind=mood]
[VPREFNEG1] → **ti** [VPREF_NEG1 Neg=polarity1]
[VPREFSP] → **n** [VPREF_SPM1S Spm1s=agrmt1]
[VPREFTM] → **ka** [VPREF_TM5 Tm=Tense5]
[VPREFOP] → **mu** [VPREF_OPM3 Opm3=objectprefix3]
[VEXT] → **er** [VSUFF_APPL1 Appl1=applicative1]
[POSTV1] → **ho** [VSUFF_POST1 Post=postverbal1]
[POSTV2] → **ga** [VSUFF_POST2 Post2=postverbal2]

Fig 4. Sample rules of the verb grammar (Non-terminals are enclosed in square brackets: [VPREFNEG1] = verb prefix negative1; [VPREFSP] = verb subject prefix; [VPREFTM] = verb prefix tense marker; [VPREFOP] = verb prefix object marker; [VROOT] = verb root; [VEXT] = verb extension; [VEND] = verb end; [POSTV1] = Verb suffix post

verbal1; and [POSTV2] = Verb suffix post verbal2. Symbols after morphs in bold-face indicate categorical information. | means disjunction.)

The grammar fragment in Fig. 4 applies to verb forms such as *gyenda* 'go', *zina* 'dance', *gamba* 'talk' and *shoma* 'read'. However, we need also to account for *tambura* 'walk', *rya* 'eat' etc, which are not provided for in the fragment. The grammar fragment is simplified, since it would be computationally too expensive to include the complete set of Runyakitara verb roots, which would result in grammars with tens of thousands of rules. We therefore subdivided the set of verb roots into ten equivalence classes, each class containing all the verb roots that participate in the same word-grammatical constructions and represented by a unique symbol in the grammar. After compiling the word grammar into a finite-state acceptor ***AG***, a final processing step then substitutes each symbol denoting an equivalence class by the set of its corresponding verb roots. This also simplifies the addition of new verb roots, since the grammar automaton remains unchanged and only the final substitution has to be recomputed. Nevertheless, the compilation of a grammar with approx. 330 rules and subsequent substitution takes less than a quarter of a second on a modern CPU, resulting in a finite-state acceptor with ≈ 800 states and ≈ 1,200 transitions.

It should be noted that the subdivision of verb roots and subsequent substitution was only carried out for verbs and nouns because they involve a complex morphology. For other word categories, roots were included using the '#Include' statement, also provided by *fsm2*.

The output generated by the grammar is still a set of morpheme concatenations that form strings although some are merely abstract concatenations (morphotactics) without proper phonological and orthographical representation. Fig. 5 presents some of the strings described by the grammar.

ziaakugambira: **zi**[VP_SPM10 Spm10=agrmt10]
aa[PERF Perf=perfective]
ku[VP_OPM2S Opm2s=agrt2s]
gamb[VROOT_SIMPLE]
ir[
a[VEND End=indicative]

kuaagyenda: **ku**[VP_SPM15 Spm15=agrmt15]
aa[PERFPerf=perfective]
gyend[VROOT_SIMPLE]
a[VERB_END End=indicative]

yae : **ya**[POS_PRON_PREF_9S9s=ppref9s]
e[PRON_ROOT9 Root9=class9]

omuana : omu[NPREF_1S 1s=npref1s]
ana[NROOT_PS Ps=class1]

Fig. 5 Grammar output

zi-aa-ku-gamb-ir-a, ku-aa-gyend-a, ya-e and **omu-ana** are valid underlying forms in Runyakitara, that represent correct grammatical information but are not correctly spelt and pronounced words. The grammatical forms are *zaakugambira, kwagyenda, ye*, and *omwana*. This calls for a change in some characters and deletion of others. To deal with this kind of allomorphic variation, we switch from the Item-and-Arrangement model, inherent in the grammar approach, to an Item-and-Process model (see Hockett 1954 for a more details).

4.7.3 Context-Dependent Rewriting Rules:

Rewriting rules relate to morpho-phonological and orthographical variation and are of the abstract form:

$$\alpha \rightarrow \beta \,/\, \gamma \_ \,\delta$$

This means that an instance denoted by α is replaced by an instance β, if α is preceded by a γ and followed by a δ. It is well-known (Johnson 1972, Kaplan and Kay 1994) that rules of this kind stay within the realm of regular devices if certain conditions apply:

[i] α, β, γ and δ must denote regular languages and [ii] rules are not allowed to apply to their own output. For example, the replacement rule:

i → [] / _ [VP_SPM10 Spm10=agrmt10] **aa** [PERF Perf=perfective] states that **i** is deleted (i.e., replaced by **nothing)**, whenever **i** [a verb prefix marker for class 10] occurs before **aa** [verb prefix marker for perfective]. This kind of rule will change **zi-aa-ku-gamb-ir-a** to *z-aa-ku-gamb-ir-a*, a well formed Runyakitara word.

The rules in this category are able to delete and substitute symbols in the string as long as the context is clearly defined. Each replacement rule ***Ri*** – which corresponds to an infinite *regular relation* (Kaplan and Kay 1994) – is included in a finite-state transducer, and all resulting rule transducers are in turn combined to produce one big transducer representing all the rules simultaneously as illustrated below:

$$\boldsymbol{RR} = def\ \boldsymbol{R1} \circ \boldsymbol{R2} \ldots \circ \ldots \boldsymbol{Rk} \quad (\circ \text{ denotes composition})$$

To apply the replacement rules to the strings generated by the grammar, the two finite-state machines are composed:

$$\boldsymbol{AG} \circ \boldsymbol{RR}$$

All the allomorphic changes performed by the combined rule transducer ***RR*** manifest themselves in the output of ***AG*** o ***RR***. But these changes have to occur at the surface, input level. If we wish to analyse surface forms, we achieve the desired effect by *inverting* the transducer, which is accomplished by switching input and output tape. But before doing so, we have to get rid of the categorical information (introduced in the root

and affix lexicons) still present on both tapes of the transducer. For that purpose, we define a simple unconditional rewriting rule which replaces each category by ε, the empty string, effectively deleting all categories:

[<Category>] → ε

Here <Category> is a special meta-symbol denoting all the grammatical categories defined in the symbol signature. The transducer for the morphology of each Runyakitara word category is then defined as follows:

[***AG*** o ***RR*** o [[<Category>] → ε]] -1 [-1 denotes inversion]

This transducer maps Runyakitara word forms (incorporating all the allomorphic changes) to sequences of underlying forms alternating with categorical information about these morphemes. The alternation is due to the various complementation operations intended to restrict the replacements to the correct contexts (*P-iff-S*-operator, see Kaplan and Kay 1994).

When the transducer of each word category is complete, all transducers are combined into one comprehensive morphological system using a union operator.

4.8 Results and discussion

The first notable result of RUMORPH is that a regular verb can output all the millions of forms noted in Chap. 2 when all phenomena are taken into account. This result was demonstrated by testing the verb *kureeba* 'to see' in the grammar sub-module. Such a finding is not documented elsewhere, and we deem it to be worth noting. It may be illustrated by the following ten-line output sample:

```
ariteere:[VERB_PREF_SPM3SSpm3s=agrmt3s]ri[VERB_PREF_OPM5Opm5=agrt5]teer[VERB_ROOT_SIMPLE
Simple=simpleverb]e[VERB_END_SUBJ Subj=mood2]
ariteerese:    a[VERB_PREF_SPM3S Spm3s=agrmt3s]ri[VERB_PREF_OPM5
Opm5=agrt5]teer[VERB_ROOT_SIMPLE Simple=simpleverb] es[VERB_EXT_CAUS
Caus=true]e[VERB_END_SUBJ Subj=mood2]
ariteerere : a[VERB_PREF_SPM3S Spm3s=agrmt3s]ri[VERB_PREF_OPM5
Opm5=agrt5]teer[VERB_ROOT_SIMPLE Simple=simpleverb] er[VERB_EXT_LOC
Loc=prep]e[VERB_END_SUBJ Subj=mood2]
ariteerere : a[VERB_PREF_SPM3S Spm3s=agrmt3s]ri[VERB_PREF_OPM5
Opm5=agrt5]teer[VERB_ROOT_SIMPLE Simple=simpleverb] er[VERB_EXT_APPL
Appl=prep]e[VERB_END_SUBJ Subj=mood2]
ariteererere : a[VERB_PREF_SPM3S Spm3s=agrmt3s]ri[VERB_PREF_OPM5
Opm5=agrt5]teer[VERB_ROOT_SIMPLE Simple=simpleverb]erer[VERB_EXT_INT
Int=degree]e[VERB_END_SUBJ Subj=mood2]
ariteerane : a[VERB_PREF_SPM3S Spm3s=agrmt3s]ri[VERB_PREF_OPM5
Opm5=agrt5]teer[VERB_ROOT_SIMPLE Simple=simpleverb]an[VERB_EXT_REC
Rec=assoc]e[VERB_END_SUBJ Subj=mood2]
anteere : a[VERB_PREF_SPM3S Spm3s=agrmt3s]n[VERB_PREF_OPM1S
Opm1s=agrts]teer[VERB_ROOT_SIMPLE Simple=simpleverb]e[VERB_END_SUBJ Subj=mood2]
anteerese : a[VERB_PREF_SPM3S Spm3s=agrmt3s]n[VERB_PREF_OPM1S
Opm1s=agrts]teer[VERB_ROOT_SIMPLE Simple=simpleverb]es[VERB_EXT_CAUS
Caus=true]e[VERB_END_SUBJ Subj=mood2]
anteerere : a[VERB_PREF_SPM3S Spm3s=agrmt3s]n[VERB_PREF_OPM1S
Opm1s=agrts]teer[VERB_ROOT_SIMPLE Simple=simpleverb]er[VERB_EXT_LOC
Loc=prep]e[VERB_END_SUBJ Subj=mood2]
anteerere :    a[VERB_PREF_SPM3S Spm3s=agrmt3s]n[VERB_PREF_OPM1S
Opm1s=agrts]teer[VERB_ROOT_SIMPLE Simple=simpleverb]er[VERB_EXT_APPL
Appl=prep]e[VERB_END_SUBJ Subj=mood2]
```

Fig. 6 - Sample verb output

The overall output of the system includes morphemes, their categories and features. The linguistic tags seem to be longer than normal, but the excessive length is maintained until the time that it becomes necessary to shorten them. A sample of RUMORPH output is provided below:

```
atwekire :  a[VERB_PREF_SPM3SSpm3s=agrmt3s]twek[VERB_ROOT_SIMPLE
Simple=simpleverb]ire[VERB_END_PAST Pend=nearpast]
ou :        ou[DEM_PR_CLASS12/14]
Isingiro :  Isingiro[PNAME]
mbwenu :    mbwenu[ADVERB]
ogamba:     o[VERB_PREF_SPM2SSpm2s=agrmt2s][VERB_PREF_PRESENT
Present=habitual]gamb[VERB_ROOT_SIMPLESimple=simpleverb]a[VERB_END_IND
Ind=mood]
aba :       aba[DEM_PR_CLASS12/14]
kwonka :    kwonka[ADJECTIVE_ROOT]
kwonka :    kwonka[ADVERB]
akabi :     aka[ADJECTIVE_PREF_12S12s=apref12s]bi[ADJECTIVE_ROOT12P
12p=class12]
iwe :       iwe[DEM_PR_CLASS12/14]
```

```
eki :       eki[DEM_PR_CLASS16]
orikuza :   o[VERB_PREF_SPM2SSpm2s=agrmt2s][VERB_PREF_PRESENT
Present=habitual]ri[VERB_PREF_OPM5       Opm5=agrt5]kuz[VERB_ROOT_SIMPLE
Simple=simpleverb]a[VERB_END_IND Ind=mood]
orikuza:    o[VERB_PREF_SPM2SSpm2s=agrmt2s]ri[VERB_PREF_OPM5
Opm5=agrt5]kuz[VERB_ROOT_SIMPLE         Simple=simpleverb]a[VERB_END_IND
Ind=mood]
manya :     many[VERB_ROOT_SIMPLESimple=simpleverb]a[VERB_END_IND
Ind=mood]
manya :     manya[ADJECTIVE_PREF_12S
omushaija : omu[NOUN_PREF_1S 1s=npref1s]shaija[NOUN_ROOT_PS Ps=class1]
mukuru :    mu[ADJECTIVE_PREF_1S1s=apref1s]kuru[ADJECTIVE_ROOT1S
1s=class1]
```

Fig. 7 -Sample output from RUMORPH

4.8.1 Testing and evaluation

During development, the Runyakitara system was tested on word lists for individual word classes. Such testing is insufficient, however. In this paper, we present results from corpora that were not used during the development process. Note that Runyakitara is a poorly documented group of languages and has no compiled corpora from which one might draw material for testing. Therefore, developers had to compile material for testing from different sources. Corpora for two languages were compiled: two for Runyankore-Rukiga (where the data for developing the system was drawn) and another for Runyoro-Rutooro. In the former case, we compiled a corpus from the online newspaper *Orumuri* and another from a story book entitled *Ishe Katabaazi*. Runyoro-Rutooro had no newspaper at the time, so we only compiled corpus from a story book. This third corpus was created for the following reasons:

a) To establish the extent to which a morphological system developed with Runyankore-Rukiga material can analyze Runyoro-Rutooro corpus.
b) To have a factual base on which to draw in order to determine if the four languages can be regarded and used as one language.

All corpora were pre-processed, which means that punctuations marks were removed and the corpora were tokenized and otherwise made ready for testing. We employ recall and precision as measures of our output (Beesley & Karttunen 2003). We considered both tokens and types.[6] Table 6 presents the results in relation to the number of word types and tokens in each corpus, along with their recall and precision.

[6] Tokens and types in computational morphology have technical usage. Tokens are words as are used in real world, that is, with duplicates. Types are words minus duplicates.

Corpus	No of words		Recall(%)		Precision (%)	
Runyankore-Rukiga	**Tokens**	**Types**	**Tokens**	**Types**	**Tokens**	**Types**
Newspaper	6933	3223	62	58	94	92
Stories	4314	1740	75	60	98	96
Runyoro-Rutooro						
Stories	2187	1326	51	48	94	96

Table 6: Results

The above results indicate that the precision was higher than recall across the board. A considerable number of strings were not analyzed; thus, there was low coverage/recall. However, what was analyzed was correctly analyzed (precision was high). However, there were many points with respect to which coverage is incomplete which resulted in the system having low recall. Sub-sections 4.2 and 4.3 discuss these issues in detail.

4.8.2 Error analysis

A set of 100 non-analyzed words was picked at random from the story-book corpora of the two language clusters. We only considered 'types' in order to determine the type of errors and thus to account for the non-analysis of the strings. The analysis of the sampled set of items that were not analyzed is presented in the following table:

Error type	% of Runyankore-Rukiga types	% of Runyoro-Rutooro types
Familiar strings but not yet included in the system	76	24
Out of vocabulary	06	03
Proper names	05	11
Contractions	04	-
Foreign words	04	03
Spelling errors	05	13
Phonological and Orthographical differences	-	46
Total	**100**	**100**

Table 7: error analysis

The table above clearly indicates that most of the unanalyzed strings (76%) in the Runyankore-Rukiga corpus were common strings but were rejected by the analyzer because they were not yet included in the morphological system. A close look at these strings reveals that, although the unanalysed items came from all the word categories, most of them were inflected and extended verb forms which require more replacement rules. Therefore, more rules need to be written to account for this error category.

The second considerable number of unanalyzed strings came from the Runyoro-Rutooro corpus (46%) and resulted from morpho-phonemic differences from Runyankore-Rukiga. Here, the word may have a similar morphological and semantic form, but a different morpho-phonemic form. For example, the English phrase 'to go' is represented as *kugyenda* in Runyankore-Rukiga and *kugenda* in Runyoro-Rutooro with only *gye* and *ge* differing in both languages. This means that the system built with Runyankore-Rukiga lexicon failed to analyze some Runyoro-Rutooro strings of a similar nature. This issue can be solved by making allowance for Runyoro-Rutooro variations, but we postpone this until later work.

4.8.3 General issues

There are a number of issues related to the successful development and performance of a morphological analyzer of Runyakitara and these are elaborated below.

a) Corpora
There is no organized corpus in any of the Runyakitara languages. The newspaper archive which acted as a fallback for this study has a number of limitations:

- **language usage:** journalists have their own style of writing where many of the words are jargon and not standard Runyakitara words.
- **typographical errors**: most journalists have no formal instruction in Runyakitara and therefore do not fully know the orthographical rules governing the language.
- **writing style** (e.g. *aha kihandiiko – ahakihandiiko*): Separate words are written as compounds or vice versa.

The lack of organized corpora in Runyakitara represents a significant obstacle and has impacted heavily on the development and testing of the morphological system.

b) Writing system
The orthographies that exist are from the 1950s and have never been revised. There are a number of issues that have arisen with regard to orthography and that have not yet been resolved. Therefore, most writers spell words as they see fit without referring to the standard orthographies of the 1950s. Related to this issue, the languages in Runyakitara were for a long time not taught in schools and were subsequently introduced at university level. As a result, people writing in this language group do not necessarily master the written language but have developed expertise in other fields. Setting this issue aside, the writing system of Runyakitara has many contractions, (i.e. two words are contracted to form one word by inserting an apostrophe and omitting some characters, (e.g. 'at the old woman's home/house' is written as *ow'omukaikuru*, while its uncontracted version is *owa omukaikuru***,** the **a** of **owa** being replaced by the apostrophe). For instance, all the possessive pronouns in a text may be contracted with nouns and adjectives. Given a system that conducts analyses at word level, such contractions cannot currently be taken into account.

c) Nature of Runyakitara

As mentioned earlier, Runyakitara is a name that applies to two language clusters, while each language cluster has various dialects. Although there are standard orthographies, the languages have many phonological variations. A successful single morphological system for all these variations requires a large set of replacement rules to account for the phonological variations. This may be achievable, as mathematical solutions are already available. However, given the time we have for this study, we cannot consider all the Runyakitara phonological variations at this time. Comprehensiveness will be the product of a long-term task requiring substantial resources and, especially, time.

d) Foreign words

Many foreign words from Luganda, English and other languages have entered into Runyakitara. These imports are mostly manifested in the language used in the media, where, mostly urban based writers assume that a certain English word will be understood simply because they understand English. Out of the 100 words that we sampled from a newspaper corpus, 20 were foreign words, where as the storybook corpus contained only 4 foreign words. This difference indicates that not all the foreign words found in the newspaper corpus may be analysed by the current morphological analyzer as Runyakitara words.

4.9. Conclusion and future work

This study reports on the performance of RUMORPH, the comprehensive morphological analyzer of Runyakitara: a group of four closely-related languages. The automatic morphological system is based on freely available finite state tools and, specifically the *fsm2* interpreter. Language specific knowledge and insight have been applied to classify and describe the morphological structure of the language group, and quasi context-free and rewriting rules have been written to account for the grammatical words of Runyakitara.

The above-described results represent an effort at building a comprehensive morphological analyzer applicable to all four languages of Runyakitara. RUMORPH, which results from the combination of the *Item-and-Arrangement* and *Item-and-Process* models proposed by Hockett, (1954; 1958), shows how these models may be applied to Runyakitara morphology.

To sum up, this study has provided:

i. The first computational description of the entire morphology of Runyakitara.
ii. Proof that the *fsm2*-based approach (context-free grammar + rewriting rules) is applicable to a morphologically complex group of Bantu languages such as Runyakitara. 90%+ precision is a good indicator of the initial success of RUMORPH.
iii. Proof that a single verb in Runyakitara can have myriad inflected forms.

iv. Identification of various linguistic knowledge gaps in Runyakitara have been identified and, where possible, means of bridging these gaps, as in the case of the noun classification system and the verb morphological template of Runyakitara.

v. Evidence that the current use of Runyakitara outside the classroom environment does not sufficiently represent all the Runyakitara languages. This is particularly in reference to language corpora obtained from the local newspaper.

vi. Evidence that computation of a single morphological system covering all four languages of Runyakitara is possible.

vii. Phenomena in Bantu morphology such as reduplication, infixation and verb extension, which are challenging in finite state morphology (Beesley & Karttunnen 2003) are accounted for by context free grammar framework and context sensitive replacement rules in fsm2 driven model.

4.10. Future work

The plan for the current morphological analyzer of Runyakitara is to apply it to the further development of a language learning system for Runyakitara. This is why even the noun transducer by itself would be sufficient for our purposes. The form in which morphological analyzer is currently cast is enough for our purpose. Although we realize that the work on a morphological system specific to Runyakitara may be a life-long project, the immediate need is to increase the coverage and performance of the system by:

- including more morphemes in the lexicon
- finding solutions for contractions
- working on derived nouns not accounted for in the noun lexicon

Other researchers might take up a task of writing phonological rules that apply to Runyoro-Rutooro phonology that differs from the phonology of Runyankore-Rukiga.

Chapter 5

Language Teaching and Learning in Uganda: situation analysis and the need for Computer Assisted Language Learning (CALL)

(*An earlier version of this chapter was presented at the 8th Strathmore ICT Conference 2-3rd September 2011, Nairobi, Kenya as: Fridah Katushemererwe, Rehema Baguma & Irina Zlotnikova: Language Teaching and Learning in Uganda – situation analysis and the need for Computer Assisted Language Learning (CALL)*

Abstract

This paper analyzes the situation of language teaching and learning in a multilingual country – Uganda – and the benefits of employing Computer-Assisted Language Learning (CALL) systems in such a context. Indigenous languages are not widely taught in Uganda, even though the majority of Ugandans do not understand the official language, English, well. To address this issue, the paper reports results from a small scale survey that was carried out to establish the extent to which Computer-Assisted Language Learning systems (CALL) can help in the teaching and learning of local Ugandan languages. The results from the mini-survey indicate that, at the time of the survey, many survey participants had virtually no experience with CALL, but that there was great interest in using CALL if it ever became available. The paper concludes that Uganda should think of integrating CALL in all efforts of language teaching and learning for effectiveness and efficiency in of the teaching and learning process.

5.1. Introduction

In the current chapter we investigate the potential need for an interest in CALL software for supporting the learning of local languages. We therefore examine the overall language situation in Uganda in the first section of the chapter, including which foreign languages are taught in the schools and universities and which local languages. In a second section we briefly examine the preferred methods for teaching and learning languages, noting the lack of computer-supported methods and techniques. In a third and final section we poll a group of thirty people in Western Uganda to understand the potential interest in supporting the learning of local languages, and in particular via CALL.

5.2. Languages, language teaching and learning in Uganda

Uganda is officially known to have 45 languages, 43 of which are living languages and 2 of which are believed to have no known speakers (Lewis 2009). A full list of these languages can be found in Ethnologue (2009). The multilingual environment has arisen partly because of the colonial era in African history, when colonizers carved up the continent into new political entities without regard for political, cultural and linguistic frontiers. Post-colonial Africa has not subsequently been able to redefine political boundaries (Badejo 1989) in ways that might be more ethnologically understandable. The result is the existence of many small ethnic groups – commonly known as tribes – which

speak many different languages. Because of the linguistic complexity of multilingual societies, English has continuously improved its position as the best choice for the official language in countries such as Uganda.

Uganda has two official languages, English and Kiswahili, but English is used in administration, courts of law, education and trade. This enhances the status of English and makes it an important language to be taught and learnt by everyone in Uganda. Kiswahili, which is used in such cases as are felt important to the parties involved, took a long time to take root in Uganda because of its historical association with the soldiers who tortured people in the 1970s during the time of Idi Amin. People's attitudes towards Kiswahili have only recently begun to change for the better with the advent of the East African Community, and there is now hope that, sooner or later, it will be used by many Ugandans.

Although English is the prevalent official language of Uganda, a large number of Ugandans do not understand or speak it (Tembe & Norton 2008). In many cases, English is referred to as a language of the elite and has for a long time failed to serve as a language of mass communication. In courts of law, for example, interpreters may be required to provide translations/interpretation services from English into a local language that is understood by the party/parties involved. This circumstance would appear to indicate official recognition of the fact that many Ugandans do not understand English.

Children who attend school in cities and townships and interact with others from different ethnic groups tend to communicate more in English. In addition, families resulting from intermarriages also often prefer to use English. As a result, members of the young generation growing up in towns are commonly no longer proficient in the (local) native language of their parents and, are in most cases ignorant of its culture and heritage. This situation is not a good one for Uganda, as such children tend to grow up without any defined (cultural) identity. A population of culture-less citizens sharing only a foreign language as a means of communication weakens the spirit of nationalism and national development (Prah 2008). The problem of children not learning their parents' language is well known from the literature on language endangerment and reversing language shift (Fishman 2000, 2001). It is particularly common in situations where families have migrated to areas where another language is spoken, e.g. in emigration situations. But in Uganda there is a large migration from the rural areas to the large cities that likewise results in families living in areas where their language is relatively unknown. We return to this problem in Chap. 7 (below).

Given the multilingual situation in Uganda, the teaching and learning of languages in this African country must meet many challenges. One challenge is the need to ensure access to and success in education at all levels. Acquiring knowledge in foreign languages in Uganda has always favoured a few privileged people who can afford good schools and facilities for learning. The diversity of languages in Uganda may seem a challenge to those who call for the use of local languages in teaching and learning, since after all there are many local languages and it might seem that finding qualified teachers will be difficult. Others might be worried that providing some education in local languages may have an adverse effect on the status of English and Kiswahili, as it might depress abilities in these official languages. But,

in fact, there is consensus that primary education would be best conducted in a local language that children understand (Alidou et al 2006). Children are easily discouraged by any failure to understand a language used for instruction. Providing instruction via local languages is likely therefore to improved overall educational levels.

At present, many Ugandans cannot effectively read and write in their native languages even when they are educated. They can hold simple conversations in these languages but cannot express themselves freely or precisely in their first languages, and they cannot write correct sentences, even though there are situations which call for the use of local languages in Uganda.

This calls for strategies that would motivate and help Ugandans to learn their local languages. Providing and improving (computer-assisted) language learning resources may have a substantial impact on improving the existing literacy situation with respect to some language, including local languages. This has the advantage of facilitating citizens' access to written material crucial to governmental services. Language proficiency is a key instrument for bringing about a common understanding among the citizens of any country, especially those wishing to enjoy its rich cultural heritage (Prah 2008).

Literature on the language situation in Uganda, such as the studies of Mukama (1991), Kwesiga (1994), Tembe & Norton (2008) or Namyalo (2010), does not discuss the situation of language learning and teaching at all educational levels in Uganda or the manner in which computers can assist in the existing language situation. This chapter therefore analyzes the current language teaching and learning situation in Uganda and the extent to which computer-assisted language teaching and learning systems might improve the situation. The central questions addressed in this paper include:

i)What is the current situation of language teaching and learning in Uganda?

a) What languages are taught at all levels of learning in Uganda?

b) What is the best language policy for the teaching and learning of languages in Uganda?

c)What are the methods used at all levels of language teaching and learning?

ii)What role are computers currently playing in language teaching and learning?

a) What is the current state of Computer-Assisted Language Learning (CALL) in Uganda?
b) What are the achievements, challenges and unexploited opportunities of the current state of CALL in Uganda?
c) How can the challenges be addressed using the unexploited opportunities to improve the state of computer-assisted language teaching and learning in Uganda?

5.2.1 Foreign language teaching and learning in Uganda: Policy and practice

We refer to non-local languages, including Kiswahili, as "foreign languages". Strictly speaking, Kiswahili is actually a local language for a very small number of people. But

its unpopularity ensures that most Ugandans view it as a foreign language, much as they view English and other European languages and also Chinese.

5.2.2 Primary, secondary and tertiary levels (excluding university)

English, as an official language of Uganda, is taught at all educational levels: kindergarten, primary, secondary and tertiary. Tertiary institutions here include teacher training colleges, business colleges and technical colleges. English enjoys a special status in the education system of Uganda both as the language of instruction from primary four (fourth year in primary) through university and as a subject taught from kindergarten to university as well. In urban areas, English is, in fact used as the language of instruction from kindergarten on, as well as being one of the subjects taught at all educational levels.

Kiswahili, the second official language of Uganda, is offered as an option by schools at all levels. There is currently no policy that makes Kiswahili compulsory at any level of education, and it is mainly taught and examined at secondary level. There is no primary leaving examination for Kiswahili.

Another foreign language taught at secondary and tertiary level is French. French has become a popular subject of study among Ugandans because of the recent stability in Rwanda and the Democratic Republic of Congo. German is taught and examined at secondary level but only in a few selected schools mainly in urban areas. It is not compulsory at any level.

To sum up, foreign languages in Ugandan primary, secondary and tertiary institutions include English, Swahili, French and German.

5.2.3 University level

Uganda has 5 public and 25 private universities. In view of the programs/courses that most private universities offer, it is clear that languages have relatively low priority in university curricula. This perception was verified by a review of 6 university programmes (three private and three public) in order to acquire a general overview of language course offerings at the university level. It is also important to note that all universities in Uganda use English as the language of instruction.

Makerere University, the oldest of all the universities in Uganda, has two departments responsible for language studies and instruction: the Makerere Institute of Languages (MIL) and the Department of Language Education at the School of Education. The Institute of Languages (MIL) is responsible for instruction and research in all languages, both African and foreign languages, while the department of Language Education is primarily concerned with language teaching methods. Methods in language teaching applicable to both foreign and local languages are taught to students aspiring to become teachers of humanities subjects in secondary schools. To provide an overall picture in tertiary (post-secondary) education, the table below shows the actual state of language

teaching and learning in mainstream university programs and short courses at selected universities in 2010:

Table 1: Foreign Languages offered at University level in Uganda (2010)

University	Language	Comment
Makerere University	French	Degree course for beginner and advanced students[7]
	Arabic	Degree course for beginners and advanced students
	Kiswahili	Degree course for Beginners and advanced students
	English	English Language Studies
	German	Degree course for beginners and advanced students
	Spanish	Elective[8]
	Italian	Elective
	Japanese	Elective
	Russian	Elective
	Chinese	Elective
Kyambogo University	English	English Language Studies
	Kiswahili	Degree course for Beginners and advanced students
	French	Beginners and advanced
Uganda Christian University (UCU)	English	All language courses at UCU are part of language education
	Kiswahili	
	French	
Uganda Martyrs University	English	Communication Skills
Islamic University of Uganda (IUIU)	English	
	Arabic	
	French	
Gulu University	English	

[7] Advanced students come with knowledge of a language while beginner students have no prior knowledge of the language at all.
[8]Not part of the main University Curriculum.

Apart from the regular education system, there are many language learning centres in Uganda that offer different foreign and local languages for people who need them. These are private learning centres which specialize in language for specific purposes, and they teach a language in response to demand.

5.2.4 Local language teaching and learning in Uganda

As already stated in the introduction, Uganda has many local languages (45) but those with officially defined orthographies and language learning materials are quite few in number. In fact, it is reported that, as of 2006, fewer than ten languages had developed/written materials such as dictionaries, grammar books, children's storybooks, the bible, etc (Bukenya 2008).

Prior to the implementation of an educational language policy in Uganda in 2006, the teaching of local languages was permissible but not compulsory at all levels and in many language learning schools. Some universities and schools acknowledged the need for or importance of local languages by teaching and promoting them, but they were few in number.

The current policy, adopted in 2006 and implemented in 2007, requires the language of instruction in the primary school years (1-3) to be a local language and English to be taught as one of the primary school subjects. After the third year at school, English then becomes the language of instruction. The policy has only been effectively implemented in rural areas, however. In urban areas and specifically in Kampala, schools have continued to use English as the language of instruction from the first school year onward. To justify this contravention of policy, reference is usually made to the multi-lingual situation in urban areas.

The policy of teaching local languages in Uganda is not new, but has suffered a number of setbacks over the years. In 1992, the government's policy on English and Kiswahili required them to be taught to all pupils at primary level, while local languages were to be taught as subjects at the same level (Government of Uganda 1992). By 2005, the reality in the classroom had, however, become quite different (Majola 2006). According to Majola, neither Kiswahili nor local languages were being taught in any of the schools that she visited in Kampala in 2004. This observation confirms Lodhi's statement (1993) that language policies in Africa are generally not implemented or enforced, so that decrees or directives from ministries of education requiring the use of a particular language or languages of instruction at different levels of the educational system often have little effect.

Local languages, unlike foreign languages, have not been a focus of attention for the Ugandan government, educators or even the native people/speakers themselves (Majola 2006). In fact when the policy for instruction in local languages was adopted in Uganda, some parents and other stakeholders publicly expressed their concerns about the possible detrimental effects that the learning in the local language may have on their children's success (Tembe & Norton 2008). These groups failed to understand that English often poses an obstacle to childhood learning in Uganda. Majola (2006) urged governments to recognize the fact that, for education to bring about transformation and social

development in Uganda, it must be rooted in the culture and language of their population. Prah (2002) noted that language is a key challenge to African development. The dominance of "metro-languages" deprives the majority of Africans of access to education and prevents them from participating in national politics and decision-making processes.

Despite all the documentation supporting such views, few people in Uganda appreciate the importance of local languages. Bukenya (2008), director of language education at the Ugandan National Curriculum Development Centre (NCDC), specifically emphasizes the importance of local native languages in education, particularly in the Ugandan context.

- Local languages are tools for socialization that help to shape people's relationship with the their environment and neighbouring cultural groups. This enhances participation in the classroom by promoting the tolerance and collaboration that are necessary for effective learning.

- Local languages create confidence in one's own language. This forms the basis for learning other languages. In Uganda, it is known that the people who used local languages in lower levels of school in the 1960s or before, have a better command of the first and second languages because of this background.

- Local languages provide a bridge between the home (native) and the school (proto-metropolitan) environment. This complies with an approach to teaching that emphasizes a trajectory of learning based on the familiar and becoming increasingly more accepting of the unfamiliar. The development of this learning trajectory helps the child to relate the domestic (childhood) environment to the (proto-adult) knowledge and learning acquired at school (and eventually to the realities of life in metropolitan Africa).

Whereas there is critical evidence in both theory and practice that knowledge and skills gained in one's mother tongue can transfer across languages (Klaus 2003; Obondo 2007), some Ugandans are still unaware of the benefits of learning one's native language at a more advanced (educated) level (Asiimwe 2008).

Universities, especially Makerere University, have been trying to cultivate the teaching, learning and development of local languages for years. Makerere University, whose language and education departments have become entirely convinced of the benefits of local-language education from both theoretical and practical points of view, introduced the teaching of local languages in the early 1990s. These programmes have continued up to the present day, despite the fact that government support for these languages has been limited.

To provide some idea about the situation of local-language instruction at university level in Uganda, the following table indicates the languages taught at selected universities in 2010:

Table 2: Local Languages offered at selected Universities in Uganda (2010).

University	Language	Comment
Makerere University	Luganda	Advanced[9]
	Runyakitara	Advanced
	Luo	Advanced
	Lusoga	Elective
	Lumasaaba	Elective
Kyambogo University	Luganda	Language education
Gulu University	Luo	Advanced
Uganda Christian University	Luganda	All languages at UCU are part of language education
	Runyakitara	
Uganda Martyrs University	None	
Islamic University of Uganda (IUIU)	None	

Apart from Makerere University, local language teaching and learning in other universities is minimal. This situation would indicate that local language teaching and learning is not popular in Uganda. Scholars have categorized the reasons for this unpopularity as political, economic and social (Obondo 2007). Although these issues are relevant to the adoption of CALL in Uganda, their complexity and volatility places them beyond the scope of this initial study.

5.3. Methods used in language teaching and learning in Uganda

The methods and principles used at all the universities listed above and all other learning levels involve face-to-face instruction/delivery and textbooks. In such a context, instruction is provided by teachers, who presumably know and understand the language, to learners, who are then allowed ample opportunities to rehearse their learning by engaging in meaningful discourse with other students and teachers using the language being learned.

Radio and television programming is sometimes used to supplement face-to-face instruction and the use of textbooks. However, such electronic media are not commonly used in classroom situations, but only as extramural support for classroom language learning.

[9] Advanced here means that students have knowledge of the language and are not beginners.

It is important to note that to date, Computer Assisted Language Learning (CALL) is not used at any university, education institution or language-learning centre in Uganda. Several reasons account for its absence notably the unavailability and inaccessibility of computer hardware, an insufficient and unreliable electricity/power supply, underdeveloped IT skills, lack of knowledge about existing CALL software, etc.

5.4. Computer Assisted Language Learning (CALL)

Researchers concerned with CALL define it as a group of computer systems/technologies designed to help people teach and learn languages (Nerbonne 2003). CALL is a form of computer-based learning which has two important features: individualized learning and bidirectional (interactive) learning. It is neither a method nor a medium, but rather a collection of methods and media intended to bring about computer-mediated language learning.

Many educators – including Shaalan (2005), Warschauer (1998), Ma and Kelly (2006), and Jager (2009) – indicate that using computer technology in language learning is very important in accomplishing the following objectives:

> **Repeated exposure:** The learner receives repeated exposure to the same material, which is beneficial or even essential for language learning. A computer is ideal for carrying out iteration drills, since a machine does not become bored or tired of presenting the same material and can provide immediate nonjudgmental feedback.
>
> **Individualization:** CALL allows learners to have non-sequential learning habits. This means that they are allowed to determine the skills to develop and the method(s) to be used. In addition, a computer can present such material on an individualized basis, allowing learners to proceed at their own pace and freeing classroom time for other activities. The process of finding the right answer involves a fair amount of student choice, control and interaction. The computer can create a realistic learning environment, since listening can be combined with seeing, just as in the real world.
>
> **Variety and motivation:** Multimedia and hypermedia technologies allow a variety of media (text, graphics, sound, animation and video) to be accessed on a single machine. All the language skills can, consequently, be easily integrated, since the variety of media makes it natural to combine reading, writing, speaking and listening into a single activity.
>
> **Accessibility:** Internet technology facilitates access to language learning resources world-wide and enables remote communications between teachers and the language learners (narrow-casting). It therefore allows an individual teacher or student to share a message with a small group, the entire class, a partner class or an international discussion list of hundreds or even thousands of people. In

addition, there is optimal use of learning time, where the importance of flexible learning is stressed: anywhere, anytime and anything a learner wants.

There is no doubt, therefore, that the computer has proven useful as a medium and technology for language learning and has potential to improve language teaching and learning in multilingual Uganda.

However, scholars have also noted that CALL also has its limitations, especially in developing countries such as Uganda. It benefits the few individuals, groups or organizations that can afford the technology. The necessary capital outlay means that educational costs are initially increased over the short run. There is a lack of trained teachers able to make effective use of CALL resources. Finally, there are structural problems such as inadequate access to the internet or even to a reliable supply of electricity.

Nevertheless, the advantages of CALL are enormous and undoubtedly outweigh the limitations. Prices of computers and internet connectivity would prevent extensive use of CALL in Uganda at the moment. But these prices have fallen regularly over the past three decades, so that it is reasonable to expect them to continue to fall,[10] meaning that CALL will be within the reach of African educators before too long. Educators and other stakeholders in Uganda therefore need to investigate the possibilities of integrating CALL into mainstream language learning as a means of addressing some of the challenges facing the language learning situation in Uganda.

5.5. Applications of CALL

Nerbonne (2003) identifies three structured areas where CALL might be applied, namely in schools, universities and industry, as well as for self-study. He stresses that the fact language education is a task officially assigned to schools and universities. In these institutions, CALL can effectively assist teachers and learners both inside and outside the classroom. Government, industry and other parts of the private sector may organize their own language courses, usually at considerable cost. Including CALL in their arrangements will ultimately reduce some of these expenses since language teaching professionals' time is often scarce and therefore expensive. As for self-study, a number of people study languages without the benefit of the formal instruction provided by their employers or educational institutions. Obviously, CALL can be used to provide structure and guidance in aid of language learners learning on their own.

5.6. CALL in Uganda

Language learning in Uganda, whether second language learning, foreign language learning or local language learning, is mainly conducted in classrooms where learning content is provided by teachers with the assistance of textbooks. There is limited

[10] A 1.500 rupee (€25) tablet has been announced in India, where the government has announced plans to subsidize its use by students. http://www.youtube.com/watch?v=fxbrRm54U2s (validated Feb. 8, 2013).

involvement of electronic language resources, such as television and radio, and there is no evidence of CALL. As a consequence, Uganda is missing out on the intellectual, social and developmental benefits of increased linguistic proficiency with the aid of CALL.

5.7. Untapped opportunities for CALL in Uganda

Reaffirming the point made by Klaus (2003), knowledge acquired using a native language is readily transferable to other languages. There is a great deal of psychological evidence indicating that knowledge and understanding are better and more easily acquired when communicated in a native language. Furthermore a child who masters his or her first language has fewer problems in subsequently acquiring knowledge via other languages (Klaus 2003). Applying CALL to the teaching of indigenous languages in Uganda will enable many children born in the semi-urban areas of Uganda to obtain this beneficial proficiency in their local languages before being introduced to English in schools.

There are also Ugandan children born outside Uganda. Once CALL systems for local languages are developed, such children can benefit from the facilities to learn the language of their Ugandan heritage, and their parents may also share something with their children in this respect.

There are a number of unique tourist centres in western Uganda, namely impenetrable forests, national parks, rift valleys and lakes, to mention just a few. These centres are visited by foreigners generally lacking any knowledge of the local language. Developing CALL systems for local languages in Uganda will help stimulate the interest of (prospective) visitors by providing remote access to the local languages, enabling people to learn them abroad and, subsequently, to communicate with native speakers on their visits to Uganda, at least at a basic level.

Primary school children in villages have difficulties learning English. Implementing CALL systems for local languages in Uganda will indirectly help to improve the English language proficiency of most Ugandans by first improving the proficiency in native languages. CALL systems for multi-lingual learning may ease the transfer of learning from local languages to foreign languages.

The above considerations led us to survey a number of Ugandans in order to gauge their familiarity with CALL and the appreciation of its potential value in Uganda, specifically focusing on the Runyakitara language group. Runyakitara is a name given to four languages spoken by approximately 6 million people in western Uganda. This group was selected because it would be a good candidate for language instruction if policies change. Although it is not commonly taught nor well documented, it has a large number of speakers, is used in the broadcast media of western Uganda and occurs in regional communications for commercial purposes (e.g. advertising).

5.8. Needs assessment for Runyakitara CALL

The need for any local language learning system, such as a system for learning Runyakitara, is uncertain. Specific issues concern the number of potential users of the system (learners), access to information and communication technology and the existence of sufficient interest in learning local languages online/using computers. For this reason we investigated the perceived value of CALL software for supporting the learning of local languages such as Runyakitara. As should be evident from remarks above, we are convinced that local languages deserve educational support, we wished to understand others' attitudes to this.s.

5.8.1 Objectives

The needs assessment for the computer-assisted language learning of Runyakitara had the following objectives:

i. To identify the learners' interest/need for a CALL system;
ii. To determine IT/Internet accessibility for potential learners who might use the CALL software;
iii. To learn about the learners' experience with IT and language learning; and
iv. To identify the requirements that a Runyakitara language learning system should satisfy.

5.8.2 Study design

The general objective of this "marketing" study was to establish the need and possible requirements for the Computer-Assisted Language Learning (CALL) of less documented and less taught languages, particularly Runyakitara. The area of study, participants and methods of data collection were chosen based on the specific objectives mentioned above.

5.8.3 Study area and participants/subjects

The study was mainly carried out in Western Uganda, where the languages of Runyakitara are spoken. Four districts were selected based on regional representation and the important activities in which language plays a developmental role. The districts selected in Western Uganda were Mbarara, Kabale, Bushenyi and Kabalore. They were chosen because these districts contained institutions involving people with diverse cultures and languages who might be interested in learning their native or other local languages if facilities were available. Makerere University in Kampala was later included in the study because of its role in teaching Runyakitara.

Initially, the study targeted people who are not native speakers of any Runyakitaran language, but who work in areas where Runyakitara is spoken. These were lecturers, doctors, administrators etc. However, the responses from this category indicated that they had no time to dedicate to local language learning. Although this was disappointing to learn, it is valuable as we seek support for local language education. The professionals

were not opposed to support for local languages in the educational system, but it is clear that they do not see themselves as the primary beneficiaries of local language education.

The focus then was redirected to teachers and students of Runyakitara in universities and teacher training colleges. A total of 33 respondents participated in the study. The following table shows the number and composition of the group approached to participate:

Area	**Number of people**	**Students**	**Language Teachers**	**Others: (lecturers, doctors, administrators, etc)**
Mbarara	5	-	-	5
Kabale	6	4	-	2
Bushenyi	8	4	2	2
Kabalore	4	3	1	-
Kampala	10	6	-	4
Total	**33**	**17**	**3**	**13**
	100%	52%	09%	39%

5.8.4 Data collection and analysis

We administered a structured questionnaire to 33 respondents in areas where Runyakitara is spoken as well as at Makerere University, where Runyakitara is taught. The questions covered the learners' interest in CALL, IT/Internet access and their experience with IT/CALL as well as identifying the participants' CALL needs. All the questionnaires were returned, providing us with a response rate of 100%. Clarification of issues that were not clear in the written questionnaire was provided orally when the questionnaires were returned. The data collected was cross tabulated and analyzed manually.

5.8.5 Presentation and discussion of results

a) Participants' interest in Runyakitaran CALL

One of the objectives was to determine the level of the participants' interest in a potential CALL system for Runyakitara. Results indicate that 28 out of 33 participants (85%) would be interested in Runyakitara CALL, if available. The following table indicates the responses per district:

Area	**Interested**	**No interest**	**Total**
Mbarara	-	5	5
Kabale	6	-	6
Bushenyi	8	-	8
Kabalore	4	-	4
Kampala	10	-	10
Total	**28**	**5**	**33**
Percentage	**85**	**15**	**100**

The table above reveals that, apart from the respondents in Mbarara who had no interest in learning Runyakitara online, 85% of respondents were potentially interested in the computer-assisted language learning of Runyakitara. Respondents with no interest were mainly non-Runyakitaran-speaking administrators and lecturers who indicated that they had too little time to devote to online learning because of their busy schedules.

The above results suggest that there is interest in CALL systems for local languages, especially among students and teachers wishing to obtain learning support in these languages, to provide comprehensive learning content for their students and to benefit from the opportunities of learning local languages anywhere, anytime. It also suggests that the primary target group should be educators and certainly not busy professionals.

b) IT/Internet access of respondents

The second objective relates to the accessibility and availability of computer hardware, software and/or internet services. The success of CALL systems heavily depends on the availability and accessibility of computer hardware, software and, in some cases, internet connections. Mbarara, Kabale and Kampala respondents reported the availability of IT/internet access, while respondents from Bushenyi and Kabalore had no IT/Internet access at the time of the survey. This finding is further detailed in the following table:

Area	**Access**	**No access**	**Total**
Mbarara	5	-	5
Kabale	6	-	6
Bushenyi	-	8	8
Kabalore	-	4	4
Kampala	10		10
Total	**21**	**12**	**33**
Percentage	**64**	**36**	**100**

The results indicate that 64% of the study subjects have access to IT/Internet services, while 36% do not. Bushenyi respondents indicated that they can access a computer and internet only in townships and only when they urgently need to communicate. At the Primary Teachers College, in Kabalore where study participants from that district were recruited, there was no internet access at all at the time of study, and students could only connect to the internet in townships when they needed to communicate.

The limited access to IT/Internet in Uganda is not only a problem in Bushenyi and Kabalore, but most other districts as well. This should not stop the development of CALL systems for local languages learning. The software should be developed and used by those who have access; others will obtain access as the country's internet connectivity improves.

However, we note from the above finding that standalone systems would be preferred, at least for the near future.

c) Learners' experience with language learning systems

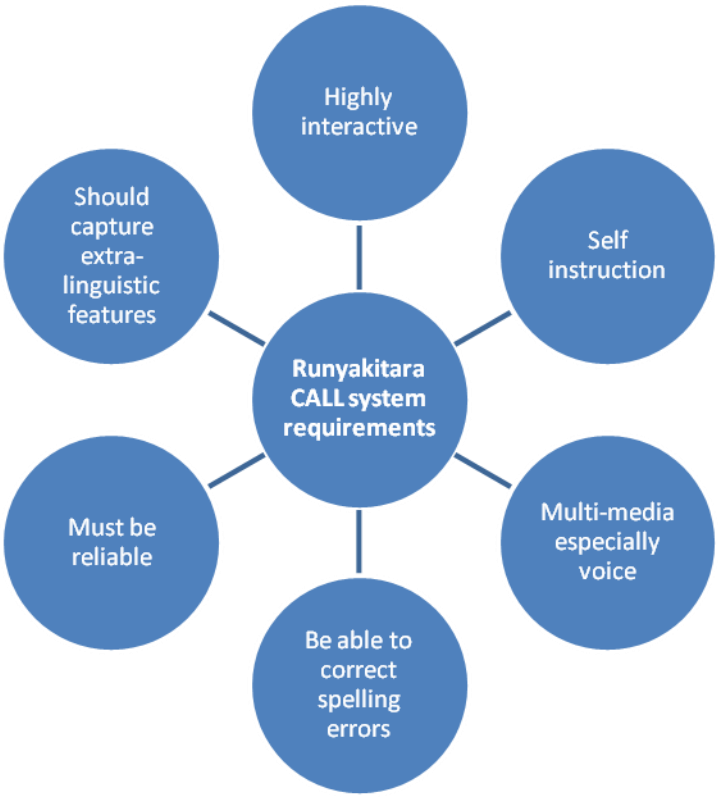
The third objective was intended to reveal if study participants had any experience in using CALL software or, more specifically, if they have ever used CALL to teach or learn any language. Based on the responses, we were able to conclude that none of the participants in any of the five sites had experience with language software apart from word processing.

The fact that none of the respondents had any experience with computer-assisted language learning systems prompted other questions about the ways in which participants used computers. These questions were administered through interviews. When orally interviewed on the use made of computers, all responded that they used computers for word processing. Further questioning revealed that most had used the thesaurus and spell-checker embedded in Microsoft Word. Therefore, they have used a computer mainly to check spelling and find synonyms in English. Since the study's focus was on Runyakitara, the conclusion is that no respondent had interactive experience with a language learning system for Runyakitara. Respondents from Kabale District indicated that they had learnt certain Rukiga vocabulary from broadcasts on radio 'Kigezi'. However, such experiences constitute electronically supported language learning and not CALL.

CALL is relatively unknown in Uganda, and it is not the case that people have tried it and discarded it. This implies that there is a knowledge gap that needs to be filled. It is important to provide CALL systems for local languages so that people can benefit from online language learning environments. They need to be able to continue to actively learn the language even in the absence of expert human teaching.

d) CALL system requirements according to survey participants

Regarding the features that learners would like to find in a computer assisted language learning system for Runyakitara or any other language, the following diagram summarizes the responses given:



All interested respondents indicated that they would like a system that is highly interactive, which means a system can be used with ease and that provides a great deal of feedback. They indicated that the user interface for the system should be friendly enough to support language dialogues.

Participants also indicated that they would like a system that would allow them to learn independently. Even when there is no tutor, the system should be able to help them to learn the elements of the language in which they are interested.

They would also like a multi-media system, particularly one that is voice-enabled. Respondents were very interested in pronunciation. This was especially emphasized by teachers, who wanted their students to learn how Runyakitara words are pronounced

Responses also revealed that study participants were interested in a system that is able to correct spelling errors. The orthography of Runyakitara is a great challenge because the Runyakitara languages have not been taught at lower educational levels, such as primary school, for a long time. A system that is able to correct spelling errors is very important to the learners for that reason.

Every respondent indicated a preference for a system that was reliable, meaning a system that remained stable without ever losing data when it is used. Some respondents also indicated that they would prefer a system that also dealt with extra-linguistic features such as cultural factors. This feature is important to the extent that it allows non-native speakers to understand and appreciate why certain language features are the way that they are.

The survey also revealed that the language usage should focus on academic and pedagogical language where national policy has already intervened. In effect, there is a need for a system to support students in their curricula. Such students need additional support to enhance their teachers' efforts to improve their language skills and knowledge. Language for general purposes should be targeted when electricity and internet coverage in Uganda improves.

5.9. Conclusion

On the basis of the analysis of language learning and teaching in Uganda, we conclude that the situation of both local and foreign languages is still difficult. While the situation of foreign language teaching and learning is not the best, the situation for local languages is far worse. English is promoted by policy and is generally felt to be of paramount importance, but it has failed to satisfy all of Uganda's language needs. As a result, some children drop out of school because they fail to cope with English and therefore fail to acquire the knowledge and skills delivered in that foreign language – which they do not understand sufficiently. Given the advantages of CALL, the situation of both foreign and local language teaching and learning might be improved in Uganda if CALL were to be deployed more broadly.

5.9.1. Focused summary

- The current situation of language teaching and learning in Uganda is not well documented in any literature that we have been able to obtain. Various studies dealing with the language situation in Uganda (e.g. Mukama 1991; Namyalo 2010) have not discussed teaching and learning. This chapter has discussed the dangers of this knowledge gap and how it can be filled.

- Different methods are currently employed in Uganda to teach and learn languages. This corrects statements made by some researchers that computers have influenced almost every aspect of life (Negroponte 1995). In Uganda, the computer has not yet influenced language teaching and learning.

- Computer-assisted language learning can be useful for the multi-lingual situation in Uganda;

- There is a perceived need for CALL, specifically for local language teaching and learning in Uganda. The perception is strongest among educators, while professionals definitely do not see themselves as future users and beneficiaries.

- CALL software for supporting local language learning should not at the moment rely on good internet connectivity, but rather should be capable of being used in a stand-alone fashion.

- CALL has improved the teaching and learning of languages in many countries of the world, we hope that, when it is deployed in Uganda, both foreign and local language teaching and learning can be improved.

Chapter 6

Computer-Assisted Language Learning of Runyakitara: A Pilot Study

(An earlier version of this chapter was presented at the 8th Strathmore ICT Conference 2-3rd September 2011, Nairobi, Kenya as: Fridah Katushemererwe & John Nerbonne: Computer Assisted Grammar Learning: An E-learning environment for Runyakitara.)

Abstract

This research reports on a pilot study on the effectiveness of an e-learning environment for Runyakitara grammar aimed at university students of the language. The objectives of this study were threefold: to examine the accuracy of the output from the morphological analyzer with respect to the intended application in computer-assisted language learning (CALL), to get a sense of whether background skills would be present to a sufficient degree, and to assess in a rough and ready fashion one realization of a Runyakitara CALL system for its suitability for use in language instruction. Content was designed and delivered using an existing electronic language learning environment (HOLOGRAM). The results indicate that developing electronic content for the instruction of Runyakitara grammar enhances learners' knowledge of word forms, motivates the learners and can effectively supplement the efforts of teachers.

6.1. Introduction

Language instruction in electronic learning environments has a long history, and teaching grammar in this manner has enjoyed a great deal of popularity, especially in the early years of computer technology. According to studies, popular grammar programs developed in the 1960s were mainly drill and practice programs implementing Behaviourist ideas in vogue at the time (Warschauer 1996; Nerbonne et al 1998). With the rapid development of computer technology, language learning programs based on other theories were also developed, resulting in a trend that moved away from behaviourist towards communicative and now integrative computer-assisted language learning (CALL). It should be noted, right from the start, that the reporting on this trend only includes very little research on Bantu languages and no literature at all on Runyakitara. We would like to remedy this scholarly gap within the line of research reported on here.

Complementing investigations regarding the use of computers in foreign language teaching, a growing body of research since the 90s has established that awareness of language categories, forms and rules is important for an adult learner's ability to successfully acquire and master a language. However, given the limited amount of time an instructor can spend with students, there are few opportunities in the classroom to foster linguistic awareness of a language's formal features and to provide individual feedback on errors. Jager (2004, 2009) documents that foreign language teachers still want their students to practice grammar through exercises even though they have long embraced the pedagogical wisdom of emphasizing communicative abilities as the objective of their instruction. CALL may then support these teachers' wishes to provide

practice opportunities using structured exercises so that valuable classroom time may be devoted to more communicatively oriented tasks. Given the results of Chap. 5, suggesting that our CALL efforts for Runyakitara should focus on its use in formal education (as opposed to use in self study), we find it valuable to follow Jager's suggested framework. We note one additional peculiarity of the Runyakitara situation, however. As a language for which there is little written (grammatical) description and few resources, the instructor's (and the students') access to written language materials is also limited and in some cases virtually non-existent. In this sort of situation CALL may become especially useful.

The instruction of grammar has received wide attention in research. It has been argued that grammar plays an essential role in the success or failure of formal communication (Ismail 2010). Grammar is also frequently regarded as the basis of the four language skills - listening, speaking, reading and writing. In listening and speaking, grammar plays a crucial part in the formulation and interpretation of oral expression (Widodo 2006). Written expression also demands the grammatically correct use of language, especially in formal situations. The ultimate goal of grammar instruction is to provide learners with knowledge of the way that utterances can be constructed in a language so that when they listen, speak, read and write the language, they have no trouble communicating in the language that they have learned. Language teachers are, therefore, challenged to find creative and innovative methods of teaching grammar. There is, however, a further question concerning the grammatical content to be taught.

For Gasser (2009), the content of grammar instruction depends on the language. In a predominantly analytic language such as English, syntax should have a major focus, as emphasis needs to be placed on the techniques of forming questions, the structure of active/passive sentences, relative clauses, etc. However, Bantu languages, such as Swahili or Runyakitara, require a significant amount of time to be dedicated to morphology because of its complexity and relationship to syntax.

There is a vast body of knowledge on both the theory and practice of teaching and learning the grammar of languages. Most of the theories have been tested and empirical results are available from many languages; however limited empirical research is available for Bantu grammar learning, especially in electronic learning environments.

This paper presents the results of a pilot study on an e-learning environment for Runyakitara morphology. The major objective of this study was to understand how to design a comprehensive Runyakitara grammar learning system. Specifically, the study set out to:

i) Determine the suitability (and deficiencies) of the Runyakitara morphological analyzer's output content for the instruction of Runyakitara grammar;
ii) Identify the required user skills, such as computer skills;
iii) Collect some experience on presenting Runyakitara grammatical material within a CALL system.

The next section reviews some aspects of Runyakitara morphological structure and its challenges for learning. The rest of the chapter discusses the design of the content in Hologram (Jager 1998), user reactions and the conclusions we draw.

6.2. Runyakitara morphological structure and how it challenges language learners

This section repeats material from Chap. 1, Section "Runyakitara Morphology". It was needed for the independent publication of this chapter but may safely be skipped by those who have read the earlier section.

Like the morphology of any language, the morphology of Runyakitara employs a categorization of words into word classes. Following the linguistic typology in Comrie (1989), Runyakitara can be characterized as a synthetic, agglutinative language. However, just as most languages cannot be placed exclusively in one class, Runyakitara exhibits some features of fusion and its verbs are highly inflected. Below is an overview of the Runyakitara morphological complexity that is of interest to this study:

a) Agglutination: Runyakitara is an agglutinative language in which words are formed from a process of combining morphemes, each contributing some meaning to the whole. For example, the verb root **shutam** 'sit' can be combined with morphemes as indicated in I below:

I) *shutam-**a*** 'sit!'
ku**-shutam-**a 'to sit'
ku**-shutam-**ir**-**a 'to sit on'
n**-shutam-**a 'I sit'

In the example in I above, the following morphemes have been added to the root: ***a*** (an indicative mood marker), ***ku*** (in the above example, an infinitive marker like the English 'to', but it also has other functions), ***ir*** (an applicative marker with the meaning of 'on') and ***n*** (a first-person singular subject marker). Each of the mentioned morphemes adds meaning to the root **shutam.**

b) Reduplication: In Runyakitara, reduplication is productive, insofar as words are formed by copying/doubling a part or a whole word. The following are examples of reduplication in Runyakitara:

II)	*kureeba* 'to see'	*kureeba-reeba*	'seeing (extremely)'
	kutema 'to cut'	*kutema-tema*	'to cut into smaller pieces'
	emwe 'one'	*emwe-emwe*	'one by one'
	omuntu 'a person'	*omuntu-ntu*	'a 'stupid' person'
	ezo 'those'	*ezo-ezo*	'those ones'

Note: verbs and nouns duplicate roots (in most cases), while others such as pronouns and numerals can duplicate the whole word.

c) Inflection: a Runyakitara verb can be inflected for negation, subject, tense, aspect, object, mood and adverbial content. The examples below illustrate the different forms of inflection of a Runyakitara verb, which are indicated in bold face:

vi) Mood – *reeb-**a*** 'see'
vii) Tense – *mureeb-**ire*** 'I saw him/her' – there are 7 tenses in Runyakitara, each with a different tense morpheme.
viii) Aspect – *n-**aa**-mureeb-ire* 'I have seen him/her'
ix) Negation – ***ti**-naamureeba* 'I have not seen him/her'
x) Adverbial marker(s) *mureeb-**er**-e* 'see for him/her'

d) Allomorphy: Runyakitara has various allomorphs, which is to say that a single morpheme can be realized in two or more different ways. A case in point is provided by the causative morpheme, which has six different realizations [**es/is/iz/ez/sy/y**]. The following forms illustrate the possibilites:

*reeb**es**a* 'cause … to see'
*kwat**is**a* 'cause … to touch'
*guru**sy**a* 'cause … to jump'
*r**iz**a* 'cause … to cry'
*te**ez**a* 'cause … to beat'
*ham**y**a* 'make … firm'

There are seven verb extensions that have two or more allomorphs (e.g. applicative, passive, stative and reversive morphemes).

e) Noun classification: Runyakitara has a noun class system, just like the other Bantu languages. Nouns of Runyakitara are categorized into noun classes most of them including in both singular and plural forms, but some of them occurring as either singular or plural. Noun classes have a great influence on other linguistic features, such as syntax and even morphology. This is because they determine three major linguistic properties: a) the noun **class itself**, b) number of a noun, i.e. whether a noun is singular or plural, and c) the concord system. The concord system is beyond the scope of this study.

Despite the complex nature of Runyakitara morphology, there is limited literature available to help learners to understand the different components of the words in Runyakitara and the components from which they are formed. The latest study available (Taylor 1985) does not clearly explain all the elements of Runyakitara morphology. Additionally, the way Taylor (1985) divides words into their components is sometimes misleading and confusing. For example, the verb *nooyenda* 'you want' is analyzed by Taylor as *n-oo-yenda* instead of *ni-o-end-a*. He also reads *rw-a-hend[ek]a*, instead of *ru-a-hend-ek-a 'It is broken'*. Taylor's interpretation of *nooyenda* is wrong because there is no morpheme **-oo-** instead it is **o** for 'you', He uses ***n-*** to mark a progressive aspect instead of ***ni***. In Runyakitara verb morphology, **n-** marks a subject and not an aspect. Taylor's system is confusing because it is difficult to account for the morphemes in a

given word. It would therefore appear that the knowledge gap in this area has not yet been filled.

Utilization of an already existing morphological analyzer for Runyakitara, which precisely indicates the word forms and their linguistic information, could fill the existing knowledge gap. Presumably, this output can provide accurate content with which to learn the morphological/grammatical structures of the language.

Based on the above assumption, we designed content on Runyakitara word formation for use in our pilot study. It is our belief that such content will enable us to achieve our research objectives.

6.3. Morphological analyzers as aids for the learning of morphology

Computational morphology has been applied in language instruction, particularly in second language teaching. It has been used as a means of analyzing the unfamiliar words encountered by students during reading (Nerbonne, Dokter, & Smit 1998; Shaalan 2005), locating words in a corpus that match a grammatical description provided by the student and generating word form exercises directly from a morphological analysis of morphologically complex languages (Gasser 2009).

We intend to describe another possible application of a morphological analyzer, one in which a learner directly interacts with the exercises developed on the basis of morphological analyzer output. We tested the idea using an existing e-learning environment – Hologram.

6.4 Runyakitara morphology instruction in Hologram

Runyakitara content was developed for delivery in Hologram, an e-learning environment designed by the University of Groningen in 1998 as a vehicle for providing practical exercises in English grammar (Jager 2009). Hologram was freely available and adaptable for use in the learning of Runyakitara grammar.

There are several reasons why Hologram was selected for this study:

To begin, Hologram strongly relies on linguistic structure as a defining element in pedagogy, and such a feature is important for Runyakitara grammar instruction. The target knowledge in this system is morphological competence and grammatical accuracy, both objectives that align well with Hologram programme structure.

The design of the our Hologram programme was also pedagogically motivated, as it aimed to stimulate more active engagement of the learners with the subject matter to be learnt. The programme offers flexibility in terms of content, time and resources. Learners are free to choose a topic during the process of learning while working at a convenient time and at their own pace; an alternative resource is offered for classroom teaching and traditional textbook forms.

The targeted group comprises adult learners, as the aim of the primary learning environment is to serve university students, who are regarded as adult learners. There is evidence that adult learners are more focused than school-age learners: the former have a better understanding about what is required for the learning process (Cross 1981).

Although many theories state that the main objective of learning grammar is to apply it in communicative contexts (Harmer, 2003), knowledge of grammar may have benefits for second language learners (L2).[11] One potential benefit is that the abstract understanding of language may facilitate the use of various grammatical structures in tasks of linguistic analysis. Students may also consciously improve on their grammatical prowess and become better and more cognizant performers in the classroom. The grammatical material chosen for instruction in the Hologram environment was selected with these objectives in mind.

The following figure illustrates the implementation of Runyakitara grammatical content in the design of a Hologram learning environment:

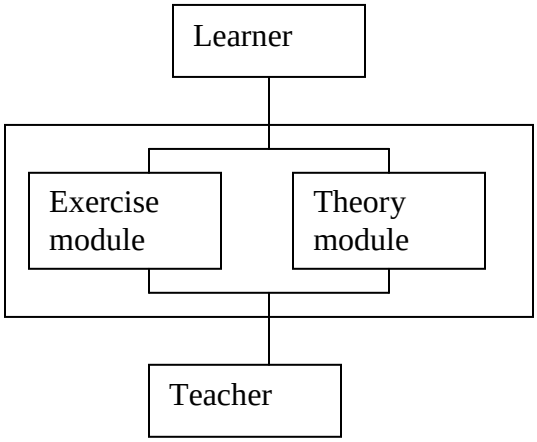


Fig. 1 Runyakitara-Hologram architecture

6.5 Development

Before turning to our intended users and their learning objectives, we note one technical point which became important as we wished to experiment further. Although we should have preferred to incorporate our morphological analysis software directly into the CALL application, this turned out to be infeasible, which meant that we implemented only a fairly small version of the program we wished to develop. Naturally we need to keep this in mind as more ambitious implementations are undertaken.

[11] But see the following chapter as well on the question of whether the learners we encountered should be regarded as second language learners or perhaps as students re-learning their first language. We naturally intend our work to serve the needs of various sorts of learners.

There are two categories of users in a Hologram system: teachers and learners. Teachers ensure that the exercises and theoretical content are developed and made available to students for learning. There are many steps that teachers follow to develop both content and exercises, but a detailed discussion of this process is not relevant to this chapter. Learners are supposed to access the system and learn. They do this by performing the following actions in Hologram: (1) retrieval of a relevant theory topic, (2) answering a presented question, (3) verification of the answer and (4) obtaining feedback. According to the Hologram developers (Jager 1998), none of these steps are mandatory. A learner can start answering questions without accessing theory or vice versa.

The **exercise module** consists of three kinds of exercises: cloze, multiple choice and drag and drop. All kinds of exercises have been utilized for Runyakitara grammar purposes. The following figure illustrates a multiple-choice exercise:

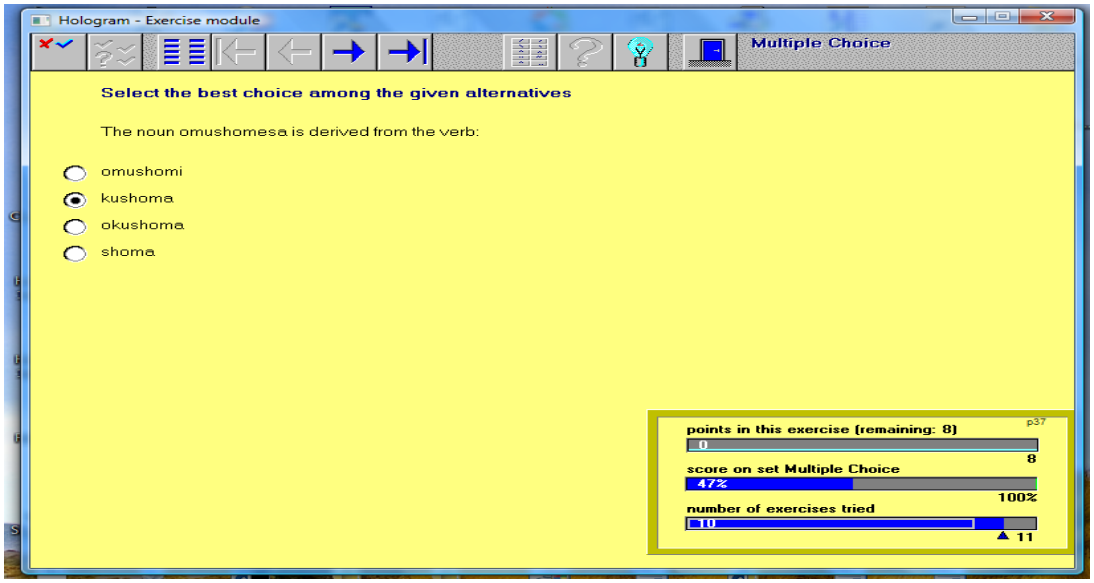


Fig. 2 User view containing a multiple choice exercise

Theory module

This module consists of the theoretical knowledge of Runyakitara grammar divided into two topics: the first topic deals with word formation in Runyakitara, specifically the combination of roots and affixes, as well as the noun classification system. The second theme relates to grammatical structures such as the concord system and simplified word order. The following is an example of a user view from the theory module:

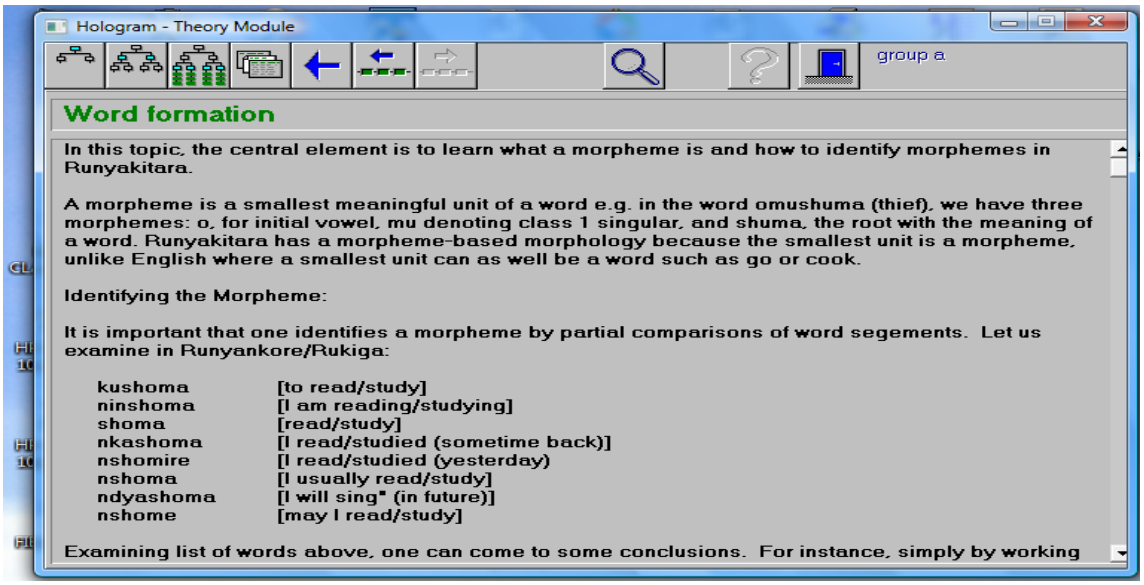


Fig. 3 – User view of theory

6.6. User study

To achieve our objectives, it was important to empirically evaluate the e-learning environment of Hologram-Runyakitara. The following questions guided the study:

i) How well can we support grammar instruction using the Runyakitara morphological analyzer?
ii) Are users able to use computers or do they need prior training?
iii) What can users – teachers and learners – tell us about using a CALL system for Runyakitara to gain morphological skills? We can collect evidence with respect to this point partly by simply asking how learners found the e-learning environment.

a) Study participants/subjects

Initially, two groups of Runyakitara learners were to be selected. One group (A) was to be randomly selected irrespective of the year of study and the language or dialect spoken. This group was particularly interesting for us because we wished to see which of them would benefit from moprhological exercises. Learners speaking any form of Runyakitara were to participate. The second group (B) was to be consist only of Runyoro-Rutooro speakers, in order for us to see how well they might deal with the content of Runyankore-Rukiga, i.e. the "other" major Runyakitara variety.

During the consultation meeting with lecturers of Runyakitara, we learned that not all students of Runyakitara were able to understand its morphology and grammar. Only students in their final year (i.e. third year of study) had sufficient knowledge of the subject targeted in this study. Therefore, we had to re-focus on only third year students.

The class consisted of only six students, four of whom agreed to participate in the study. This was disappointing, as the small number of participants meant that we would not be able to draw very reliable conclusions. But we were engaged in a pilot study in order to collect impressions of the software in use and gain insights for more ambitious implementations, so we plowed on. We emphasize that we definitely do not claim to be doing more in this pilot study.

b) Methods and Instruments

Different methods were used for different objectives. Participatory group discussion among language teachers was first used to determine the language content to be implemented in the analyzer system. User computer skills were identified by observation. A pre- and post-experiment test and a questionnaire were used to assess the skills acquired and to learn how the users experienced the system. The post-test was administered by means of software (Hologram). The purpose of the pre-test was to gauge grammatical knowledge of learners prior to exposure to the digital learning content. The pre- and post-tests consisted of word formation queries in the form of multiple-choice and cloze questions. The post-experiment questionnaire was intended to elicit user views on the Hologram-Runyakitara environment and consisted of objective and open ended questions.

c) Teachers' reactions with respect to system accuracy

There was a 3-hour participatory session for Runyakitara teachers who are native speakers to examine 300 words from the morphological analyzer of Runyakitara. The purpose of this session was to discover if the output of the morphological analyzer was accurate enough (from a learner's point of view) to be used directly for learning purposes. We relied here on the expertise of the experienced, native-language teachers for this.

The teachers were shown the results of analyzing 300 words from a Runyakitara corpus, which consisted of analyzed and non-analyzed forms. They were not asked for their evaluations of the pedagogical virtues (or vices) of the system, which we felt was too immature for valuable criticism at that point. After seeing the teachers' comments, we undertook a more detailed analysis of the existing system to ascertain which word categories were analysed to a sufficiently accurate degree for learning purposes. We manually examined the analyzer output to ascertain exactly which word categories were not accurate. The following table illustrates how the examination of 300 forms turned out:

Word class	**Noun**	**Verb**	**Pronoun**	**Adv.**	**Adj.**	**Conj.**	**Prep.**	**Interjection**
Output	96	100	20	26	44	5	4	5
Correct	94	76	20	26	42	5	4	5
Errors	2	24	00	00	2	0	0	0

Table 1– Accuracy of word forms

The above table reveals that verbs had the highest number of what we categorized as 'errors' (i.e. insufficient accuracy for learning purposes). An example of such a verb was *tuaashoma* 'we have read' which is a correct morphotactic form but which requires the further application of phonological rules to obtain the correctly pronounced form *twashoma*. These "errors" required that we implement more replacement rules in the system, which was not immediately possible due to time constraints. The other word categories were relatively accurate. We therefore decided not to include verbs in the learning applications until further development of the underlying morphological analyzer. It was also noted that even without verbs, it was possible to develop interesting and challenging content for learning, such as word formation in Runyakitara, noun classes, agreement, etc.

This achieves our first goal, determining the suitability (and deficiencies) of the Runyakitara morphological analyzer's output content for the instruction of Runyakitara grammar. We learned that it is best to continue CALL experiments that do not rely on verbal morphology, and that nouns, pronouns, adverbs and adjectives are suitable focus points for use with learners.

d) Learners' reactions

The learning programme was installed on the computers of the Institute of Languages at Makerere University. Before interacting with the electronic learning system, a pre-test was administered on students in which they were asked to answer questions on provided topics. All students were subsequently introduced to the learning material in Hologram, which they used to learn and do exercises at their own pace and at any time of their choosing. They were, however, allowed a period of 10 days to complete the material. The post-test and post-learning questionnaires were administered after this time.

e) User skills

Our second goal in this work was to see whether users were able to use a computer or whether they would require training before e-learning. Three of the four participants had the requisite computer skills and keyboard control, but one was computer illiterate. Efforts were undertaken to train him in the shortest time possible, so that he could use the computer to interact with Runyakitara exercises in Hologram. However, his learning outcome was lower than the others, and he attributed this reduced success to his limited computer skills. This finding has a lot of implications for the fully fledged study we were planning. Some students at Makerere University, especially among those studying languages and the humanities, are more or less computer illiterate, which means that any programme requiring computer use must first offer training in computer skills before e-language learning. This achieved our second goal in the study.

f) Data collection and analysis

Data collected from the participatory session was tabulated and analyzed manually. Three data sets were collected from each student during the learning experiment: individual

scores from a pre-experiment test, individual scores registered by the Hologram system and feedback from a post-experiment questionnaire. Due to the small number of participants involved, we decided to analyze results manually and present results in tabular form. A rating scale was used to analyze the feedback from the questionnaire, and we also used a grading scale for pre- and post-test results: 80 – 100 (excellent), 70 – 79 (very good), 60 – 69 (good), 50 – 59 (fairly good) and <50 (fail).

6.7. Results and discussion

We present the results thematically, based on the research questions that were posed.

a) Pre-test results for word forms

The aim was to get a sense of the extent to which a Hologram-Runyakitara learning environment might enable learners to gain morphological knowledge. The results from a pre-test indicated that students had some knowledge of the subject, with the highest score being 83% and the lowest 63%. However, such scores also indicated that the students needed to improve their knowledge. The following were scores obtained by each participant:

Participant	**Score (%)**
1	**70**
2	**73**
3	**63**
4	**83**

We had previously set a passing mark at 50% in accordance with the Makerere University standard. According to the students' scores indicated above, everybody passed the pre-test. This result suggested that students had some knowledge of the subject matter before the pilot study.

The post-test was done using Hologram. Scores registered by Runyakitara-Hologram system indicated improvement on the part of students. The following table illustrates the difference between results of pre- and post-tests

Participant	**Pre-learning Scores (%)**	**Experimental exercises scores (%)**
1	70	81
2	73	78
3	63	72
4	83	90

Scores registered by the study participants after the post-test indicate a positive change from the scores of pre-learning test. Student 1, for example, registered an improvement

from 70 to 81. We of course would like to attribute the improvement in scores to the e-learning material/content that was availed to learners all the time, but we realize that we cannot claim to have shown this on the basis of four students' improvement. We interpret the scores as an encouraging indication, but emphatically not a proof that the system is supporting morphological learning. The learning outcomes indicated above also reveal the excellent motivation of the learners, who needed to devote time to the study.

b) Post-experiment questions

We also queried the views and opinions of users regarding the Hologram-Runyakitara e-learning environment using questionnaires administered after a post-learning test. The purpose was to get a sense of the usefulness and comprehensibility of the content, as well as to assess the learning environment in relation to a classroom environment.

c) Usefulness of the learning environment

All four members of the experimental group gave the highest possible score (5, very useful) implying that they experienced the learning environment as very useful for grammar learning. The predominant view was that the system enhances learning. They emphasized that the system was also useful in supporting teachers' instructional materials. Given that the Hologram-Runyakitara system is designed for self-study either independently or in conjunction with a group language course, it is essential that students find it useful and that they are motivated to use it.

d) Assessing the programme in terms of usability

The group indicated that the system is user friendly, enjoyable and convenient. Convenient in this case meant that it was available throughout the ten days that the learners needed it. The group also noted that the learning environment was as good as classroom-based instruction. Users also emphasized that the system was user friendly and enjoyable. The insight we would promote is that CALL exercises and classroom instruction might complement each other.

The reactions and abilities of the students meant that we achieved our third and final goal, that of obtaining experience on presenting Runyakitara grammatical material within a CALL system. We learned that users found the system useful and we were encouraged by the fact that they improved in their morphological skills after using it.

6.8. Conclusion

The major objective of this study was to enable us to plan a more comprehensive CALL programme for Runyakitara using the morphological analyzer. As far as this was concerned, we achieved our objectives. The major points that were clarified by the study were:

- Verb analysis was not ready for learning purposes at the time of the pilot study, but the analyses of other parts of speech were quite accurate. We therefore focus on nonverbal morphology in later work (see following chapters).
- Computer skills are not "common knowledge" for everyone at university level, contrary to what many might assume.
- Results from the user study indicated (but definitely díd not prove) a learning effect, suggesting that the students learned because they used CALL. A questionnaire also indicated a receptive attitude toward the technology, suggesting that it might be adopted without resistance, perhaps even enthusiastically.

All in all, the use of electronic formats to present and practice Runyakitara grammar was a successful pilot activity for researchers and a beneficial experience for students at Makerere. Learners enjoyed the environment and requested that it should be made available to them in order to provide assistance for their learning. E-language learning has several advantages and is therefore worth an investment.

Chapter 7

Computer Assisted Language Learning (CALL) in support of (re)-learning native languages: the case of Runyakitara

(The paper from this chapter has been accepted for publication in Computer Assisted Language Learning (CALL) Journal. Acceptance letter was received on 19th March 2013.)

This study presents the results from a CALL system for Runyakitara (RU_CALL). The major objective was to provide an electronic language learning environment that can enable learners with mother tongue deficiencies to enhance their knowledge of grammar and acquire writing skills in Runyakitara. The system currently focuses on nouns and employs natural language processing in order to generate a large base of exercise material without extensive tuning by teachers. Language learners used the system over ten sessions, and their improvements were charted. Besides this empirical evaluation, we also sought the opinions of Runyakitara experts about the system (as a judgmental evaluation). Results from the evaluation study indicate that RU_CALL has the ability to assess users' knowledge of Runyitara and to enhance grammar and writing skills in the language. This computational resource can be utilized by other interested learners of Runyakitara, and the idea can be extended to other indigenous languages with emigrant populations who wish to maintain their language skills.

Keywords: CALL, re-learning native languages, Runyakitara

7.1. Introduction

This chapter presents a computer-assisted language learning (CALL) system that provides exercise material to learners of Runyakitara, a Bantu language (group) spoken in western Uganda. The system focuses on morphology, a notoriously difficult system in Bantu languages in general (Taylor, 1985), which is also difficult in Runykitara. In order to obviate the need to specify morphological forms one by one, the system employs a morphological analysis system developed with techniques from natural language processing (Nerbonne, 2002), in particular, finite-state morphology (Beesley & Karttunen, 2003).

The intended users of the system constitute an unusual target group for CALL. They are neither high-school or college (or university) language students nor do they need to learn the language for their work. They are likewise not tourists who wish to learn enough of a language to function in basic ways while traveling. Instead our intended learners are the children of native speakers who have emigrated from areas where Runyakitara is spoken natively. The children of migrants often fail to learn their parents' language in their new communities, and parents often see little value in passing their language on to their children (Ohiri-Aniche, 1997; Landweer, 2000). As Joshua Fishman (2000:5) put it "People who speak a language don't necessarily transmit it, and that is *the* problem [emphasis in original]". The children of Runyakitara migrants then have only very basic skills in the language (in an essay with another focus Nancy Dorian has dubbed such

individuals "semi-speakers", Dorian, 1977), but, as they grow older, they may be motivated to improve their abilities in order to become literate, to function more inconspicuously in their (extended) families, and to keep the option open of moving back to areas where the language is normally used in all facets of life. We aim therefore to support (re-)learning. We use the term '(re-)learning' with the 're-' in parentheses in order to be studiedly vague about the degree to which the students ever were competent speakers. The students have some limited competence in the target language, Runyakitara, but it is unclear whether they once knew it well. We envision supporting not only this unusual, but sizable group of learners, but also playing a role in more traditional settings for language learning. In school settings, for example, teaching literacy skills in native languages can aid in their preservation by increasing respect for them and providing a larger group of speakers with educated skills in the language. We draw attention to this unusual group of learners, who generally have little access to formal teaching, because CALL facilities may be especially important to them.

We also report on an evaluation of the system which consisted of comments from experts in the language and the analysis of a set of ten lessons in which users' abilities were tracked. The experts were positive, and the users systematically improved in their ability to recognize and to produce very complex Runyakitara forms.

The following section elaborates on our argument that the group of users we target is both unusual but also worth the effort involved in system development.

7.2. Motivation

Uganda is linguistically diverse with 43 living languages (Lewis, 2009). Great ethnolinguistic diversity means that English (the language of the former colonizer) had to remain the official language after independence. Today English is spoken by approximately 5% of the population which has a literacy rate of around 50% (Buttery *et al*, 2009). Although English is the official language of Uganda, a large number of Ugandans do not understand or speak it at all (Tembe & Norton, 2008).

Runyakitara, a name given to four languages, namely Runyankore, Rukiga, Runyoro and Rutooro, is spoken by about 6 million people in western Uganda. Other speakers can also be traced in Tanzania (Haya, Kerewe, Nyambo, etc) and Democratic Republic of Congo (Tuku, Hema, etc.). Having said that, let us hasten to add that the learners we target are in no sense acquiring a standard language on the basis of a mastery of a dialect (see below).

Even though Ugandans are not in general capable in English, local languages such as Runyakitara are not well documented or well known, not even to all their native speakers! Presently, some Ugandans cannot effectively read or write in their first languages even when they are educated, simply because they are encouraged to use English from childhood on and take pride in using it in daily communication. This means that individuals are often motivated later to (re-)learn local languages in order to function socially and economically in different places of residence.

Uganda as a country recognizes an obligation to provide information to its citizens in the languages they understand well, and to encourage the development, preservation and enrichment of all Ugandan languages (Constitution of Uganda, 1995). Therefore, Uganda

supports newspapers and radios in local languages. But, because literate speakers of local languages are scarce, government documents in local languages are full of typographical and grammatical mistakes. The people employed are not proficient enough in the different local languages. This means that there is also an officially recognized need to support proficiency in local languages.
We therefore aim to support Ugandans in learning their local languages, in particular by providing CALL systems designed for this purpose. Specialized systems may have an impact on the existing language situation by improving the general level of proficiency.

This research targeted most specifically a group of learners that has not been widely considered, i.e., people not proficient in their own first language. These learners may have suffered from language attrition (Schmid & de Bot 2004; Schmid et al. 2004), but it is likely that many of them never learned their parents language well, just as many migrant children fail to learn their children's language well, as many contributions to Fishman's (2001) collection document (see especially M.Clyne's contribution on Australian immigrant languages). While we envision a larger potential group of beneficiaries for the system we present and evaluate below, we focus in our evaluation on a group of learners who had acquired some ability in Runyakitara from their native-speaker parents, who had moved from the Runyakitara-speaking area to Kampala. The parents often spoke Runyakitara to each other but not to their children, leaving the children with little proficiency in their first language.

Given these circumstances, such people need help in their *own* first language (Fillmore, 2000). In most cases, such people shy away and do not participate where language proficiency is required. As Halliday (1968) states: 'A speaker who is made ashamed of his own language habits suffers a basic injury as a human being: to make anyone, especially a child, feel so ashamed is as indefensible as to make him feel ashamed of the color of his skin'.

7.3. Related research

Extensive research has been done in CALL and also in Intelligent Computer Assisted Language Learning (ICALL) (Warschauer & Healey, 1998; Gamper & Knapp, 2002). There is also a wealth of research on teaching morphology using CALL (Antoniadis et la., 2005; Shaalan, 2005; Blanchard et al., 2009; Nagata, 2009; Dickinson, 2010; Esit 2011; Amaral & Meurers, 2011). This section does not attempt to review CALL and ICALL generally, but focuses instead on literature on CALL systems for learning morphology and on systems for native African languages.

Warschauer and Healey (1996) observed that recent years had shown an explosion of interest in using computers for language teaching and learning. They describe the role of computers in CALL, a brief history of CALL, and various desgn philosophies, including Behaviorist CALL, Communicative CALL and Integrative CALL. The authors further predict that the future of CALL will heavily rely on the ability of learners and teachers to find, evaluate, and critically interpret net-based information. Their insights informed our research with respect to the history and future directions of CALL.

GLOSSER (Nerbonne & Dokter, 1999) is an early system that extensively utilizes a morphological analyzer in language learning. The major components of this system include a morphological analyzer for French, a part-of-speech disambiguation system, a bilingual dictionary, and aligned bilingual corpora. The system provided intelligent assistance to Dutch students learning to read French. The system's strength lay in its individualized instruction and its facilitation of access to additional learning resources (see above). The focus, however, is the learning of vocabulary that needs to be acquired separately from reading exercises.

Gamper and Knapp (2002) provide an overview on intelligent computer-assisted language learning (ICALL) systems. The most advanced systems were investigated and classified along five dimensions: supported languages, Artificial Intelligence techniques, language skills, language elements, and availability. The authors also discuss outstanding problems which still need further research in order to exploit the full potential of intelligent technologies in modern language learning environments. This review of literature provided a framework for the practical, empirical research that we aimed at.

Amaral and Meurers (2011) present the motivation and prerequisites of a successful integration of ICALL tools into current foreign language teaching and learning (FLTL) practice. The authors focused on (i) the relationship between activity design and restrictions needed to make natural language processing tractable and reliable, and (ii) pedagogical considerations and the influence of activity design choices on the integration of ICALL systems into FLTL practice. We profited from their insights while focusing on the task of supporting the (re-)learning of a first language.

Dickinson and Herring (2008) employed the **TAGARELA** framework developed by Amaral and Meurers (2006) to develop online ICALL exercises for Russian. Their system aims to teach basic grammar to learners of Russian, and its strength derives *inter alia* from audio and video exercises that enable the observation of language situations outside the classroom and life-like listening practice. Their system is internet-based, facilitating learning anytime and anywhere. Their exercises have fixed content, however, thus limiting learners to the content the developer put in the exercise.

Shaalan (2005) developed an ICALL system for Arabic learners. His system employs a morphological analyzer, sentence analyzer, reference material, feedback analysis and multi-media exercises. The aim was mainly to teach Arabic grammar to primary school children and learners of Arabic as a second/foreign language. The strength of this system lies in its multi-media and detailed feedback. In addition, learners are encouraged to produce sentences freely in various situations and contexts. The weakness of this system is that it follows a strict primary school curriculum, which may not be suitable for adolescent and adult learners of foreign languages.

Nagata (2009) presents a new version of Robo-Sensei's NLP (Natural Language Processing) system which updates the version currently available as the software package *ROBO-SENSEI: Personal Japanese Tutor.* According to Nagata (2009) the new system can analyze all of the grammatical structures introduced in a standard 2- to 3-year Japanese curriculum. It is supposed to serve as the backbone of a new, online CALL Japanese textbook capable of providing immediate, personalized feedback in response to errors produced by students in full-sentence-production exercises. The research focuses on strategies for error detection and feedback generation and describes how these

strategies are integrated into Robo-Sensei's NLP system, what types of errors are detected, and what kinds of feedback messages are generated.
Hurskainen (2009) presents a UNIX-based ICALL system for Kiswahili learners. The system trains word order and concord patterns. It is based on a morphological analyzer of Kiswahili and does not limit the learner with respect to vocabulary. No evaluation or user study is presented. Katushemererwe and Hurskainen (2011) discuss an idea for a Runyakitara ICALL system. The system involves an implementation of rules for learning word order, concord and vocabulary in Runyakitara. No testing or evaluation of the system was done. In addition, the target group was different from the group targeted in the present study because the system targeted advanced students of Runyakitara at university level and teachers of Runyakitara in primary teachers' colleges.

Odejobi and Beaumont (2003), Oyelami (2008), Hamwedi and Dalvit (2012) and Van Huyssteen (2007) report on CALL systems for Yoruba, Igbo, Oshikwanyama, and eleven (!) South African languages, respectively, focusing on the children of emigrants, and second and foreign language learners. They are therefore different in focus from the present paper.

Despite some interest in CALL for African languages, it is evident that more research needs to be done. From the literature reviewed, the focus of our study remains different from other studies reported in the following ways:

i) We focus on Runyakitara, a less documented and not commonly taught language.

ii) We target "re-learners", including learners who have only basic, passive abilities in Runyakitara, a group unlike those in most other studies.

iii) We provide exercises derived from a natural language processing system, unlike in other learning systems where a morphological analyzer is used to analyze the learners' answers (Shalaan, 2005), or as aid in providing morphological knowledge or dictionary access (Nerbonne & Dokter, 1998; Amaral, 2007). We utilized the morphological analyzer to develop exercises for learning.

iv) We report the results of evaluating an implemented system. Learners experimented with the system, and their experience (including their learning) is analyzed later in this paper.

7.4. Highlights of Runyakitara noun morphology and consideration for RU_CALL

We have focused on noun morphology in RU_CALL to-date because it is difficult to learn as already stressed by some Bantu language learners: "One of the most difficult aspects of learning Swahili is its system of nouns…"[12] Naturally, a more complete system

[12] www.transparent.com/learn-swahili/overview.html

would have to include exercises for verbs as well. Table 1 illustrates singular and plural morphology in Runyakitara:

Table 1. Examples of noun forms in Runyakitara.

Class 1/2	**Singular**	**Plural**
(people class)	O**mu**kazi (a woman)	a**ba**kazi (women)
	Mukazi (woman)	**ba**kazi (women)
	O**mw**ana (child)	a**ba**ana (children)
	Swenkuru ((my)grandfather)	**baa**shwenkuru ((my)grandfathers)
	----------	abaryakamwe (people symbolizing oneness)
	O**mu**hangi (creator)	----------
Class 9/10	e**n**te (cow)	e**n**te (cows)
	E**m**buzi (goat)	e**m**buzi (goats
	Ebaafu (basin)	ebaafu (basins)
	Baasi (bus)	**zaa**baasi (buses)

Table 1 shows examples from only two declension classes of nouns, including class 1/2 containing the greatest number of forms. In total there are 18 declension classes in Runyakitara, all of which are instantiated extensively in RU_CALL, each with two or more forms for singular vs. plural. These are complex and challenging to learners, as we have argued above in section 4.4.2. The complexity stems from the fact that they are not phonologically motivated, but rather must be learned lexeme by lexeme.

We focused on nominal morphology not only for its complexity, but also because the noun is an important word category in Runyakitara. The noun class of a given noun influences other nominal constituents such as pronouns, adjectives and verbs which must agree with the nouns they form constructions with (or represent anaphorically). For example, in the phrase *a**ba**ana **ba**to **ba**ija* (‘young children have come’), the noun class plural marker **ba** appears in a noun (*a**ba**ana*), an adjective (***ba**to*) and a verb (***ba**ija*).

Nouns in Runyakitara are associated with an initial vowel which serves as a pre-prefix to the root or stem. These vowels are specific. They include: **a,** (***aba**ntu* ‘people’) **e**, (***eki**tookye* ‘banana’) and **o**, (***omu**ntu* ‘person’) as presented by Ndoleriire and Oriikiriza (1990). There are rules that govern the occurrence of the initial vowel. If the noun class prefix contains the vowel **a**, e.g. **ba** or **ma**, the initial vowel will be **a,** thus, ***ama**te* ‘milk’ ***aba**kazi* ‘women’. When the noun prefix has **i** or -, the initial vowel is **e** for example, ***e**ki**i**tookye*, ***e**m**i**ti*, etc. The initial vowel is **o** when the noun class prefix has **u**, as in ***o**m**u**ntu* ‘person’ or, ***o**m**u**ti* ‘tree’. When a noun is preceded by a preposition such as **omu** ‘in’ or **aha** ‘at’, the initial vowel is dropped e.g. *omu muti* ‘in the tree’.

Once the noun morphology has been mastered, the learner has less trouble in phrase and sentence construction in Runyakitara. We pursue this further in Chap. 8.

7.5. RU_CALL: design and implementation

RU_CALL is a drill and practice system as well as a testing system. Although we are aware of language teachers’ preferences for communicatively oriented language teaching, we also note that many of the same teachers frequently assign CALL drills and exercises for use outside the classroom (Jager, 2009). Specific objectives for designing RU_CALL were:

i) To act as a testing tool of the learners' knowledge of vocabulary of their own first language;

ii) To test learners' knowledge of grammar, that is, whether they can identify a given noun as either singular or plural, and whether, given one form, they can produce another with contrasting number, e.g., plural when shown singular.

iii) To act as an evaluation tool by providing scores which will aid the teacher to evaluate learners of the language.

iv) To provide grammatical (morphological) exercises for students of Runyakitara.

To achieve the above objectives, the following was devised as a conceptual design:

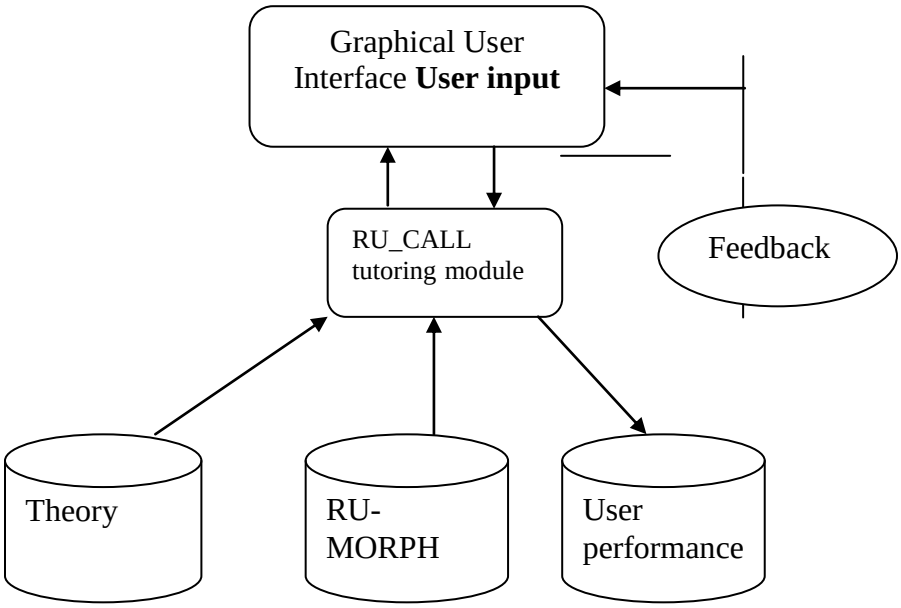


Figure 1. A Simplified RU_CALL Architecture

We chose to develop a stand-alone system rather than a web-based system in order to benefit communities in Uganda where there is little or no Internet connectivity, including therefore the large majority of places where it is limited or unreliable. RU_CALL provides the learner the opportunity of learning at his convenience in terms of time and medium.

7.5.1. RU_MORPH (The Morphological Analyzer of Runyakitara)

The linguistic knowledge in this learning system is derived from a morphological analyzer of Runyakitara, which was developed using Natural Language Processing (NLP) techniques (Jurafsky & Martin, 2008). NLP techniques have been identified as instrumental in developing pedagogically sound language learning applications (Nerbonne, 2002) and computationally tractable (Amaral & Meurers, 2011). The

morphological analyzer of Runyakiara specifically utilized Finite State Automata (Beesley & Karttunen, 2003; Hanneforth, 2009).

Because the the work on the morphological analyzer for Runyakitara nouns has been published in a journal (Katushemererwe & Hanneforth, 2010), we refer the reader to that article for technical details. But we summarize here that the 4274 nouns used were extracted from a Runyankore-Rukiga dictionary, *Kashoboorozi* (Oriikiriza 2007), which, according to Oriikiriza (2007), incorporates the material from all the Runyankore-Rukiga dictionaries published earlier. In addition, *Kashoboorozi* was the most recent dictionary available at the time. We note nonetheless that *Kashoboorozi* does not cover all the nouns in all the four Runyakitara languages. Since, however, the four Runyakitara languages are judged to be 80% mutually intelligible (Lewis, 2009), we expect the coverage to be quite adequate for all four languages. The software was also tested at various levels of development and it presently analyzes newspaper corpora at 78% recall, and 72% precision. In addition to measuring based on newspaper text, we asked lecturers on the Runyakitara language (see below for more detail on these lecturers) to evaluate the coverage of the the nouns in RU_CALL by the recall and precision searching for 100 nouns s/he knew. They reported that 90% of the nouns they sought were in the system. We interpret this to mean that the nouns most commonly known and used are covered by RU_CALL.

The following is the sample test output from the morphological analyzer of Runyakitara:

Table 2. Linguistic Information from the Morphological Analysis System.

```
aheeru :            aheeru[ADJECTIVE_ROOT15S
ahi :               ahi[DEM_PR_CLASS16]
ahu :               ahu[DEM_PR_CLASS16]

ahurira :           a[VERB_PREF_SPM3S Spm3s=agrmt3s][VERB_PREF_PRESENT
                    Present=habitual]hurir[VERB_ROOT_SIMPLE
                    Simple=simpleverb]a[VERB_END_IND Ind=mood]

ebijwaro :          ebi[NOUN_PREF_8P 8s=npref8p]jwaro[NOUN_ROOT_IT
It=class7]

naagamwaraguza : n[VERB_PREF_SPM1S Spm1s=agrmt1s]aa[VERB_PREF_ASPECT2
                    Aspect2=perfective]ga[VERB_PREF_OPM6 Opm6=agrt6]
                    mwaraguz[VERB_ROOT_SIMPLE
                    Simple=simpleverb]a[VERB_END_IND Ind=mood]
```

All word categories are described in the morphological analyzer of Runyakitara as illustrated above. For the purposes of the RU_CALL system, the following word categories were exploited:

Word category	**Class**	**Number of forms**
Nouns – classes	1-18	12,480
Demonstrative pronouns	1-18	72
Adjectives - classes	1-18	1,546

7.5.2. RU_CALL tutoring module

RU_CALL comprises learning content, tutoring and feedback control. As noted above, we offer grammar exercises. Awareness of language forms and rules is important in

language learning (Amaral & Meurers, 2011). As noted above, Jager (2009) further elaborates that many teachers pursue a communicative philosophy in class but assign grammar-oriented CALL exercises.

7.5.3. Theory

The system has supplementary material in form of grammatical explanations. This content is not part of the morphological analyzer, but can be accessed by the learner when he/she accesses the system. Grammatical content is organized in topics and sub-topics which should be easy for the learner to understand. We do not elaborate on this here as it is not innovative.

7.5.4. Learner Performance Monitoring

We maintain a database containing learners' identification (name and/or student number), date of learning, content already covered and scores the learners obtained. In addition, a search facility was designed to allow teachers to search for the scores of a given learner in case the number of learners grows.

7.5.5. Feedback

After each input from the learner there is feedback. The importance of feedback in enhancing learning has been demonstrated often (Sauro, 2009). There are three types of feedback included in our system: corrective, motivational, and directive feedback. When the input is correct, feedback is motivational, i.e., the learner is informed that the input is correct and directed to the next course of action. When the input is incorrect, the learner is also informed accordingly and normally asked to try again or to consult the theory module. With respect to corrective feedback, the learner is given the correct answer after a number of attempts. The learner is also guided to consult theory just in case s/he wants to learn more about the word/phrase.

7.6. The RU_CALL system

RU_CALL system may be described from different perspectives: a user's view of the system, RU_CALL tutoring, assessment, morphological analyzer and theory.

7.6.1 User's view of the system

An interface provides a means of communication between the user and the RU_CALL system. It is used to present lessons, allow the learner to submit input and to obtain feedback.

7.6.2. Learner

To access the system, the learner must first register to allow the system to recognize the learner profile and be able to store his or her scores. Once the learner is logged on, s/he

performs an exercise, including the following: i) answering the multiple choice questions, ii) providing alternative singular/plural words and phrases as prompted, and iii) getting feedback. The learner can also ask for an answer in case s/he does not have any clue. The learner is also free to invoke theory if s/he needs it either before, during or after learning. None of these steps are mandatory. One can start answering questions without accessing theory or vice versa. One can also ask for a correct spelling without answering the question. We walk through one exercise item in the next section.

7.6.3 RU_CALL tutoring module

This module controls the sequence and selection of the subject matter presented to the learner. In addition, it has a response mechanism to answer learner's questions with appropriate answers. This module also tracks the learner's level of proficiency in the exercises.

RU_CALL implements two types of lessons covering plural formation in Runyakitara. The first consists of individual nouns, while the second consists of noun phrases. A learner is required to identify whether the material – word or phrase – is singular or plural and then go on to provide the appropriate alternative (singular/plural). For example, if a learner selects a word as plural (correct form), the system prompts the learner to then provide additionally its singular form. Figure 3 shows the interactive interface with the learner:

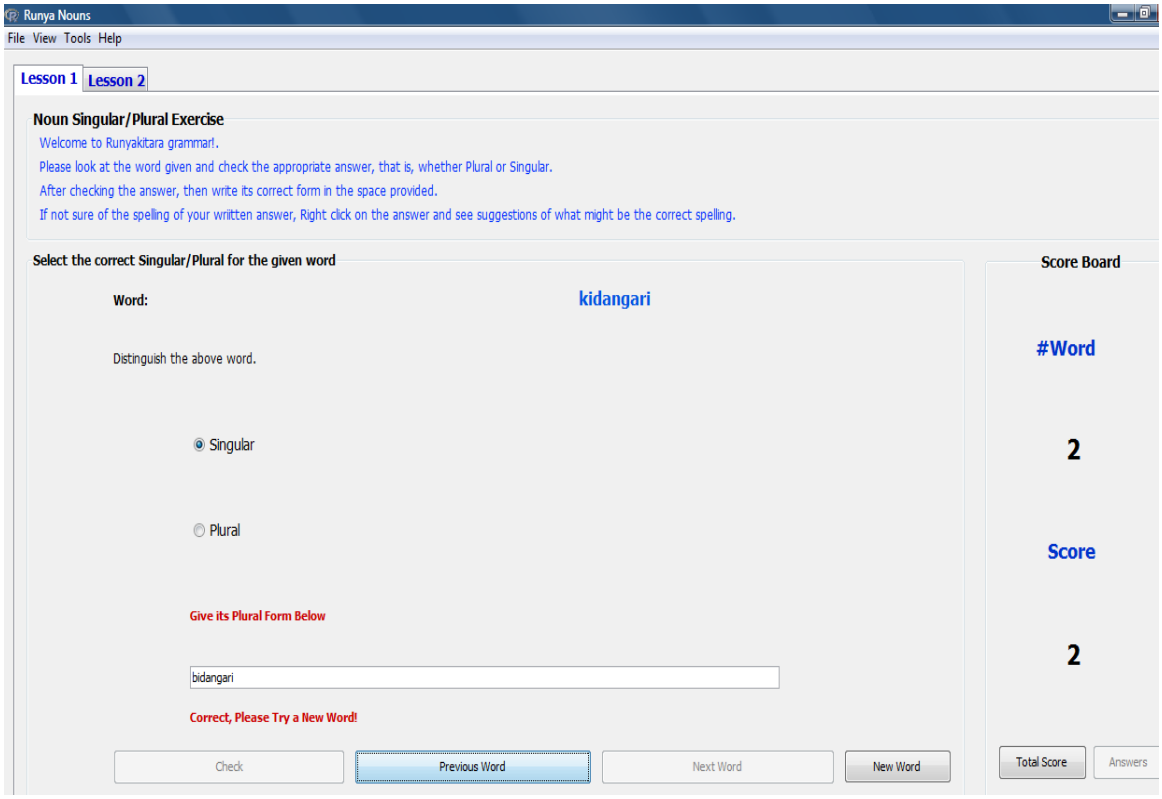


Fig. 2: RU_CALL learning interface

Feedback, as part of the RU_CALL tutoring module, was implemented motivationally as '*please try again'*, correctively as '*the right answer is …*' and directively 'next'. The following exercises illustrate the steps the user takes to interact with the system using the example in fig. 3. Section a illustrates a correct input, while section b a wrong input.

a)	System:	select the correct singular/plural form of the given word - **kidangari**
	User:	singular
	System:	Correct, please give its plural form below.
	User:	bidangari
	System:	Correct, please try a new word.
b)	System:	select the correct singular/plural for the given word - **kidangari**
	User:	plural
	System:	Incorrect. Please try again
	User:	singular
	System:	Correct, please give its plural form
	User:	kidanga
	System:	Incorrect, please try again
	User:	kidangariri
	System:	Incorrect, please try again
	User:	kidangari
	System:	Incorrect, the correct form is bidangari

Table 3: user interaction-system exercise

These are not simple tasks given the learners and the nature of the language. In the first place, the task requires knowledge of both words and phrases. First, if the learner does not know the word, (as in the case of b) s/he has no ability to identify its grammatical number. Second, the task requires writing skill of the learner. By requiring a written singular or plural form, productive competence and writing skills are being acquired and tested.

7.6.4. User Performance

The module keeps track of every learner with respect to individual lesson(s) and the date, time and success of learning, and uses the data to compile statistics and provide feedback to the learner and the teacher. The statistics compiled are the total score and the percentages for each lesson. The system displays performance in two ways: to the learner, the score board is displayed immediately after login. To the teacher, the system compiles a list of all learners who are registered together with their scores, and is able to display it on request. Figure 4 illustrates the scores interface:

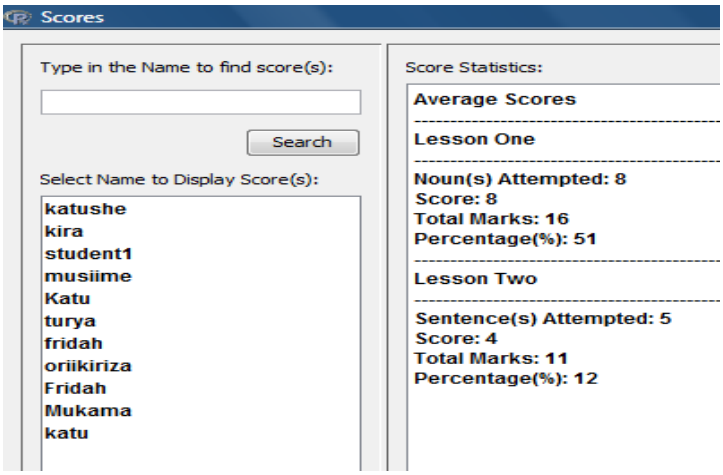


Fig. 3: Scores interface of RU_CALL

7.6.5. Interface to morphological analyzer

Rather than require that the (rather complex) morphological analyzer be invoked during use, we compiled its output for several thousand nouns and nominal phrases and stored this in a database, Noun Property. Noun Property has a list of all nouns, a display window, and a search facility. When you click on a particular noun, properties of that noun are displayed on the noun property window on the right. The purpose of the search facility is to find nouns not visible on the list. This is illustrated in Figure 4 below:

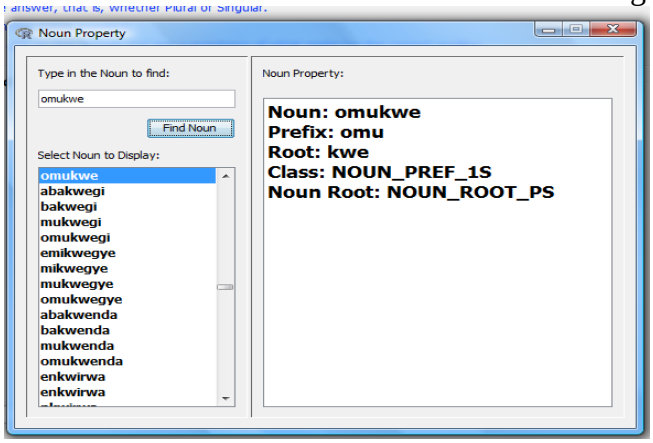


Fig. 4: A noun property view from the morphological analyzer.

7.7. Evaluation of RU_CALL

Traditionally, CALL evaluation took a comparative framework in which the learning outcome of CALL activities was measured through experimental or quasi-experimental design and compared with non-CALL activities. It is now commonly agreed that this type of evaluation itself is outmoded and not very revealing, largely due to the methodological limitations associated with making comparison between CALL and non-CALL activities (Ma, 2008). This means that CALL can be evaluated without comparing with non-CALL activities. What is needed is to improve CALL to make it more efficient and effective.

Charpelle (2004) provides a set of 5 principles for evaluating CALL as summarised in Hubbard (2006). We followed the principles to evaluate RU_CALL. These are summarized below as follows:
1) CALL evaluation is situation specific;
2) CALL should be evaluated both judegmentally and empririically;
3) CALL evaluation criteria should come from instructed SLA theory and research;
4) The criteria should be applied relative to the purpose of the CALL task; and;
5) The central consideration should be language learning potential
The most convincing way of CALL efficiency/inefficiency is to measure the learning outcome. Learning outcome is interpreted by CALL researchers as learning potential, that is, how well the linguistic forms are mastered after CALL use (Chapelle 2001). This usually involves identifying the targeted language learning objectives, (e.g. grammatical/lexical knowledge, reading comprehension, or writing competence), and designing the corresponding tests to measure the learning of these objectives – usually using a pre and post test. Our target was to test the feasibility of RU_CALL in measuring the learning outcome.

7.7.1. Study design

Evaluation was carried out in terms of the learning outcome, system appropriateness and users' general views about the RU_CALL system, keeping in mind that it was their first experience. The following were the more specific research questions:
How do experts evaluate the appropriateness of the system with regard to:

- Learner fit, as described by Hubbard (2006): What is the quality of the opportunity for engagement with language under conditions appropriate for the learners?
- The accuracy of the learning that is stimulated?

How well have users mastered the forms of Runyakitara, focusing on specific aspects of grammar, vocabulary and writing?

- Can a learner recognize/understand the meaning of a given word?
- To what extent can a learner distinguish a given noun as singular or plural?
- To what extent can a learner write the alternative *number form* of a noun accurately?

What is the learning outcome of the digital Runyakitara learning environment?

- To what extent will the digital learning environment help Runyakitara learners enhance their knowledge of grammar?

How do learners evaluate CALL system for Runyakitara?

- What unique aspects do learners discover in this learning environment?
- Do they find the system to be useful?
- How do they compare it with classroom controlled learning?

a) Study participants/subjects. The study used two categories of respondents: experts and learners. Experts were included to judge the appropriateness and accuracy of the system,

learners were essential for gauging the effectiveness of the system empirically. Three experts were employed, all university lecturers of Runyakitara. Runyakitara has a limited number of experts; therefore, only three were available to take part in the study.

Learner respondents were students entering university and were of Runyakitara heritage. This particular group of students was randomly selected to participate in the study. Some CALL authorities suggest that between 20 and 30 participants are appropriate for user studies (Ma and Kelly, 2006). We targeted 30 learners, but only 26 participated in the study.

We should have preferred to conduct the study using a second, control group, but there is not tranditional self-study material available for Runyakitara, i.e. language text book (or draft materials) with paper-based exercises. Developing that material solely for the purpose of comparison with CALL material would have been prohibitively expensive.

b) Instruments. A checklist and also a questionnaire were designed to obtain judgmental responses from experts. The checklist required 'yes' or 'no' answers, while the questionnaire comprised both structured and open-ended questions.

For learners, a pre-experiment test and a post-experiment test together with an evaluation questionnaire were designed. The pre-test comprised 100 fill-in-the-blank questions involving nouns and nominal morphology. The post-learning test was administered after the software (RU_CALL) was used to ascertain whether there were gains in grammar and spelling. The post-test was constructed in the same fashion as the pre-test. The purpose of the pre-test was to gauge vocabulary, spelling and grammatical knowledge of students before the digital learning content exposure. The post-experiment questionnaire was intended for acquiring information concerning the learners' views on the learning environment.

c) Procedure. The entire experiment for learners followed a three-step procedure: pre-learning test, learning experiment and post-learning questionnaire. The learning program was installed on Makerere University (School of Computing) computers. Before interacting with the electronic learning system, a pre-test was administered on paper. All learners were then exposed to the learning material in RU_CALL, to learn and do exercises at their own pace, two hours a day, so that the overall time of the experiment was ten hours, spread across five days. Given that the learners had had passive exposure to Runyakitara, we hypothesized that ten hours of continuous grammatical exercises would be sufficient to demonstrate enhanced command of the language. Detailed instructions were given to learners regarding system access, use, and the entire learning procedure was fully explained.

7.8. Results and Discussion

7.8.1 Results from experts

We asked experts to evaluate RU_CALL system with respect to the following dimensions: effectiveness, coverage, accuracy and selection of content for learning.

System effectiveness. The three experts agreed that RU_CALL would be able to achieve its intended objectives. We interpreted this to imply that RU_CALL was ready to be empirically evaluated.
Coverage. The system was intended to cover all Runyakitara nouns, and the experts were satisfied that over 90% of the nouns learners were likely to encounter would be covered. One also pointed out some missing common nouns, which we took to indicate that the system must be updated from time to time. The nouns which were missing at the time of evaluation were later included, since the system is easily expandable.
Content accuracy. The noun forms in the system were intended to be accurate and familiar to the experts of Runyakitara, because they were from a 2007 dictionary of Runyankore-Rukiga. In the experts' opinion, nouns were mostly familiar, but they also noted a few cases where nouns seemed foreign. For example, none of them knew the meaning of *ebyangato*, even though it is from a dictionary. Perhaps this shows only that not even experts know all the words in the dictionary.
Random selection of content for learning. Regarding the pedagogical aspect of selecting content for the learner, the experts were all dissatisfied with the random selection of nouns as a good method of selecting content for learning. They suggested that nouns should be systematically presented (arranged under topics) and selected so that learners would be likely to understand them. Our assumption had been that learners should focus on grammar in these exercises rather than on vocabulary. We concede, however, that it would be preferable to group nouns in order to synchronize the morphological learning with other parts of language courses which may systematically vary the situation in which a language is used.

7.8.2 Results from learners

At the beginning of this study, it was not clear whether the assumption we had about learners was true. The basic assumption was that students of Runyakitara heritage raised in a non-Runyakitara area would have limited knowledge of the Runyakitara language. We therefore tested the extent to which they knew Runyakitara vocabulary, grammar and writing. Table 4 below presents the mean scores for the pre-test, broken down into vocabulary and grammar scores. (We examine scores for improvement below):
Table 4: Mean Scores and Standard Deviations for the Pre-test.

	Pre-test experiment (N=26)		
	Vocabulary	**Grammar**	**Grammar + writing**
Mean	60.0	63.5	54.8
Standard deviation	16.9	18.2	16.5

The pre-test results indicate that participants had fair knowledge of vocabulary, indicating that the average learner could provide an English equivalent for 60 out of 100 words. Every Runyakitara speaker would like to improve his or her vocabulary knowledge.

With respect to grammar, we tested only whether the participants could identify a word as plural or singular. Knowledge of grammar and writing resulted in an average of 55.1,

where 25% (identifying singular vs. plural) would represent a chance level. We note in passing here that these low early scores indicate level of ability that would be low for native speakers, but not for "semi-speakers" who need to re-learner their first language. In this exercise, learners were required to specify the correct *number* of a word, that is, singular or plural and to provide an alternative form, meaning that spelling was also tested. The scores in Table 4 show that participants indeed had considerable knowledge of their language, even if they clearly do not have native-speaker levels of ability.

Grammar improvement. After the pre-test (manual exercise), learners were given the RU_CALL system to learn and complete exercises. Table 5 shows That performance clearly improved once learners used the system.

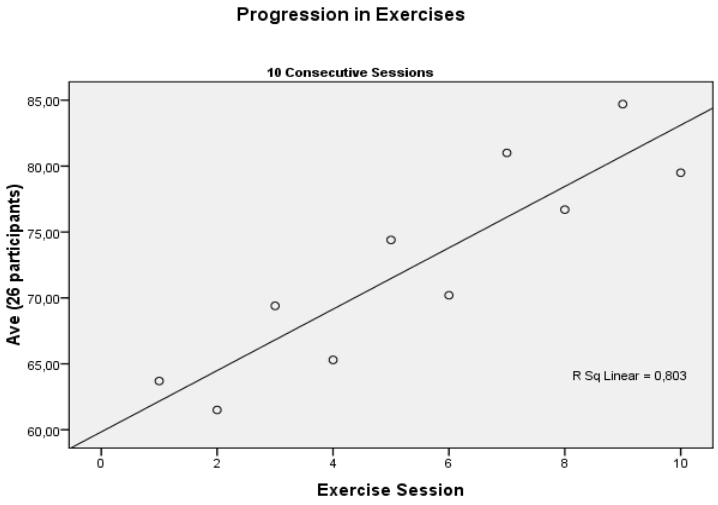
Table 5: Before and After Scores for Learners.

Variable	**Learners**	**Mean score**	**Standard deviation**
Pre-test	26	59.73	17.4
Post –test	26	74.61	9.17
t-value (paired differences)	**Degrees of freedom**		**Probability**
7.413	25		<.001

Table 5 indicates that there is a statistically significant difference between the mean grammar scores for the pre- and post-tests for the study participants ($t(25)=7.413$, $p < .001$). In other words, after using the software the participants had mastered nominal morphology better than they had in the pre-test. The digital learning environment appears to help in learning Runyakitara.

To confirm that students are indeed improving as they follow instruction, we conducted a regression analysis using the average session score as a dependent variable and the session number as an independent variable (the first session had the value 1, the second 2, etc.). This confirmed that we see a significant and steady learning effect ($r=0.89$, $p<0.001$). As the students used the system, their daily scores improved (See Figure 5).

Figure 5. Progression in Exercises.



The scatterplot also indicates that the average scores of students on even-numbered lessons (e.g., lesson 2) were consistently lower than those in the previous odd-numbered lesson (e.g., lesson 1). This happened because lesson 1, etc. focused on words, while lesson 2, etc. focused on phrases. The pattern indicates that words were easier to learn than phrases.

7.8.3 Learners' evaluation of RU_CALL

This section examines learners' views regarding the usefulness and of RU_CALL and its perceived advantages and disadvantages when compared with classroom learning. Results discussed in this sub-section are from the rating scale questions and the open-ended questions.

*Perceived RU_CALL usefulness***.** Learners rated RU_CALL on a Likkert scale of 1 (very useless) to 5 (very useful). Their ratings used only the categories 5 (very useful) and 4 (useful). Table 5 summarizes their responses:

Scale rate	**5**	**4**
Number of respondents	20	6
Percentage	77	23

Table 6. Usefulness of RU_CALL

The fact that none of the participants used the lower or even the middle section of the scale implies that RU_CALL was appreciated for its role in enhancing participants' grammar and spelling. The system's usefulness could also be seen in the comments learners made about using the software: *All* twenty-six learners indicated that they will continue using the software.

Unique aspects found in RU_CALL. The learners also found that they had understood the instruction and content provided by the digital learning environment for Runyakitara, and they remarked on how it was flexible in allowing them to revise their answers and to find correct answers. Some found the system good for documentation, and others indicated that it was convenient and enjoyable. Most indicated that the assessment part was unique and interesting to them because it was their first time to learn and get real time feedback.

7.9. Conclusion and pointers to future research

This study has presented a CALL system of Runyakitara, including a review of its design and implementation and an evaluation of its effectiveness. Our main objective has been to provide a digital learning environment that enables learners to enhance their grammatical mastery of this difficult language and to support the acquisition of writing skills. We applied both judgmental and empirical evaluation.

The results from the evaluation are positive. We confirmed that our targeted learners had basic, but limited knowledge of vocabulary and grammar in Runyakitara so that they needed to improve if they wished to function smoothly in Runyakitara.

The system also led to enhanced grammar abilities, which was the most important goal of the development effort. Learners improved regularly and substantially. The system

facilitated the learning of Runyakitara, the opportunity to use CALL software was motivational for the participants, most of whom admitted that their first interaction with the software (day 1) was a challenge, which motivated them to work hard to benefit from it. Some reported that they had been accustomed to consulting dictionaries, and others, native speakers in order to acquire information on the language.
With respect to the learners' subjective evaluation of software, results are quite satisfactory, with majority of learners reporting that they would like to continue using it.

Future practical steps should be to include other grammatical structures in the system, especially verbs and their tense, aspect, and topic morphology, which are essential to effective language use. Future directions to this research might be to include the morphological facilities in more natural exercises such as choosing the correct forms of words already embedded in texts.

Acknowledgments

We are grateful to Dr. Geoffrey Andogah, Gulu University, and his students for implementing the user interface to RU_CALL. We further gratefully acknowledge useful criticism and suggestions from Ad Backus, Rehema Baguma, Kees de Bot, Eve Clark, Sake Jager and Irina Zlotnikova.

Chapter 8

Toward a CALL system for Runyakitaran syntax

(A paper was extracted from this chapter and presented at the 7th International Conference of Computing and ICT Research, August 7-9, 2011 at Makerere University Kampala. The bibliographic information include: Fridah Katushemererwe & Arvi Hurskainen (2011). Intelligent Language Learning Model: Implementation on Runyakitara, in Kizza M. J., Lynch C. & Nath R. Special Topics in Computing & ICT Research: Strenghthening the role of ICT in development, Vol.7, Fountain Publishers, Kampala, Uganda. It is available at cit.mak.ac.ug/iccir/?p=iccir_11.)

Abstract

This chapter describes the design and implementation of an language learning system for some aspects of Runyakitaran syntax. The objective of this work is to demonstrate that the linguistic knowledge coded in the Runyakitara morphological analyzer provides a sound basis for CALL systems treating elementary aspects of syntax, in particular syntactic concord (grammatical agreement) and word order. The system makes use of a morphological parser, disambiguation and an extensive lexicon of Runyakitara. The strength of proposed system, which has been implemented in prototype fashion, is that the learner's use of vocabulary is not restricted to predefined simulations simplified for the sake of learning. The model builds on the ideas of the independent language learning approach proposed by Hurskainen (2009) for the learning of complex language structures.

Key words: Language Learning, (Intelligent) Computer-Assisted Language Learning (ICALL), Runyakitara, Grammatical Agreement, Word Order

8.1. Introduction

In the chapters above we have emphasized the importance and difficulty of Runyakitaran morphology and the need for software implementing its analysis in CALL applications. We have also remarked that computational morphology provides a useful base from which to launch more sophisticated learning facilities. In this chapter we wish to demonstrate more concretely what sorts of more sophisticated facilities come within reach, given the availability of a computational morphology. We shall not attempt to provide this in a user-friendly CALL application, as the focus is on the sorts of linguistic structure which can be checked automatically and therefore serve as the linguistic infrastructure to a CALL system.

The sorts of system we have in mind for support are often dubbed "Intelligent Computer Assisted Language Learning" (ICALL) systems, and they have been championed as useful tools in language instruction, helping learners to understand the forms and rules of

a language (Amaral & Meurers 2006). They are empowered by deep linguistic knowledge, such as the knowledge encoded in the computational morphology described in Chap. 2-4 of this thesis, and their benefit derives on the one hand from the additional practice outside the formal classroom that they are able to provide and on the other hand from their focus on grammatical forms. Given that computers have become more powerful, faster, easier to use, more convenient and cheaper, and that they can process, store and transfer much more data than ever before, the modern PC provides abundant possibilities for developing powerful language learning systems, which may even be applied to less documented and studied languages.

Educators recognize that utilizing computer technology and the language learning programmes developed for it can facilitate the creation of independent and collaborative learning environments, while providing students with specific language experiences that adapt to the needs of a student moving through the various stages of second language acquisition (Kung 2002).

Lai & Kritsonis (2006) discuss the advantages of computer technology in ICALL for second language acquisition, specifically pointing out that computer-based instruction modules can free language learners from classroom confinement, allowing them the freedom to learn wherever and whenever they want. Furthermore, the computerized language learning programmes make it possible for students to practise while undergoing a process of experimental learning. These learning tools motivate learners, enhance student achievement, increase authentic materials for study, and encourage greater interaction between students, teachers and peers. They also emphasize individual needs, provide independence from a single source of information and enable understanding and knowledge sharing from around the world.

The increasing globalization of life makes language learning more valuable and therefore creates a demand for more language learning systems, even for those languages that are not well documented.

Shaalan (2005) notes that by far the majority of language learning systems have been developed for English, followed by Japanese, French and German. Most African languages have not been part of the ICALL development. Shaalan therefore calls for more research that combines natural language processing techniques with language learning systems. Bantu languages have received too little attention with regard to CALL, and very little indeed that focus on structure.

This paper will describe an ICALL system tailored to facilitate the learning of Runyakitara language structures, specifically the concord structure and word order. Runyakitara is a name given to four closely related languages: Runyankore, Rukiga, Runyoro and Rutooro, with similar language structure and word order. We reviewed related research in Chap. 7, Sec.2 and refer the reader back to that section for relevant information.

We summarize here just that our brief overview of ICALL systems and our review of published literature in Chap.7 shows that there have been relatively few system which have exploited natural language processing (NLP) to improve CALL exercises (Amaral & Meurers 2006; Dickinson & Herring 2008) and also that there have been few applications of NLP technology for the development of ICALL system for the Bantu languages, which is the reason for the current study. Only one intelligent system for Bantu language learning has been recorded in a publication (Hurskainen 2009a), this regarding an application to Kiswahili. Hurskainen's (2009a) proposal has been partly employed in the Runyakitara model.

8.2 Highlights of Runyakitara grammar focused on here

As earlier noted, Runyakitara is composed of four languages; therefore, its grammar is somewhat complex. In this section, we will concentrate on word order on the one hand and on phrases involving grammatical agreement, or concord, on the other. These include nouns, possessive pronouns, demonstrative pronouns, adjectives and verbs. Concord patterns and word order are important in these phrases, as the class of the noun defines the concord pattern of the other constituents of the phrase.

a) Nouns and their classification system in Runyakitara

Just as all Bantu languages, Runyakitara has a noun class system. Researchers in Bantu languages agree that noun class features are determined by grammatical number, semantics, (that is, whether they are human, animal, vegetable, or inanimate) and, in some cases arbitrarily (Aikhenvald 2006; Katamba, 2003). Although Bantu languages have a general noun classification system, each language has its own unique set of sub-classifications. Ndoleriire & Oriikiriza (1990) determined that the Runyakitara noun classification system has twenty noun classes. This system was revised by Katushemererwe & Hanneforth (2010), who provide a detailed description accounting for the numbering system.

Important to this discussion is that nouns in Runyakitara are associated with initial vowels as pre-prefixes to the noun prefix. These are **a** (***a-ba-****ntu* 'people'), **e** (***e-ki-****tookye* 'banana') and **o (*o-mu-****ntu* 'person'). As discussed by Ndoleriire & Oriikiriza (1990), there are rules that govern the occurrence of the initial vowel, although it has other syntactic functions. If the noun class prefix has the vowel **a** (e.g. **ba**, **ma)**, the initial vowel will be **a,** thus, ***a-ma-****te* 'milk', ***a-ba-****kazi* (women). When the noun prefix has **i** or **-** , the initial vowel is **e,** for example ***e-ki-****tookye* 'banana', ***e-mi-****ti* 'trees' etc. The initial vowel is **o** when the noun class prefix has **u**, ***o-mu-****ntu* 'person', ***o-mu-****ti* 'tree'. At morphological level, the initial vowel does not have any other role other than to indicate the class of prefixes that it combines with. The initial vowel also plays a role at syntactic level. For example, when a noun is preceded by a preposition such as *omu* (in) *aha* (at), the initial vowel is dropped in phrase and syntactic operations e.g. *omu muti* 'in the tree', and not **omu omuti*. These facts shed some light on the manner in which the initial vowel

in nouns should be understood, an issue that learners will encounter in the learning process.

Although the nominal pre-prefix is rule-governed and has certain functions in syntactic structures, the morphological analyzer of Runyakitara on which we are basing the learning system interprets a pre-prefix and a prefix as one unit, called a noun prefix. Therefore, a noun like *abantu* is taken as **aba**[NPREF1/2] **ntu**[NROOT]. This means that *aba* is a noun prefix belonging to class 1-2 and **ntu** the noun root. This more simplified method of noun classification should be kept in mind when trying to recognize and understand concord patterns.

b) Concord patterns in Runyakitara

Concord patterns in Bantu languages have been extensively discussed in the pioneering work of Meeussen (1967), although he based most of his discussion on Kiswahili. According to Nurse and Phillipson (2003) noun class prefixes are at the heart of an extensive system of concord in these languages. The head noun takes a prefix marking its class, and other structurally-associated words obtain an appropriate matching prefix. And this is also a practice that is more or less followed in the Runyakitara.

Although there is no detailed description of concord patterns in Runyakitara, the syntactic description given in Taylor (1985) and insights from other Bantu literature (Hurskainen2009) can provide some assistance in understanding the concord patterns of this group. In Runyakitara, all the constituents of the noun phrase, such as adjectives, numerals, verbs and pronouns, are given a class prefix in accordance with the class of the noun. Specifically, the agreement is in one or all of the following:

i) The possessive pronoun prefix agrees with the noun prefix
*o-**mw**-ana **wa**-ngye* 'my child'
*a-**ba**-ana **ba**-ngye* 'my children'

ii) The adjective prefix agrees with the noun prefix
*o-**mw**-ana **mu**-kuru* 'a big child'
*a-**ba**-ana **ba**-kuru* 'old children'

iii) The subject prefix of a verb agrees with the noun prefix
*o-**mw**-ana **a**-rya* 'a child eats'
*a-**ba**-ana **ba**-rya* 'children eat'

iv) Noun, possessive pronoun and adjective prefixes all agree
*a-**ba**-ana **ba**-ngye a-**ba**-kuru* 'my big children'

v) Noun, possessive pronoun, adjective and verb prefixes all agree:
*a-**ba**-ana **ba**-ngye a**ba**-kuru **ba**-rya* 'my big children eat'

vi) Concord in a long sentence

*A-**ba**-ana **ba**-ngye **ba**-ahika aha **ba**-gambire **ba**-ije **ba**-ndeebe.*
'When my children arrive here, tell them to come and see me'.[13]

There is evidence that concord or agreement patterns in Bantu language are difficult for non-Bantu language speakers to learn, yet these patterns are essential for communicating in the languages. A Swahili learner observed: "One of the most difficult aspects of learning Swahili is its system of nouns…"[14]

Informal observations from (non-Bantu) Luo speakers in Uganda confirm that non-Bantu speakers find the grammatical agreement patterns difficult to learn. It is therefore worthwhile to develop a language learning model to help learners learn the grammatical agreement patterns in Runyakitara, which are difficult to learn.

c) Word-order in Runyakitara

There is extensive literature on word order in Bantu languages (Nurse & Phillipson 2003; Martenet al 2007 and Mchombo 2004). According to these authors, the dominant word order is SVO (Subject Verb Object), but there are also languages with SOV, VSO and OVS.

In Runyakitara, the unmarked word order has been reported as SVO (Morris & Kirwan 1972). Although Taylor (1985) does not give a specific general order for all words, his approach to word order is preferred because it deals with specific constituents of specific word classes. In our view, word order is flexible in Runyakitara, and this is mainly caused by the argument structure, emphasis and topicalization. For example, the words in the following simple sentence may change their order as follows:

i) *Omwegi yaashoma ekitabo* 'a student read a book' (SVO)
ii) *Omwegi ekitabo yaakishoma* 'a student indeed read a book'(SOV)
iii) *Yaakishoma ekitabo omwegi* 'a student *read* a book' (VOS)
iv) *Yaakishoma omwegi ekitabo* 'a student *read a book*' (VSO)

In the last two sentences we have tried to suggest the grammatical import of the alternative word words by suggesting that they might be translated using emphatic stress. The above illustration demonstrates the flexibility in the word order of Runyakitara with respect to major constituents, subjects, verbs and objects. In this study, we will follow the guidelines provided in Taylor (1985) regarding the word order of these specific constituents. Our notion of word order also covers the combinations of constituents in noun phrases. However, we exclude noun phrases where the noun is not the first member of the phrase. Examples of the relevant phrases which we *do* cover are:

– Noun and demonstrative pronoun – *omuntu ogu* 'this person'.

---

[13] For details about all concord markers of some word categories with all noun classes of Runyakitara, refer to Appendix A.

[14] www.transparent.com/learn-swahili/overview.html

- Noun and a possessive pronoun – *omuntu wangye* ‘my person’.
- Noun and adjective – *omuntu murungi* ‘a good person’.
- Noun and verb – *omuntu areeba* ‘a person sees’.
- Noun, possessive pronoun and demonstrative pronoun – *omuntu wangye ogu* ‘this person of mine’.
- Noun, possessive pronoun, demonstrative pronoun, and adjective – *omuntu wangye ogu omurungi.* ‘this good person of mine’
- Noun, possessive pronoun, demonstrative pronoun, adjective and verb – *omuntu wangye ogu omurungi areeba.* ‘this good person of mine sees’

Note: The order of the above constructions can also change, such as the phrase in (i) above, which may be *ogu muntu* or *owangye omuntu*. Such changes are usually context-related, especially in response to pragmatic issues. In this study, we treat the basic order as indicated above.

8.3. Design of the learning system

Our aim in designing the learning system is to help the learner (a) to identify spelling errors, (b) to formulate the correct word order in phrases and (c) to ensure that the grammatical concord in phrases is correct. These three items should function globally, so that any words of a given word class could be used for instruction.

In addition to the non-guided learning, we have also implemented a series of so-called guided tours, where in each learning phase the learner is shown how to continue. These tours were implemented as a means to help students learn the concord patterns of each noun class. Below is a diagram of the design model that was implemented:

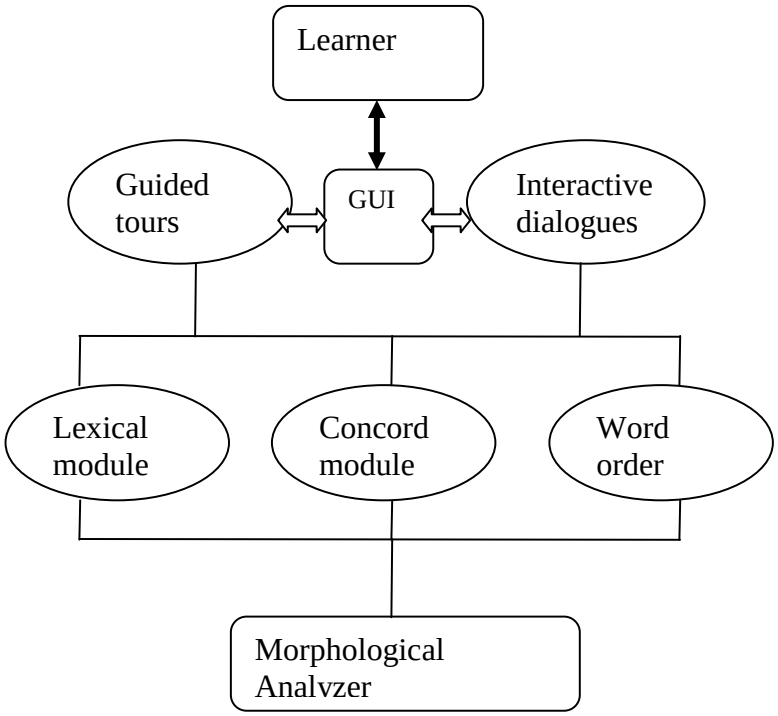


Fig. 1: Architecture of Runyakitara ICALL system

8.3.1 Morphological Analyzer

The basic tool in the learning system is a morphological analyzer of Runyakitara. The analyzer for Runyakiara was developed using finite state automata (Hanneforth 2009). As we hope to demonstrate, the morphological analyzer also provides information about the grammatical categories used in the affixes in a word, and these are the basis component of syntactic agreement systems. The morphological analyzer also provides information on the syntactic category of the entire, moprhologically complex word, and this determines the order of the words in phrases and sentences.

The morphological analyzer was tested at various levels of development, and it currently analyzes a newspaper corpus at 78% recall and 72% precision. Some results regarding the Runyakitara morphological system can be found in Katushemererwe & Hanneforth (2010). Normally, morphological analyzers are designed to include all linguistically significant information about the word. In various applications, however, only part of this information is needed. In our case, we removed all the tags not required either to program the learning system or to provide useful information for the learner. In addition to removing certain tags, we also abbreviated some long annotations to make the output shorter. All this was done using a reformatting tool. The reformatted output of the Runyakitara morphological analyzer is illustrated below:

amaisho : ama[N-5/6P] isho [eyes]
amakye : ama[ADJ-6] kye[small/few]
rireeba : ri[V-5][VERB-PREF] reeba[IND][it sees]

All word categories were described in the morphological analyzer but, for purposes of the learning system, only the following were considered:

- Nouns – classes 1-18
- Possessive pronouns
- Demonstrative pronouns
- Adjectives
- Verbs

There are at least two methods of exploiting the morphological analyzer in the learning system. In one method, the analyzer is an integral part of the runtime learning system, so that each time the learner enters a text string, the system performs a morphological analysis of that string. In another method, which is also the method that we have applied here, a list of the word forms is analyzed and stored off-line with the relevant annotations about the grammatical agreement information in its component affixes and the syntactic category of the complex word forms. This file is then used as a basis for constructing the learning system. The first type of system is more comprehensive but, at the same time, prone to functional errors. In the latter approach, it is necessary to restrict the number of verb forms for practical reasons, as they may run into the millions.

A more modest number of verb forms is sufficient for a learning system, provided that the extracted list contains the most commonly occurring verb forms, together with their analyses. We compiled such a list by extracting all the verb forms from a newspaper corpus, analyzing them and including them in a list of analyzed word-forms.

To make use of this morphologically analyzed word list in the interactive learning system, we constructed a pattern matching system that enriches the keyed-in string with analysis tags. When the learner enters words from the keyboard, they are plain words without analysis. These words are matched with similar words in the morphologically analyzed lexicon. For example, *omuntu* is matched with *omuntu : omu[N_1/2]ntu {person}.*

8.3.2 Dealing with ambiguity

There are several types of ambiguity in the analysis. The most important types to be solved in the learning system are noun class ambiguity and part-of-speech ambiguity. Examples of noun class ambiguity are illustrated in (1).

(1)
yangye : ya[POSS_PRON_4]ngye ‘mine’ (this is possessive pronoun for class 4)
yangye : ya[POSS_PRON_9]ngye ‘mine’ (possessive pronoun for class 9)

egi : egi[DEM_PRON_4] ‘this’ (Demonstrative pronoun for class 4)

egi : egi[DEM_PRON_9] ‘this’ (Demonstrative pronoun for class 9)

nungi : mi[ADJ_4]rungi ‘good’ (Adjective in class 4)
nungi : n[ADJ_9]rungi ‘good’ (Adjective in class 9)

These are forms of the classes 4 and 9 which are ambiguous. The ambiguity here concerns two different classes with a similar output. In class 4 *yangye*, the speaker is referring to trees or plants, while in class 9 *yangye*, s/he is referring to animals. Therefore, these forms can be combined with nouns of the classes 4 and 9, as in (2).

(2)
emiti yangye
emi [N_3/4P]ti ya[POSS_PRON_4]ngye
my trees

ente yangye
en[N_9/10S]te ya[POSS_PRON_9]ngye :
my cow

When such ambiguous word-forms are entered into the learning system, we get both interpretations, as in (3).

(3)
emiti yangye
emi[N_3/4P]ti ya[POSS_PRON_4]ngye
my trees

ente yangye
en[N_9/10S]te ya[POSS_PRON_9]ngye
my cow

Note that ambiguous readings of each word-form belong to the same part-of-speech category. There are two ways of dealing with the problem. Either we disambiguate the output and select the correct interpretation (or alternatively remove the wrong one), or we under-specify the interpretation of class-ambiguous readings. In our system, we have chosen the latter alternative. When using under-specified marking, we get the result as in (4).

(4)
emiti yangye
emi[N_3/4P]ti ya[POSS_PRON_4/9]ngye

ente yangye
en[N_9/10S]te ya[POSS_PRON_4/9]ngye

Note that *yangye* has been described as POSS_PRON_4 and not POSS_PRON_9. This kind of output makes it possible to write concord rules without selecting or deleting any output.

Another type of ambiguity arises in the part-of-speech category. In this case, the word-form belongs to two or more word classes. Examples are in (5).

(5)
amahango : ama[N_6]hango{special branches}
amahango : ama[ADJ_6]hango{big ones}

ebyago : ebi[N_8]ago{spirit}
ebyago : ebi[POSS_PRON_8]ago{theirs}

mwenda : mu[N_3]enda{cloth}
mwenda : mu[NUM_3]enda{nine}

kiniga : ki[N_7]niga{anger}
kiniga : ki[V_7]niga{it strangles}

To handle these ambiguities, we have chosen to use disambiguation rules instead of under-specification. Because the learning system is restricted to noun phrases, it is possible to use the correct word order as a criterion for disambiguation rules. Examples of such rule types include:

- *Remove the noun reading if a non-ambiguous noun is on the left.*
- *Remove the verb reading, if it is not the last member of the phrase.*
- *Remove the noun reading, if it is not the first member of the phrase, with the exception of the demonstrative pronoun that can be before the noun.*
- *Remove POSS_PRON reading, if followed by a non-ambiguous POSS_PRON.*

8.3.3. Detection of spelling errors

The flagging of spelling errors has been implemented in such as way that, if a word has no output, it is considered wrongly spelled. Any correctly spelled words that are not in the system are thus treated as spelling errors. Examples of words detected as misspelled are provided in (6). We recognize that this is not full fledged spelling correction, but we suggest that users may benefit from having access to.

[6]
ontu??

ente en[N_9/10SP]te zaany??

omuntu ogu
omu[N_1/2S]ntu wange?? ogu[DEM_PRON_1/3]
N+DEM_WO

Please recall that we are illustrating the sort of information we are positioned to provide to language learners. We would not propose to mark spelling errors tersely with double question marks, but it is convenient in this exposition to keep the annotations brief.

8.3.4. Correction of word order

The morphological analyzer also identifies the parts of speech or basic syntactic categories of the words it analyzes, identifying words e.g. as nouns, adjectives, verbs, prepositions, determiners, etc. It thus identifies the fundamental categories in terms of which word order constraints are formulated. This enables us to check the word order in learner input in fairly simple fashion, and we illustrate how this is done in the present section.

As noted above, word order in Runyakitara is quite flexible, making it difficult to construct a comprehensive system of word order rules. There is a set of core rules that cannot be violated. In addition, there are several cases, where word order depends on stress and other prosodic features. A large variety of acceptable word orders can be implemented in the learning system, while learning priorities should determine which usages should be learned first and which ones at later stages of learning.

We adopted the basic word order that should be followed in normal language use. For example, a modifier of the noun, such as an adjective, possessive pronoun, demonstrative pronoun and numeral, follows the noun in the noun phrase. If more than one modifier is attached to the noun, these modifiers follow the noun in a certain sequence, as shown in (7).

(7)
Noun+Poss-Pron+Dem-Pron+Adj+Verb

omuntu	wangye	ogu	murungi
omu[N_1/2S]ntu	wa[POSS_PRON_1/3]angye	ogu[DEM_PRON_1/3]	mu[ADJ_1/3]rungi

ashoma
a[V_1][VERB_PREF_PR]shoma[IND]

N+POSS+DEM+A+V_WO
This good person of mine reads

An alternative sequence is in (8):

(8)
Noun+Poss-Pron+Adj+Dem-Pron+Verb

omuntu	wangye	murungi	ogu
omu[N_1/2S]ntu	wa[POSS_PRON_1/3]angye	mu[ADJ_1/3]rungi	

ogu[DEM_PRON_1/3]
ashoma
a[V_1][VERB_PREF_PR]shoma[IND]**_WO**
This good person of mine reads

The tag '_WO' is added to the last element of annotated glosses to indicate that the word order is correct. As examples (7) and (8) illustrate in the line of annotated glosses, our morphological analyzer identifies the parts of speech of the input words. Using this information we can check the sequence of parts of speech against a set of templates we have coded to then give the learner feedback about word order. We emphasize that we are technically in a position to do this only due to the morphological analyzer. Incidentally, the next to last word is glossed as a demonstrative pronoun, but it might also function as a preposition, while the other constituents would not be ambiguous with respect to syntax in normal language use.

The rules for checking word order were implemented in two phases. In the first phase, only the correctness of the word order is checked. If the word order is correct, the rule issues the '_WO' tag for correct word order, as in (9) (see the string of annotations after the last word).

(9)
amaisho gangye aga amakye gareeba
ama[N_5/6P]isho ga[POSS_PRON_6]angye aga[DEM_PRON_6] ama[ADJ_6]kye ga[V_6][VERB_PREF_PR]
reeba[IND] **N+POSS+DEM+A+V_WO**

If the word order is wrong, the output is indicated by the '_WO!' tag shown in (10).

(10)
amaisho aga gangye
ama[N_5/6P]isho aga[DEM_PRON_6] ga[POSS_PRON_6]angye **DEM+POSS_!WO**

Note that the module for checking word order produces the word order tag, as shown in examples (8) and (9) above. If the word order is correct, the tag ends in '_WO'. If the word order is wrong, the tag ends in '_!WO', as in (10). On the basis of these word order tags, it is then possible to provide appropriate feedback to the learner. This is the second phase in the word-order check: correction. Examples are in (11).

(11)

Word order is correct!
amaisho gangye aga amakye gareeba
ama[N_5/6P]isho ga[POSS_PRON_6]angye aga[DEM_PRON_6] ama[ADJ_6]kye ga[V_6][VERB_PREF_PR]
reeba[IND] **N+POSS+DEM+A+V_WO**

When you input "amaisho aga gangye" the feedback is:

Demonstrative pronoun cannot be before a possessive pronoun!
amaisho aga gangye
ama[N_5/6P]isho aga[DEM_PRON_6] ga[POSS_PRON_6]angye **DEM+POSS_!WO**

We emphasize that the exact wording of the feedback to the learner is not an issue for us in (11). Our point is the morphological system is informationally rich enough to support feedback on word order, i.e. on a syntactic topic.

We note, too, that we have not evaluated how effectively we can test for correct word order. The examples illustrated above are genuine, and they are analyzed in the system implemented exactly as shown, but we would have to evaluate the system a lot more before being confident that it would support real pedagogical use. We conjecture that an evaluation would show that the system is insufficiently exact and comprehensive for use in a grammar checker, but that it nonetheless would suffice for careful use in a CALL program.

8.3.5. Correction of concord (Concord Module)

In languages with noun classes such as Runyakitara, learning the correct concord patterns for all the noun classes requires a lot of practice. With the help of grammar books, it is possible, but troublesome, to identify grammatical patterns and to attempt to internalize them through practice. But unfortunately, Runyakitara does not have any grammar books specifically concerned with concord patterns. A learning program that identifies errors and provides appropriate feedback would therefore be useful.

The morphological analyzer identifies the noun class of the various affixes in nouns, verbs, adjectives and determiners that must agree with one another for a noun phrase to be syntactically well formed. This enables us to check automatically whether noun phrases (and simple sentences with only subjects) are well formed with respect to classifier agreement. In order to demonstrate the usefulness of the morphological analyzer, we implemented a system for checking concord patterns in two phases. In the first phase, the concord of each constituent is checked, and if all constituents have a common correct concord tag, the system outputs a corresponding summary tag. For example, if the structure has five constituents and each constituent has the correct concord, the output is CONC5. This is demonstrated in (12).

(12)
Inputting "omuntu wangye ogu murungi ashoma" brings feedback as:
Word order and concord are correct!

omuntu	wangye	ogu	murungi	
omu[N_1/2S]ntu	wa[POSS_PRON_1/3]angye	ogu[DEM_PRON_1/3]	mu[ADJ_1/3]rungi	
	ashoma			
	a[V_1][VERB_PREF_PR]shoma[IND]	N+POSS+DEM+A+V_WO	**CONC5**	

On the other hand, if the concord is wrong, the learner is warned about it. Reporting on the mistakes can be implemented in various ways. One method is to give the same warning message for all types of mistakes. An example is in (13). If no concord tag is produced, the concord is wrong.

(13)
Inputting “omuntu zangye ogu murungi ashoma” outputs:
Word order is correct but concord is not!

omuntu zangye ogu murungi
omu[N_1/2S]ntu zi[POSS_PRON_4/9]angye ogu[DEM_PRON_1/3] mu[ADJ_1/3]rungi
ashoma
a[V_1][VERB_PREF_PR]shoma[IND] N+POSS+DEM+A+V_WO

In addition to this simple warning system, we also implemented a system that gives more detailed information and shows the words where mistakes lie. First, for each word with wrong concord, a tag indicating wrong concord pattern is produced. Consider the examples in (14).

(14)
omuntu bangye
omu[N_1/2S]ntu ba[POSS_PRON_2]angye N+POSS_WO **CONC_!POSS**

omuntu bangye aba
omu[N_1/2S]ntu ba[POSS_PRON_2]angye aba[DEM_PRON_2] N+POSS+DEM_WO **CONC_!POSS CONC_!DEM**

omuntu bangye aba murungi
omu[N_1/2S]ntu ba[POSS_PRON_2]angye aba[DEM_PRON_2] mu[ADJ_1/3]rungi bashoma
ba[V_2][VERB_PREF_PR]shoma[IND] N+POSS+DEM+A+V_WO **CONC_!POSS CONC_!DEM CONC_!VERB**

We see that for each word that does not agree with the noun, a tag indicating a mistake is produced. In the first example (14) the pronoun is marked as in construction with a noun in class ‘2’, while the noun itself is class ‘1’. This mistake is repeated in the second sentence, which additionally includes a possessive marker and a verb which are also incorrectly marked for noun class.

On the basis of these tags and their combinations, it is then possible to give appropriate feedback to the learner. The examples in (14) are reproduced in (15) with appropriate sorts of warning messages.

(15)
omuntu bangye.
Concord of possessive pronoun is incorrect!
omuntu bangye
omu[N_1/2S]ntu ba[POSS_PRON_2]angye N+POSS_WO **CONC_!POSS**

omuntu bangye aba.
Concord of possessive pronoun and demonstrative pronoun is not correct!
omuntu bangye aba
omu[N_1/2S]ntu ba[POSS_PRON_2]angye aba[DEM_PRON_2] N+POSS+DEM_WO **CONC_!POSS CONC_!DEM**

omuntu bangye aba murungi bashoma.
Concord of possessive pronoun, demonstrative pronoun and verb is not correct!
omuntu bangye aba murungi
omu[N_1/2S]ntu ba[POSS_PRON_2]angye aba[DEM_PRON_2] mu[ADJ_1/3]rungi
bashoma
ba[V_2][VERB_PREF_PR]shoma[IND] N+POSS+DEM+A+V_WO **CONC_!POSS CONC_!DEM CONC_!VERB**

We should wish to emphasize that we do not propose that the exact wording of the feedback to the learner is optimal in the examples in (15), only that the system is capable of supporting such feedback automatically. An instructor might wish to tailor the feedback to his or her own manner of teaching, and we would be in favor of supporting that sort of flexibility.

A second important qualification is to note that we have not rigorously tested how effectively we can test for correct concord. The examples illustrated above work in the system implemented, but we should need to evaluate the coverage of the system thoroughly before recommending that it be adopted for pedagogical use. We suspect that a strict evaluation would again show that the system fails to check for correctness to the degree that one would wish to have in a grammar checker, but that it nonetheless would suffice for careful use in a CALL program.

Our major programmatic point has been stated above, namely that the morphological analyzer is useful not only in CALL programs for learners of Runyakitara morphology, that it may also play a useful role in CALL programs for learners of Runyakitara syntax. We hope to have shown this with respect to word order and classifier agreement.

8.4. Learning applications

A learning system based on morphological analysis makes it possible to develop several kinds of learning applications. The tags included in the resulting analyses range from low-level tags (e.g. word lemma) to high-level tags (e.g. part-of-speech). These tags enable the developer to construct a whole range of learning applications. We shall take the liberty here of suggesting a two sorts of exercises that we should be in a position to support. We do not claim to demonstrate that these exercises are superior, but we are at times impatient about the sorts of exercises we see implemented in CALL systems, which all too often resemble the pencil and paper exercises we know from language learning textbooks before the CALL era. To simplify matters for the sake of this paper, we will describe here only two types of applications, free interactive dialogues and guided tours.

8.4.1. Interactive dialogues

The system makes it possible to use the whole lexicon of the language being learned. Any words in any of the language's part-of-speech categories can be used in the instruction programme, provided that the words are listed in the lexicon. The system detects three kinds of mistakes, as described above: spelling errors, word order errors and concord

errors. An error in any of these categories prompts a relevant feedback message. In case of word order errors, detailed guidance is given, showing the place where the mistake occurs. In case of concord errors, the provision of detailed feedback is more difficult, mostly because of the ambiguities in the analyzed results. If the noun reading is non-ambiguous or if it can be disambiguated, the decision is easier, because the noun determines the noun class for the agreement pattern. In the current learning system, the ambiguity of nouns can normally be resolved. Therefore, detailed feedback can be also provided for concord errors.

8.4.2. Guided tours

Although a learning system with almost unlimited vocabulary may seem desirable, the learner also needs some guidance, so that relevant topics will be covered. Such guidance can be organized at various levels. At the most elementary level, the student is coached in the construction of noun phrases for each noun class. This coaching can be sufficiently detailed, requiring the student to add a given word after each step. Such guided learning exercises can be expanded so that, instead of specific words, the student is asked to use a word from a given selection. By using various vocabulary groups, the student will learn the vocabulary, and at the same time receive practice and guidance in constructing noun phrases. The overall learning process is schematically illustrated below:

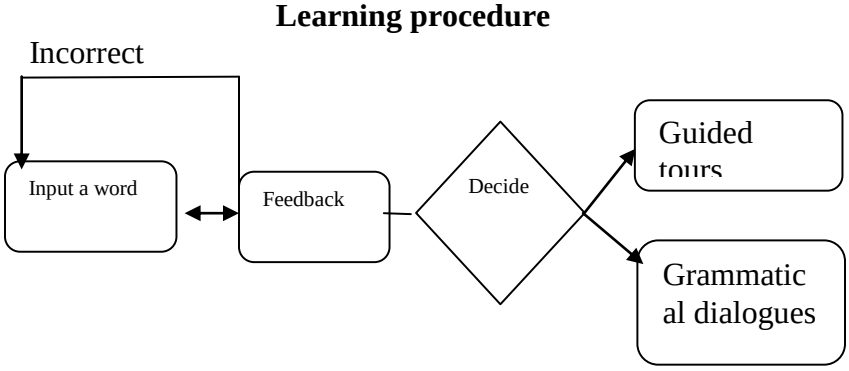


Fig. 2: The learning procedure of Runyakitara ICALL

The learning sequence on the system is as follows: the learner first inputs one of the words from those that s/he knows as nouns, pronouns, adjectives and verbs. The learner then receives feedback indicating whether a word is correct or incorrect. The learner is then asked to decide whether to undergo learning by means of a guided tour or grammatical dialogues. In either case, the learner is provided with feedback on the correctness of the concord patterns, word order and spelling.

In order to illustrate these ideas more effectively, we turn to some illustrations that will perhaps be enlightening through their concreteness.

8.4.3. Learning by means of interactive grammatical dialogue

1. When the learner inputs a **word**, the system checks it to determine if the word is correct, and then issues a report and an invitation to practice some more as follows:

   $ echo omukazi.
   This is a correct word in Runyakitara.
   What would you like to learn?
   If you wish to have guided practice of various noun classes, type one of the following:
   omuntu, abantu, omuti, emiti, eihamba, amaisho, ekimuri, ebimuri, ente, orutaro, entaro, akabaaho, obubaaho, oturo, oburo, okuguro.

   Each of these words is an example of a certain noun class.
   If you wish to use your own vocabulary, please go ahead.

   In this case, you will receive reports on the correctness of three criteria: spelling, word order and concord.
   omukazi omu[N_1/2S]kazi N

2. **Noun and adjective**: If the learner chooses the interactive dialogue, s/he is asked to start inputting phrases of his/her choice. The following is an example of feedback for concord and word order after a noun and an adjective have been entered:

   ```
   $ echo omukazi murungi.
   Word order and concord are correct!
   omukazi murungi omu[N_1/2S]kazi mu[ADJ_1/3]rungi N+ADJ_WO CONC2
   ```

Note that, in the text below, the feedback contains phrases with their tags, 'N' for noun and ' POSS_PRON_' for possessive pronoun: If the noun and possessive pronoun are correct, the same feedback is reported.

$ echo omukazi wangye. | rlearn
Word order and concord are correct!
omukazi wangye omu[N_1/2S]kazi wa[POSS_PRON_1/3]angye N+POSS_WO CONC2

3. **Noun and demonstrative pronoun:**

$ echo omukazi ogu.
Word order and concord are correct!
omukazi ogu
omu[N_1/2S]kazi ogu[DEM_PRON_1/3] N+DEM_WO CONC2

4. **Noun, possessive pronoun and demonstrative pronoun:**

$ echo omukazi wangye ogu.
Word order and concord are correct!
omukazi wangye ogu
omu[N_1/2S]kazi wa[POSS_PRON_1/3]angye ogu[DEM_PRON_1/3] N+POSS+DEM_WO CONC3

5. Noun, possessive pronoun, demonstrative pronoun and adjective:

```
$ echo omukazi wangye ogu murungi.
Word order and concord are correct!
omukazi          wangye                    ogu                    murungi
omu[N_1/2S]kazi wa[POSS_PRON_1/3]angye ogu[DEM_PRON_1/3]
          mu[ADJ_1/3]rungi N+POSS+DEM+A_WO CONC4
```

6. A wrong verb *shoma* instead of *ashoma* is included:

```
$ echo omukazi wangye ogu murungi shoma.
Please check spelling!
omukazi          wangye                    ogu                    murungi
omu[N_1/2S]kazi wa[POSS_PRON_1/3]angye ogu[DEM_PRON_1/3]
          mu[ADJ_1/3]rungi
          shoma??
```

7. Noun, possessive pronoun, demonstrative pronoun, adjective and verb:

```
$ echo omukazi wangye ogu murungi ashoma.
Word order and concord are correct!
omukazi          wangye                    ogu                    murungi
omu[N_1/2S]kazi wa[POSS_PRON_1/3]angye ogu[DEM_PRON_1/3]
          mu[ADJ_1/3]rungi
          ashoma
          a[V_1][VERB_PREF_PR]shoma[IND] N+POSS+DEM+A+V_WO CONC5
```

8. Wrong input:

```
$ echo wangye omukazi ogu murungi ashoma.
Possessive pronoun cannot be before a noun except in certain contexts!
wangye                   omukazi          ogu                    murungi
wa[POSS_PRON_1/3]angye omu[N_1/2S]kazi ogu[DEM_PRON_1/3]
          mu[ADJ_1/3]rungi
          ashoma
          a[V_1][VERB_PREF_PR]shoma[IND] N+DEM+A+V_WO POSS+N_!WO
```

The above input (8) automatically invokes appropriate grammatical feedback from which the learner is supposed to learn that, under normal circumstances, a possessive pronoun does not precede a noun.

The grammatical dialogue exercises occur on an individual basis, where the learner interacts with the system and learns from the feedback. We present this here, not as a polished CALL system, but rather as a prototype intended to prove the feasibility of supporting syntactic CALL exercises on the basis of the morphological expertise elaborated on in previous chapters.

8.4.4. Learning from guided tours

The learner is asked to input a word as instructed by the system, and then to follow the guidance that is given to him/her on the screen. The following is an example of a learner in a guided tour:

d) $ echo omuti.
OK. Add to this the possessive pronoun 'angye'! Please add a full pronoun.
omuti
omu[N_3/4S]ti N_INIT_EXE_3_4

e) $ echo omuti gwangye.
OK. Add to this string the demonstrative pronoun 'ogu'!
omuti gwangye
omu[N_3/4S]ti gu[POSS_PRON_1/3]angye N+POSS_WO CONC2

f) $ echo omuti gwangye ogu.
OK. Add to this the adjective 'hango'!
omuti gwangye ogu
omu[N_3/4S]ti gu[POSS_PRON_1/3]angye ogu[DEM_PRON_1/3]
N+POSS+DEM_WO CONC3

g) $ echo omuti gwangye ogu muhanga.
Please check spelling!
omuti gwangye ogu
omu[N_3/4S]ti gu[POSS_PRON_1/3]angye ogu[DEM_PRON_1/3] muhanga??

h) $ echo omuti gwangye ogu muhango.
OK. Add to this the verb 'kura'!
omuti gwangye ogu muhango
omu[N_3/4S]ti gu[POSS_PRON_1/3]angye ogu[DEM_PRON_1/3]
mu[ADJ_1/3]hango N+POSS+DEM+A_WO CONC4

i) $ echo omuti gwangye ogu muhango gukura.
OK. This is a full sentence with a long noun phrase. Now do the same exercise using plural forms.
Continue by typing 'emiti'!
omuti gwangye ogu muhango
omu[N_3/4S]ti gu[POSS_PRON_1/3]angye ogu[DEM_PRON_1/3]
mu[ADJ_1/3]hango gukura
gu[V_3][VERB_PREF_PR]kura[IND] N+POSS+DEM+A+V_WO CONC5

In step (d) of the guided learning, the learner made an error in typing **muhanga* instead of *muhango*. The system prompted the learner to check the spelling by putting two question marks on either side of the incorrect word. This alerts the learner about the need to check the word and correct the spelling, as shown in step (e).

8.5. Conclusion

We have shown how the morphological analyzer can provide a great deal of learning input for a language learning system, indicating that it may function as a source of important information in the morpho-syntactic learning of the Bantu languages. We

noted first that the morphological analyzer is in an excellent position to detect spelling errors, and second that the same analyzer, with the addition of some simple pattern matching, errors, can also detect errors in word order and in agreement.

We noted that the agreement structures (also known in Bantu languages as concord patterns) are difficult to learn, as they may involve dependencies in several words simultaneously. We then illustrated the sorts of CALL applications which the morphological analyzer together with some pattern matching would be in a position to support. We offered these implemented prototypes not as proof of pedagogical effectiveness, but rather as proof of technical feasibility.

We exploited the opportunity given by illustrating CALL exercises to suggest some alternatives to the usual exercises, where vocabulary is limited to the terminology that a developer installed in a system. In our design, the learner can freely choose the vocabulary to use in learning based on his/her previous knowledge. We also intend that this sort of exercise might play the usual CALL role of complementing classroom learning, allowing learners to strengthen their knowledge by working with an interactive digital learning environment.

Guided tours in particular provide a means of learning that is an alternative to learning techniques usually implemented in exercise-based systems. The method used in developing guided tours is so clear and adaptable that many guided tours might be developed and modified further for many purposes.

8.6. Future work

Technically, future work may aim to cover the concord patterns of object constructions as well as relative constructions. As noted in the previous chapters, Runyakitara may have one or more objects in sentence construction. We intend to incorporate double object constructions which are not yet included in the discussed prototype. We also intend to develop phrases where a noun is not a first member of the noun phrase.

We intend to develop a user-friendly learning environment that not only aids enthusiastic learners, but also tries to motivate learners by enhancing the usability factors.

Chapter 9

Summary, conclusion and directions for future research

9.1. Conspectus

Computational morphology plays an important role in contemporary computer-assisted language learning (CALL), particularly in vocabulary extension, learning of morphology, dictionary access and the enhancement of reading skills. With the emergence of low-cost IT capacity, morphological analyzers can be utilized for a wide range of language learning applications. However, the languages reported on in the research literature are commonly taught, widely spoken and well documented languages – mostly English, French, German, Spanish and Japanese. Little work has been done on less commonly taught and poorly documented languages, even when they present technical challenges due to their complex morphologies. Runyakitara is such a language group and is the focus of the present work.

In this dissertation, we have designed and implemented a morphological analyzer for the Bantu languages in the Runyakitara group. We then put the analyzer to use in supporting language learning in a novel way, namely via exercises in word inflection. Earlier studies had used morphological analysis software to provide information to students learning to read foreign languages, to automate dictionary access for them and to find examples of words (of potentially different morphological form) in large collections of text. We also note some earlier work on supporting exercises using natural language processing (NLP), but the little we found (Amaral & Meurers 2006; Dickinson & Herring 2008) focused on the commonly taught languages noted above. In contrast to almost all the work reported in previous studies, we utilized the morphological analyzer to develop exercises for learning. To achieve our goal, we applied our morphological analyzer of Runyakitara to two language learning applications. We describe both the morphological analyzer and the language learning software in more detail in this thesis.

In preparing the development of the Runyakitara morphological analyzer, our analysis revealed that Runyakitara had no systematic and up-to-date morphological description or collection of material that one might use to develop and evaluate a morphological analyzer. We therefore turned to descriptions of other Bantu languages, existing studies of special topics in Runyakitara (including studies in manuscript form) and our own intuition. We formalized, designed, implemented and evaluated the first morphological analyzer of Runyakitara. We note in passing that this also meant that the CALL application we developed could not fairly be compared to sets of pencil and paper exercises from published language course. Such courses simply do not exist.
The Runyakitara morphological analyzer is presented, discussed and evaluated in Chapters 2, 3 and 4 of this dissertation. The Runyakitara morphological analyzer is now functional and can be used for a variety of purposes. The results were adequate for use in

language-learning software, although the system's recall (degree of coverage) ought to be improved for more ambitious (more advanced) courses. The software was designed and implemented as a proof of concept that NLP could serve CALL in more effective ways than had been demonstrated to date.
Although morphological analyzers have been utilized in CALL with some success, they involved different languages, different uses to which the morphological analysis was put, and entirely outside the Ugandan context. So this thesis broke completely new ground with respect to the language we supported in CALL and it is one of few pioneers in demonstrating the utility of NLP techniques in supporting the development of CALL exercises.

We also conducted a small-scale survey and pilot to further our understanding of the situation of language teaching and learning in Uganda. This study was mainly carried out to establish the need and constraints of CALL in a Ugandan context. Chapters 5 and 6 report on this research. The most notable result was that participants were sharply divided with respect to their interest in using CALL software for learning local languages in Uganda, with busy professionals showing no interest, and educators showing a great deal. This sharpened our focus in development, and confirmed Jager's (2009) thesis that language education professionals are the key stakeholders in determining the acceptance of CALL. We noted in addition that it would be unwise to rely on the Internet to deliver course material, as all the participants in Western Uganda noted that it was very unreliable. The latter point meant that we focused on developing an application that might stand by itself and did not rely on the Internet. It also turned out that none of our participants had any experience with CALL, although some of them used computers on a daily basis and were language teachers. The educators were also all eager to experiment with CALL as soon as it became available.

In examining the course participants at a course on Runyakitara given at Makerere University, we identified a special group interested in (re-)learning Runyakitara. These Ugandans have Runyakitara speaking parents but had moved outside of the Runyakitara area (to Kampala, where Luganda is spoken locally). The parents never used their native language to speak to their children, but the children had acquired some passive abilities from overhearing their parents and other (extended) family members speak. As some literature research confirmed, the children of emigrants often fail to learn their parents' language (Dorian, 1977; Fishman 1991, 2000; Ohiri-Aniche, 1997; Landweer, 2000). But importantly, the students in the Makerere course were motivated to learn the language in order to maintain family ties and, in some cases, to seek work in Western Uganda where Runyakitara is spoken. Based on these observations, we developed a language learning application and evaluated it empirically. Detailed results of the study are discussed in chapter 7 of this dissertation, scoring little better than chance in grammar. We note here that the participants did poorly at the beginning of the course, confirming their "semi-speaker" status, and that there were substantial gains in language ability as a result of using the software, confirming its effectiveness. Given the total absence of competing material for learning Runyakitara, we claim that the CALL software developed has proven its potential. This implies that, given the right CALL software, the situation of the Runyakitara languages in Uganda can be improved.

To demonstrate the further potential of the morphological analyzer to generate learning content, we designed and implemented a second application to support the learning of Runyakitara syntax, specifically, concord and word order. The programme and its implementation are reported on in Chapter 8 of this dissertation.

9.2. Contributions

Given its inter-disciplinary nature, the study offers contributions to researchers and practitioners interested in computational morphology and language learning in general, as well as in CALL for Runyakitara in particular. The following table summarizes the contributions of this study:

Type		**Contribution**
Theory	**1**	Proof that the *fsm*-driven model (CFG plus local allomorphic rules) is applicable to Runyakitara
	2	The first computational description of Runyakitara morphology
	3	A template of Runyakitara verb morphology (for the first time)
	4	An improved description of the noun classification system of Runyakitara
	5	The identification of a new application area for CALL, i.e. re-learning native languages
	6	Several implemented prototypes for learning Runyakitara morphosyntax, nominal number, classifier concord and some word order constraints
Practical	**1**	The morphological analyzer can be put to practical use as a digital dictionary, spell-checker, grammar checker, etc.
	2	Development and evaluation of an ICALL system for (re-)learning Runyakitara
	3	Development of a prototype ICALL system for Runyakitaran syntax

Table 1: Contributions of this thesis

9.2.1 Contribution to theory

This dissertation has contributed to theory as summarized in Table 1 above. First, the study has demonstrated that, by using context-free grammar and re-write rules supported in the fsm2-driven model, we can account for a complex morphological system like that of Runyakitara. Secondly, the study provides a computationally implemented set of Runyakitara morphological rules available to researchers, including the template (see below). There had been no up-to-date framework for Runyakitara morphology which researchers in these languages could use to develop computer applications. These rules

may also guide research in other Bantu languages such as Luganda, Lusoga and Kiswahili.
The template of Runyakitara verb morphology is another theoretical contribution of this study. Nurse & Philipson (2003) discuss the general Bantu template, but this was not adequate for all the specifics of Runyakitara. The elaborated template improves our understanding of Bantu morphology, implying that it may be difficult to generalize certain morphological aspects.

In addition, the noun classification system of Runyankore-Rukiga proposed by Taylor (1985) was improved in this study. The improved system of noun classes now accounts for Runyakitara as a whole and constitutes a comprehensive, detailed system, which will act as a reference model for researchers and students of Runyakitara.

Two Runyakitara learning models (RU_CALL and ICALL for Runyakitara) were developed as contributions to CALL theory. RU_CALL was developed extensively enough to be evaluated empirically and has already been proven useful in language learning. It will help researchers in language learning who work on the pedagogy of teaching languages like Runyakitara, with similarly complex morphologies. The ICALL model for supporting learners who are tackling Runyakitaran syntax is important as well, since it will stimulate debate on the best techniques for learning difficult Bantu language structures such as classifier concord and word order.

9.2.2 Contribution to practice

As De Pauw & Schryver (2008) stress, finding minimal meaning bearing units that constitute a word can provide a wealth of linguistic information that becomes useful when processing the text on other levels of linguistic description such as phonology, syntax and semantics. The morphological analyzer of Runyakitara (RU_MORPH) can be used to develop many other applications outside CALL. These range from syntax and semantics analysis to providing important writing tools such as spell-checkers, grammar checkers, digital dictionaries, etc.

The study resulted in the RU_CALL system, which is ready for practical use. There will never be an adequate supply of Runyakitara teachers in all places where there are learners. An electronic version of Runyakitara learning material recommends itself for its ease of distribution and reproduction. Being the first software for learning Runyakitara, we hope that it will not only be an important contribution to the improvement of Runyakitara language learning by, in particular, the group of (re-)learners who took part in the test of the system, but also that it will inspire similar efforts for other local languages in Uganda and perhaps even in the rest of Africa.

The electronic version of Runyakitara provides immediate and relevant feedback in the way that printed materials cannot. This is in line with Jager's recommendation for CALL implementation framework (Jager 2009), that immediate, detailed and relevant feedback is one of the sound pedagogical stragegies in CALL. As Jager notes, such software may

also play an important auxiliary role in classroom course for language learning – for use outside and complementary to the classroom.
The ICALL system for Runyakitara can be used by advanced learners of Runyakitara morphology and syntax for improving and deepening their language abilities. Systems based on natural language parsing are known to return detailed specifications of linguistic problems, that is, error specific feedback in accordance with well-established pedagogical principles for language learning (Vandeventer Faltin, 2003; Menzel, 2004). As we emphasized in the ICALL chapter, we sketched and implemented a simple system for detecting a limited class of errors.

In this indirect way, we hope that the systems we have presented here may also contribute to language documentation, preservation, and revitalization. The Runyakitara group, like most indigenous Ugandan languages, is being abandoned for English and Kiswahili, which are the official languages. These languages are still healthy today in terms of the number of speakers, but developments are worrisome.

9.3. Limitations of the study and future research

Like all research, this dissertation can be improved and extended. First, we have applied the morphological analyzer of Runyakitara only to language learning, but further research can and should investigate other applications of the analyzer, particularly with regard to improved automatic analysis of syntax and semantics, as well as in terms of more practical issues, such as spell checking and/or automatic dictionary access.

Second, children were not the focus of the Runyakitara language-learning software that we proposed and evaluated. Further research can be directed towards development of applications to assist children in learning to read and write in Runyakitara. Such a development would address the lack of instructional materials for local languages currently taught in Ugandan primary schools. This would require a very different delivery, of course, informed by the relevant pedagogical research.

Third, and finally, in our short list of new directions in which to take this research would be the development of a web-based version of RU_CALL, will be designed, implemented and offered for native speakers of Runyakitara living/working in other countries both for their own use in maintaining their language skills and perhaps as an aid for their children, who might otherwise never progress far in language learning and literacy in their mother tongue.

References

Aikhenvald, A.H. (2006). Classifiers and noun classes: Semantics; In K. Brown (ed.) *Encyclopedia of language and linguistics* 2nd ed., Vol. 1, pp. 463-70. Elsevier: Oxford.

Alidou, H., Aliou B., Brock-Utne B., Diallo Y.S., Heugh, K. Wolff, E. (2006). Optimizing learning and education in Africa – the language factor. A stock-taking research on mother tongue and bilingual education in Sub-Saharan Africa. Paper presented at ADEA meeting, Libreville, Gabon. Mar. 27-31, 2006.

Amaral, L. (2007). *Designing intelligent language tutoring systems: integrating natural language processing technology into foreign language teaching*. (Doctoral dissertation, The Ohio State University, Columbus, Ohio).

Amaral, L.A. & Meurers, D. (2011). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, 23(1), 4–24.

Amaral, L. & Meurers, D. (2006). Where does ICALL fit into foreign language teaching? Paper presented at CALICO Conference. University of Hawaii, http://purl.org/net/icall/handouts/calico06-amaral-meurers.pdf

Anderson, S. (1992). *A-Morphous morphology*. Cambridge: Cambridge University Press.

Antoniadis, G. Kraif, O., Lebarbé, T., Ponton, C. & Echinard, S. (2005) Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues: La plateforme MIRTO. *ALSIC. Apprentissage des Langues et Systèmes d'Information et de Communication,* 8(2), 65-79.

Aronoff, M. (1994). *Morphology by itself*. Cambridge: MIT Press.

Asiimwe, S. (2008). Teaching in vernacular is educationally retrogressive. *New Vision.* Uganda. Mar. 2, 2008. Retrieved from http://allafrica.com/stories/200803210326.htm

Badejo, A.B. (1989). Nigeria and India relations, 1960–1985: A study in South-South cooperation. In Akinyemi, A.B. et al. (Eds.). *Nigeria since independence: The first 25 years: International relations.* Vol. X. Ibadan: Heinemann Educational Books, Ltd.

Badenhorst, J. Heerden C. Davel, M. & Etienne, B. (2009). Collecting and evaluating speech recognition corpora for nine Southern Bantu languages. *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages* (AfLaT 2009). (pp. 1–8), Athens, Greece: ACL.

Bane, M. (2008). Quantifying and measuring morphological complexity. In: C.B. Chang & H.J. Haynie (Eds.) *Proceedings of the 26th West Coast Conference on Formal Linguistics.* (pp. 69-76). Somerville: Cascadilla Proceedings Project. Avail at: http://www.lingref.com/cpp/wccfl/26/

Beesley, K. R. & Karttunen, L. (2003). *Finite state morphology.* Stanford: CSLI Publications.

Bellomo, T.S., (2009). Morphological analysis and vocabulary development: Critical criteria. *The Reading Matrix*, 9(1), 44-55.

Bernsten, J. (1998). Runyakitara, Uganda's 'new' language. *Journal of Multilingual and Multicultural Development*, 19(2), 93-107.

Blanchard, A., Kraif, O., & Ponton, C. (2009) Mastering overdetection and underdetection in learner-answer processing: Simple techniques for analysis and diagnosis. *Calico Journal*, 26(3), 592-610.

Brown, H. (2000). *Principles of language learning and teaching* (4th ed.), New York: Addison-Wesley Longman, Inc.

Bruner, J. S. (1966). *Toward a theory of instruction.* Cambridge: Harvard University Press.

Bukenya G. (2008). Teaching in local languages, good policy. *http://allafrica.com*, Last retrieved on June 2, March, 2012.

Buttery, A. Rai, I. and Beresford A. (2009). Language learning on a next-generation service platform for Africa. Paper presented at *2nd Satellite Workshop on Innovative Mobile Technology & Services for Developing Countries.*15th March 2009, Kampala, Uganda.

Casalis, S. & Louis-Alexandre M. (2000). Morphological analysis, phonological analysis and learning to read French: A longitudinal study. *Reading and Writing: An Interdisciplinary Journal*, 12, 303–335.

Chapelle, C. A. (2001). *Computer applications in second language acquisition.* Cambridge, UK: Cambridge University Press.

Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology*. 2nd ed. Oxford: Blackwell.

Cormen, T.H., Leiserson, C.E., Rivest, R.L. & Stein, C. (2001). *Introduction to algorithms*. 2nd Edition. Cambridge, Mass.: MIT Press.

Crawhall, N. (1998). Still invisible: San and Khoe in the new South Africa. *AfricaFiles*. www.africafiles.org. Retrieved on 20th June, 10 2011.

Cross, K.P. (1981). *Adults as learners.* San Francisco: Jossey-Bass.

Crystal, D. (1997). *Language. The Cambridge encyclopedia of language*. 2nd ed. Cambridge: Cambridge University Press.

De Pauw, G. & De Schryver, G.-M. (2008). Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes. *Lexikos,* 18(1), 303-318.

De Pauw, G., Wagacha, P.W., & De Schryver, G.M. (2009). The SAWA corpus a parallel corpus English-Kiswahili. *Proceedings of the First Workshop on Language Technologies for African Languages* (AfLaT '09). Athens, Greece.: ACL, 9–16.

DeKeyser, R. M., (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds). *The handbook of second language acquisition* (pp. 313-348). Oxford: Blackwell.

Demuth, K., (2000). Bantu noun class systems: Loan word and acquisition evidence of semantic productivity. In: G. Senft (ed.), *Classification systems*. (pp. 270-292). Cambridge: Cambridge University Press.

Dickinson, M. (2010) On morphological analysis for learner language, focusing on Russian. *Research on Language & Computation* 8(4), 273-298.

Dickinson, M., & Herring J., (2008). Developing online ICALL exercises for Russian. *Proceedings of the Third Workshop on Innovative Use of Natural Language Processing for Building Educational Applications.* (pp. 1-9). Columbus, Ohio: ACL. Avail. http://aclweb.org/anthology/W08-0910

Dorian, Nancy C. (1977). The problem of the semi-speaker in language death. *Linguistics,* 15(191), 23–32.

El Kholy, A. & Habash, N. (2012). Rich morphology generation using statistical machine translation. *Proceedings of the Seventh International Natural Language Generation Conference* (INLG 2012). (pp. 90-94). Utica, IL: ACL.

Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language Learning,* 54(2), 227-275.

Elwell, R. (2005). Finite-state methods for Bantu verb morphology. In: N. Gaylord, S. Hilderbrand, H. Lyu, A. Palmer & E. Ponvert (Eds.), *Texas Linguistics Society 10: Computational linguistics for less-studied languages*. (pp. 56-67). Stanford: CSLI Publications.

Eshton, E. O. (1937). The structure of Bantu language with specific reference to Swahili. *Bulletin of the School of Oriental and African Studies* 8(4), 1111-1120.

Esit, O. (2011). Your verbal zone: An intelligent computer-assisted language learning program in support of Turkish learners. *Computer Assisted Language Learning*, 24(3), 211-232.

Fillmore, L.W. (2000). Loss of family languages: Should educators be concerned? *Theory into Practice***,** 39(4), 203-210. doi: 10.1207/s15430421tip3904_3

Fishman, J. (2000) Reversing language shift: RLS theory and practice revisited. In: G.Kindell & M.P. Lewis (eds.) *Assessing ethnolinguistic vitality. Theory and practice.* (pp. 1-26) Dallas: SIL International.

Fishman, J. (Ed.) (2001). *Can threatened languages be saved?* Clevedon, Tonawanda & North York: Multilingual Matters.

Ford, N. (2001). The increasing relevance of Pask's work to modern information seeking and use. *Kybernetes*, 30 (5/6), 603-29.

Ford, N. (2004). Conversational information systems: Extending educational informatics support for the web-based learner. *Journal of Documentation*, 61(3), 362-384.

Gamper, J. & Knapp, J. (2002). A review of intelligent CALL systems. *Computer Assisted Language Learning*, 15(4), 329-342.

Gardner, S. (2008). Changing approaches to teaching grammar. *English Language Teacher Education and Development,* 11, 39-44.

Gasser, M. (2009). Computational morphology and the teaching of indigenous languages. In S.M. Coronel-Molina & J.H. McDowell (Eds.) *Proceedings of the First Symposium on Teaching Indigenous Languages of Latin America*, (STILLA 2008). (pp. 52-63). Bloomington, Indiana: Center for Latin American and Caribbean Studies.

Government of Uganda (1992). Government white paper on education policy review. *Commission Report*, Kampala, Uganda.

Gregg, K. (1984). Krashen's monitor and Occam's razor. *Applied Linguistics*, *5*, 79-100

Gumede, T. & Plauch E, M. (2009). Initial fieldwork for Lwazi, a telephone-based spoken dialogue system for rural South Africa. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (EACL 2009), 59–65.

Gündüz, N. (2005). Computer assisted language learning (CALL). *Journal of Language and Linguistic Studies*, 1(2), 193-214.

Guthrie, M. (1967-1971). *An introduction to the comparative linguistics and the pre-history of the Bantu languages.* Vol. 1-4. Farnborough: Gregg International.

Hajič, J. & Hladká, B. (1998). Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. *Proceedings of the 17th International Conference on Computational linguistics*. (COLING 17) Vol. 1, Montreal: ACL. 483-490.

Halle, M. & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In: K. Hale & S. J. Keyser (eds.) *The view from building 20*. Cambridge: *MIT Press*, 111-176.

Halliday, M. (1968). The users and uses of language. In J. Fishman (Ed.), *Readings in the sociology of language.* The Hague: Mouton.

Hamwedi, M.A. & Dalvit, L. (2012). E-learning and M-learning in African languages: A survey of Oshikwanyama students at a northern Namibian school. (pp. 78-83) *eLmL 2012: The Fourth International Conference on Mobile, Hybrid, and Online Learning, 30th – 4th Feb. 2012, Valencia, Spain.*

Hanneforth, T. (2009). *fsm2 - A scripting language for creating weighted finite-state morphologies*. In: C. Mahlow & M. Piotrowski (Eds.): *State of the art in computational morphology* (pp. 48-63). Heidelberg: Springer Berlin.

Harmer, J. (2003). Popular culture, methods and context. *ELT Journal* ,57(3), 288-294

Hartley, J. (1998). *Learning and studying. A research perspective*, London: Routledge.

Heilman, M., Collins-Thompson, K., Callan, J. & Eskenazi, M. (2006). Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension. In *Proceedings of the Ninth International Conference on Spoken Language Processing*. Pittsburgh, PA.

Hockett, C.F. (1954). Two levels of grammatical description. *Word*, 10, 210-31.

Hockett, C.F. (1958). *A course in modern linguistics*. New York: Macmillan.

Hopcroft, J. E., Motwani R., Ullman J. (2001). *Introduction to automata theory, languages and computation* (2nd ed.). Reading, Mass: Addison-Wesley.

Hubbard, P. (2006). Evaluating CALL software. In L. Ducate and N. Arnold (Eds.) *Calling on CALL: From theory and research to new directions in foreign language teaching. (pp. 313-338).* San Marcos, TX: CALICO.

Hurskainen A. (1992). A two-level computer formalism for the analysis of Bantu morphology: An application to Swahili. *Nordic Journal of African Studies*, 1(1), 87-119.

Hurskainen, A. (1996). Disambiguation of morphological analysis in Bantu languages. *Proceedings of COLING-96, Copenhagen.* 568-573.

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

Hurskainen A. (2009a). Intelligent computer-assisted language learning: Implementation to Swahili. *Technical Reports in Language Technology*, Report No 3. Avail. at *http://www.njas.helsinki.fi/salama*

Hurskainen, A. (2009b). Enriching text with tone marks: An application to Kinyarwanda language *Technical Reports in Language Technology*. Report 4. Avail. at *http://www.njas.helsinki.fi/salama*

Hyman, L. (2007). Niger Congo verb extensions: overview and discussion. In: D. L. Payne and J. Peña (eds.), *Selected Proceedings of the 37th Annual Conference on African Linguistics.* pp. 149-163. Somerville, MA: Cascadilla Proceedings Project. Avail. at http://www.lingref.com/cpp/acal/37/

Ismail, S. A. A., (2010). ESP students' views of ESL grammar learning. *GEMA Online Journal of Language Studies,* 10(3), 143-155.

Jager, S. (1998). HOLOGRAM. In S. Jager, J. Nerbonne & A. J. van Essen (Eds.), *Language teaching & language technology* (pp. 82-87). Amsterdam: Swets & Zeitlinger.

Jager, S. (2004). Learning management systems for language learning. In A. Chambers, J.E. Conacher & J. Littlemore (Eds.), *ICT and language learning: Integrating pedagogy and practice* (pp. 33-48). Birmingham: University of Birmingham Press.

Jager, S. (2009). *Towards ICT-Integrated Language learning: developing an implementation framework in terms of pedagogy, technology and environment*. Doctoral Dissertation, University of Groningen.

Johnson, C.D. (1972). *Formal aspects of phonological description.* Mouton: The Hague.

Jurafsky, D. & Martin J.H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition.* New Jersey: .2nd ed. Prentice-Hall.

Kaplan, R. M. & Kay M. (1994). Regular models of phonological rule systems. *Computational Linguistics,* 20(3), 331-378.

Karttunen, L. & Beesley, K. R., (2005). Twenty-five years of finite-state morphology. In: *Inquiries into words. A Festschrift for Kimmo Koskenniemi on his 60th birthday*. CSLI Studies in Computational Linguistics. (pp. 71-83). Stanford, CA: CSLI.

Karttunen, L. (2003). Computing with realizational morphology. In: A. Gulbekh (Ed.): *Computational linguistics and intelligent text processing*, Lecture Notes in Computer Science, 2588. (pp. 205-216). Heidelberg: Springer.

Katamba, F. (1984). A non-linear analysis of vowel harmony in Luganda. *Journal of Linguistics*, 20(2), 257-275.

Katamba, F. (2006). Bantu nominal morphology. In D. Nurse & G. Philippson (Eds.), *The Bantu languages*, Routledge Language Family Series 9, (pp. 103–120). London: Routledge.

Katushemererwe, F., & Hanneforth, T., (2010). Finite state methods in the morphological analysis of Runyakitara verbs. *Nordic Journal of African Studies*, 19(1), 1-23.

Katushemererwe, F. & Hanneforth, T. (2010a). *fsm2* and the morphological analysis of Runyakitara nouns – first experiences from Runyakitara. *International Journal of Computing and IT Research.* Special Issue, 4(1), 58-69.

Katushemererwe, F. & Hurskainen, A. (2011). Intelligent computer-assisted language learning system: Implementation on Runyakitara. In M. Kizza (Ed.) Vol. VII: 426-444, *Special Topics in Computing and ICT Research.* Kampala, Uganda: Fountain Publishers,

Kiango, J. (2005). Problems of citation forms in dictionaries of Bantu languages. *Nordic Journal of African Studies*, 14(3), 255–273.

Kieffer, M.J. & Lesaux, N.K. (2007). Breaking words down to build meaning: vocabulary, morphology and reading comprehension in the urban classroom. *The Reading Teacher*, 61, 134-144.

Kihm, A. (2002). What is in a noun: Noun classes, gender, and nounness. Ms., *CNRS, Paris, Laboratorire de Linguistique formelle*.

Klaus, D.A. (2003). The use of indigenous languages in early basic education in Papua New Guinea: A model for elsewhere? *Washington, DC: World Bank*.

Koskenniemi, K. (1983). Two-level morphology: a general computational model for word-form recognition and production. Publication No. 11. University of Helsinki: Department of General Linguistics.

Koskenniemi, K. (1984). Two level morphology: A general computational model for word form recognition and production. *Proceedings of COLING*. 178-181.

Krashen, S.D. (1981). Bilingual education and second language acquisition theory. In *Schooling and language minority students: A theoretical framework* (p.51-79). California State Department of Education.

Kung, S. C. (2002). A framework for successful key-pal programs in language learning, *CALL-EJ Online*, 3(2). Avail at http://callej.org/journal/3-2/sckung.html

Kuruoğlu, G., Algar A., Sezer E., Erol S., Jaeckel R. (2000). Language learning framework for Turkish. Report, National Council of Organisations of Less Commonly Taught Languages, UCLA

Kwesiga, J.B. (1994). Literacy and the language question: Brief experiences from Uganda. In D. Barton (Ed.), *Sustaining local literacies.* Special issue of *Language and Education, 8(1-2),* 57–63.

Ladefoged, P., et al, (1972). Language in Uganda. Ford Foundation *Language Survey Vol. 1,* London, Oxford University Press.

Lai, C., & Kritsonis, W. A. (2006). The advantages and disadvantages of computer technology in second language acquisition. *National Journal for Publishing and Mentoring Doctoral Student Research,* 3(1), 1-6.

Landweer, M. L. (2000). Endangered languages. Indicators of ethnolinguistic vitality. *Notes on Sociolinguistics* 5.1:5-22. Avail. at www.sil.org/sociolx/ndg-lg-indicators.html

Landweer, M. L. (2000). Endangered languages. Indicators of ethnolinguistic vitality. *Notes on Sociolinguistics* 5.1:5-22. Avail. at www.sil.org/sociolx/ndg-lg-indicators.html

Lazarov, M. (2006). Finite-state methods for spelling correction. *Unpublished BA thesis.*

Eberhard-Karls-Universität Tübingen

Lewis, M. Paul (ed.), 2009. *Ethnologue: Languages of the World*, Sixteenth edition. Dallas, Tex.: SIL International. *Online version*: http://www.ethnologue.com/.

Lodhi, A. Y. (1993). The language situation in Africa today. *Nordic Journal of African Studies,* 2 (1), 79-86.

Ma, Q. & Kelly P. (2006). Computer-assisted vocabulary learning: Design and evaluation. *Computer Assisted Language Learning*, 19(1), 15-45.

Ma, Q. (2008). Empirical CALL evaluation: the relationship between learning process and learning outcome. *CALICO Journal*, 26(1), 108-122.

Maho, J. F. (2007). A linear ordering of TAM/NEG markers in the Bantu languages. *SOAS working papers in linguistics*, 15, 213-225.

Majola, K. (2006). *Language and Education in Uganda: An encounter with the National Indigenous Language Forum*. Oslo: Multicultural and International Education, University of Oslo.

Marten, L. & Nancy C. K. (2007). Morpho-syntactic co-variation in Bantu, *Bantu in Bloomsbury: Special Issue on Bantu Linguistics, SOAS Working Papers in Linguistics 15*.http://www.soas.ac.uk/academics/departments/linguistics/research/workingpapers/volume-15/swpl-volume-15.html

Martens, S. (2011). *Quantifying linguistic regularity*. Ph.D. thesis, University of Leuven.

McArthur (1992). *Oxford companion to the English language*. Oxford and New York: Oxford University Press.

Mchombo S. (2004). *The syntax of Chichewa*. Cambridge: Cambridge University Press.

Meeussen, A. E. (1967) . Bantu grammatical reconstructions. *Africana Linguistica, 3, 79-121.*

Menzel, W. (2004). Errors, intentions, and explanations: Feedback generation for language tutoring systems. *Proceedings International Conference InSTIL/ICALL-2004,* (pp. 75-82), Venice, Italy.

Meriwether, N. (2001). *12 easy steps to successful research papers* (second ed.). New York: McGraw-Hill.

Mohri, M. & Pereira, F. C. N. (1998). Dynamic compilation of weighted context-free grammars. *Proceedings of the 17th international conference on Computational linguistics* -Vol. 2. Montreal: ACL. 891-897.

Morris, H. F. & Kirwan B. R. (1972). *A Runyakore Grammar*. Kampala: East African Literature Bureau.

Mugane, J. M., (1997). Learning African languages with evolving digital technologies. *Africa Today*, 44(4), 423-444.

Muhirwe, J. (2007). Computational analysis of Kinyarwanda morphology: the morphological alternations. *International Journal of computing and ICT Research,* 1(1), 85-92.

Muhirwe, J. (2010). Morphological analysis of tone-marked Kinyarwanda text. In Yli-Jyrä, A., Kornai, A., Sakarovitch, J. & Watson, B. (Eds.) *Finite-State Methods and Natural Language Processing*. Lecture Notes in Artificial Intelligence, 6062, 48-55.

Muhirwe, J. & Trosterud, T. (2008). Finite state solutions for reduplication in Kinyarwanda language. *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages* (pp. 73–80). Hyderabad, India: Asian Federation of Natural Language Processing.

Mukama, R.G. (1991). Getting Ugandans to speak a common language. In H.B. Hansen & M. Twaddle (Eds.), *Changing Uganda: The dilemmas of structural adjustment and revolutionary change* (pp. 334–350). Kampala: Fountain Publishers.

Myers, D. G., & Spencer, S. J. (2004). *Social Psychology*. 2nd Canadian Edition, McGraw Hill.

Nagata, N. (2009). Robo-Sensei. *CALICO Journal*, 26(3), 562-579.

Namyalo, S. (2010). The challenges of using mother tongue languages in education for sustainable development: the case of Uganda. Presentation at the Universität Zürich, avail at. http://www.uzh.ch/news/agenda/record.php?id=11593&group=12

Ndoleriire, O. & Oriikiriza, C. (1990). *Runyakitara studies*. Unpublished manuscript, Makerere University, Uganda.

Negroponte, N. (1995). *Being digital*. New York: Vintage Books.

Nerbonne, J. & Dokter D. (1999). An intelligent word-based language learning assistant. *Traitement Automatique de Langage,*. 40(1), 125-142.

Nerbonne, J. (2002). Computer-assisted language learning and natural language processing. In: Ruslan Mitkov (Ed.) *Handbook of computational linguistics* (pp.670-698). Oxford: Oxford University Press.

Nerbonne, J., Dokter, D. & Smit. P., (1998). Morphological processing and computer-assisted language learning. *Computer-Assisted Language Learning*. 11(5), 421-37.

Nerbonne, J. & Smit, P. (1996) GLOSSER-RuG: In support of reading. *Proceedings of the 16th conference on Computational linguistics. Vol. 2* 830-835.

Nurse, D. & Philippson, G. (2003). The Bantu languages. London: Routledge Language Series.

Obondo, M.A. (2007). Tensions between English and mother tongue teaching in post-colonial Africa. In J. Cummins, & C. Davison (Eds.), *International handbook of English language teaching*. (pp. 37–50). New York: Springer.

Odejobi, O.A. & Beaumont, T. (2003). Web-based intelligent computer-assisted language learning system for Yoruba (YiCALL). *International Association for Development of the Information Society (IADS) e-conference*. Lisbon, Portugal. June 3-6, 2003. Avail. at http://leilbadrahzaki.wordpress.com/

Ohiri-Aniche, C. (1997). Nigerian languages die. *Quarterly Review of Politics, Economics and Society*, 1(2), 73-9.

Ondari Okemwa, E. & Ng'ang'a, W. (2008). Kiswahili spell checker and morphological analyzer. Software avail at aflat.org

Oriikiriza, C. (2007). *Kashoboorozi y'Orunyankore-Rukiga*. Uganda: Fountain Publishers.

Ormrod, J. (1995). *Educational psychology: Principles and applications*. Englewood Cliffs, NJ: Prentice-Hall.

Ostler, N. (2007). *Ad infinitum: A biography of Latin*. London: Harper Press.

Oyelami, O. (2008). Development of Igbo language e-learning system. *Turkish Online Journal of Distance Education,* 9 (4), 39-52.

Pangaro, P. (2001). "THOUGHTSTICKER 1986: A personal history of conversation theory in software and its progenitor, Gordon Pask", *Kybernetes*, 30(5/6), 790-806.

Pask, G. (1975). *Conversation, cognition and learning*, Elsevier, Amsterdam.

Pijls, F., Daelemans, W. and Kempen, G. (1987). Artificial intelligence tools for grammar and spelling instruction. *Instructional Science*, 16, 319–336.

Prah, K. K. (2008). The language of instruction conundrum in Africa. Presentation at the UNESCO *Meeting on the Implications of Language for Peace and Development*. Oslo, 2-3, May, 2008.

Prah, K. K. (2002). Race and culture: Myth and reality." In N. Duncan et al,, (Eds.) *Discourses on difference. Discourses on oppression*. Cape Town: The Centre for Advanced Studies of African Society (CASAS). (Book Series No 24).

Pretorious, R., Berg, A. & Pretorious, L. (2009). Setswana tokenisation and computational verb morphology: Facing the challenge of a disjunctive orthography. *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages* – AfLaT 2009, Athens, Greece, March 31, 2009. 66–73.

Pretorius, L. & Bosch, E.S. (2003). Finite-state computational morphology: An analyzer prototype for Zulu. *Machine Translation*, 18(3), 195-216.

Roark, B. & Sproat, R. (2007). *Computational approaches to morphology and syntax*. Oxford: Oxford University Press.

Samaras, K. (2005). Indigenous Australians and the 'digital divide'. *Libri.* 55, 84-85

Sauro, S. (2009). Computer-mediated corrective feedback and the development of L2 grammar. *Language Learning and Technology*, 13(1), 96-120

Schmid, M.S. & de Bot, K. (2004). Language attrition. In: A. Davis & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 210-234). Oxford: Blackwell.

Schmid, M. S., Köpke, B., Keijzer, M., & Weilemar, L. (2004). *First language attrition: Interdisciplinary perspectives on methodological issues*. Amsterdam: John Benjamins.

Shaalan, K. (2005). An intelligent computer assisted language learning system for Arabic learners *Computer Assisted Language Learning*, 18(1-2), 81-108.

Shacham, D. & Wintner, S. (2007). Morphological disambiguation of Hebrew: a case study in classifier combination. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, June, 2007. Prague: ACL, 439-447.

Spencer, A. (1991). *Morphological theory.* Oxford & New York: Wiley-Blackwell.

Taylor, C. (1959). *A teachers' handbook of Runyankore-Rukiga orthography*. Kampala: Eagle Press.

Taylor, C. (1985). *Nkore-Kiga*. (Croom Helm Descriptive Grammars Series) Cambridge: Cambridge University Press.

Tembe, J. & Norton, B. (2008). Promoting local languages in Ugandan primary schools: The community as a stakeholder. *Canadian Modern Language Review*, 65(1), 33-60.

Uganda Bureau of Statistics (2002). The 2002 Uganda Population and Housing Census, Population Size and Distribution. October, 2006, Kampala, Uganda.

Uganda (1992). The Government White Paper on the Education Policy*: Review Commission Report.* Ministry of Education and Sports, Kampala, Uganda.

Van Huyssteen, (2007). Designing an e-learning system for language learning: A case study. In: M. Iskander (Ed.), *Innovations in E-learning, Instruction Technology, Assessment and Engineering Education* (pp. 105-110). Berlin: Springer.

Vandeventer Faltin, A. (2003). *Syntactic error diagnosis in the context of computer-assisted language learning*. Diss., University of Geneva.

Warschauer M. (1996). Computer assisted language learning: An introduction. In S. Fotos (Ed.) *Multimedia language teaching* (pp. 3-20). Tokyo: Logos International:

Warschauer, M. & Healey, D. (1998). Computers and language learning: an overview. *Language teaching forum*, 31, 57-71.

Widodo, H. P. (2006). Approaches and procedures for teaching grammar. *English Teaching: Practice and Critique*, 5(1), 122-141.

Wintner, S. (2008). Strengths and weaknesses of finite-state technology: A case study in morphological grammar development. *Natural Language Engineering*, 14(4), 457-469.

Wolff, H. E. (2005). The language factor in discourse on development and education in Africa. Paper presented to the *Symposium on Language for Development in Africa*, held at Moi University, Eldoret, Kenya, June 1-3, 2005.

Wu, M. (2008). Principal approaches of grammar instruction. *US-China Foreign Language*, 6(11), 29-35.

Yli-Jyra, A. (2005). Toward a widely usable finite-state morphology workbench for less studied languages – Part 1: Desiderata. *Nordic Journal of African Studies*, (14 (4), 479-491.

Yona S. & Wintner, S. (2005). A finite state morphological grammar for Hebrew. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, Mich., US: ACL. 9-16. Avail. at http://aclweb.org/anthology-new/W/W05/W05-07.pdf

Zahram, M. (2001). Instructing Kiswahili for foreign aid workers: Nordic experiences. *Africa & Asia*, 1, 79-83.

Samenvatting

Computationele morfologie speelt een belangrijke rol in hedendaags computerondersteund taalonderwijs (*computer-assisted language learning*, CALL), in het bijzonder in de verwerving van de woordenschat en de morfologie, in het toegankelijk maken van woordenboekinformatie en in het verbeteren van de leesvaardigheid. Vanwege dalende kosten in de ICT, is het mogelijk software voor morfologische analyses steeds vaker in CALL applicaties te gebruiken. Helaas worden CALL applicaties slechts voor een klein aantal talen in de wetenschappelijke literatuur gerapporteerd, en vrijwel uitsluitend voor de beter bestudeerde en gedocumenteerde talen zoals het Engels, Frans, Duits, Spaans, Russisch en Japans. Weinig literatuur wordt geweid aan minder vaak bestudeerde talen, zelfs talen die echte wetenschappelijke uitdagingen vormen vanwege hun morfologische complexiteit. Runyakitara is een groep van Bantutalen met zeer complexe morfologie en vormt de focus van dit proefschrift.

Dit proefschrift presenteert het ontwerp, de implementatie en evaluatie van een systeem voor de morfologische analyse van de Runyakitara talen. Vervolgens hebben we het systeem op een nieuwe manier in CALL ingezet, met name om oefeningen voor flexie te creëren. Tot nu toe werd in CALL morfologische software ingezet om informatie te geven aan studenten die een taal wilden leren lezen, om toegang tot woordenboeken te faciliteren en om voorbeelden van woorden (uit hetzelfde morfologisch paradigma) in grote tekstcorpora te vinden. Verder bespreken we ander werk in CALL dat oefeningen ontwikkelt door de inzet van natuurlijke taalverwerking (NLP), maar het weinige dat we konden vinden (van Amaral & Meurers 2006; Dickinson & Herring 2008) concentreert zich op de bovengenoemde goed documenteerde talen. In tegenstelling tot bijna alle literatuur van eerdere studies, hebben we software voor morfologische analyses gebruikt om de oefeningen te ontwikkelen. Om dit doel te bereiken hebben we de morfologische analyse in twee CALL applicaties ingezet. We beschrijven zowel de morfologische analyse als ook de CALL applicaties in deze dissertatie.

Gedurende de voorbereiding voor het ontwikkelen van de Runyakitara morfologische analyse, ontdekten we dat er noch systematische morfologische beschrijvingen, noch geschikte gegevensverzamelingen bestonden om te gebruiken voor het ontwikkelen en testen van de morfologische analyse. We hebben daarom beschrijvingen van andere Bantutalen gebruikt, alsmede bestaande studies over bijzondere thema's in Runyakitara (inclusief niet-gepubliceerde manuscripten) en onze eigen intuïtie. Zo hebben we het eerste morfologische analysesysteem voor Runyakitara geformaliseerd, ontworpen, geïmplementeerd en geëvalueerd. We merken hierbij op dat dit ook betekent dat de applicatie die we hebben ontwikkeld niet met ander leermateriaal voor Runyakitara vergeleken kan worden. Dit soort materiaal bestaat eenvoudig niet.

Het morfologische systeem van Runyakitara wordt in hoofdstukken 2 t/m 4 van dit proefschrift gepresenteerd, besproken en geëvalueerd. Het systeem voor de morfologische analyse van Runyakitara is intussen functioneel en kan voor verschillende

doeleinden worden ingezet. Het niveau van afdekking is voor leerdoeleinden voldoende vanwege de hoge mate van precisie, alhoewel het *recall* (de hoeveelheid verschillende vormen die geanalyseerd kunnen worden) verbeterd zou moeten worden voor meer gevorderde leerprogramma's.

Morfologische analyse werd tot nu toe in CALL ook succesvol ingezet, maar in andere talen en ook om andere specifieke leertaken te ondersteunen. Dit proefschrift bevat pionierswerk met betrekking tot de taal en in het demonstreren van het nut van NLP in de ontwikkeling en ondersteuning van CALL oefeningen. We beschouwen het succes van de ontwikkeling van de applicaties als indicatie dat NLP in CALL effectiever gebruikt kan worden dan in veel applicaties tot nog toe zichtbaar wordt.

Om de situatie omtrent taalonderwijs en taalverwerving in Oeganda beter te begrijpen hebben we ook een kleine enquête doorgevoerd. Deze studie beoogde de behoeften en beperkingen van het gebruik van CALL in Oeganda in kaart te brengen. Hoofdstuk 5 rapporteert over de resultaten van dit onderzoek. Het meest opvallende resultaat was dat de deelnemers zeer verdeeld zijn ten aanzien van het gebruik van CALL software om locale talen in Oeganda effectiever te laten leren. Drukke managers hadden geen belangstelling, terwijl onderwijskundigen wel degelijk geïnteresseerd waren. Omdat Jager (2009) onderstreept dat taalonderwijskundigen de grootste belanghebbenden zijn in beslissingen over taalonderwijsmethoden, verplaatsten deze resultaten de focus van onze softwareontwikkeling. We hielden voortaan de inzet in formeel onderwijs in het oog en trokken verder de conclusie dat we de CALL applicatie onafhankelijk van internet moesten maken, want internet bleek in vele Oegandese steden onbetrouwbaar en soms zelfs onbeschikbaar te zijn. Geen van de ondervraagden in onze enquête hadden ervaring met CALL, alhoewel sommige deelnemers taaldocent waren die computers dagelijks gebruikten. Alle taaldocenten waren positief over de mogelijkheden met CALL te experimenteren in hun onderwijs. Hoofdstuk 6 rapporteert over een pilotstudy ten behoeve van de inzet van het morfologische software in een CALL applicatie binnen de HOLOGRAM software. We hebben uiteindelijk vanwege technische redenen deze studie niet lang voortgezet, maar de opzet van HOLOGRAM, met oefeningen en naslagwerk in één applicatie, hebben we wel aangehouden.

Om de inzet van de ontwikkelde software beter te begrijpen en te evalueren, hebben we een tweede pilotstudy uitgevoerd, waarover hoofdstuk 7 verslag uitbrengt. In deze studie is ons opgevallen dat we met een bijzondere groep te doen hadden, met name Oegandese studenten aan de Makerere Universiteit uit het gebied waar Runyakitara wordt gesproken en wiens ouders uit dat gebied weg waren getrokken (naar Kampala, waar Luganda wordt gesproken). De ouders spraken Runyakitara en hun kinderen (nu studenten) hadden ook passieve kennis van de taal, ze konden één en ander verstaan, maar ze konden het niet goed spreken en helemaal niet schrijven. Literatuuronderzoek bevestigt dat de kinderen van emigranten vaak de taal van hun ouders niet overnemen (Dorian, 1977; Fishman, 1991, 2000; Ohiri-Aniche, 1997; Landweer, 2000). De studenten van de cursus aan Makerere Universiteit wilden de taal leren om in de familie (gedeeltelijk nog in het gebied waar Runyakitara wordt gesproken) beter mee te draaien en soms om hun arbeidskansen in het westen van Oeganda—waar Runyakitara wordt gesproken—te

verbeteren. We hebben een applicatie voor het leren van nominale morfologie ontwikkeld om vervolgens deze studenten ermee te laten werken. De resultaten van deze studie worden eveneens in hoofdstuk 7 besproken. Aan het begin van de cursus beheersten de deelnemers de morfologie niet, hetgeen hun status als “semi-sprekers” bevestigde, maar er traden substantiële verbeteringen in de loop van de tiendaagse cursus, hetgeen de effectiviteit van het ontwikkelde materiaal bevestigde. Gezien concurrerend leermateriaal voor Runyakitara niet bestaat, beweren we dat het hier ontwikkelde CALL software zijn potentieel hierdoor bewezen heeft. Dit betekent ook dat men door middel van dit of vergelijkbaar leermateriaal de positie van locale talen zoals Runyakitara kan verbeteren.

Om het verdere nut van de morfologische analyse voor het ontwikkelen van inhoud voor taallessen te demonstreren hebben we in hoofdstuk 8 een tweede applicatie ontworpen en geïmplementeerd, maar nu vooral gericht op het leren van de syntaxis, in het bijzonder congruentie en woordvolgorde.

Curriculum Vitae

Fridah Katushemererwe was born in Rubaga, Kampala in 1972. She attended several Primary Schools and St. Michael High School, Bushenyi for both Ordinary and Advanced levels of secondary education.

In 1995, she joined Makerere University for her undergraduate studies. In 1998, she received a Bachelor of Arts degree (Economics and Runyakitara). In 2001, Fridah received a Master of Science (Information Science) and in 2002 she received a Post-graduate Diploma in Computer Science all from Makerere University, Kampala, Uganda.

Fridah embarked on her PhD studies in 2007. In 2009, she received a DAAD research grant and stayed at the University of Potsdam, Germany as a visiting research student for the period of 6 months. From 2010 - 2013, Fridah was sponsored under an NFP project entitled: Strengthening ICT Training and Research Capacity in the Four Public Universities in Uganda.

Fridah is an Assistant Lecturer in the department of Linguistics, English Language Studies and Communication Skills. She is able to work and carry out research on her own initiative and as part of a team. She is able to analyse issues keenly using acquired analytical skills, but most importantly, she is dedicated to maintaining quality and academic excellence.

Groningen dissertations in linguistics(GRODIL)

---

1. Henriëtte de Swart (1991). *Adverbs of Quantification: A Generalized Quantifier Approach*.
2. Eric Hoekstra (1991). *Licensing Conditions on Phrase Structure*.
3. Dicky Gilbers (1992). *Phonological Networks. A Theory of Segment Representation.*
4. Helen de Hoop (1992). *Case Configuration and Noun Phrase Interpretation*.
5. Gosse Bouma (1993). *Nonmonotonicity and Categorial Unification Grammar*.
6. Peter I. Blok (1993). *The Interpretation of Focus*.
7. Roelien Bastiaanse (1993). *Studies in Aphasia*.
8. Bert Bos (1993). *Rapid User Interface Development with the Script Language Gist*.
9. Wim Kosmeijer (1993). *Barriers and Licensing*.
10. Jan-Wouter Zwart (1993). *Dutch Syntax: A Minimalist Approach*.
11. Mark Kas (1993). *Essays on Boolean Functions and Negative Polarity*.
12. Ton van der Wouden (1994). *Negative Contexts*.
13. Joop Houtman (1994). *Coordination and Constituency: A Study in Categorial Grammar*.
14. Petra Hendriks (1995). *Comparatives and Categorial Grammar*.
15. Maarten de Wind (1995). *Inversion in French.*
16. Jelly Julia de Jong (1996). *The Case of Bound Pronouns in Peripheral Romance*.
17. Sjoukje van der Wal (1996). *Negative Polarity Items and Negation: Tandem Acquisition*.
18. Anastasia Giannakidou (1997). *The Landscape of Polarity Items*.
19. Karen Lattewitz (1997). *Adjacency in Dutch and German*.
20. Edith Kaan (1997). *Processing Subject-Object Ambiguities in Dutch*.
21. Henny Klein (1997). *Adverbs of Degree in Dutch*.
22. Leonie Bosveld-de Smet (1998). *On Mass and Plural Quantification: The case of French 'des'/'du'-NPs*.
23. Rita Landeweerd (1998). *Discourse semantics of perspective and temporal structure*.
24. Mettina Veenstra (1998). *Formalizing the Minimalist Program*.
25. Roel Jonkers (1998). *Comprehension and Production of Verbs in aphasic Speakers*.
26. Erik F. Tjong Kim Sang (1998). *Machine Learning of Phonotactics*.
27. Paulien Rijkhoek (1998). *On Degree Phrases and Result Clauses*.

28. Jan de Jong (1999). *Specific Language Impairment in Dutch: Inflectional Morphology and Argument Structure.*
29. H. Wee (1999). *Definite Focus.*
30. Eun-Hee Lee (2000). *Dynamic and Stative Information in Temporal Reasoning: Korean tense and aspect in discourse.*
31. Ivilin P. Stoianov (2001). *Connectionist Lexical Processing.*
32. Klarien van der Linde (2001). *Sonority substitutions.*
33. Monique Lamers (2001). *Sentence processing: using syntactic, semantic, and thematic information.*
34. Shalom Zuckerman (2001). *The Acquisition of "Optional" Movement.*
35. Rob Koeling (2001). *Dialogue-Based Disambiguation: Using Dialogue Status to Improve Speech Understanding.*
36. Esther Ruigendijk (2002). *Case assignment in Agrammatism: a cross-linguistic study.*
37. Tony Mullen (2002). *An Investigation into Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection.*
38. Nanette Bienfait (2002). *Grammatica-onderwijs aan allochtone jongeren.*
39. Dirk-Bart den Ouden (2002). *Phonology in Aphasia: Syllables and segments in level-specific deficits.*
40. Rienk Withaar (2002). *The Role of the Phonological Loop in Sentence Comprehension.*
41. Kim Sauter (2002). *Transfer and Access to Universal Grammar in Adult Second Language Acquisition.*
42. Laura Sabourin (2003). *Grammatical Gender and Second Language Processing: An ERP Study.*
43. Hein van Schie (2003). *Visual Semantics.*
44. Lilia Schürcks-Grozeva (2003). *Binding and Bulgarian.*
45. Stasinos Konstantopoulos (2003). *Using ILP to Learn Local Linguistic* S*tructures*.
46. Wilbert Heeringa (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*.
47. Wouter Jansen (2004). *Laryngeal Contrast and Phonetic Voicing: ALaboratory Phonology*.
48. Judith Rispens (2004). *Syntactic and phonological processing indevelopmentaldyslexia*.
49. Danielle Bougaïré (2004). *L'approche communicative des campagnes de sensibilisation en santé publique au Burkina Faso: Les cas de la planification familiale, du sida et de l'excision*.
50. Tanja Gaustad (2004). *Linguistic Knowledge and Word Sense Disambiguation.*

51. Susanne Schoof (2004). *An HPSG Account of Nonfinite Verbal Complements in Latin.*
52. M. Begoña Villada Moirón (2005). *Data-driven identification of fixed expressions and their modifiability.*
53. Robbert Prins (2005). *Finite-State Pre-Processing for Natural Language Analysis.*
54. Leonoor van der Beek (2005) *Topics in Corpus-Based Dutch Syntax*
55. Keiko Yoshioka (2005). *Linguistic and gestural introduction and tracking of referents in L1 and L2 discourse.*
56. Sible Andringa (2005). *Form-focused instruction and the development of second language proficiency*.
57. Joanneke Prenger (2005). *Taal telt! Een onderzoek naar de rol van taalvaardigheid en tekstbegrip in het realistisch wiskundeonderwijs.*
58. Neslihan Kansu-Yetkiner (2006). *Blood, Shame and Fear: Self-Presentation Strategies of Turkish Women's Talk about their Health and Sexuality*.
59. Mónika Z. Zempléni (2006). *Functional imaging of the hemispheric contribution to language processing.*
60. Maartje Schreuder (2006). *Prosodic Processes in Language and Music.*
61. Hidetoshi Shiraishi (2006). *Topics in Nivkh Phonology.*
62. Tamás Biró (2006). *Finding the Right Words: Implementing Optimality Theory with Simulated Annealing.*
63. Dieuwke de Goede (2006). *Verbs in Spoken Sentence Processing: Unraveling the Activation Pattern of the Matrix Verb.*
64. Eleonora Rossi (2007). *Clitic production in Italian agrammatism.*
65. Holger Hopp (2007). *Ultimate Attainment at the Interfaces in Second Language Acquisition: Grammar and Processing.*
66. Gerlof Bouma (2008). *Starting a Sentence in Dutch: A corpus study of subject- and object-fronting.*
67. Julia Klitsch (2008). *Open your eyes and listen carefully. Auditory and audiovisual speech perception and the McGurk effect in Dutch speakers with and without aphasia.*
68. Janneke ter Beek (2008). *Restructuring and Infinitival Complements in Dutch.*
69. Jori Mur (2008). *Off-line Answer Extraction for Question Answering.*
70. Lonneke van der Plas (2008). *Automatic Lexico-Semantic Acquisition for Question Answering.*
71. Arjen Versloot (2008). *Mechanisms of Language Change: Vowel reduction in 15th century West Frisian.*

72. Ismail Fahmi (2009). *Automatic term and Relation Extraction for Medical Question Answering System.*
73. Tuba Yarbay Duman (2009). *Turkish Agrammatic Aphasia: Word Order, Time Reference and Case.*
74. Maria Trofimova (2009). *Case Assignment by Prepositions in Russian Aphasia.*
75. Rasmus Steinkrauss (2009). *Frequency and Function in WH Question Acquisition. A Usage-Based Case Study of German L1 Acquisition.*
76. Marjolein Deunk (2009). *Discourse Practices in Preschool. Young Children's Participation in Everyday Classroom Activities.*
77. Sake Jager (2009). *Towards ICT-Integrated Language Learning: Developing an Implementation Framework in terms of Pedagogy, Technology and Environment.*
78. Francisco Dellatorre Borges (2010). *Parse Selection with Support Vector Machines.*
79. Geoffrey Andogah (2010). *Geographically Constrained Information Retrieval.*
80. Jacqueline van Kruiningen (2010). *Onderwijsontwerp als conversatie. Probleemoplossing in interprofessioneel overleg.*
81. Robert G. Shackleton (2010). *Quantitative Assessment of English-American Speech Relationships*.
82. Tim Van de Cruys (2010). *Mining for Meaning: The Extraction of Lexico-semantic Knowledge from Text.*
83. Therese Leinonen (2010). *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects.*
84. Erik-Jan Smits (2010). *Acquiring Quantification. How Children Use Semantics and Pragmatics to Constrain Meaning*.
85. Tal Caspi (2010). *A Dynamic Perspective on Second Language Development.*
86. Teodora Mehotcheva (2010). *After the fiesta is over. Foreign language attrition of Spanish in Dutch and German Erasmus Student.*
87. Xiaoyan Xu (2010). *English language attrition and retention in Chinese and Dutch university students.*
88. Jelena Prokić (2010). *Families and Resemblances*.
89. Radek Šimík (2011). *Modal existential wh-constructions*.
90. Katrien Colman (2011). *Behavioral and neuroimaging studies on language processing in Dutch speakers with Parkinson's disease.*
91. Siti Mina Tamah (2011). *A Study on Student Interaction in the Implementation of the Jigsaw Technique in Language Teaching.*
92. Aletta Kwant (2011).*Geraakt door prentenboeken. Effecten van het gebruik van prentenboeken op de sociaal-emotionele ontwikkeling van kleuters.*
93. Marlies Kluck (2011). *Sentence amalgamation.*

94. Anja Schüppert (2011). *Origin of asymmetry: Mutual intelligibility of spoken Danish and Swedish.*
95. Peter Nabende (2011).*Applying Dynamic Bayesian Networks in Transliteration Detection and Generation.*
96. Barbara Plank (2011). *Domain Adaptation for Parsing.*
97. Cagri Coltekin (2011).*Catching Words in a Stream of Speech: Computational simulations of segmenting transcribed child-directed speech.*
98. Dörte Hessler (2011).*Audiovisual Processing in Aphasic and Non-Brain-Damaged Listeners: The Whole is More than the Sum of its Parts.*
99. Herman Heringa (2012). *Appositional constructions.*
100. Diana Dimitrova (2012). *Neural Correlates of Prosody and Information Structure.*
101. Harwintha Anjarningsih (2012).*Time Reference in Standard Indonesian Agrammatic Aphasia*.
102. Myrte Gosen (2012). *Tracing learning in interaction. An analysis of shared reading of picture books at kindergarten.*
103. Martijn Wieling (2012). *A Quantitative Approach to Social and Geographical Dialect Variation*.
104. Gisi Cannizzaro (2012). *Early word order and animacy*.
105. Kostadin Cholakov (2012). *Lexical Acquisition for Computational Grammars.A Unified Model.*
106. Karin Beijering (2012). *Expressions of epistemic modality in Mainland Scandinavian.A study into the lexicalization-grammaticalization-pragmaticalization interface.*
107. Veerle Baaijen (2012). *The development of understanding through writing.*
108. Jacolien van Rij (2012).*Pronoun processing: Computational, behavioral, and psychophysiological studies in children and adults.*
109. Ankelien Schippers (2012). *Variation and change in Germanic long-distance dependencies.*
110. Hanneke Loerts (2012).*Uncommon gender: Eyes and brains, native and second language learners,& grammatical gender.*
111. Marjoleine Sloos (2013). *Frequency and phonological grammar: An integrated approach. Evidence from German, Indonesian, and Japanese*.
112. Aysa Arylova. (2013) *Possession in the Russian clause.Towards dynamicity in syntax.*
113. Daniël de Kok (2013). *Reversible Stochastic Attribute-Value Grammars.*
114. Gideon Kotzé (2013). *Complementary approaches to tree alignment: Combining statistical and rule-based methods*.

115. Fridah Katushemererwe (2013). *Computational Morphology and Bantu Language Learning: An Implementation for Runyakitara.*

GRODIL
Center for Language and Cognition Groningen (CLCG)
P.O. Box 716
9700 AS Groningen
The Netherlands