Wiring Kenyan Languages for the Global Virtual Age: An audit of the Human Language Technology Resources

Edward O. Ombui[1] and Lawrence Muchemi[2]
[1] Computer Science Dept. Africa Nazarene University (Kenya)
[2] School of Computing and Informatics, University of Nairobi (Kenya)

Abstract

Whereas we recognize the advancement of computing and internet technologies over the years and its impact in the areas of health, education, government, etc., there is increasing cognizance that the languages used in these technologies will have a far reaching impact in terms of accessibility and usability by a wider audience. European languages and specifically English is considered the lingua franca of computing and the Internet due to the vast amount of language resources available in these languages. Does this therefore exacerbate the language and technology gap, especially in regards to African languages?

This research is motivated by this question and begins to tackle a strand of the overarching language technology issue by auditing the human language technologies for Kenyan languages. The research uses the Basic Language Resource Kit (BLARK) to do the inventory. This method has been successfully used to conduct language resources surveys in other countries.

Introduction

Discourse over the use of local languages as the primary instructional language for early childhood education and generally pre-primary education has incrementally taken center-stage in Education policy matters in Kenya with renewed advocacy coming from the ministry of education (Omwenga, 2014). Moreover, research indicates that the use of local languages as a medium of instruction in the formative years offers many advantages because it ensures the pupils have seamless learning at the home and school environments. (Piper, 2010). The use of local languages as media of instruction in education and development in Africa is widely acknowledged by several African governments (Prah, 2002: 9). However, implementation has not been without challenges, including the lack of national policy on language and technology, and competition from highly resourced languages like English, French, Portuguese, Arabic, etc. Besides, English is the most highly resourced language and it is considered the lingua-franca of internet technology. These languages have been well documented over the years with properly defined grammars, unlike the local languages most of which have poorly defined and undocumented grammars and dictionaries (Ombui & Wagacha, 2007).

Advances in technology over the years including higher processing speeds and larger memory capacity of computers have facilitated further research in the application of computers to support decision-making in various disciplines and domains ranging from health, security, agriculture, education, governance etc. In all these areas, data capturing, processing, storage and final display must be in a language that is meaningful and understood by human beings. Otherwise, there will be a gap between the computing technology and its users. This language gap in technology has led to more specific research under the Human Computer Interaction discipline to address challenges in hardware and software technologies in respect to the use of human languages.

Language technologies research basically falls under natural language processing which is subsequently informed mainly by computer science and linguistic efforts. However, these two efforts are often sub optimized from the beginning as each group works independently in pushing their

agenda at the expense of the overarching objective of advancing language resources that are informed by and interdependent on both efforts (Grover *et al*., 2009, Binnenpoorte et al., 2002) . These resources include but are not limited to text and speech corpora, lexicons, machine translation, machine learning, speech recognition and synthesis, text summarization, part-of-speech taggers, parsers, annotators, data collection and management tools.

In order to get a good gist of the current work undergoing in terms of Human Language Technology (HLT), a research into existing projects was inevitable. A HLT audit was considered a good starting point for both ongoing and future research into HLT in Kenya.

This research therefore explored the existing research efforts in the area of natural language processing technologies for Kenyan languages and the existing local content on the Internet.

This paper starts by indicating the related work, the methods and tools used to conduct the audit, the results obtained, a conclusion and recommendation for the way forward in regard to the future of HLT in Kenya.

Related Work

Research efforts into existing Human Language Technologies are not new because some countries have already done their Human Language Technology-audits. Some of the well documented HLT audits include: the Dutch HLT survey (Binnenpoorte et al., 2002), Bulgarian (Simov et al., 2004), Arabic HLT (Maegaard et al.: 2006, 2009), Swedish (Elenius et al., 2008), and South African HLT (Grover et al., 2010).

The Dutch HLT survey, just like the other surveys, used the Basic Language Resource Kit (BLaRK) to do its language resource inventory. Their BLaRK had two main categories i.e. Language technology and Speech technology components each of which had two subcategories consisting of Modules and Data resources which were further subjected to an evaluation criteria. The criteria consisted of a checklist that included: availability of the components on the Internet, on Literature and information from actors in the HLT field; reusability, adaptability, extendibility, and compatibility with standards, among others.

In the South African case, the evaluation criteria for HLT was based on the amount of research, resources and technologies developed and available in each of the eleven official languages that were studied. Moreover, this resulted in the identification of inventory gaps which subsequently informed strategic planning in regards to funding research into specific HLT applications and prioritizing under-resourced languages.

A common goal of HLT audits across these countries was to identify the gaps in BLARK and recommend priority components to be developed to complete the BLARK. In addition, the HLT platforms in these countries serve to strengthen the research, development, and sharing of HLT resources for their languages.

Methodology

For the work we are reporting a case study of two universities: University of Nairobi (public) and Africa Nazarene University (private) were used to evaluate the research done in HLT over the past ten years. Apparently, a purposive sampling criterion was used in selecting the public and private universities using the criteria of the oldest and most active computer science programs in the country. Languages: Bantu (6), Nilotic (4), Cushitic (3), and Semitic (1)

A Library survey of HLT projects in the two universities was done at the respective university's department libraries whereby theses and final-year project documents were perused and the language technologies recorded on an excel spreadsheet.

The research adopted a descriptive approach at the beginning whereby an inventory of the available components was recorded as found in the existing literature. Later on, a prescriptive approach was taken whereby the inventory was mapped onto the BLARK instrument (Krauwer, 1998) informed by the South African and Dutch HLT indices for modules and data.

Given the limited funding and time available for the research, the only criterion used on the BLARK was the availability of the HLT components from the existing literature in the thesis and project documents and from Internet sources. This is as shown in Table 1.

An Internet survey was conducted by searching for HLT projects in the country using a web search engine (Google). Here, key words informed by the Basic Language Resource Kit (BLARK) modules were used to do the google search. Moreover, internet sites with data in the local languages were recorded.

Data Analysis was done using MS. Excel spreadsheet whereby the list of modules and data of the existing inventory was entered into the BLaRK instrument developed on the excel spreadsheet and consequently used to identify the gaps in the BLaRK.

Results

Internet search results revealed that approximately 63% of the data on Kenyan local language on the Internet is religious texts, particularly Christian. Parallel text across all the sampled local languages was found with the four spiritual laws and rosary recitation. Moreover, there were Bible chapters in several local languages and some websites that contained basic language learning content. Besides, there were a few encyclopedia projects and dictionaries in some of the local languages.

The modes of the content retrieved from the Internet search were basically HTML texts, PDF documents and audio files. Most of the bible chapters were in PDF format and some in audio files. Moreover, there were video files, especially songs in some of the local languages, retrieved from you-tube.

The library research for language technology projects in the sampled universities indicated that the earliest HLT projects and research were published in 2002 with an upsurge in 2005 and 2006. Majority of these projects were done at the school of computing and informatics, university of Nairobi. These are as shown in Figure 1.

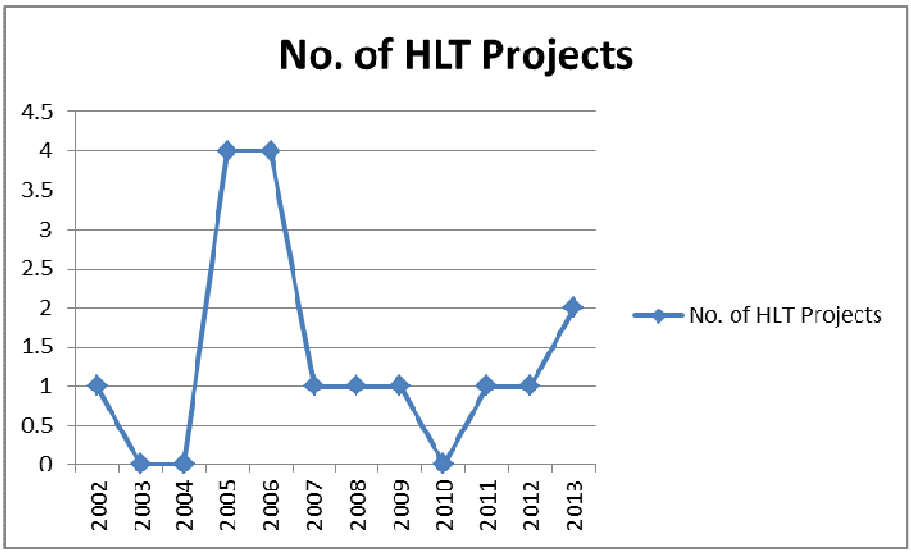

Figure 1: Number of Human Language Technology Projects in UoN and ANU

Apparently, HLT projects have not been a favorite of student projects over the years although the figure indicates a turn-around of revived efforts from 2011.

Language Technology Projects in the Sampled Universities mapped to 7 categories on the BLaRK. Generally, these language tools and applications fulfilled 33.3% of BLaRK. This is as indicated on figure 2. These included a sentence Interpreter system, Text-to-Speech systems for Swahili, Ekegusii, Gikuyu, Dholuo and Kalenjin, a Voice browser for local languages, a Swahili Morphological analyzer, an automatic Correction of Gikuyu language diacritics, a Swahili Dictation System, Machine Translation systems for Ekegusii, Kamba, Lingala, Gikuyu, and Swahili, a Part Of Speech Tagger for Kamba, and a compiler for Swahili.

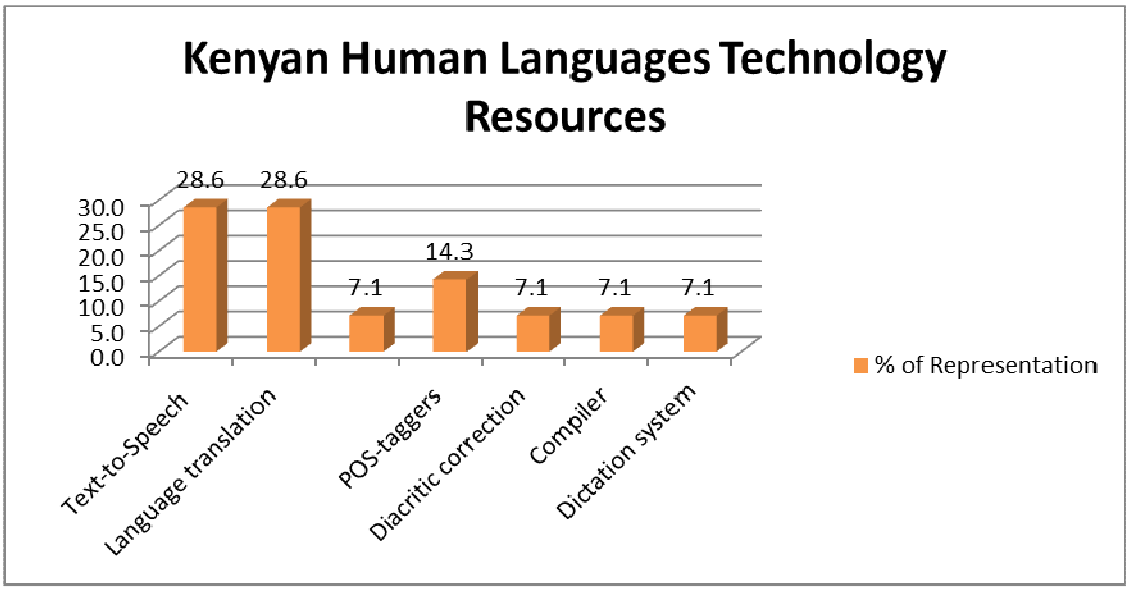

Figure 2. HLT projects mapping on BLaRK.

Further Internet research indicated five active HLT projects in the continent of Africa. These included the African Network for Localization (ANLoc,www.africanlocalisation.net) headed by Donald Z. Osborn, USA; the Zuza Software Foundation (www.translate.org.za) in South Africa headed by Mr. Dwayne Bailey; the Kamusi Project International, headed by Dr. Martin Benjamin at Switzerland, the African Languages Technology Initiative headed by Dr. Tunde Adegbola from Lagos, Nigeria, and the African Language Technologies (AfLaT) project headed by Dr. Guy De Pauw from the University of Antwerp, Belgium. Apparently, there was no internet search result for any HLT forum for Kenyan languages.

MODULE/Tool	**AVAILABILITY**	**LANGUAGES**
Speech Recognition, LMS	1	
Text-to-Speech	4	Gikuyu, Dholuo, Swahili
Language Translation	4	Ekegusii, Swahili, Akamba, Gikuyu
Part-of-Speech tagging	2	Akamba
Diacritic correction	1	Gikuyu
Compiler	1	Swahili
Dictation System	1	Swahili
Morphological analysis	1	Swahili
Parsers and Grammars	1	Swahili
Word-meaning disambiguation	1	Swahili
DATA	**AVAILABILITY**	**LANGUAGES**
Text Corpora	3	Ekegusii, Gikuyu, Dholuo
Speech Corpora	Segments	Swahili, Gikuyu
Parallel Corpora	Select Bible Chapters	Swahili, Ekegusii, Gikuyu, Dholuo, Luhya(Bukusu), Kimeru, Kigiriama, Kalenjin, Maasai, Turkana, Somali, Kamba, Oromo
Online Dictionaries & Thesaurus	3	Ekegusii, Swahili, Gikuyu

Table 1: HLT Component Index for Modules and Data

Conclusion

Generally, there were lots of data resources about the local Kenyan languages including history, culture, economic activities, religious activities, education etc. However, most of the websites surveyed generally had minimal content in the Kenyan languages themselves while using English. Moreover, research projects into Kenyan and generally African languages were not only done by

universities and researchers from the West, but also solely funded by grants from the west. Apparently, every funder seeks and expects the research to primarily meet their specific interests which may not be of fundamental interest to the subjects of the research. It is therefore important that research into Kenyan languages and generally African languages should not only be primarily investigated by African researchers but heavily funded by African governments to serve specific interest and needs of African languages and people.

The level of basic language resources and technologies for Kenyan languages is very low as indicated by the 33.3% of BLaRK. Further, this indicates the low number of research activities on human language technologies in Kenyan universities. Moreover, a lack of online HLT forums is a clear indicator of little and segmented research on HLT in the country.

This research is the first of its kind based on BLARK for the Kenyan languages. Publication of these results will offer an excellent platform for researchers to not only be informed of what has been done but more importantly to offer collaboration opportunities to do further research and avoid duplication in HLT research and development efforts.

Generally, the most resourced local language in Kenya, apart from Swahili, is the Gikuyu with several online resources including corpora, part-of-speech tagger, text-to-speech modules, automatic diacritic applications, machine translation, and video content. The other languages covered in this research had less than five HLT components available.

Apparently, most of the published HLT research on African languages has been primarily sponsored and headed by non-Africans. This should be a big challenge for African governments to fund research and development of HLT efforts to address their own unique challenges and interests, just as any funder would. Moreover, the African governments could therefore help prioritize research and development efforts on the basic language and technology resources for their vast local languages to enable national development.

Recommendations

The field on HLT is multidisciplinary and requires concerted effort of Computer Science, Linguistics and other disciplines like Psychology in order to be better understood and consequently address the existing HLT challenges exhaustively. Unfortunately the predominant culture of operating in silos in higher education research exacerbates these efforts. Researcher collaboration across these disciplines is imperative to the success of HLT efforts. Besides, the HLT platforms can help bridge the academia-industry relationships and expediting research and development of well informed and sustainable HLT solutions. Moreover, there is need to establish a HLT online forum to help coordinate, prioritize and reduce duplication of work in the HLT field in Kenya.

With the rise of new, faster and affordable collaboration and social networking technologies, it is about time that volunteer models e.g. Crowd-sourcing technologies be explored for data collection to add to the low data warehouse of Kenyan and African language content i.e. Data or content developed by the natives and in the native languages.

Despite the limited research funding, personnel and time expended in this research, it did break the ground in efforts towards an inventory of HLT in the country. Nonetheless, there is need for a large scale research on HLT in the country that will include field surveys , questionnaires and other methods to cover more if not all Kenyan languages. Moreover, the research should include more HLT participants from industry and academia in order to help build consensus on the country's component index that will build an exhaustive BLARK audit. Besides, the relevant government

agencies need to get involved in order to benefit from the results of the audit exercise which will inform policy making on matters related to HLT in the country.

References

Abiud Ogechi (2005) NLP: Text-to-Speech in local language: Ekegusii . (BSc final year project) . School of Computing and Informatics. University of Nairobi

Benson Kituku (2011). Kamba Part Of Speech Tagger. (MSc Thesis project) . School of Computing and Informatics. University of Nairobi

Bible chapters: http://gospelgo.com

Binnenpoorte, D., De Friend, F., Sturm, J., Daelemans, W.,Strik, H., and Cucchinari, C. (2002). A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch. In: Proc. LREC 2002, Spain.

Daniel Wafula (2009). Swahili compiler (BSc final-year project). Computer and Information Technology department. Africa Nazarene University.

Dorothy Abade (2006) .Text-to-Speech system with visemes for Dholuo. (BSc final year project) . School of Computing and Informatics. University of Nairobi

Edith Wanjiru (2002) . A sentence Interpreter system (BSc final year project) . School of Computing and Informatics. University of Nairobi

Edward Ombui (2007). Interlingual Machine Translation for Ekegusii-Swahili-English. (MSc Thesis project) . School of Computing and Informatics. University of Nairobi

Elenius, K., Forsborm, E., and Megyesi, B. (2008). Language Resources and Tools for Swedish: A Survey. In: Proc. LREC 2008, Marrakesh, Morocco.

Evans Miriti (2006).Kiswahili Dictation System: Swa-Speak (MSc Thesis project) . School of Computing and Informatics. University of Nairobi

Geoffrey Gakundi (2006). Automatic Correction of Gikuyu language diacritics. (BSc final year project) . School of Computing and Informatics. University of Nairobi

Infrastructure for Bulgarian. In: Proc LREC 2004, Lisbon, Portugal: 1685-1688.

Judy Muthee (2013). Kamba-English translation system. (BSc final-year project). Computer and Information Technology department. Africa Nazarene University.

Karen Waithera (2014). Kikuyu –English translation system . (BSc final-year project). Computer and Information Technology department. Africa Nazarene University.

Kennedy Odero (2005) Voice browser for local languages (BSc final year project) . School of Computing and Informatics. University of Nairobi

Language and Speech Technology. [Online]. Available: http://www.medar.info/MEDAR_BLARK_I.pdf(Accessed June 2009)

Maegaard, B., Krauwer, S., and Choukri, K. (2009). BLaRK for Arabic. MEDAR – Mediterranean Arabic

Maegaard, B., Krauwer, S., Choukri, K., and Jørgensen, L. (2006). The BLARK concept and BLARK for Arabic. In: Proc. LREC 2006, Genova: 773-778.

Michel Kabonga (2013). Lingala-Swahili translation system . (BSc final-year project). Computer and Information Technology department. Africa Nazarene University.

Milgo, k. (2008). Text to Speech: Kalenjin. (MSc Thesis project) . School of Computing and Informatics. University of Nairobi

Nicholas Owino (2005).Swahili Morphological analyzer (BSc final year project) . School of Computing and Informatics. University of Nairobi

Omwenga, G. (2014, March 4). Medium of Instruction. *The Daily Nation*, pp.4.
Online encyclopedia http://www.omniglot.com

Pauline Githinji (2006). Text-to-Speech system for Gikuyu Language with Mobile phone. (BSc final year project) . School of Computing and Informatics. University of Nairobi

Piper, B. (2010). Uganda Early Grade Reading Assessment Findings Report: Literacy

Prah, K.K. (2002). Researching African Languages for Scientific and Technological Development: The CASAS Experience. In Speaking African . Cape Town: The Centre for Advanced Studies of African Society.

Robert Njuguna (2005). Kiswahili Text-to-Speech system (BSc final year project) . School of Computing and Informatics. University of Nairobi

Rosary prayers: www.marysrosaries.com

Simov, K., Osenova, P., Kolkovska, S., Balabanova, E., and Doikoff D. (2004). A Language Resource