Contents lists available at www.innovativejournal.in

Asian Journal of Computer Science and Information Technology

Journal homepage: http://www.innovativejournal.in/index.php/ajcsit

ISSN 2249 - 5126

SWAHILI TEXT AND SPEECH CORPUS: A REVIEW

Aaron M. Oirere*, Ratnadeep R. Deshmukh, Pukhraj P. Shrishrimal, Vishal B. Waghmare

Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India

ARTICLE INFO

Corresponding Author:
Aaron M. Oirere
Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad-431004 (MS), India

Keywords: Swahili, Sentence Selection, Phonetics, Text Corpus and Speech Corpus, Speech Recognition.

ABSTRACT

This paper explores the review of Swahili text and speech databases/corpus in different dimensions i.e. word, sentence and phrases selections with their phonetics and selection of speakers. It also reviles the availability of the corpus. Text and Speech database is needed for various purposes like speech synthesis, speech recognition, speech detection; corpus based text-to-speech (TTS) synthesis and machine translation. This work is done for the purpose of developing a speech corpus in Swahili language.

INTRODUCTION

Speech is the most prominent and natural form of communication between humans. There are various spoken languages throughout the world. Communication among the human being is dominated by spoken language, therefore it is natural for people to expect speech interfaces with computer [1].

Speech has potential of being used as a mode of interaction with computer. Human beings have been motivated to create computer that can understand and talk like human. In this direction, researchers have tried to develop system for analysis and classification of the speech signals. Since, 1960s computer scientists have been trying to find out different ways and means to make computer record, interpret and understand human speech.

Speech processing has become increasingly important in daily life, as the number of web enabled mobile phone users in rural as well as in urban area is increasing. Most of the research efforts in the field are of natural language processing (NLP) for African Language where Swahili is included has majorly rooted in the rule-based paradigm. The rule based approach has some merits as well as demerits. The merit of Swahili is in term of its design transparency and demerits are being high language dependent and costly to develop as it typically involves a lot of manual effort of experts from Natural Language Processing field.

The system are decidedly competence-based which is often tweaked and tuned towards a small sets of ideal sample words or sentences neglecting the real-world language technology application. In Language technologies for many African languages the researchers are getting tired of publication on real-world data or reports.

Currently with the increased need of digital resource usage in the continent of Africa, there is a great need for more empirical approaches such as data driven and corpus based approach for language technologies. The main advantages of these approaches are: language impendence, development speed, robustness and empiricism.

There is scarcity of sources in the sense that the digital text resources are few. The recent effort on the same is carefully selected procedure for Swahili [2, 3].

For language technology applications such as speech recognition system, text-to-speech synthesis, machine aided translation and web related issues there is a great need for translation and usability of the Swahili language. There is a great need for work to be done in semantics and syntactic of Swahili language as the biggest online web Text resources which are available on Google, Yahoo and Wikipedia. The major need is the extraction of information which enhances and refocuses on embarking on Swahili as a language, the corpus availability needs to be syntactically and semantically correct. Unsupervised approach can be used to bootstrap annotation of resource-scarce languages by automatically finding linguistic pattern in large amount of raw text.

This paper describes the development of text and speech corpora / database for Swahili language. The details of the Swahili language are described in the Section 2. Section 3 and 4 describes the Swahili text and Speech Corpora. The comparative study of Swahili text and speech corpora is performed in section 5. The Conclusion and the Future work are discussed in Section 6.

I. About Swahili language

The basic phone set of Swahili comprises of 5 vowels and 27 consonants [4]. Swahili is currently written in a slightly defective orthography using the Roman alphabet. Swahili has no diphthongs; in vowel combinations, each letter is pronounced separately. The language had previously been written in the Arabic script. Unlike adaptations of the Arabic script for other languages, relatively little accommodation was made for Swahili.

Swahili is a Bantu language that serves as a second language to various groups traditionally inhabiting in the parts of the East African coast. Some Swahili vocabulary is derived from Arabic through more than twelve centuries of contact with Arabic-speaking inhabitants of the coast of Zanj. It has also incorporated Persian, German, Portuguese, English and French words into its vocabulary through contact during the last five centuries. Swahili has become a second language spoken by millions of people in three countries in African Continent, Tanzania, Kenya, and Congo (Democratic Republic of Congo), where it is one of the national languages out of four recognized national languages [5]. The neighboring nation of Uganda made Swahili a required subject in primary schools from 1992, although this mandate has not been well implemented and declared it as an official language in 2005 as preparation for the East African Federation. Swahili, or other closely related languages, are spoken by nearly the entire population of the Comoros and by relatively small numbers of people in Burundi, Rwanda, Malawi, Northern Zambia and Mozambique. The language is still understood in the southern ports of the Red Sea and along the coasts of Southern Arabia and the Persian Gulf in the twentieth century [3]. In the Guthrie non-genetic classification of Bantu languages, Swahili is included under Zone G.

The earliest known documents written in Swahili are letters written in Kilwa in 1711, in the Arabic alphabet. They were sent to the Portuguese of Mozambique and their local allies which are now preserved in the Historical Archives of Goa, India [6, 7]. Another ancient written document is an epic poem in the Arabic script titled *Utendi wa Tambuka* (*The History of Tambuka*); it is dated back in 1728. The Latin alphabet has become standard under the influence of European colonial powers [8, 9].

Swahili is unusual among sub-Saharan languages having lost the features of lexical tone (with the exception of the numerically important Mvita dialect, the dialect of Kenya's second city, the Indian Ocean port of Mombasa).

II. Swahili Text corpora

In this section we have mentioned the various text corpuses that have been developed in the various universities and research labs around the world for Swahili language. We have gone through some of the text corpuses developed by the different researchers.

The researchers at the University of Antwerp Belgium, University of Nairobi, Kenya and University of the Western Cape, South Africa developed a morphological database of 97,000 entries which were extracted from the Helsinki Corpus of Swahili. The developed corpus is been used for Machine Learning. The researchers had found that computational morphological analysis is an important first step in the automatic treatment of natural language and a useful lexicographic tool. So the researchers have extracted the morphological database [2].

The researchers at University of Nairobi, Kenya and University of the Western Cape, South Africa jointly created a text of Swahili Language. The researchers in the universities developed a parallel corpus for Swahili and English language. The said corpus has been developed for the Machine Translation system. The SAWA corpus consists of the text selected from Bible, Quran, Politics, Kamusi.org, Movie Subtitles, Local translator and investment reports. The total sentences in SAWA corpus is 73.7 thousand. The total number of English tokens in the corpus is 1.463 million and for Swahili 1.201 million tokens [10].

The researchers at the University of Helsinki, Finland developed a Text corpus for Swahili language. The corpus has been developed for automatic lexical acquisition method to learn semantic properties of Kiswahili words directly from the data. The machine learning component was being implemented by using the Self Organizing map. The data was selected from the Helsinki Corpus of Swahili developed at the University of Helsinki, Finland. The selected data consists of 35 Verbs and 60 Nouns for the machine learning from a huge text corpus [11].

The Researchers at University of Dar-es-Salaam, Tanzania, developed a text corpus for Swahili language. The developed corpus was domain specific corpus for health care. The data was collected from the books and journals which were scanned and edited using the tool EMACS text editor. The total edited text consisted of 92,285 words. The words were saved in two text files named pattern-describing text file consisting of 50,877 words and pattern-testing text file consisting of 92,285 words [12].

The researchers at the Carnegie Mellon University, USA developed a Text corpus for the Swahili Language for the problem of Named Entity Recognition (NER) system. The developed system was known as SYNERGY. The SYNERGY addresses NER problem for a new language by breaking it into three relatively easier problems those are Machine Translation to English, English NER and word alignment between English and the new language. The Named Entity (NE) labeled data is scarce for Swahili. For Swahili, the researchers used slightly more than 27,000 test set of words, which was selected from the Helsinki Corpus of Swahili [13].

The researchers at the Università degli Studi di Napoli L'Orientale, Italy and University of Dar es Salaam, Tanzania jointly developed a text corpus for the Swahili language. The researchers used corpus data to support autonomous learning of kiswahili by italian speakers. By generalized usage patterns meant for assisting second language Swahili learners in appropriate use of the amba (ambapo, ambako, ambapo)- locatives by applying corpus-based discovery procedures. The major aim was the researchers intended to show the use of the raw data extracted from a corpus, structured on the basis of detected crucial features, can reinforce discriminative learning in teaching situations but especially in acquiring a working ability to communicate in a second language. The authors involved in teaching Kiswahili had experienced difficulties in providing adequate description patterns and efficient prescription rules concerning the usage of amba- forms (respectively ambapo, ambako and ambamo). The selected corpus is made up of about 500,000 tokens and 55,000 types, with a type/token ratio of 8.98. A random control check of other Swahili texts and consultation with native Kiswahili speakers has also been performed [14].

University of Antwerp Belgium, University of Nairobi, Kenya and University of the Western Cape, Bellville, Republic of South Africa jointly did the survey of a corpus-based of four electronic Swahili – English bilingual dictionaries in order to develop the machine translation system and attempt to consolidate the dictionaries into a unified lexicographic database and compare the coverage to that of its composite parts. The total number Swahili entries were 93,600 from all dictionaries and from them it was found a total of 21000 number of orthographically distinct lemmas in all dictionaries (not taking into account homographs with different morphosyntactic or semantic features) and a

12,150 lemmas were exclusive (i.e. unique) also considered other sources to compute the coverage of the dictionaries, including: The Helsinki Corpus of Swahili, HCS (Hurskainen 2004a) consisting of more than 9 million words,the TshwaneDJe Kiswahili Internet Corpus, TeDJe-KIC (De Schryver and Joffe 2009) of more than 20 million words, the Swahili part of the parallel SAWA corpus (De Pauw et al. 2009), containing about 0.5 million words and Wikipedia in Swahili: almost 12 000 Internet pages, good for more than 1 million words. The four dictionaries used for developing the Text corpus are described in the Table 1[15].

Table 1. Showing the name of the dictionary along with the Text Corpus available

Dictionary	No. of text corpus available
The Internet Living Swahili Dictionary [ILSD]	60 000 entries.
The Freedict Swahili–English Dictionary [Freedict]	2 600 entries
The TshwaneDJe Swahili–English Dictionary [TeDJe-SED]	16 000 entries
The TUKI Swahili–English Dictionary [TUKI]	14 500 entries

CNTS -University of Antwerp, Belgium, Ghent University, Belgium, University of the Western Cape, South Africa, University of Nairobi, Kenya jointly did the Data-Driven Part-of-Speech Tagging of Kiswahili with the help of the corpus from the Helsinki Corpus of Swahili using the four of the current state-of-the-art data-driven taggers, TnT (Trigrams'n'Tags) MXPOST (Maximum Entropy Modeling), MBT (Memory-Based Learning) and SVM Tool: Support Vector Machines. The researchers had a corpus of 3,656,821 words (169,702 sentences) after clear up and disposal of duplicate sections. The corpus was randomly divided into an 80% training set (2,927,846 words), a 10% validation set (362,866 words) on which the optimal parameters of the algorithms could be established, and finally a 10% blind test set (366,109 words) for evaluation on unseen text [16].

University of Helsinki Finland have developed the corpus known as Helsinki Corpus of Swahili containing standard Swahili corpus which have been annotated with SALAMA a multi-purpose language manage environment, corpus contains information of such features as word (lemma), part of speech and morphology including noun class affliction and verb morphology. The corpus size is of about 12.5 million words. It is available for scientific research without charge. [17, 18]

III. Audio/ Speech Corpus

In this section we have mentioned the various speech corpuses that have been developed in the various universities and research labs around the world for Swahili language. We have gone through some of the speech corpuses developed by the different researchers.

University of California at Los Angeles conducted the acquisition of Swahili verbal morphology where the data was collected over a period of 11 months in Nairobi, Kenya. The recording was done twice a week of naturalistic speech in the home of four children of different ages. Due to social and economic reasons all the children were unable to remain in the study for the complete duration of the project. The table 2 shows the details of the subjects:

Table 2: Details of the Speakers with recording details

Child	Age range	No of recordings	MLU
Haw	2;2-2;6	7	1.54-2.46
Mus	2;0-2;11	23	1.52-3.57
Fau	1;8-2;2	10	2.97-3.93
Has	2;10-3;1	5	3.15-4.23

The purpose of the data collection was for the use of children verb morphology and it was focused on the so-called Root Infinitive (RI) phenomenon, where children in languages use infinitive verbs in root context. The similar type of work has been done in the other languages such as German, Dutch, French, Swedish, Russian, Italian, Spanish, Catalan and English [19].

University of Nairobi, United states international University, Kenya and Outside echo ltd., Chepstow, U. K. jointly developed the text to speech synthesis system for Swahili language. It consisted of 10,558 sentences. It was done from many sources to ensure that they have captured maximum language features as possible. The designing of their database used the Database selection tool [20] which takes the units to be selected and chooses the minimum number of sentences that contain the units by comparing the text corpus and the transcribed text, it contained phonetically balance sentences. They selected 1,997 units out of 3,725 units for sentence selection and 414 sentences were selected as the minimum number of sentences to contain the possible units found in the corpus. A phone count was then carried out both in the text corpus (10,558 sentences) and in the selected sentences (414 sentences) and an almost 100% correlation was achieved. The selected sentences were phonetically balanced and therefore represented Kiswahili sound system. Professional speakers who were familiar with the language features were selected for recording. Recording was done in the professional studio at Kenya broadcasting corporation. The database was hand-labeled and later the segmentation and annotation was carried out using Festival engine [21].

Laboratoire Dynamique Du Langage-France and Laboratoire Informatique de Grenoble, - France developed a speech database for Automatic Speech Recognition. The text corpus was developed by extraction of the sentences from news websites broadcast. The native speakers read the extracted sentences. The researchers used mining web broadcast news speech which had an advantage of massive and directly available. The quality of each radio differs i.e. studio: both music and speech was of high quality, telephone had low quality; some audio was without noise and others with out door noise were the audio quality was bad. In order to quickly provide the transcription to audio corpus they used web crowd sourcing tool: Amazon's Mechanical Turk (MTurk), the sentences read by speaker during recording were used as gold standard to compare with the transcription obtained by MTurk. The acoustic model trained using MTurk transcription was quite similar to one trained using researchers. The transcription was 38.5% and 38.0% WER (word error rates) respectively on small 82 sentences test set [22].

IV. Comparison

The overall paper describes the various text and Speech corpus developed for Swahili language. Till date lot of work has been done on Text Corpus but not for Speech application. The Text corpuses developed are for Machine learning or machine translation purpose very few of them have been used for speech recognition or speech synthesis system.

In Section 3 we have briefly described the text corpus developed in Swahili language. The Text corpuses are compared on the basis of the resources used, number of units present and the purpose of the Text Corpus. The Table 3 shows the comparison of the studied text corpuses.

Table 3. Comparison of the Text Corpora

Sr. No	Database developed at	Resources used for selection of text	No. of units	Application
01	University of Antwerp Belgium, University of Nairobi, Kenya and University of the Western Cape, South Africa	Helsinki Corpus of Swahili (HCS)	97 000 entries	Machine Learning
02	University of Nairobi, Kenya and University of the Western Cape, South Africa	Religious Books, Movies subtitles, kamusi.org, Investment report, local translator	73.7 Thousand	Machine Translation
03	University of Helsinki, Finland	HCS, news papers articles, parliamentary proceeding & standard books texts	95 different words used	Machine Learning
04	University of Dar-es-Salaam, Tanzania	Books and journals	92,285 words	Text Compilation
05	Language Technologies Institute, School of Computer Science, Carnegie Mellon University, USA	HCS	27000 tokens	Machine Translation (online)
06	Università degli Studi di Napoli L'Orientale, Italy, University of Dar es Salaam, Tanzania	novels and a few short stories from Zanzibar and continental authors	± 500,000 tokens and 55,000 types,	Machine Learning
07	University of Antwerp Belgium. University of Nairobi, Kenya. University of the Western Cape, Bellville, Republic of South Africa	Dictionaries, HCS, Wikipedia.com	93600 entries	Machine Translation & Machine Learning
08	University of Dar es Salaam, Tanzania	The TUKI Dictionary [TUKI]	14 500	Machine Translation
09	CNTS -University of Antwerp, Belgium, Ghent University, Belgium, University of the Western Cape, South Africa, University of Nairobi, Kenya	Helsinki Corpus of Swahili	3,656,821 words	Machine Learning
10	University of Helsinki Finland (Helsinki Corpus of Swahili) www.aakkl.helsinki.fi www.csc.fi	Swahili newspapers, extracts from books: prose text, including fiction, education and sciences.	12,500,000	Multi purpose

When we compare all the 10 text corpuses that are studied we observed that 5 Text corpuses are being developed for Machine Learning purpose, 3 are developed for Machine Translation, 1 corpus is developed a multipurpose database and 1 is for Text Compilation in Health Care Domain. The developed 5 Text corpuses have used the Helsinki Corpus of Swahili (HCS) as a Resource while developing. We observed that the work for Swahili language is done with collaboration with other universities or by other universities. Out off the studied 10 text corpus 2 Text corpuses are been developed by the University of Dar es Salaam, Tanzania, individually. Other corpuses are developed in collaboration with other universities; it shows that the work for Swahili is not done by the native researchers on their own they are lagging behind in the development of technology.

In Section 4 we have briefly described the speech corpus developed in Swahili language. The speech corpuses are compared on the basis of the recording environment, amount of corpus, no of speakers, recording device and the application of the said corpus. The Table 4 shows the comparison of the studied speech corpuses.

We studied 3 speech corpuses developed for the Swahili language. During the comparison we observed that the speech corpus developed are for different purpose. The Corpus developed for Automatic Speech Recognition consists of only 82 sentences. The corpus for TTS system consists of 10,558 sentences. A very little work has been done. For speech corpus development is also not carried by the resident university. The work is done by researchers at other universities.

Table 4. Swahili Speech Corpus

Sr. No.	Database Developed at	Recording Environment	Amount of Corpus	No of Speakers	Recording Device Used	Application of the Database
01	University of California, Los Angeles	Home	Not mention	4 children, Native speakers	Not mentioned	Study of Root Infinitive Phenomenon
02	University of Nairobi , Kenya, United States International University, Nairobi, Kenya, Outside echo ltd., Chepstow, U.K.	Studio	10,558 sentences	1 Professional Male Native speaker	High quality mic and speaker	Text – To – Speech Synthesis System
03	Laboratoire Dynamique Du Langage, CNRS - Universit´e de Lyon, France, Laboratoire Informatique de Grenoble, CNRS - Universit´e Joseph Fourier Grenoble 1, France	Studio and Outdoor with back ground Noise	82 sentence	Not mentioned, Native speakers	Telephone, Standard Mic	ASR

CONCLUSION

In this paper we have studied the various text and speech corpus developed for Swahili language. The research for the development of speech application is very less. The text corpora have been developed mostly for machine learning and Machine translation.

Most of the researchers are concentrating on machine translation or machine learning. The researchers that are working are not doing the research at the place where Swahili is the native language. The native researchers should actively participate in the development of language technology. The developed text corpus can be utilized for the development of speech corpus which is not done.

The researchers in the future should try to utilize the developed text corpus and develop language technology. This study has helped us to understand the current status of development of the language technology in Swahili language. After this study we have been motivated for carrying out the work for Swahili language. We will try to utilize the developed text corpus for developing a Automatic Speech Recognition System in Swahili.

ACKNOWLEDGEMENT

The authors would like to thank the University Authorities for providing the infrastructure to carry out the research.

REFERENCES

1. Pukhraj Shrishrimal, R R Deshmukh, Vishal Waghmare, , (2012) “Indian Language Speech Database: A Review”. Intenational Journal of Computer Application (IJCA), Vol 47, No. 5, pp. 17-21.
2. Guy De Pauw and Gilles-Maurice de Schryver , (2008) “Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes” Lexikos 18 (AFRILEX-reeks/series 18:): 303-318.
3. G. De Pauw, G.-M. de Schryver, and P.W. Wagacha. (2006). “Data-driven part-of-speech tagging of Kiswahili”. In P. Sojka, I. Kopeˇcek, and K. Pala, editors, Proceedings of Text, Speech and Dialogue, 9th International Conference, volume 4188 of Lecture Notes in Computer Science, pages 197–204, Berlin, Germany. Springer Verlag.
4. Gakuru, Mucemi Iraki, Frederick K. Tucker, Roger Shalonova, Ksenia Ngugi, Kamanda, (2005) “Development of a Kiswahili text to speech system”, In INTERSPEECH, 1481-1484.
5. http://en.wikipedia.org/wiki/Languages_of_the_Democratic_Republic_of_the_Congo dated 27/06/2012
6. E.A. Alpers, , (1975)“Ivory and Slaves in East Central Africa”, London, pp. 98–99 ;
7. T. Vernet, (2002)“Les cités-Etats Swahili et la puissance omanaise” (1650–1720), Journal des Africanistes, 72(2), pp. 102–105.
8. “Ethnologue list of countries where Swahili is spoken” Thomas J. Hinnebusch,(1992),"Swahili", International Encyclopedia of Linguistics, Oxford, pp. 99–106
9. David Dalby, (1999/2000) “The Linguasphere Register of the World's Languages and Speech Communities”, Linguasphere Press, Volume Two, pg. 733–735.
10. Guy De Pauw, Peter Waiganjo Wagacha, Gilles-Maurice de Schryver, (2011) “Exploring the SAWA corpus: collection and deployment of a parallel corpus English—Swahili”, International Journal of Lang Resources & Evaluation, Springer Verlag, vol 45, pp 331-344.
11. Wanjiku Ng’ang’a, (2003) “Semantic Analysis Of Kiswahili Words Using The Self Organizing Map”, Nordic Journal of African Studies vol. 12, Issue 3, pp 405-423.
12. Seleman S. Sewangi, (2000) “Tapping the Neglected Resource in Kiswahili Terminology: Automatic Compilation of the Domain-Specific Terms from Corpus”, Nordic Journal of African Studies, Vol. 9, issues 2, pp 60-84.
13. R. Shah, B. Lin, A. Gershman, and R. Frederking, (2010) “Synergy: A named entity recognition system for resourcescarce languages such as swahili using online machine translation”, in Proceedings of the Second Workshop on African Language Technology (AfLaT 2010), , pp.21–26.
14. Maddalena Toscano, Simon Sewangi, (2005) “Discovering Usage Patterns for the Swahili amba-Relative Forms cl. 16, 17, 18: Using Corpus Data to Support Autonomous Learning of Kiswahili by Italian Speakers”, Nordic Journal of African Studies volume 14, issue 3, pp:274–317.
15. Guy De Pauw, Gilles-Maurice de Schryver and Peter Waiganjo Wagacha, (2009) “ A Corpus-based Survey of Four Electronic Swahili–English Bilingual Dictionaries”, Lexikos volume19 (AFRILEX-reeks/series 19:), pp: 340-352
16. Guy De Pauw1, Gilles-Maurice de Schryver, and Peter W. Wagacha (2006) “Data-Driven Part-of-Speech Tagging of Kiswahili” Petr Sojka, Ivan Kopeˇcek and Karel Pala (Eds.): TSD, LNAI volume 4188, Springer-Verlag Berlin Heidelberg 2006, pp. 197–204,
17. Arvi Hurskainen (2004) “Helsinki Corpus of Swahili. Compilers”: Institute for Asian and African Studies (University of Helsinki) and CSC.
18. Arvi Hurskainen, (2004) “Swahili Language Manager: A Storehouse for Developing Multiple Computational Applications”. Nordic Journal of African Studies volume 13, 363-397.
19. Deen, Kamil Ud, (2002c) “The acquisition of Swahili verbal morphology”, Palmela, Portugal. Costa, Joao & Freitas, Maria (Eds), In the proceedings to G.A.L.A conference pp.41-48.
20. Talukdar, P., "Optimal Text Selection Module Version 0.2", available from http:// www. llsti.org/downloadstools.htm
21. Gakuru, Mucemi , Frederick K. Iraki, Roger Tucker, Ksenia Shalonova, Kamanda Ngugi, (2005)"Development of a Kiswahili text to speech system", In INTERSPEECH, pp1481-1484.
22. Hadrien Gelas, Laurent Besacier, F. Pellegrino, (2012) “Developments of Swahili resources for an automatic speech recognition system”, SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages, Cape-Town, South Africa
23. .