Fridah Katushemererwe*, Andrew Caines and Paula Buttery

# Building natural language processing tools for Runyakitara

https://doi.org/10.1515/applirev-2020-2004

**Abstract:** This paper describes an endeavour to build natural language processing (NLP) tools for Runyakitara, a group of four closely related Bantu languages spoken in western Uganda. In contrast with major world languages such as English, for which corpora are comparatively abundant and NLP tools are well developed, computational linguistic resources for Runyakitara are in short supply. First therefore, we need to collect corpora for these languages, before we can proceed to the design of a spell-checker, grammar-checker and applications for computer-assisted language learning (CALL). We explain how we are collecting primary data for a new Runya Corpus of speech and writing, we outline the design of a morphological analyser, and discuss how we can use these new resources to build NLP tools. We are initially working with Runyankore–Rukiga, a closely-related pair of Runyakitara languages, and we frame our project in the context of NLP for low-resource languages, as well as CALL for the preservation of endangered languages. We put our project forward as a test case for the revitalization of endangered languages through education and technology.

**Keywords:** natural language processing, endangered languages, language corpus, morphological analyser, CALL

## 1 Introduction

We describe our endeavour to build natural language processing (NLP) tools for a group of west Ugandan Bantu languages, Runyakitara. Orthographies for these languages were formalized no more than 150 years ago, and we have found that spelling errors are common in written Runyakitara, even by students in higher

***Corresponding author: Fridah Katushemererwe**, Department of Linguistics, Makerere University, Kampala, Uganda, E-mail: katu@chuss.mak.ac.ug
**Andrew Caines:** Department of Theoretical & Applied Linguistics, University of Cambridge, Cambridge, UK, E-mail: apc38@cam.ac.uk
**Paula Buttery:** Computer Laboratory, University of Cambridge, Cambridge, UK, E-mail: paula.buttery@cl.cam.ac.uk

education and ‘professional’ users such as journalists. We therefore plan to build a spell-checker, RunyaSpeller, for Runyakitara which will help to improve writing accuracy, underpin general literacy, and support government policies in favour of indigenous language education in Uganda.

Despite healthy speaker numbers (13% of the Ugandan population are thought to speak a Runyakitara language; Lewis et al. 2015) we observe that only a minority of schools in western Uganda provide instruction in Runyakitara, a reality that falls short of the government’s ‘mother tongue’ policy that all schools should do so, and furthermore, adult native speakers of Runyakitara do not necessarily pass on their native language to the next generation. In terms of domains of use, Runyakitara is mostly an ‘at home’ language which on the one hand does not penetrate into official circles but on the other has seen the emergence of newspapers and radio stations. We therefore propose that although speaker numbers may mark the language group as ‘safe’ on any scale, the pictures of intergenerational transmission and domains of use are less healthy. On the scale published by the Catalogue of Endangered Languages[1], for example, Runyakitara would be labelled ‘threatened’ or ‘endangered’ in these two dimensions. The negative trends have been compounded by an acceleration in migration to urban areas, especially the capital, Kampala, and by the continuing dominance of the official languages, Kiswahili and English.

We therefore see a second purpose to this project, which is to develop language learning products to aid the revitalization of endangered languages. These products, referred to as computer-assisted language learning (CALL) systems in the linguistic field, would give Runyakitara native and non-native speakers the opportunity to access teaching materials at any time and in any location, offering autonomous learning and regular feedback, which would either support classroom teaching or indeed stand alone as a mechanism to aid immersive learning. We describe below our previous efforts designing pilot CALL products, and the linguistic resources required to further develop a ‘LearnRunya’ system.

If the outlook for Runyakitara is uncertain, what is more certain is that the Runyakitara languages are under-resourced in terms of NLP resources, and in comparison with the major world languages – above all English. As the initial phase in the project and as a prerequisite for any NLP technology, we are currently collecting a corpus of Runyakitara speech and writing – the ‘RunyaCorpus’ – subsections of which we will make freely available to other researchers. In parallel we will continue to develop our morphological analyser for Runyakitara – ‘RunyaMorph’ (Katushemererwe 2013) – which allows us to parse the internal workings of words in these morphologically rich languages. Both RunyaSpeller and LearnRunya will be built on RunyaMorph as an underlying technology and the

[1] Source: http://www.ethnosproject.org/scale-of-endangerment; accessed 2016-02-17.

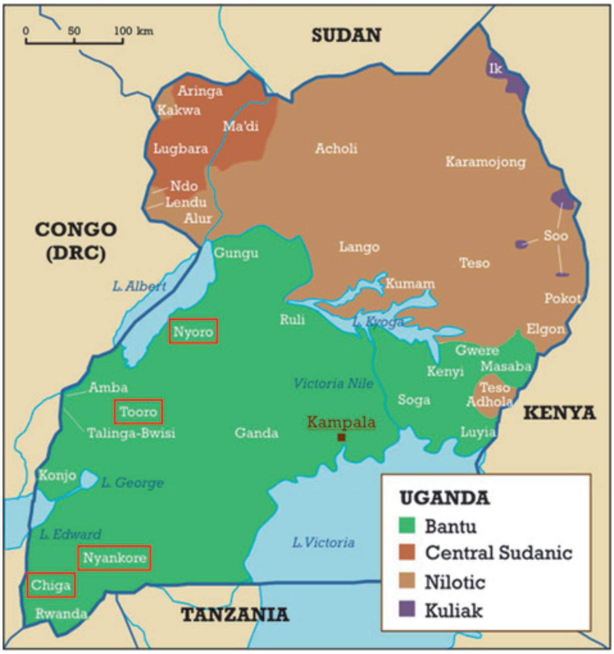

**Figure 1:** Language map of Uganda (source: Wikimedia[2]; annotations added).

RunyaCorpus as a seed set of attested language data, and in a circular design, will in turn feedback corpus and error data to our research project. We hope here to demonstrate, in the spirit of work by others motivating such approaches (Hugo 2015; Ward and Genabith 2003), how computational technology may be applied to a low-resource language for its future survival and the benefit of its users.

## 2 Runyakitara

The name ‘Runyakitara’ does not refer to a single language, but is instead the collective name for a group of four closely-related Bantu varieties spoken in western Uganda – namely, Runyankore, Rukiga, Runyoro and Rutooro (also known as Nyankore, Chiga, Nyoro, and Tooro; see Figure 1). Runyoro was the seed language from which the four distinct varieties grew. The separation into distinct varieties was crystallized by missionaries working in the area, who used different orthographies to publish the Bible in each dialect.

[2] Created by Mark Dingemanse and released under the Creative Commons Attribution 2.5 license: http://commons.wikimedia.org/wiki/File:Languages_of_Uganda.png.

Many missionaries wrote religious materials and spread the gospel in Luganda – the main spoken language in the central Ugandan region, including the capital Kampala. This move encouraged Ugandans to associate Luganda with *The Bible*. This certainly disadvantaged other indigenous languages such as the Runyakitara group (Runyoro as it was then). Subsequently, the arrival of the British and the colonial period (until independence in 1962) encouraged the literacy and culture of indigenous languages, while at the same time further establishing English at the administrative level (Namyalo and Nakayiza 2014). In this period the orthographies of Runyankore and Rukiga were brought together, not without some resistance, with the same being done for Runyoro and Rutooro (Bernsten 1998: 98).

Although the education policies during colonialism focussed on English as the official language, they allowed a bilingual mode of instruction (English alongside the local language) in the early years of primary schooling. The language situation remained diglossic but nevertheless supportive of indigenous varieties until the 1970s, at which point indigenous languages were excluded from schools. Teachers were left to find surreptitious ways of teaching the local languages (while continuing to use English). Students were taught in English throughout their schooling, a factor that has slowed the development of educational materials for Runyakitara, which did not return to the curriculum until the 1990s. However, even now we know little as to the actual implementation of language education policies in Uganda (Ojijo 2012).

Whether or not 'the four Rs' actually developed into distinct languages, rather than 'only' being dialects, has been debated elsewhere and we will not discuss it at length here[3]. Suffice it to say that a clear split was in place at the point where a movement toward re-consolidation began. At two conferences in the 1990s at Makerere University in Kampala, the name 'Runyakitara' was selected for the language group on the basis that its territory spans the former lands of the great medieval-to-modern Kitara Empire. In addition, an eventually successful campaign was begun calling for the teaching of western Ugandan languages at university level.

Although currently comprised of four varieties, there are other closely-related languages that might in future be considered part of Runyakitara group: Haya and Nyambo, spoken in Tanzania, and Hema, spoken in the Democratic Republic of Congo[4]. The work we report on here is based on Runyankore-Rukiga.

[3] For references, see Bernsten 1998.

[4] It should be noted that there is an ongoing debate in Uganda about the use and meaning of the name 'Runyakitara'. Some people believe that the name Runyakitara is outlandish, meaning that it cannot represent the four language varieties appropriately. Others believe that since Bunyoro Kitara was the largest kingdom that ever ruled the land where the languages are spoken, then, the name Runyakitara fits.

**Table 1:** Speaker counts, Runyakitara language group, Uganda.

| Variety | Speakers |
|---|---:|
| Runyankore | 2,330,000[5] |
| Rukiga | 1,580,000[5] |
| Runyoro | 6,67,000[5] |
| Rutooro | 4,88,000[6] |
| *Total* | *5,065,000* |

We intend in due course to extend the project to the Runyoro-Rutooro pairing as well.

The status of the Runyakitara group in terms of written and digital resources is negligible. In 2002 the Runyakitara languages were thought to be spoken by approximately five million people (Table 1), which is more than 20% of the Ugandan population at the time (25 million[7]). If we assume a similar ratio speak the languages now, then recent population figures of over 40 million[8] imply that currently the Runyakitara speech community numbers about eight million. Runyakitara is therefore a sizeable language group in terms of speaker numbers, but it remains under-studied and under-resourced compared to the major European languages and Bantu languages such as Kiswahili and isiZulu.

The most recent descriptive grammars for Runyankore-Rukiga and Runyoro-Rutooro were published in the late twentieth century (Rubongoya 1999; Taylor 1985). These remain the most recent linguistic publications on the Runyakitara language group, to the best of our knowledge, previous to our own body of work presenting computational linguistic analyses of these languages (Katushemererwe 2013; Katushemererwe and Hanneforth 2010; Katushemererwe and Nerbonne 2015).

Runyakitara, like many other indigenous languages in Africa, remains of lower social status than English and Kiswahili, the official languages of Uganda. In the shadow of these major, well-resourced languages, Runyakitara is perceived as a non-literary, vernacular language group, one which does not carry the same economic benefits or social prestige as English and Kiswahili. It is therefore difficult to convince speakers of the benefit of acquiring literacy in Runyakitara, especially in the face of the global spread and cultural dominance of English.

[5] Source: 2002 population census, Uganda Bureau of Statistics.

[6] Source: 1991 population census, Uganda Bureau of Statistics.

[7] Source: World Bank http://data.worldbank.org/indicator/SP.POP.TOTL; accessed 2016-12-15.

[8] Source: http://data.un.org; accessed 2016-12-15.

The present “mother tongue” education policy (Kateeba 2009) – according to which children from primary 1 to 3 (5-8 years) are to be instructed in their local language – is faced with many challenges. One major obstacle in its implementation is a lack of teaching and learning materials in Uganda’s indigenous languages. We aim to bridge this gap for Runyakitara by providing both online and offline language resources for the teaching, learning and general use of Runyakitara.

The language group has interesting linguistic features, which have as yet received only limited attention from linguists barring a few valuable contributions (e. g. Ndoleriire and Oriikiriza 1990; Rubongoya 1999; Taylor 1985). Two such features are a very productive morphology (Katushemererwe 2013) and flexible word order (Ndoleriire and Oriikiriza 1990). Thus the first motivation for this project is simply to obtain better descriptions of this group of languages, which will develop alongside our work preparing the NLP tools.

## 3 Language endangerment and language revitalization

Whilst estimates as to the number of endangered languages vary[9] (let alone the total number of languages in the world[10]), the prospect of future language extinction is clearly a worldwide threat. Many mourn the cultural and scientific losses that would result, and the trend towards a small number of globally-dominant languages is not universally welcomed. It has previously been observed that CALL systems may play a role in language preservation and revitalization (Hugo 2015; Ward 2004; Ward and Genabith 2003), and that is the motivation for our project of building NLP tools for Runyakitara.

Various criteria for what counts as an endangered language are in use, but number of speakers is common to most schemas. By *Ethnologue*’s reckoning, on their extended 13-level version of Fishman’s original “graded intergenerational disruption scale” (EGIDS, see Appendix A; Lewis and Simons 2010), the Runyakitara languages are “educational” or “developing”, meaning they are “in vigorous use” with some recognizable signs of standardization, education and/or literature (Lewis et al. 2015). This places Runyakitara above the level of a

[9] The UNESCO Atlas refers to a ‘generally accepted’ figure of 3000 endangered languages (Moseley 2010). Others paint a more alarming picture: “the coming century will see either the death or doom of 90% of mankind’s languages” (Krauss 1992: 7).

[10] Krauss (1992: 5) accepts an estimate of around 6000.

"threatened" or "shifting" language status by virtue of speaker counts and use in education.

However, speaker counts are not the only determining factor in the health of a language: contexts of *use* – the social or communicative functions for which it is used (Lewis et al. 2015) – are a key consideration, and shifts in domains of use are a concomitant of gradual language shift. For example, there are concerns that Cantonese, though it has 60 million speakers, may become endangered within two generations, as today's children speak only Mandarin at school. Cantonese is generally sidelined at an institutional level by the Chinese government's policy of spreading Putonghua (Mandarin) across the entire country[11].

A language may be potentially "endangered" in more than a numerical sense, therefore. In Africa, many regions continue to suffer the linguistic consequences of European expansion and colonialism. A language may be endangered in more than a numerical sense, therefore. Katushemererwe and Nerbonne (2015) observed that the children of Runyakitara native speakers are not necessarily speaking Runyakitara. It was also noted by Fishman (2000) that people who know a language do not necessarily transmit it.

In the case of Runyakitara, two noticeable groups are emerging: (i) children of native speakers who have emigrated from Runyakitara-speaking regions (Figure 1); (ii) children of native speakers who live in urban and peri-urban centres of the former Kitara region, in places such as Mbarara, Kabale and Rukungiri. The former have only basic knowledge, if any, of speaking, but are generally illiterate in Runyakitara. The latter are unable to learn Runyakitara to a highly proficient level due to the schooling system and home environment which prioritise English. Such a scenario raises questions as to whether Runyakitara will survive in 20 years from now if these trends continue.

We maintain that for Runyakitara speaker numbers may be relatively healthy, but the cultural allure and economic benefit of learning English and Kiswahili are evident and increasing. This fits in with the scenario described by Ward (2004) in which endangerment arises through loss of "social role":

> If the language is not spoken in public and in the education system, its sphere of influence (and usage) is diminished. If children are educated in the locally dominant language (as opposed to their own language), they may not become literate in their own language and come to regard it as not being 'important enough' for education (and other formal activities) (Ward 2004: 346).

[11] Source: http://www.scmp.com/lifestyle/family-education/article/1450856/hongkongers-take-steps-preserve-their-language-and; accessed 2016-12-15.

The threats we perceive to Runyakitara are as follows:

- It is not widely taught at schools. Although it has been introduced to the curriculum in government-funded schools at primary school levels P1–P3 (5–8 years), in private schools local languages are perceived to be detrimental to learning in a system where examinations are administered in English, and in these schools English is generally the medium of instruction. At secondary level, only three schools of more than two hundred approached agreed to pilot the teaching of Runyakitara, and in those schools, few students enrol in language classes.
- There is very little written material to support its use in more formal registers and in contexts other than informal communication among friends and family.
- Speakers' perceptions about its low status mean that younger generations are beginning to abandon it in favour of dominant global languages such as English and Kiswahili (Katushemererwe and Nerbonne 2015).

These problems are especially acute in Uganda, as they are threats particularly associated with urbanization, and Uganda was found to be in the top 10 most rapidly urbanizing countries of the world over the past five years[12]. The proportion of the population who could speak English as a second language was estimated as 10% in the early 2000s (Crystal 2003) but we believe that proportion will have grown in the intervening decade-and-a-half, given the way that the statistic was calculated and the increasing pull of English as a global language[13].

We would therefore counter that *Ethnologue*'s assessment of Runyakitara at "educational" or "developing"status is overly optimistic, if educational use is all that distinguishes these levels from "vigorous" or "threatened" below them (Lewis and Simons 2010, Appendix A). The situation Abidi described in the late 1980s persists:

[12] Source: *The Economist Pocket World in* Figures 2016 *Edition*. London: Profile Books.

[13] Crystal estimated 2.5 million English L2 speakers (2003: 65) and the population of Uganda was 24 million in 2001. Crystal listed a number of sources for a table of global English statistics without stating which (combination) was used for which country: the *UNESCO statistical yearbook, The Encyclopaedia Britannica, Ethnologue,* census reports, or a fallback method of the population over 25 who'd completed secondary schooling for those countries in which English is an official language. We can rule out *Ethnologue* because it cites Crystal (2003) for its English speaker count. To the best of our knowledge, neither *UNESCO* nor *Britannica* have access to an authoritative count of English speakers, and the 2002 population census did not include a question relating to knowledge of English. Therefore we presume that the age-over-25 method was used.

> School children learn more about foreign countries than their own. They have no books in local languages. There is no institution working for the development of Uganda's local language publications. Schools have no priorities to teach local tradition and culture. In this situation, we are preparing our children to think in a foreign language (Abidi 1989: 47).

However, the scenario for Runyakitara is not entirely bleak. The agreement to unify these four western Bantu varieties under the umbrella of "Runyakitara"stemmed from a post-independence movement "toward merger rather than separation" (Bernsten 1998: 98). The people of west Uganda "have a group identity based on their shared history and linguistic code" (Bernsten 1998: 102). Their efforts to strengthen their linguistic identity matches an established definition of 'language revitalization' by "imparting new vigour to a language still in limited or restricted use, most commonly by increased use through the expansion of domains" (Paulston 1994: 92).

Nonetheless a lack of literacy and training continues to affect the development of Runyakitara. The domains of journalism, law and commerce remain closed for the most part to languages other than English and Kiswahili. We propose that CALL applications for Runyakitara can help address these problems by – (a) providing educational tools in view of the lack of school teaching, (b) demonstrating the written form, (c) associating Runyakitara with digital technology thereby increasing its cultural prestige, and (d) reaching out to a global audience. Academic research and the availability of digital technology increase the prestige of languages (Ward and Genabith 2003), hence our project to develop computational tools for Runyakitara may help to counter the sometimes negative views attached to these languages.

Whether or not one agrees that Runyakitara is endangered, it is evidently a low-resource language in terms of computational linguistic research: very few texts have been digitized, and there are no existing corpora, to the best of our knowledge. One aim of this project is therefore to provide some computational linguistic resources for Runyakitara; the other is to demonstrate how such resources might be used for CALL projects.

We aim also to strengthen efforts in place for revitalizing Runyakitara. There are several governmental and non-governmental projects to produce newspapers (e. g. *Orumuri* and *Entatsi*), books, and other literary materials in the Runyakitara languages. There are now Runyakitara-language radio stations such as Radio West and Voice of Kigezi, as well as museums of Kitara culture such as Igongo Cultural Museum, Mbarara. Others meanwhile have used Runyakitara for music composition and drama (Muranga 2009). However, most of these are individual initiatives which would benefit from assistive technology for the accurate use of Runyakitara in writing – namely, spelling and grammar checkers. As a stepping stone to full

CALL systems, we aim to produce such tools using NLP techniques, as outlined below.

## 4 Natural language processing and language learning

NLP is an umbrella term for the practice of passing texts to computers for some purpose, whether that be manipulation of the form of the text, analysis of linguistic features, information extraction, or any combination of the above. Such operations may be used to then present the text in a new way, to 'understand' a text and summarise or respond to it, or simply to report the statistical properties of language in a given domain.

Note that by *natural* language we mean to restrict our focus to the set of all human languages used day-to-day, a rather heterogenous grouping of old and new communication systems, emerging through collaboration and interaction over time, and full of arbitrary – at times mystifying – rules and patterns of use. In contrast, artificial languages as designed for computer programming, mathematic notation, or transmission by a medium other than sound (e. g. Morse code), are usually well defined, relatively unchanging and regular. Hence, given its subject matter, NLP is a great challenge with many problems as yet unsolved.

We also note here that we are dealing with the written form, whether the source of any given text was originally written authorship or transcription from speech. NLP interacts and overlaps with fields which deal with other modalities – automatic speech recognition as addressed by engineers, and the recent development of video corpora in linguistics, to name just two examples – but it is presently concerned with the written form exclusively and indeed works best with texts authored in a canonical register. How NLP tools may be adapted to non-canonical registers such as learner language, dialects, or spoken transcription, is an area of ongoing research.

NLP underpins many familiar and somewhat ubiquitous tools of the computer age, from Internet search engines to predictive typing on mobile phones. Any such applications rely on a core of mature NLP technology. Some examples are given below.

- *Tokenisation and lemmatisation*: often the first step in any NLP pipeline, tokenisation involves splitting a text up into *tokens*, something akin to *words* (but sometimes more/less than what we think of as a word, for various reasons). In Latin alphabet languages, for instance, the task is relatively straightforward thanks to the conventional use of whitespace between words (*cf*. Chinese). Even so, compounds, proper names, and abbreviations cause headaches and require principled decision in designing a tokeniser: for example, is “I’m” one token or two? How about “ice cream”, “touch type” and “big brother”? Or the institutions, fictional or otherwise, known as “Houses of Parliament”, “Big Brother” and “Heart of Midlothian”? These latter examples touch on the general task known as ‘named entity recognition’, an NLP field of its own. Lemmatisation is another fundamental task that involves returning any given token to its basic state: *i. e.* its lemma. The resulting form may or may not resemble a real word: thus “cats” and “cat” would be lemmatised to the stem “cat”, while “written”, “writes”, “write” and “writing” would all be lemmatised to the stem “writ”.

- *Tagging and parsing*: these are often essential annotation tasks, whether to have a better description of a language, or as foundations on which to build further applications. Tagging involves associating each token in a text with a part-of-speech label. Various ‘tagsets’ have been designed, mainly for English, with variation as to their granularity and flexibility. For instance, frequent irregular verbs such as “have” may be allocated a distinct subset of verb tags (*e. g.* VHB: base form of ‘have’), or they may be subsumed under a more general verb class (*e. g.* VB: verb in base form). Parsing is the process of identifying morpho-syntactic structure in texts. For English, parsing usually involves syntactic analysis only, but for morphologically-rich languages such as Turkish, Hungarian and Korean, analysis of word-internal structure is essential.
- *Error detection and correction*: automatic error detection underpins the spell-checker and auto-correct tools that most readers will be familiar with from their use of personal computers and mobile phones. Detection often goes hand-in-hand with an attempt at automated error correction. Such technologies rely on a large corpus of high quality (*gold standard*) data – which may be a dictionary, or a very accurate collection of texts. To a certain extent error detection and correction requires a superficial ‘understanding’ of language and ability to disambiguate word senses, in that context and register may be crucial factors. For example, one might wish to correct the use of “yeah” to “yes” in a newspaper article, whereas we might leave it unaltered in fictional dialogue. Similarly, the spelling error “hael” is more likely to stand for “hall” if the surrounding text concerns interior design, but “hail” if the context is meteorological.
- *Language model*: a common task in NLP is to build a language model, a computational description of a given language which may be used to assign probabilities to an unseen strings of words, or to predict the next word given a truncated string. This technology underpins text message typing, or famously,

> Professor Stephen Hawking's communication system[14], and is thought to have correlates with human language processing (Hale 2001; Levy 2008; Yngve 1960). The most common type of language model, to such an extent that it is often referred to as *the* language model, is the *n*-gram model, which models language as chunks of words *n* items long. The reader may well be familiar with the Google Books Ngram Viewer, for example – an interactive web application that allows the user to visualise trends in *n*-gram frequencies over time[15].

These are the NLP tasks we work with in this project and will refer to below. Note that there are a multitude of other computational linguistic tasks involving NLP technology. A selection include the process of converting a text from one language to another (machine translation), the identification of positive or negative import in texts (sentiment analysis), and the ability to group texts into discrete groups according to some given criteria (document classification).

What all these tasks have in common, along with the empirical study of linguistics in general, is that we want to know something about language in use. We stop short of saying that NLP technology is *objective*, as the biases, conscious or otherwise, of the author(s) may well be engrained in the programming of the software. However, what NLP technology instead can contribute is a systematic and consistent way of running linguistic analyses over large numbers of texts. Those interested in reading more about computational linguistics as a discipline, and the NLP tools it gives rise to, are encouraged to refer to the textbooks by Manning and Schütze (1999), and by Jurafsky and Martin (2009).

We envisage the use and development of NLP tools for Runyakitara as a way to enable improved description of the languages, and an opportunity to support language preservation and revitalization through computational access to writing support and CALL systems. These systems support a wide range of language learning activities that enhance listening, speaking, reading and writing. A CALL system presents the learner with language materials in an appropriate medium, regularly assesses the learner's progress, and ideally adapts subsequent teaching to address the learner's shortcomings.

CALL systems can therefore generate useful immediate feedback for the learner (Meurers 2012; Nagata 2009) and provide instrumental writers' aids such as spell-checkers and grammar checkers (Shalaan 2005). Given all these benefits of NLP to language learning, it is clear that NLP tools ranging from spell-checkers to

[14] Source: http://blog.swiftkey.com/swiftkey-reveals-role-professor-stephen-hawkings-communication-system; accessed 2016-12-15.

[15] See: http://books.google.com/ngrams.

fully-developed CALL systems will be invaluable to the Runyakitara language group as a way to support classroom teaching and to reach learners who would not otherwise enrol on a language course.

## 5 The Runya Corpus

There are a number of good reasons to build a corpus for linguistic research. First there is the opportunity to inspect how language is used in various contexts and, sometimes, over many occasions. For instance, take the Runyankore-Rukiga pairing (RN-RK) preposition *omu*. From a concordance of instances of its use we can see that this single form stands in the place of four English prepositions (1)–(4).

(1) omu kitabo eki
‘in this book’

(2) endwara ezirikuruga omu burofa
‘diseases that come from dirt’

(3) n’okukoragye omu bantu baitu
‘doing well among our people’

(4) omu mpuku
‘inside the cave’

Secondly, we may obtain frequency statistics for a given language in the domain and communication mode represented by the corpus. On the basis of these first two points, if carried out comprehensively, one may compile grammatical descriptions and pedagogical materials (Barlow 1996; Leech 2000; O’Keeffe et al. 2007). Textbook authors may create exercises based on actual usage, whilst for classroom activities, a corpus is a tool for hands-on language analysis activities, enabling learners to draw their own conclusions about language use.

Finally, and as is the case with the project described here, one may apply computational linguistic methods to build a language model for use in NLP applications (Manning and Schütze 1999). There is no ready-made corpus of the Runyakitara languages, to the best of our knowledge. For this reason, a prerequisite of our work is to build such a corpus, keeping in mind that we are initially working with the RN-RK rather than Runyoro-Rutooro (RN-RT).

The size of our corpus is unbounded and collection efforts will be ongoing: in answer to the question, ‘how much is enough?’, the easy answer is, ‘the more the better’, whilst a more nuanced response takes into account the purpose for which the corpus is needed. For example, our first task is to build a spell-checker, and for this one may set to work with a relatively small corpus; much more data is required for a fully-functional CALL system, an outcome that remains our long-term aim. We are building the Runya Corpus, which we eventually envisage as a collection of spoken and written Runyakitara languages, and from which we will make as much data freely available as copyright restrictions allow.

Initially, we are collecting RN-RK written texts from a range of sources: newspapers, publicly available official documents, student essays, and *The Bible*. Each source is described in greater detail below. Eventually, whilst the other text types fall under copyright restrictions, we plan to make the public documents and student essays freely available to other researchers. The essays will be available in two forms: the original texts and with error annotations. Meanwhile, plans to collect and transcribe speech recordings are in place.

Our immediate intention is to use the corpus to construct spelling and grammar checkers, with the corpus acting as an immediate lexicon on which to base the spell-checker. We may then use the uncorrected student essays in the corpus as the test items with which to evaluate the accuracy of our spell-checker. We also hope that researchers will make use of the Runya Corpus for general tasks in the area of low-resource computational linguistics (Abney and Bird 2010; Agic et al. 2015; Allwood et al. 2010; Emerson et al. 2014).

## 5.1 Public documents

The corpus includes a number of public documents issued in one of Uganda's official languages (English, Swahili) and translated into RN-RK by Makerere University's Center for Language and Communication Services. The documents include an advice booklet for farmers, a nutrition guide, a brochure on higher education, legal documents, and manuals disseminated by non-governmental organisations (NGOs). These documents are in the public domain and are aimed at a general readership.

## 5.2 Student essays

We collected 14 essays written in RN-RK by students who completed the undergraduate course in Runyakitara at Makerere University. Essays were on the topic of their interest, mainly concerning their everyday experiences. The aim was to collect samples of RN-RK as currently used by the Banyakitara ('people of Kitara').

The essays have been annotated according to the XML schema designed for the Cambridge Learner Corpus (henceforth, 'CLC'; Nicholls 2003), such that, for example, errors of morphological form are marked 'F' (5), spelling errors are indicated 'S' (6), and missing tokens are an error of type 'M' (7). Error zones are demarcated by <NS> tags, whilst <i> tags wrap around the error (where it exists) and <c> tags show the correction (where necessary).

(5) <NS type="F"><i>bweeye</i><c>bwe</c></NS>
['bwe' here means 'his'; 'bweeye' is a non-standard dialect form]

(6) <NS type="S"><i>omujaija</i><c>omushaija</c></NS>
['omushaija' means 'man']

(7) <NS type="M"><c>aho</c></NS>
['aho' means 'there', but in this context plays a conjunctive role]

Nicholls (2003) describes the list of possible error types in full, many more than the three exemplified here. In addition, we introduce an additional error type 'O' for out-of-vocabulary tokens which are not known in RN-RK and for which a correction is unclear (8).

(8) <NS type="O"><i>amatukyire</i></NS>

After the release of a CLC subset (Yannakoudakis et al. 2011), its use has been demonstrated for various computational linguistic projects, not least among them the CoNLL-2014[16] shared task on automatic error detection and correction (Ng et al. 2014). We hope that making an error-annotated corpus of Runyakitara available will allow similar projects to take shape around RN-RK.

## 5.3 The Bible

The Bible was translated into RN-RK by the Bible Society of Uganda. We intend to transform as much of the hard copy as possible into digital form. To do so is a rather laborious process involving a scanner and optical character recognition (OCR) software. We have begun with two books of *The Bible*: one from the Old Testament, one from the New – namely, *Genesis,* and *The Gospel according to Mark* – and we are using the Tesseract OCR engine (Smith 2007). Texts from *The Bible* have been targeted because they are among the most reliable in grammatical accuracy for the Runyakitara languages.

[16] Conference on Natural Language Learning 2014.

### 5.4 Newspaper articles

*Orumuri* and *Entatsi* publish news articles in RN-RK on current affairs topics such as politics, business, public health and sports. *Orumuri* is a government newspaper published weekly by the Vision Group. *Entatsi* is published by Entatsi Publications, a private registered company. We will select newspaper articles for inclusion in the corpus so that there is a representative sample of topic types. Hard copy pages will be scanned and the text extracted using Tesseract OCR, as with *The Bible*.

### 5.5 Speech transcriptions

Speech transcriptions will come from recordings of RN-RK speakers in day-to-day communication: meetings, narratives and interviews, for example. The recordings will be transcribed in the RN-RK orthography, but we will record equal numbers of Runyankore and Rukiga speakers, a fact that allows for future typological analyses.

## 6 RunyaMorph: a morphological analyser for Runyakitara

One of the tasks in this project is to utilise an existing morphological analyser of Runyakitara (‘RunyaMorph’, formerly known as ‘RUMORPH’; Katushemererwe 2013), developed with the RN-RK lexicon, to build natural language processing tools for Runyakitara. RunyaMorph accepts a Runyakitara word as input, analyzes its internal structure using a finite-state machine[17] (Hanneforth 2009), and outputs a statement of linguistic and morphological information, as shown in examples (9)–(16).

(9) atwekire : a[VERB_PREF_SPM3SSpm3s=agrmt3s]twek[VERB_ROOT_SIMPLE
Simple=simpleverb]ire[VERB_END_PAST Pend=nearpast]
‘he/she impregnated’

[17] A machine that can be in one of a finite number of states at any given time, with transitions between those states being determined by triggers from the input; for more information, see Kornai (1999).

(10) ou : ou[DEM_PR_CLASS12/14]
'the one'

(11) Isingiro : Isingiro[PNAME]
'Isingiro'

(12) mbwenu : mbwenu[ADVERB]
'now'

(13) ogamba: o[VERB_PREF_SPM2SSpm2s=agrmt2s][VERB_PREF_PRESENT Present=habitual]gamb[VERB_ROOT_SIMPLESimple=simpleverb]a[VERB_END_IND Ind=mood]
'you say'

(14) aba : aba[DEM_PR_CLASS12/14]
'these'

(15) kwonka : kwonka[ADJECTIVE_ROOT]
'it alone'

(16) kwonka : kwonka[ADVERB]
'only'

In the analyser, a lot of information is embedded which can be used in other levels of linguistic analysis. For instance, a[VERB_PREF_SPM3SSpm3s = agrmt3s] means that *a* is a verb prefix marking subject of class 3 (VERB_PREF_SPM3SSpm3s), as well as an agreement marker, in this case of singular number (agrmt3s). This information can be used for part-of-speech tagging and noun class tagging, for example. In this project, we plan to use RunyaMorph in corpus annotation.

The complexity of Runyakitara morphology merits special attention. Our own study revealed that the morphology of RN-RK is complex, demonstrated by the regular verb *shoma* ('read') which was found to yield 40 million possible forms thanks to its sequential combination of root plus various verb extensions (10), endings (3), post-final affixes (3), affixes (2), tense-aspect markers (5), polarity markers (2), aspect markers (2), one further polarity slot (2), and three possible noun phrase complement classifiers – of which there are 18, therefore $18^3$ (Katushemererwe 2013).

The fact that a single verb paradigm is so large implies that the 4000 verbs we have extracted from an RN-RK dictionary could potentially yield 160 billion forms, without even accounting for the reduplication feature common to RN-RK verbs. It is unfeasible to consider a human compiling so much information. Even working 24 h a day all year round without break, rest or distraction over a career of, say, 30 years, would entail compiling more than six hundred thousand verb forms per hour. Evidently this is an unlikely scenario, and hence we have created RunyaMorph in order to annotate large numbers of verb forms in a relatively short space of time. RunyaMorph recognises 70% of real world texts, with a precision of 90% (Katushemererwe 2013)[18]. The system provides a starting point for Runyakitara NLP tools, in building spell-checking and language learning systems.

## 6.1 RunyaSpeller

We will make use of RunyaMorph to in turn develop a spell-checker for RN-RK, 'RunyaSpeller'. The intention is broadly to combine the output from RunyaMorph, of the form shown in (9)–(16), with an *n*-gram language model extracted from the RunyaCorpus of speech and writing. This means that for a given sequence of words we can attempt to automatically identify errors and, based on our corpus and morphological data, suggest corrections such that the incorrect chunk becomes an acceptable *n*-gram.

Naturally, various complications arise from the considerable morphological richness of RN-RK whereby morphology affects almost every other linguistic aspect. For instance, the word *baitu* in the phrase, *abaana baitu* ('our children') is pronounced /bi:tu/. Thus in some texts we find it written phonetically as *biitu*. However, because the prefix *ba-* has a unique meaning ('people' class), changing it to *bi-* ('objects' class) causes the word's meaning to change. This and many other morphologically-conditioned examples mean that RunyaMorph will be an invaluable resource for spell-checking.

One complication is that the RN-RK writing system features a phenomenon that RunyaMorph does not cater for: contraction. In most cases, prepositions and conjunctions are contracted with nouns to form a single word. For instance, *ni* and *omwana* are combined as *n'omwana* ('is a child'), whilst *na* and *omwana* are written as *n'omwana* ('and a/the child'). But RunyaMorph analyses only words, not phrases, and so would not be able to distinguish between these homographic phrases. Instead, context of use is required. This is where the distributional statistics contained in our *n*-gram language model will predict which sense is more likely in any given context.

We have demonstrated that neither the corpus nor morphological analyser are able to provide a reliable RN-RK spell-checker independently. The corpus does not have all the possible word formations that may result from the conjugations of

[18] Limitations of space do not allow a full discussion here, but we refer the interested reader to Katushemererwe and Baguma (2012) for an analysis of RunyaMorph's non-recognition of text.

roots and affixes understood by the analyser, while the morphological parser is not able to perceive the underlying form of contracted phrases. The two resources therefore complement each other and together constitute a system more than the sum of its parts.

One of the objectives of this project is to use NLP to support the learning of Runyakitara. By building a spell-checker – and eventually a grammar checker – we aim to directly support learners and writers of RN-RK. The essays collected already for the RunyaCorpus reveal that the long vowel in RN-RK causes some problems. Consider the short and long vowel pairs in (17)–(19):

(17) a. okuhiga, 'to motivate';
b. okuhiiga, 'to hunt'.

(18) a. okubuza, 'to lose (something)';
b. okubuuza, 'to ask'.

(19) a. okurira, 'to cry';
b. okuriira, 'to eat from'.

From what we have observed, essay-writers confuse these short and long vowel alternates. RunyaSpeller will be able to correct such errors, given enough context, and for this reason it is the first step along the road to a full-blown language teaching system.

## 6.2 LearnRunya

Two pilot systems for CALL have already been developed using RunyaMorph as a cornerstone. One, RU-CALL, was intended for speakers to learn (or indeed re-learn) RN-RK (Katushemererwe and Nerbonne 2015) and the other, an 'intelligent' CALL system, or 'iCALL', was aimed at more advanced learners of RN-RK (Katushemererwe and Hurskainen 2011). Our main objective in creating RU-CALL was to provide a digital learning environment to enable learners to enhance their grammatical mastery and the acquisition of writing skills. We had in mind the many children of Runyakitara heritage living in non-Runyakitara speaking areas who have very limited proficiency in the heritage language.

In trials of RU-CALL it was found that the system facilitated the learning of Runyakitara writing skills. It also provided an opportunity for Runyakitara learners to use CALL software for the first time, an experience which they found to be motivating. The major weakness of RU-CALL is that it does not have English translations, meaning that learners who have completely lost their mother tongues cannot benefit because of the lack of an accessible meta-language.

Previously we had developed iCALL for Runyakitara (Katushemererwe and Hurskainen 2011) to demonstrate that linguistic knowledge encoded in the Runyakitara morphological analyser provides a sound basis for CALL systems treating aspects of syntax – especially syntactic concord and word order. Attempting to model Runyakitaran syntax using RunyaMorph indeed taught us two key lessons. First, we were able to detect the different forms of ambiguity in RunyaMorph: class ambiguity and part-of-speech ambiguity, as in (20)–(23).

(20) yangye : ya[POSS_PRON_4]ngye
'mine' (possessive pronoun class 4)

(21) yangye : ya[POSS_PRON_9]ngye
'mine' (possessive pronoun class 9)

(22) kiniga : ki[N_7]niga
'anger'

(23) kiniga : ki[V_7]niga
'it strangles'

The RunyaMorph output in (20)–(21) shows the class ambiguity for *yangye* between noun class 4 and class 9. Meanwhile (22) and (23) demonstrate part-of-speech ambiguity with *kiniga* being either a noun or verb.

Secondly, agreement in Runyakitara, as with other Bantu languages, is not a straightforward phenomenon. All noun constituents such as pronouns, verbs, adjectives and numerals must agree with the head noun. For instance, in the RN-RK phrase, *abaana bangye babiri bariho* ('my two children are present'), *ba* is an agreement marker for the head noun *abaana* ('children'). Thirdly, word order in Runyakitara is quite flexible, making it difficult to construct a comprehensive system of word order rules. We elected to adopt the basic RN-RK word order as used in normal language. For example, a modifier of the noun, such as an adjective, possessive pronoun, demonstrative pronoun and numeral, follows the noun in the noun phrase (see Katushemererwe and Hurskainen (2011) for description of the guided tutor and interactive dialogues on agreement and word order).

Despite the difficulties encountered, the major lesson drawn is that the morphological analyser 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 Bantu languages and specifically the Runyakitara group.

Our ongoing intention is to develop a further CALL application, 'LearnRunya', based on what we have learned from the pilot systems, and with which we would target a broader spectrum of learners. In particular, we would trial English translations alongside Runyakitara content, anticipating a dual benefit: Banyakitara (members of the cultural group) whose first language is now English can benefit from learning Runyakitara, while Banyakitara who do not know English can, with guidance, learn by comparing the L1 material with the English translations. The intention is that students can use LearnRunya in a self-guided fashion – without intervention from a teacher. Nevertheless the system could still be employed in the classroom, enabling 'blended learning' – the combination of digital media and traditional teaching methods (McCarthy 2016) – to take place. We believe that this design would be of great appeal in the Ugandan context which is highly multilingual with English as the official language.

# 7 Summary

In this article we have given an overview of our project to build NLP tools for Runyakitara. We explained why we perceive the Runyakitara languages to be at risk, despite relatively large speaker populations, and the ways in which our efforts might help to support revitalization of the language group. Specifically, we intend to build a spell-checking tool – RunyaSpeller – and a language teaching system – LearnRunya. Both of these will be freely available and designed to be mobile-friendly, thereby adapting to the culture of high mobile ownership and poor broadband infrastructure in Uganda (Rice et al. 2009). For reasons detailed above, both of these computational systems require a corpus of Runyakitara data – the RunyaCorpus – which we are currently constructing from spoken and written sources, and both will benefit from our previous work in creating RunyaMorph, a morphological analyser for these highly inflected languages.

Our project has begun with the Runyankore-Rukiga language pairing, and we leave open the possibility of extending our work to Runyoro-Rutooro in future. To a certain extent, the initial RN-RK work involved in this project allows for relatively straightforward transfer to RN-RT. For instance, the overlap between orthographies means that RunyaMorph, designed for RN-RK, can already analyse 50% of RN-RT words (Katushemererwe 2013). With a little adaptation, it could be made more reliable and in turn used to build NLP tools for RN-RT specifically.

Finally, we emphasise that our work may also be taken as a test case in the creation of NLP tools for low-resource languages, and for the benefit of those languages believed to be endangered by the continuing dominance of the world's major languages. We very much welcome contact and interaction with others working on similar tasks, and believe that our efforts can only be made more streamlined and successful with many eyes looking at the same problems.

# Appendix A: Expanded graded intergenerational disruption scale (EGIDS), from Lewis and Simons (2010)

| Level | Label | Description | UNESCO |
|---|---|---|---|
| 0 | International | The language is used internationally for a broad range of functions. | Safe |
| 1 | National | The language is used in education, work, mass media, government at the nationwide level. | Safe |
| 2 | Regional | The language is used for local and regional mass media and governmental services. | Safe |
| 3 | Trade | The language is used for local and regional work by both insiders and outsiders. | Safe |
| 4 | Educational | Literacy in the language is being transmitted through a system of public education. | Safe |
| 5 | Written | The language is used orally by all generations and is effectively used in written form in parts of the community. | Safe |
| 6a | Vigorous | The language is used orally by all generations and is being learned by children as their first language. | Safe |
| 6b | Threatened | The language is used orally by all generations but only some of the child-bearing generation are transmitting it to their children. | Vulnerable |
| 7 | Shifting | The child-bearing generation knows the language well enough to use it among themselves but none are transmitting it to their children. | Definitely endangered |
| 8a | Moribund | The only remaining active speakers of the language are members of the grandparent generation. | Severely endangered |
| 8b | Nearly extinct | The only remaining speakers of the language are members of the grandparent generation or older who have little opportunity to use the language. | Critically endangered |
| 9 | Dormant | The language serves as a reminder of heritage identity for an ethnic community. No one has more than symbolic proficiency. | Extinct |
| 10 | Extinct | No one retains a sense of ethnic identity associated with the language, even for symbolic purposes. | Extinct |

# References

Abidi, Syed. 1989. Modern communication and national identity: An issue in East African context. In Jude J. Ongong'a & Kenneth R. Gray (eds.), *Bottlenecks to national identity: Ethnic cooperation towards nation building*. Nairobi: Professor World Peace Academy of Kenya.

Abney, Steven & Steven Bird. 2010. The human language project: Building a universal corpus of the world's languages. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, 88–97. Uppsala, Sweden.

Agic, Zeljko, Dirk Hovy & Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd annual meeting of the association for computational linguistics*, 268–272. Beijing, China.

Allwood, Jens, Harald Hammarström, Andries Hendrikse, Mtholeni N. Ngcobo, Nozibele Nomdebevana, Laurette Pretorius & Mac van der Merwe. 2010. Work on spoken (multimodal) language corpora in South Africa. In *Proceedings of the seventh international conference on language resources and evaluation*, 885–889. Valletta, Malta.

Barlow, Michael. 1996. Corpora for theory and practice. *International Journal of Corpus Linguistics* 1(1). 1–37.

Bernsten, Jan. 1998. Runyakitara: Uganda's 'new' language. *Journal of Multilingual and Multicultural Development* 19(2). 93–107.

Crystal, David. 2003. *English as a global language*, 2nd edn. Cambridge: Cambridge University Press.

Emerson, Guy, Liling Tan, Susanne Fertmann, Alexis Palmer & Michaela Regneri. 2014. Seedling: Building and using a seed corpus for the human language project. In *Proceedings of the 2014 workshop on the use of computational methods in the study of endangered languages*, 77–85. Baltimore, MD.

Fishman, Joshua. 2000. The status agenda in corpus planning. In Richard D. Lambert & Elana Shohamy (eds.), *Language policy and pedagogy*, 43–52. Philadelphia: John Benjamins.

Hale, John. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the association for computational linguistics (NAACL)*.

Hugo, Russell. 2015. Endangered languages, technology and learning: Immediate applications and long-term considerations. In Mari Jones (ed.), *Endangered languages and new technologies*, 95–112. Cambridge: Cambridge University Press.

Jurafsky, Daniel & James H. Martin. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, 2nd edn. Upper Saddle River, NJ: Prentice Hall.

Kateeba, Connie. 2009. *The thematic curriculum: Implications for mother tongue education in Uganda*. Kyambogo, Uganda: National Curriculum Development Centre.

Katushemererwe, Fridah. 2013. *Computational morphology and Bantu language learning: An implementation for Runyakitara*: University of Groningen PhD thesis, Groningen, The Netherlands.

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

Katushemererwe, Fridah & John Nerbonne. 2015. Computer-assisted language learning (CALL) in support of (re-)learning native languages: The case of Runyakitara. *Computer Assisted Language Learning* 28(2). 112–129.

Katushemererwe, Fridah & Rehema Baguma. 2012. RUMORPH: The morphological analyzer of Runyakitara: Approach, results and issues. In *Proceedings of the 8th annual international conference on computing & ICT research*, 269–294. Kampala, Uganda. http://cit.mak.ac.ug/iccir/?p=iccir_12 (accessed 15 December 2016).

Katushemererwe, Fridah & Arvi Hurskainen. 2011. Intelligent language learning model: Implementation on Runyakitara. In *Proceedings of the 7th annual international conference on computing & ICT research*, 426–444. Kampala, Uganda. http://cit.mak.ac.ug/iccir/?p=iccir_11 (accessed 15 Dec 2016).

Kornai, Andras (ed.). 1999. *Extended finite state models of language*. Cambridge: Cambridge University Press.

Krauss, Michael. 1992. The world's languages in crisis. *Language* 68(1). 4–10.

Leech, Geoffrey. 2000. Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning* 50(4). 675–724.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3), 1126–1177.

Lewis, M. Paul & Gary F. Simons. 2010. Assessing endangerment: Expanding fishman's GIDS. *Revue Roumaine de Linguistique* 55(2). 103–120.

Lewis, M. Paul, Gary F. Simons & Charles D. Fennig (eds.). 2015. *Ethnologue: languages of the world*, 18th edn. Dallas: SIL International. http://www.ethnologue.com (accessed 2016-12-15).

Manning, Christopher & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

McCarthy, Michael (ed.). 2016. *The Cambridge guide to blended learning for language teaching*. Cambridge: Cambridge University Press.

Meurers, Detmar. 2012. Natural language processing and language learning. In Carol A. Chapelle (ed.), *Encyclopedia of applied linguistics*. Oxford: Blackwell.

Moseley, Christopher (ed.). 2010. *Atlas of the world's languages in danger*, 3rd edn. Paris: UNESCO Publishing. http://www.unesco.org/culture/en/endangeredlanguages/atlas (accessed 15 December 2016).

Muranga, Manuel. 2009. *'What about our mother tongues? Linguistic patriotism and non-patriotism in Uganda: Some observations, reflections and recommendations'. Inaugural professorial lecture*. Kampala: Makerere University.

Nagata, Noriko. 2009. Robo-Sensei. *CALICO Journal* 26(3). 562–579.

Namyalo, Saudah & Judith Nakayiza. 2014. Dilemmas in implementing language rights in multilingual Uganda. *Current Issues in Language Planning* 16(4). 409–424.

Ndoleriire, Oswald & Celestino Oriikiriza. 1990. *Runyakitara studies*. Unpublished manuscript. Uganda: Makerere University.

Ng, Hwee Tou, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto & Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error detection. In *Proceedings of the 18th conference on computational natural language learning*, 1–14. Baltimore, MD.

Nicholls, Diane. 2003. The Cambridge learner corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the corpus linguistics 2003 conference*, 572–581. UCREL technical paper number 16: Lancaster University.

O'Keeffe, Anne, Michael McCarthy & Ronald Carter. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Ojijo, Pascal. 2012. Review of education policy in Uganda. Paper submitted to young leaders' think tank on policy alternatives in Uganda. http://www.slideshare.net/ojijop/review-of-education-policy-in-uganda (accessed Dec 15, 2016).

Paulston, Christina. 1994. *Linguistic minorities in multilingual settings: Implications for language policies*. Amsterdam: John Benjamins.

Rice, Andrew, Paula Buttery, Idris A. Rai & Alastair Beresford. 2009. *Language learning on a next-generation service platform for Africa. Africa perspective on the role of mobile technologies in fostering social and economic development*. Maputo, Mozambique: Worldwide Web Consortium Workshop.

Rubongoya, L. T. 1999. *A modern runyoro-rutooro grammar*. Cologne: Rüdiger Köppe Verlag.

Shaalan, Khaled. 2005. An intelligent computer-assisted language learning system for Arabic learners. *Computer Assisted Language Learning* 18(1-2). 81–108.

Smith, Ray. 2007. An overview of the tesseract OCR engine. In *Proceedings of the 9th IEEE international conference on document analysis and recognition*, 629–633. Brazil: Parana.

Taylor, Charles. 1985. *Nkore-Kiga*. London: Croom Helm.

Ward, Monica. 2004. The additional uses of CALL in the endangered language context. *ReCALL* 16(2). 345–359.

Ward, Monica & Joseph van Genabith. 2003. CALL for endangered languages: Challenges and rewards. *CALL Journal* 16(2-3). 233–258.

Yannakoudakis, Helen, Ted Briscoe & Ben Medlock. 2011. A new dataset and method for automatically grading ESOL text. In *Proceedings of the 49th annual meeting of the association for computational linguistics*, 180–189. Portland, Oregon.

Yngve, Victor. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5). 444–466.