Journal of Literature, Languages and Linguistics www.iiste.org
ISSN 2422-8435 An International Peer-reviewed Journal
Vol.10, 2015

A Simple Data Driven Yoruba Language Dictionary

O. Osunade[1] D. Dawodu[1] O. F. Phillips[2]
1. Department of Computer Science, University of Ibadan, Ibadan
2. Department of General Studies, Ladoke Akintola University of Technology, Ogbomoso

Abstract
The language of a people is an integral part of their lives, because it is synonymous with their identity, culture and environment. Traditionally, the language people speak tells others about their identity but in a country like Nigeria where a lingua franca, English has been adopted, the identity of the people is being suppressed such that the first language of some Nigerians is English instead of their mother tongue and in some extreme cases the indigenous language has been lost. The use of a lingua franca, globalization and civilization should not bring about the death of our indigenous languages, instead, amid all these, Information Technology as the bedrock of our time can be harnessed to propagate our indigenous languages. This work focused on the development of an electronic Yoruba language dictionary that is data driven. The tools and techniques used in this work produced results.
Keywords: Yoruba, dictionary, database, language translation, machine learning

1. Introduction

The world is undergoing a transformation from industrial economies to an information economy, in which value is shifting from material to non-material resources. This transformation has been rightly described as a revolution because of its pace and intensity. At the root of the information revolution is the development of digital technology which has brought about a major shift in the way we conceptualize, describe and anticipate our world. One of the salient social imperatives of the information revolution is the need for humans to communicate through and with machines. This has brought about the need to make machines capable of handling natural languages used by humans as against formal language used by machines.

The field of Human Language Technology (HLT) was developed to provide the necessary knowledge and skills that will enhance the effectiveness and efficiency with which machines mediate communication between humans as well as facilitate communication between humans and machines. So far, developments in HLT have not sufficiently addressed African languages. The purpose of language technology (LT) is to apply technology to language such that the outcome may further the knowledge base of the user, whether that is by providing understanding of a non-native language, perfecting the grammatical use of a written language, encompassing auditory usage or purely quizzical in nature.

Languages with large number of speakers like Yoruba can nonetheless be in danger. Brenzinger (1998) had earlier noted this when he said “even Yoruba, with over 20 million speakers, has been called ‘deprived’ because of the way it has come to be dominated by English in higher education”. Section 53 of the 1999 constitution of the Federal Republic of Nigeria recognizes English as the official language. Moreover, the suppressive effects of English over the Yoruba language and other Nigerian languages are too overwhelming and suicidal.

The need for a Yoruba language dictionary which will facilitate the use of such a language on a large scale cannot be over emphasized. The development of the Yoruba language dictionary will be one of the important by-products of the information revolution. The dictionary will help to promote the level of knowledge and skills in African language. Its usage will be within the frameworks of knowledge production for an industrial society while the information age dawns. Thus, the aim of this work is to develop a digitized Yoruba Language dictionary.

2. LITERATURE REVIEW

The application of language technology to African languages is relatively new and most efforts seem to be incidental. The most consistent efforts motivated and guided by national policy come from South Africa while projects in other countries are based primarily on private initiatives. In a report on HLT(Human Language Technology) development in Sub-Saharan Africa, Roux (2008) reported nine organizations involved in HLT activities in South Africa, one organization in West Africa and two in East Africa. Seven out of the nine organizations in South Africa are based in universities, one is a Semi-Government institution and one is an agency of the Government.

There are individual efforts in some universities. These include Dr. Odetunji Odejobi working on Text to Speech Synproject at Obafemi Awolowo University, Ile Ife, Nigeria, and Dr. Wanjiku Ng'ang'a working on Machine Translation and Dr. Peter Wagacha working on Machine Learning, both at the at the University of Nairobi, Kenya. Apart from these organizations and individuals that are strictly located in Africa, there are a

number of other efforts in various parts of the world that addresses HLT for Africa languages, usually in cooperation with some organizations in Africa. Examples include:

- Local Language Speech Technology Initiative (LLSTI), a project of Outside Echo in the UK
- West African Language Documentation, a project of the University of Bielefeld, Germany in collaboration with the University of Uyo, Nigeria and the University of Cocody, Cote D'Ivoire.

Also, there are significant short-term activities on language technology for African languages both within and outside Africa which have not been sufficiently publicized. For example, in 2002, there was an undergraduate project in Yoruba-English machine translation at the St Mary's College of Maryland, USA.
Machine translation (MT), first conceived in 1949, has come a long way from being one of the first non-numerical applications for computers. The first proposals for machine translation using computers were put forward by Warren Weaver, a researcher at the Rockefeller Foundation, in his July, 1949 memorandum (Weaver, 1949).

The Georgetown experiment in 1954 involved fully automatic translation of more than sixty Russian sentences into English. The experiment was a great success and ushered in an era of significant funding for machine translation research in the United States (Hutchins, 2005). Starting in the late 1980s, as computational power increased and became less expensive, more interest began to be shown in statistical models for machine translation.

On 7 January 1954, the Georgetown-IBM experiment, the first public demonstration of an MT system, was held in New York at the head office of IBM. The demonstration was widely reported in the newspapers and received much public interest. The system had 250 words and translated 49 carefully selected Russian sentences into English — mainly in the field of chemistry. Nevertheless it encouraged the view that machine translation was imminent — and in particular stimulated the financing of the research, not just in the US but worldwide (Hutchins, 2005).

Machine translation research suffered a setback in funding in 1966 with the publication of the ALPAC report whereby it was concluded that machine translation was more expensive, less accurate and slower than human translation, and that despite the expenses, machine translation was not likely to reach the quality of a human translator in the near future.

While research in the 1960s concentrated on limited language pairs and input, demand in the 1970s was for low-cost systems that could translate a range of technical and commercial documents. This demand was spurred by the increase of globalization and the demand for translation in Canada, Europe, and Japan.

By the 1980s, both the diversity and the number of installed systems for machine translation had increased. A number of systems relying on mainframe technology were in use, such as Systran, Logos, Ariane-G5, and Metal. As a result of the improved availability of microcomputers, there was a market for lower-end machine translation systems. Many companies took advantage of this in Europe, Japan, and the USA. Systems were also brought onto the market in China, Eastern Europe, Korea, and the Soviet Union.

At the end of the 1980s there was a large surge in a number of novel methods for machine translation. One system was developed at IBM that was based on statistical methods. Makoto Nagao and his group used methods based on large numbers of example translations, a technique which is now termed example-based machine translation(Lopez, 2007).

There was significant growth in the use of machine translation as a result of the advent of low-cost and more powerful computers. It was in the early 1990s that machine translation began to make the transition away from large mainframe computers toward personal computers and workstations. Two companies that led the PC market for a time were Globalink and MicroTac, following which a merger of the two companies (in December 1994) was found to be in the corporate interest of both. Intergraph and Systran also began to offer PC versions around this time. Sites also became available on the Internet, such as AltaVista's Babel Fish (using Systran technology) and Google Language Tools (also initially using Systran technology exclusively).

The field of machine translation has seen major changes in the last few years. Currently a large amount of research is being done into statistical machine translation and example-based machine translation. In the area of speech translation, research has focused on moving from domain-limited systems to domain-unlimited translation systems. More recently, the French-German project Quaero investigates possibilities to make use of machine translations for a multi-lingual Internet. The project seeks to translate not only webpages, but also videos and audio files found on the Internet.

Today, only a few companies use statistical machine translation commercially, e.g. Asia Online, SDL International / Language Weaver (sells translation products and services), Google (uses their proprietary statistical MT system for some language combinations in Google's language tools), Microsoft (uses their proprietary statistical MT system to translate knowledge base articles), and Ta with you (offers a domain-adapted machine translation solution based on statistical MT with some linguistic knowledge).

Today there is still no system that provides the holy grail of "fully automatic high quality translation of unrestricted text" (FAHQUT) (Lopez, 2007). However, there are many programs now available that are capable

of providing useful output within strict constraints; several of them are available online, such as Google Translate and the SYSTRAN system which powers AltaVista's BabelFish.

The Yoruba language (natively *èdè Yorùbá*) is a Niger–Congo language spoken in West Africa. The number of speakers of Yoruba was estimated at around 20 million in the 1990s. 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, Lucumi, from olukunmi is used as the liturgical language of the Santeria religion of Cuba, Puerto Rico, and the Dominican Republic. It is most closely related to the Owo Itsekiri language spoken in the Niger-Delta and Igala spoken in central Nigeria.

The Yoruba dialect continuum itself consists of several dialects. The various Yoruba dialects in the Yoruba land of Nigeria can be classified into three major dialect areas: Northwest, Central, and Southeast. Of course, clear boundaries can never be drawn and peripheral areas of dialectal regions often have some similarities to adjoining dialects.

- North-West Yoruba (NWY): Abẹokuta, Ibadan, Ọyọ, Ogun and Lagos (Eko) areas
- Central Yoruba (CY): Igbomina, Yagba, Ilésà, Ifẹ, Ekiti, Akurẹ, Ẹfọn, and Ijẹbu areas
- South-East Yoruba (SEY): Okitipupa, Ilaje, Ondo, Ọwọ, Ikarẹ, Ṣagamu, and parts of Ijẹbu

Awofolu and Malita (2002) provided a machine translator for the English and Yoruba languages that would be beneficial both practically and as a pacesetter for other West African languages in computational linguistics using classic syntactic and semantic analyzing algorithms.

Folajimi and Omonayin (2012) describe the Statistical Machine Translation (SMT) system that translates English sentences to Yoruba sentences. The resulting software provides tools to tackle the problem of language translation between Yoruba (Nigeria language) and English language. The software employs a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora.

The work of Afolabi et al (2013) gave an account of Yoruba text-to-speech (TTS) system development using concatenation method. The paper describes the design, evaluation and the analysis of the result shows that 70% respondents accepted its usability.

According to Ninan and Odejobi(2013), it is well known that narratives are influenced by cultural,linguistic, and cognitive factors. They identified and defined entities, elements, and relations necessary for the adequate description of Yorùbá narratives. It was also discussed that theoretical issues in the context of designing a formal framework that are amenable to computational modelling.

The development of Human Language Technology (HLT) is one of the important by-products of the information revolution. However, the level of knowledge and skills in HLT for African languages remain unfortunately low as most scholars continue to work within the frameworks of knowledge production for an industrial society while the information age dawns. Adegbola (2009) reported the work of African Languages Technology Initiative (Alt-i) over a five-year period, and thereby presents a proposal for the acceleration of the development of knowledge and skills in HLT for African languages.

3. METHODOLOGY

This is an experimental work involving software development. The logical design for the software is shown in the flowchart of Figure 1 below.

The following computer hardware specification were used:

- Pentium Processor II
- 128MB RAM
- 1024x768 Screen resolution
- 2GB Hard Disk space

The software development was done in the following software environment

- Windows XP
- Microsoft Office 2003 especially Access 2003.

The coding of the software work was done by programming in C# programming language. The database to store the content of the dictionary was implemented with Microsoft Access**.** The entity-relationship diagram of the database is shown in Figure 2 below.

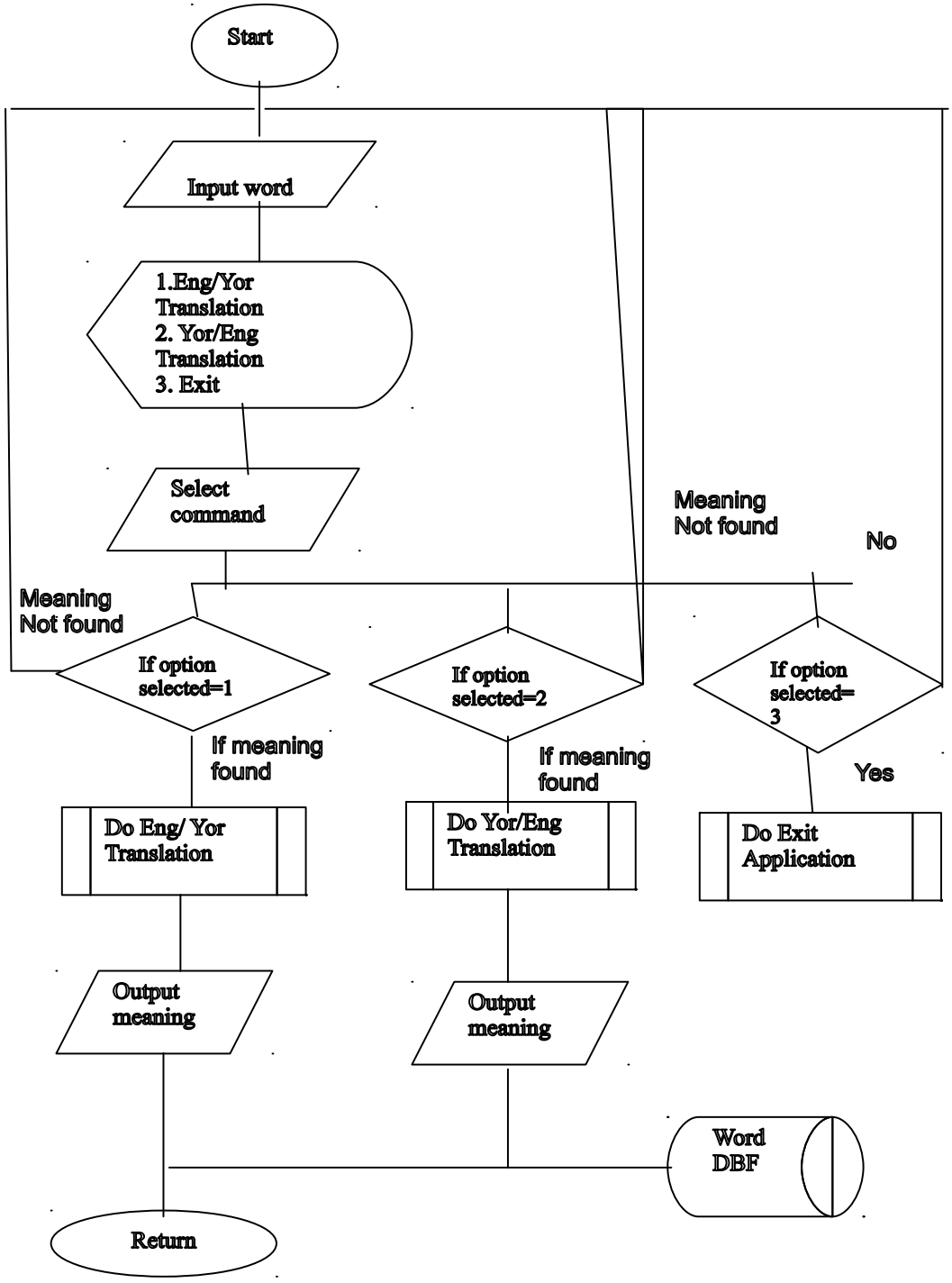


Figure 1: Flowchart for Data-driven Yoruba Dictionary

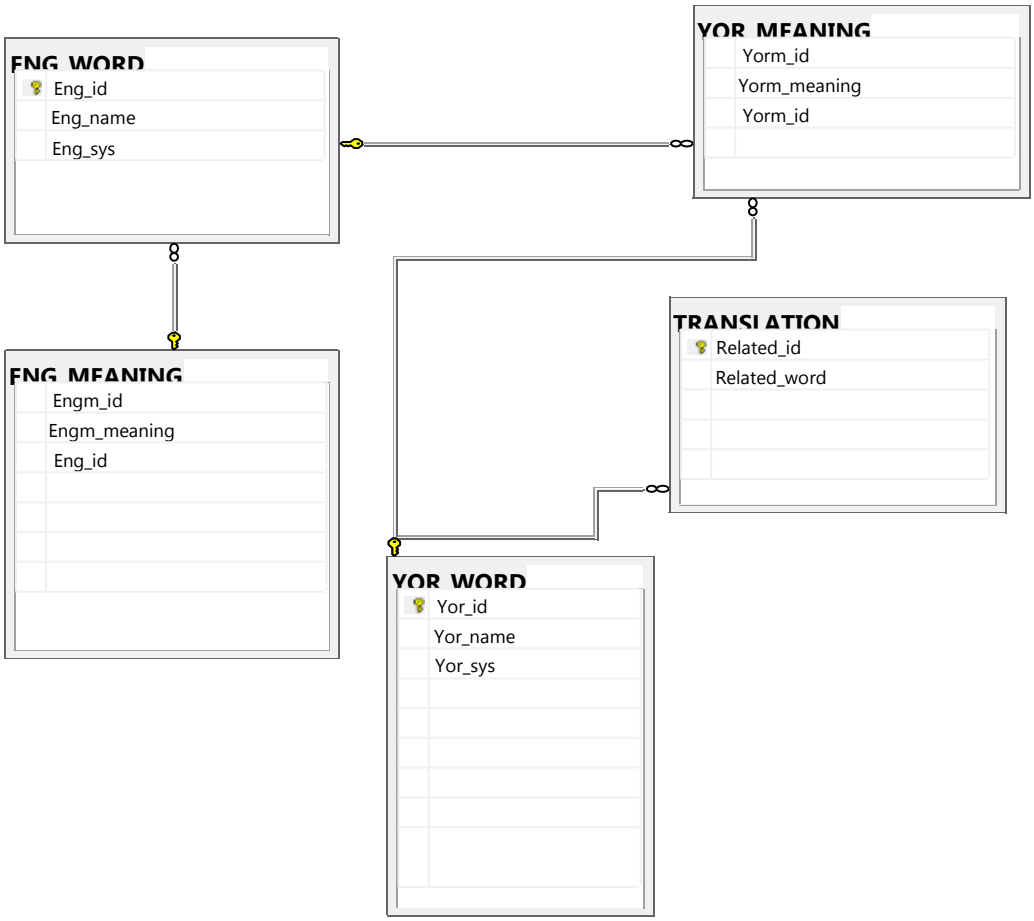


Figure 2: Entity-relationship diagram for database

4. RESULTS AND DISCUSSION

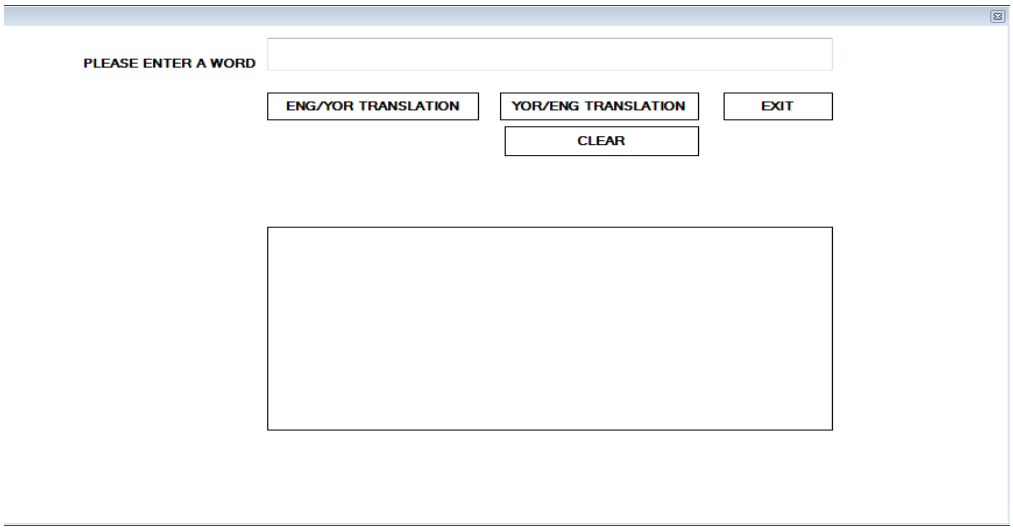
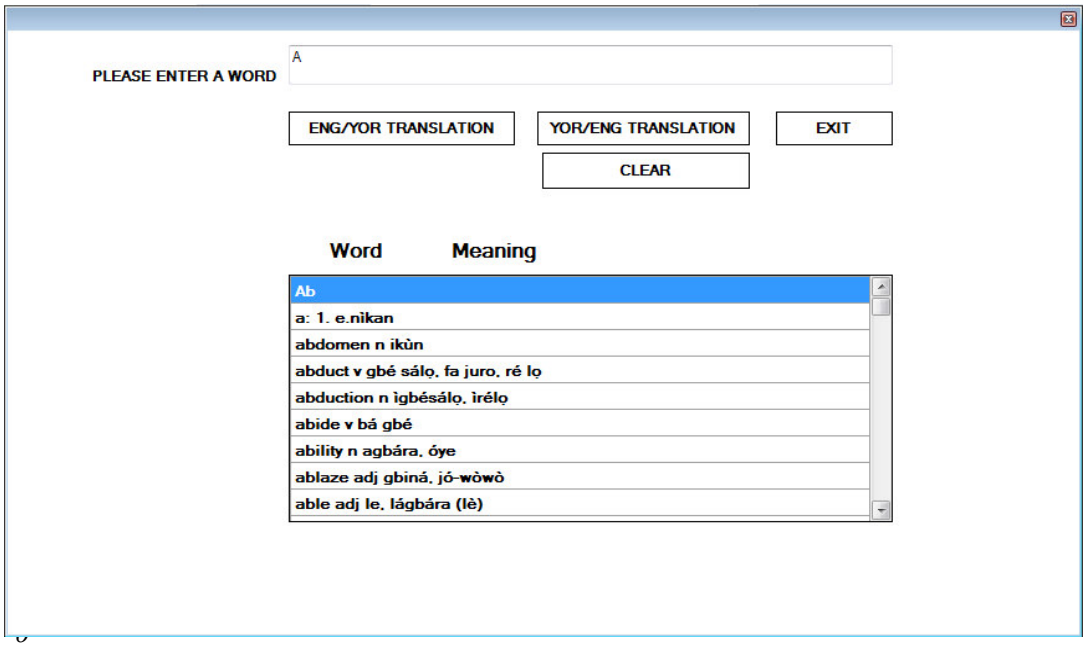


Figure 3: Home page of the translation program

If a user input a word and clicked on Eng/Yor translation button, it will output the corresponding meaning.
If user input a word and clicked on Yor/Eng translation button, it will output the corresponding meaning and a new back button appears, but if the user wishes to switch to Eng/Yor translation, while on Yor/Eng translation, the user need to just clicked on the back button, and the back button will be hidden.

Then the user can then enter an English word and it will output the corresponding meaning.



Yoruba translation interface window

Figure 4 shows that an alphabet 'a' is typed into the 'Please enter a word' button, then 'Eng/Yor Translation' button is clicked. This gave an output that shows all the English words starting with letter 'a' and their corresponding Yoruba meaning.

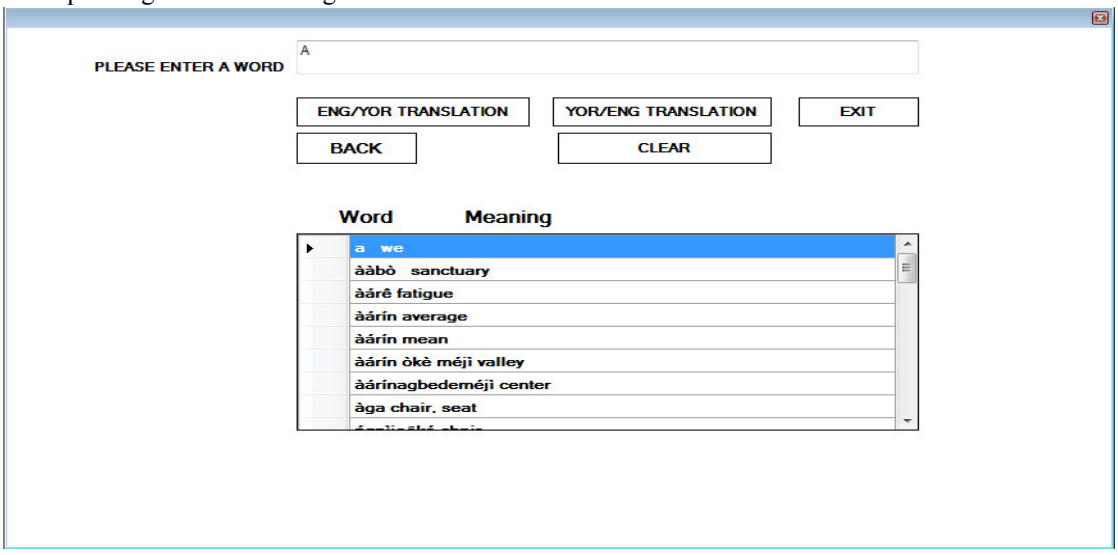


Figure 5: Yoruba to English translation interface window

Figure 5 shows that an alphabet 'A' is typed into the 'Please enter a word' button, then 'Yor/Eng Translation' button is clicked. This gave an output that shows all the Yoruba words starting with letter 'A' and their corresponding English meaning.

5. CONCLUSION

The data driven Yoruba dictionary operates in two phases - preparation of the database and the creation of an application interface. The system functions by taking input (words, in form of text) in Yoruba language from the user, removes punctuation marks like comma, semi colon, colon and blank space(s), looks into the database for the word and returns the meaning of such words. It also displays the corresponding English synonym of words while indicating the corresponding part of speech in form of text.

REFERENCES

Adegbola, T. 2009. Building Capacities in Human Language Technology for African Languages. Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009, Athens,

Greece,pages 53–58. Association for Computational Linguistics

Afolabi, A.; Omidiora, E. and Arulogun, T. 2013. Development of Text to Speech System for Yoruba Language. Innovative Systems Design and Engineering, 4(9):1-8. Special Issue - 2nd InternationalConference on Engineering and Technology Research ISSN 2222-1727 (Paper) ISSN 2222-2871 (Online)

Awofolu, O. and Malita, M. 2002. The making of a Yoruba-English machine translator. Journal of Computing Sciences in Co lleges 17(6):236-237

Brezinger, M. (1998).Endangered languages in Africa. Cologne: Rüdiger Köper Verlag.Brynon, D.1977. Historical linguistics. Cambridge. Cambridge University Press.

Fakinlede, J.K., (2003). Modern Yorùbá Practical Dictionary:Yorùbá-English, English-Yorùbá. Hippocrene Books Inc.

Federal Government of Nigeria (1999). Constitution of Nigeria.

Folajimi, Y. O. and Omonayin, I. 2012. Using Statistical Machine Translation (SMT) as a Language Translation Tool for Understanding Yoruba Language. EIE's 2nd Intl' Conf.Comp., Energy, Net., Robotics and Telecom.| eieCon201

Hutchins, J. W. (2005). Machine translation: a concise history. http://ourworld.compuserve.com/homepages/WJHutchins

Lopez, A. (2007). A Survey of Statistical Machine Translation. Technical report - LAMP-TR-135, CS-TR-4831, UMIACS-TR-2006-47.
http://www.dtic.mil/dtic/tr/fulltext/u2/a466330.pdf

Olufemi D. Ninan and Odetunji A. O 2013. Theoretical Issues in the Computational Modelling of Yorùbá Narratives. Workshop on Computational Models of Narrative 2013. Editors: Mark A. Finlayson, Bernhard Fisseni, Benedikt Löwe, and Jan Christoph Meister; OpenAccess Series in Informatics Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany pp. 153–157

Roux, J. (2008). HLT Development in Sub-Saharna Africa. Report to COCOSDA/WRITE Workshop, LREC2008, Marrakesh. http://www.ilc.cnr.it/flarenet/documents/lrec2008_ cocosda-write_workshop_roux.pdf

Weaver, W. (1949). Translation. http://www.u.arizona.edu/~echan3/539/Weaver-1949.pdf

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar