Research Article

Development of Yoruba Language Spell Checker Using Interpolation Algorithm

Fasidi FO[1]* and Adebayo OT[2]

Abstract

Globally, more and more people are using computers today because the English language is dominating language in this field. The dominance of the English language is quite overwhelming in Nigeria, and the use of computers has so far been greatly restricted only to those people who have some knowledge of the English language. This has resulted in fast way of killing the major indigenous language in the country especially the Yoruba language. Yoruba language is less used among its people because their roles have been taken over by the English language. This prompted the need to develop a text editor that will give Yoruba language a public profile in the information technology (IT) world so as to provide a platform for people to appreciate the beauty of their indigenous language.

Keywords

Spell checker; Interpolation search; Yoruba language

Introduction

Yoruba (native name ede Yoruba, the Yoruba language) is a Niger- Congo language spoken in West Africa. The number of speakers of Yoruba was estimated at around twenty million in the early years. The native tongue of the Yoruba people is spoken among other languages in Nigeria, Benin, and Togo and in communities in other parts of Africa, Europe and the Americas. A variety of the language, Lucum, from "olukumi" is used as the liturgical language of the Sanitaria religion of Cuba, puer to Rico, Dominican Republic and the United States. It is most closely related to the Owo and Itsekiri language (spoken in the Niger - Delta) and igala spoken in central Nigeria. The Yoruba people originated from the western Nigeria and the places where the language is spoken are termed "ile Yoruba" Yoruba meaning the Yoruba land. It has been discovered recently that the native language like Yoruba were not been taught and spoken by the people again especially at homes even those who speak made a lot of mistakes. The few people who even write in native language like Yoruba made a lot of mistakes. The dominance of the English language in Nigeria is quite overwhelming. This can be seen in practically all domains: government and administration, education, the media, the judiciary, science and technology to mention a few. High government officials avoid using their languages in official contacts even with their own people for fear of being labeled tribal and parochial. In the

National and State Houses of Assembly, English continues to be the language of debate and record in spite of the fact that provision is made in the constitution for the use of the major indigenous Nigerian languages. The educational policy as contained in National Policy on Education (NPE), states that the medium of instruction in pre-primary and early primary education will be the mother tongue or the language of the immediate community as affirmed by Yusuf et al. [1], therefore it is very important to develop software (spell checker) which can be used to make the corrections. This project will ultimately help students and general public to have basic knowledge of learning and correct Yoruba words. A spell checkers is a program that scans the texts misspelled and in turn offers correct spelling suggestions for the misspelling. Generally, a spell checker is a computer program that compares words in a text with file of correctly spelled words in order to detect misspellings.

Literature Review

A Yoruba spell checker is a computer program that makes use of a logically organized Yoruba words that forms a dictionary, and it detects and often corrects misspelled words in a text document. Techniques used in implementation of some spell checkers include regular expression, phonemes conversion, morphology, and algorithm such as binary search algorithm, linear search algorithm, extrapolation algorithm etc. [2].

Spell checking is the process of detecting and sometimes providing suggestions for incorrectly spelled words in a text. By definition, a spell checker is a computer program that detects and often corrects misspelled words in a text document [3]. It can be stand-alone application or an add-on module integrated into an existing program such as a word processor or search engine. Fundamentally, a spell-checker is made out of three components: An error detector that detects misspelled words, candidate spellings generator that provides spelling suggestions for the detected errors, and an error correctors that chooses the best correction out of the list of candidate spellings [1-5]. All these three basic components are usually connected underneath to an internal dictionary of words that they use to validate and look-up words present in the text to be spell-checked [4]. Natural Language Processing (NLP) system may begin at the word level to determine the morphological structure, part- of- speech, meaning of the word and then may move on to the sentence level to determine the word order, grammar, meaning of the entire sentence and then to the context and the overall environment or domain [5].

Yoruba spellchecker.net (DomainOptions, 2013) developed a spell checker for Yoruba words. The system was limited due to the fact that it cannot display (predict) words as the user enters (types) the word. There is no user assistance for typing a word; it is assumed that the user should know the exact word [6].

Another work was proposed by Marek (2008) was motivated to develop a spell checker for the Esperanto language and its implementation as a dictionary (i.e. an affix file and a word list) for the Hunspell spell checker. The word list is an adaptation of word roots coming from the renowned Esperanto dictionary [7].

Ritu (2007) was motivated to develop Hindi editor with spell checker; it works for Hindi and Marathi Characters. It consists of

*Corresponding author: Fasidi FO, Department of Computer Science, Federal University of Technology, Akure, Nigeria, E-mail: flourishhappy@gmail.com;adebayoot@futa.edu.ng

Hindi interface which have a spell checker, which underlines the wrong word in the file. It also has autocorrect along with treasures giving synonyms, and antonyms [8].

In Awoyele (2008), an attempt was made to develop a computational model of Yoruba Morphology using regular expression. Rule-based approach was used for morphological analysis and finite-state automata were used to internally represent morphological corpus. The morphological analysis was performed by parsing of the input word through the fine-state network. However the corpus is restricted to Yoruba prefixes infixes, verbs and nouns. It can only give representation of the major ways of Yoruba word forms and the system is only good for a beginner [9].

Materials and Methods

The proposed system involves data collection which is the process of preparing and collecting data. The purpose of data collection is to obtain information to keep on record, to make decisions about important issues, or to pass information on to others, here, Yoruba dictionary containing organised Yoruba words will be used. Interpolation search algorithm will search through the Yoruba dictionary, interpolation search algorithm is an algorithm for searching for a given key value in an indexed array that has been ordered by the values of the key. For example, it parallels how humans search through a telephone book for a particular name, the key value by which the books entries are ordered. In each search step, it calculates where in the remaining search space the sought item might be, based on at the bounds of the search space and the value of the sought key, usually via interpolation. The key value actually found at this estimated position is then compared to the key value being sought. If it is not equal, then depending on the comparison, the remaining search space is reduced to the part before or after the estimated position. This method will only work if calculations on the size of differences between key values are sensible.

Database design is a number of steps carried out to produce a detailed data model for the system and the relationships between the different data elements.

Interpolation search algorithm is similar to extrapolation search, which is the process of estimating, beyond the original observation interval (facts beyond to area that is certain known)

the value of a variable on the basis of its relationship with another variable. Extrapolation with polynomials additional functions have been used extensively to accelerate the conveyance of discretization methods in numerical analysis [10]. Interpolation is finding values within the sample while extrapolation is finding values beyond your sample. If say the sample ranges from X=0 and X=10 inclusive this is interpolation if finding values of y beyond X=0 that is extrapolate. Interpolation algorithm is used for searching for a given key value in an indexed array that has been ordered by the values of the key. For example, it parallels how humans search through a telephone book for a particular name, the key value by which the books entries are ordered. In each search step, it calculates where in the remaining search space the sought item might be, based on at the bounds of the search space and the value of the sought key, usually via interpolation. [11-14].The key value actually found at this estimated position is then compared to the key value being sought. If it is not equal, then depending on the comparison, the remaining search space is reduced to the part before or after the estimated position. This method will only work if calculations on the size of differences between key values are sensible. The interpolation search algorithm is shown below.

Results and Discussion

There are six thousand, and forty nine (6,049) words in the dictionary. Key 'a' has 835 words, key 'b' has 348 words, key 'd' has 504 words, key 'é' has 432 words, key 'f' has 254 words, key 'g' has 72 words, key 'gb' has 139 words, key 'h' has 35 words , key 'i' has 723 words key 'j' has 134 words, key 'k' has 195 words, key 'l' has 262 words, key 'm' has 148 words key 'n' has 118 words, key 'o' has 614 words, key 'p' has 227 words, key 'r' has 119 words, key 's' has 186 words, key 't' has 290 words, key 'ú' has 6 words, key 'w' has 162 words, key 'y' has 184 words, key 'ọ' has 37 words and key 'ỵ' has 19 words.

Figures 1-3 above show the graphical representation of Yoruba words in the dictionary, as discussed above. The graph is bars chat representing the number of words for each key in the Yoruba dictionary. From the graph, it shows that key 'a', 'i' and 'o' has the highest number of words in the dictionary, while key 'h', 'u' , 'g' has the least words available in the dictionary. The diagram below shows the Yoruba spell check system using interpolation algorithm ad intellisence.
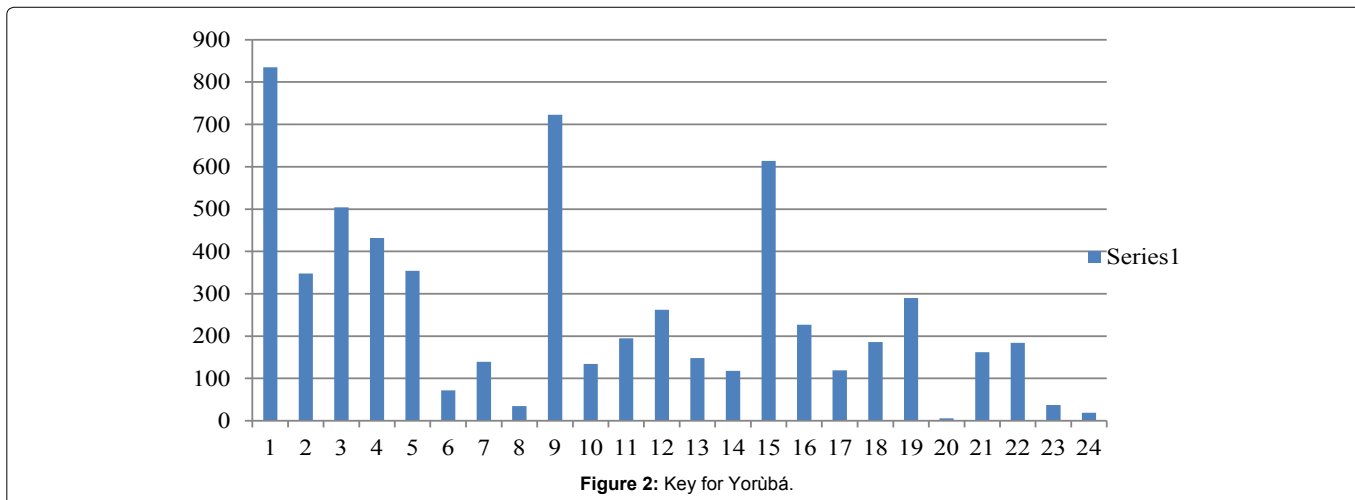

Figure 1: System Architecture.
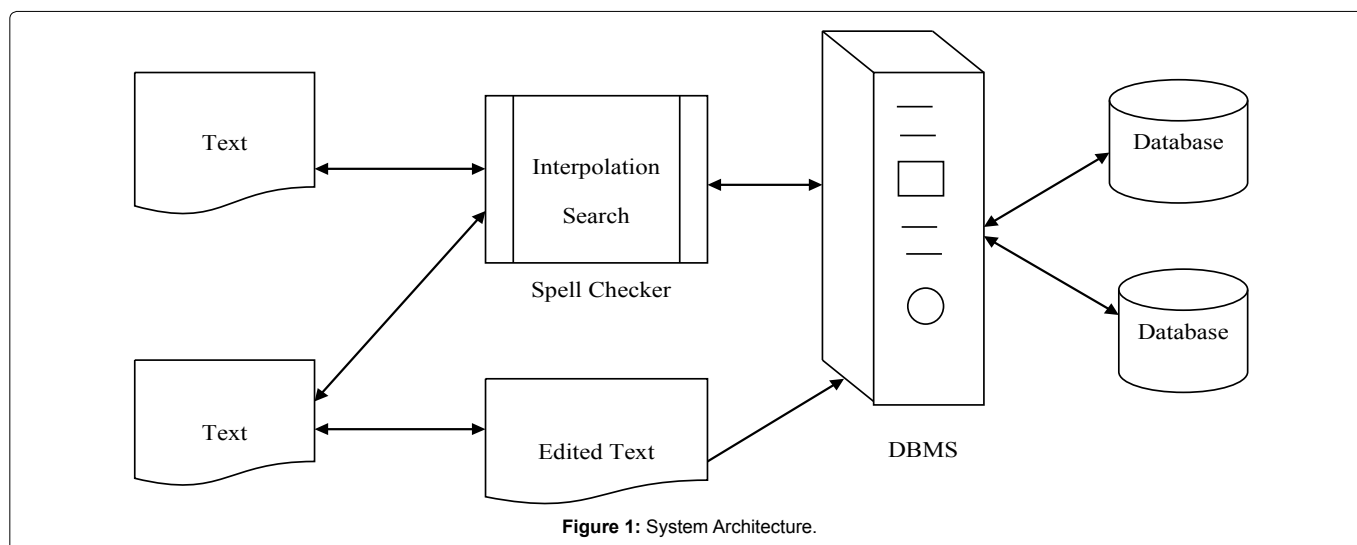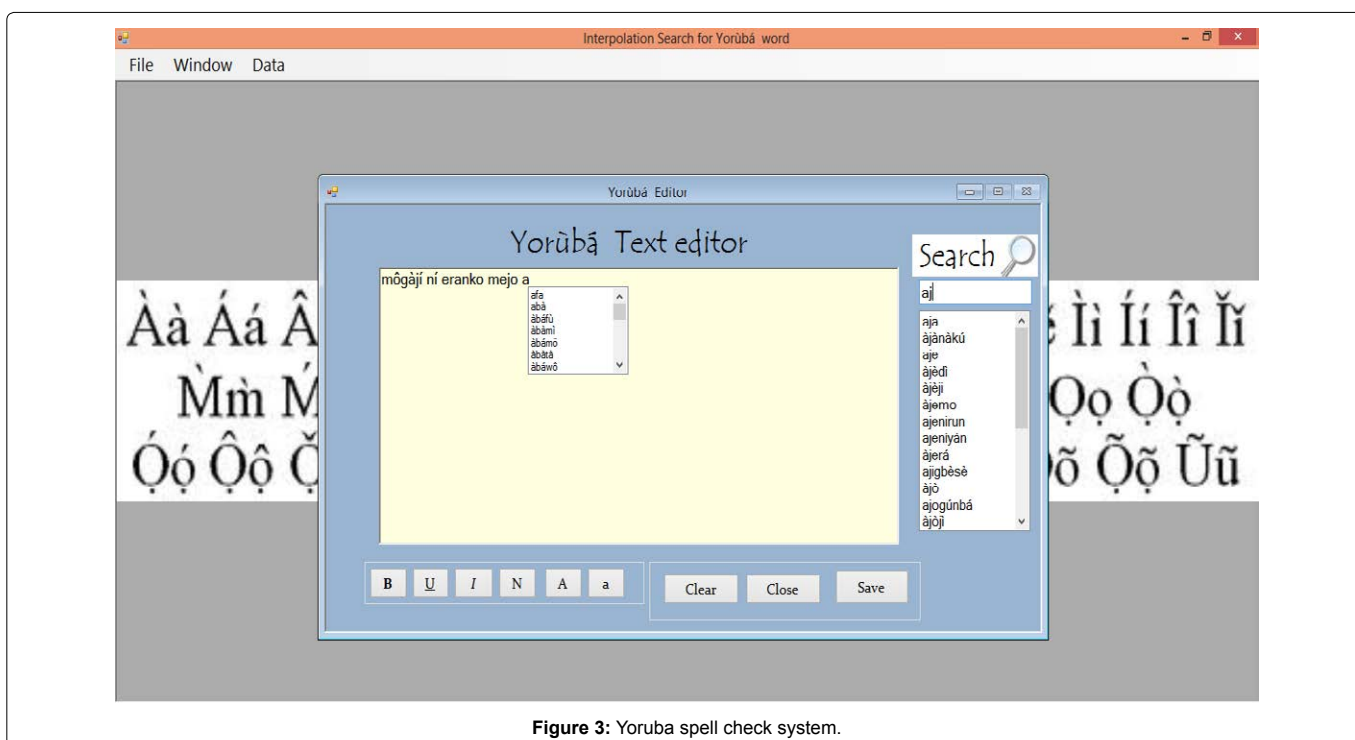
Figure 2: Key for Yorùbá.

Figure 3: Yoruba spell check system.

Conclusion, Recommendation and Limitation

Conclusion

For Nigeria to move forward there is need to appreciate the beauty of the major indigenous languages in the country, whose roles have been taken over by the English language. This is because Government no longer gives priority to the laid down policies on the use of these languages. Parents want their children to speak and learn English language straight from infancy. There are no public profiles for these languages in the information Technology (IT) world.

With the development of Yoruba spell checker system using interpolation search algorithm, it is hopeful that the system will go a long way providing a way to spell check of Yoruba words, thereby improving the use of Yoruba language.

Recommendation

I recommend that the system should be adopted by all that deal with Yoruba words for effectiveness and efficiency. The Government of this country should give priority to the laid down policies on the use of the major indigenous languages especially the Yoruba language by giving an urgent positive attitudinal re-orientation to the proper maintenance of the language.

References

1. Yusuf (2006) Basic Linguistics for Nigeria Languages Teachers Published by Linguistics Association of Nigeria in Collaboration With M And J Grand Orbit Communication Limited and Emhai Press Port-Harcourt.

2. Adeoye O, Adetunmbi AO, Fasiku AL, Olatunji KA (2014) A web-based English to Yoruba noun-phrase machine translation system. Int J English Lit 5: 71-78.

3. Igor AB, Alexander G (2004) Computational Linguistics Models, Resources and Applications 32: 186.

4. Earnest L (2011) The First three spelling checkers. Stanford University, USA.

5. Peterson JL (1980) Computer Programs for Detecting and Correcting Spelling Errors 23: 676-687.

6. Afolabi O, Olu O, Taiwo O, Bayo A (2000) Ijinle - ede at litreso yoruba.

7. Booji G (2007) The Grammar of Words an Introduction to Linguistics Morphology. Oxford University Press, New York.

8. Canon C (2008) A dictionary of the Yoruba language. University Press PLC, Nigeria.

9. Chowdbury G (2003) Natural Language Processing. University of Strathclyde, Glasgow, UK, 37: 51-89.

10. David C (2008) A Dictionary of Linguistics and Phonetics. (6Th edtn) Blackwell, Malden, USA.

11. Fabunmi AF, Akeem SS (2005) Is Yoruba endangered language? J Afr Stud 14: 391-408.

12. Guido I (2004) The interactive Multimedia Linguistics for beginners.

13. JB Sainz, Peña A (1991) Application of Extrapolation Processes to the finite element method. Elsevier.

14. Ingo P (2012) Word Formation in English. Cambridge University Press, Universitat-Gesamthochschule Siegen, Germany.

Author Affiliation

Top

[1]*Department of Computer Science, Federal University of Technology, Akure, Nigeria*

[2]*Department of Information Technology, Federal University of Technology, Akure, Nigeria*