O.-M. Ndimele, L.C. Yuka and J.F. Ilori [Eds.]

ISSUES IN CONTEMPORARY AFRICAN LINGUISTICS

A Festschrift for Ọladele Awobuluyi

M & J Grand Orbit Communications Ltd.
12/14 Okoroma (Njemanze Street)
Elechi Layout, Mile I, Diobu, Port Harcourt
Nigeria.

e-mail: mekuri01@yahoo.com
phone: 08033410255, 08052709998

ISBN: 978-33527-07-X

Published by

The Linguistic Association of Nigeria

In Collaboration with

M & J Grand Orbit Communications Ltd.
Port Harcourt, **Nigeria**

The Nigerian Linguists Festschrift Series No. 11 2013 (pp. 257-268)

18. The Peculiar Challenges of SMT to African Languages

Clement Odoje
University of Ibadan

The challenges of Machine translation (MT) have been identified and classified but the classification did not consider the peculiarities of African languages. This paper therefore explores the challenges of MT and reclassifies them in relation with the uniqueness of African languages. The study uses the Yorùbá language as a template for other African languages and identifies some of the peculiarities of African languages which include the fact that they are resource-scarce languages; dycritization, demarcation of discipline, and funding, amongst others. The paper recommends measures to overcome some of these challenges.

Keywords: *Machine Translation, African Languages, NLP Challenges, Yoruba*

1. Introduction

Awobuluyi (2010) identifies Machine Translation as one of the significant contributions of Linguistics to technology and ICT:

> … another operation which researchers would like computer to be able to perform. That operation is known as machine translation, and as its name implies, it involves getting computers to translate well-formed and fully idiomatic written expressions in one language into well-formed and equally idiomatic corresponding expression in another language … (Awobuluyi 2010: 34-35)

However, 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 (Adegbola 2009). It is unfortunate that concerted effort is not given to the development of Machine Translation (henceforth MT) in Africa like American and European countries. This may be as a result of the fact that African tribes and kingdoms have been relating with themselves before colonization, therefore translation is not seen as a serious business and does not pose a serious threat to security. However, Odoje (2010) enumerated the needs for MT particularly for Nigerian languages because of the multi-

lingual situation of the nation and the necessity of the inclusion of none speakers of English in the main stream of global activities ranging from politics, government policies, trade and commerce, to research and religion with the use of ICT.

ICT impact in a developing nation like Nigeria cannot be undermined. ICT provides developing nations with an unprecedented opportunity to meet vital development goals such as poverty reduction, basic healthcare, and education, far more effectively than before. Nations that succeed in harnessing the potential of ICT can look forward to greatly expand economic growth, dramatically improved human welfare, a stronger form of democratic government (Kamssu, Siekpe and Elizy 2004). All these are achievable only if the language of ICT is not predominantly dominated by English language and if it does MT should help to be a bridging gap. Therefore, the inclusion of African languages in the ICT or the use of MT in relation with the ICT will not just include the excluded but the majority will not be dominated by the privileged minority who could used English proficiently as observed by Bamgbose (2005: 22) that:

> ... good governance cannot be achieved unless those who make laws at all levels of government and those who implement them can function maximally in a language they are proficient in and can understand what their right and obligations are. *As long as the language of governance is accessible only to the educated elite, majority of citizens will be excluded, thereby making nonsense of participatory democracy. Consequently, the only viable alternation is for Nigerian languages to be used in many domains hitherto dominated by English language.* (*Emphasis mine*)

Though the translation of MT may not be perfect as the research in the field shows now, but at least it gives the users an idea of the content of foreign text as could be observed in the MT translations online. However, efforts have been made to develop MT for African languages exploring different kind of approaches to MT.

2. English-Yorùbá MT

Awofolu (2002) adopts rule based approach to MT because he claimed that Yoruba electronic resources were scarce. JAVA was used to code the structure of both Yorùbá and English sentences. Rowland (1969) *Teach Yourself Yoruba* was used as a base for grammatical information incorporating spellchecker and statistical analyzer on neighboring word lexical categories. The work was indeed a pacesetter using syntactic and

semantic analyzing algorithms but the concern is the structure of the sentences and linguistic feature of the lexical entries having known the deficiency of *Teach Yourself Yoruba* in the light of books like Awobuluyi (1978) Bamgbose (1990) and current syntactic findings. The report did not also discuss how structural and word ambiguities were handled by the machine since this is one of the major challenges of rule-based MT.

Another noticeable Yorùbá-English MT is Sunday (2008). The work was on interlinear translation. The total number of equivalent words is 1000 each for both Yoruba and English. The data were collected from Yorùbá Metalanguage (second edition), a dictionary of Yorùbá, Encarta eDictionary, Oxford Advanced learners' Dictionary (6th edition), a dictionary of synonyms and antonyms, and Yoruba mini multimedia dictionary. Visual Basic 6.0 was used as programming language and Ibadan Ayo SIL Charis-Alt-I font was used for inputting Yorùbá data and Microsoft served as the platform for the database design. However, the deficiency of Sunday (2008) is that the system can only translate at word level. Phrasal or sentential translation coupled with idiomatic or figurative expressions cannot be translated by the system. For instance, the system can translate (1a&b) but constructions like (1c) will not be translated by the system.

1 a. ọmọ => child
b. father => bàbá
c. ọmọdúdúnáà => the dark child

Our observation is that 2000 total lexical entries cannot represent the realities of the two languages involved. Besides morphology which was not incorporated into the system and is very important for a system of that nature for maximum output translation; we also observe that word may have more than one meaning based on the context of usage. For example University Press Nigeria PLC new bilingual dictionary (which is an ongoing project coordinated by Professor Kola Owolabi) shows the deficiency of not considering meaning in context in translation. For instance "keen" has 18 meanings against one meaning if not used in context. Consider example (2) below:

2. Keen => (i) le (ii) lile (iii) gbona (iv) gbigbona (v) mímú (vi) lóye (vii) oloye (viii) imọ̀lára (ix) ìfura (x) ní tara (xi) onítara (xii) niwàra (xiii) pípé (xiv) ní wàwára (xv) oníwàwára (xvi) jíjá fáfá (xvii) nímọ̀ (xviii) onímọ̀ as each of the words shows their meaning in the context below:

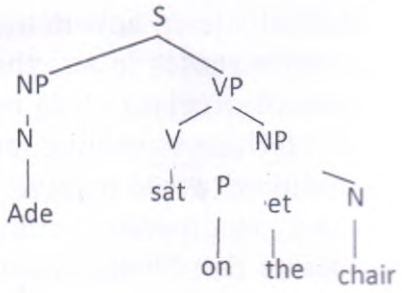
- Nigeria versus Ghana football match will be a *keen* competition => ìdíje bọ́ọ̀lù aláfẹsẹ̀gbá láàárín oríl̀è-èdè Nàìjíríà àti Gánà (Ghana) yóó jẹ́ ìdíje *líle/gbígbóná*
- The football match between Ghana and Nigeria will be very *keen* => ìdíje bọ́ọ̀lù aláfẹsẹ̀gbá láàárín Gánà àti Nàìjíríà yóó *le/gbóná* (gan-an/púpọ̀)
- Olú is not very *keen* on the idea => Olú kò *nítara/ níwà□wára/ níwàra/ lóye/ nímọ̀* púpọ̀ nípa èrò náà
- I want a *keen* razor for my shaving => Mo fẹ́ abẹ mímú kan fún fífá irungbọ̀n mi
- Olú has a *keen* sense of smell => Olú ni *ìmọ̀lára/ìfura* ìgbọ́ òórùn
- He's a *keen* supporter of opposition party => ó jẹ́ onítara/oníwà□wára/olóye/onímọ̀/oníwàra alátìlẹ́yìn (fún) ẹgbẹ́ alátakò.
- Even at the age of 90 my mother still has a keen mind => kódà bí iyá mi tìl̀è jẹ́ ẹni àádọ́rin ọdún, wọ́n ọkàn jíjá fáfá (tí ó já fáfá)/ pípé.

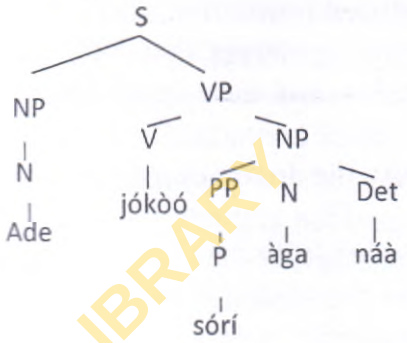
Meaning in context like the above is very important so as not to lose any meaning in context like the interlinear software of Sunday (2008) will. Though it is highly recommended that the software be adopted for technical terms storage which will serve as an aid to professional and none professional translators. The software could be upgraded also to an online databank which will accommodate new entries as they are being discovered. It could also be adopted for a project like the ongoing compilation of Yorùbá Metalanguage.

Odoje (2010) and Eludiora, Salawu, Odejobi and Agbeyangi (2011) intend to capture human language intuition in the computer for translation purposes using rule based approach. Odoje (2010) coded the sentence structure of Yoruba and English in Prolog. The system can generate sentences in each of the languages using finite rule to generate infinite sentences. It could also translate sentences though the word entries were very minimal and the word to be translated must first exist in the system dictionary. One of the major deficiencies of the work is that it is not users' friendly. The users must be able to use Prolog before they can access it; therefore there is need for an interface so that it could be users friendly. The system could not also translate idiomatic expression like *Ó ta térunípàá* the system will translate it to mean *he kicked the tarpaulin* instead of *he is dead/ he died.* Tonal manipulations which Yorùbá language is known for were not also captured.

Eludiora, Salawu, Odejobi and Agbeyangi (2011) adopted python for the rule-base MT. POS tags were used to identify the class of each lexical entry and parser parsed each sentence to its phrasal and word level categories. Just that the parsing could be highly for example consider (3) extracted from Eludiora, Salawu, Odejobi and Agbeyangi's (2011).

3.





Parsing of this type does not follow the linguistic structure of both languages and is rather confusing. What is considered as NP (object) is not headed by noun and one wonders why it is regarded as NP How can a preposition head an NP? This is mis-informing. This even violates the principle of merging though we assume the writers used Standard Theory as the base of their analysis. Consequent on the above, there is an urgent need to review either the structure (2 above) or the whole paper. It should be noted that the difference in the structural head of noun phrase was achieved in that *my father* is not translated as *mi bàbá* but *bàbá mi.*

3. Statistical Machine Translation

This is an approach to MT using human translated materials of language pairs as corpus for computer to translate new sentences using probability. Och (2006) is of the view that there has been an enormous boom in MT research in recent years. There has been not only an increase in the number of research groups in the field and in the amount of funding, but there is now also optimism for the future of the field and for achieving even better quality. The major reason for this change has been a paradigm shift away from linguistic/rule-based methods towards empirical/data-driven methods in MT. This has been made possible by the availability of large amounts of training data and large computational resources. Other important factors have been the adoption of automatic evaluation techniques, organized evaluations, and the availability of elaborated tools for building systems. Today, using a data-driven approach, it is possible for a small research lab to build with moderate effort and publicly available tools state-of-the-art

MT systems that compete with or outperform state-of-the-art commercial systems. The best systems trained on large amounts of data are able to produce MT quality that is significantly superior to classical rule-based systems – as has been shown in various recent evaluations.

To Lopez (2008) SMT treats the translation of natural language as a machine learning problem. By examining many samples of human-produced translation, SMT algorithms automatically learn how to translate. He then observes that SMT has made tremendous strides in less than two decades, and new ideas are constantly introduced. Koehn (2010) believes that SMT has actually energized the field of Machine translation bring to reality the dream of automatic language translation closer to reality.

This approach to MT does not necessarily incorporate the linguistic knowledge or detail though Hassan (2009) opines that Phrase-Based SMT lacks the capacity to produce more grammatical translations and handling long-range reordering while maintaining the grammatical structure of translation output. He integrates syntactic structures into the system to produce more fluent MT output. He was of the opinion that syntax can help Phrase-based SMT systems to produce well formed translation output by the use of syntactically-guided translation models, language models and reordering techniques.

Brown et al (1990) explain that a statistical translation system requires a method for computing language model probabilities, a method for computing translation probabilities and a method for searching among possible source sentences *S* for the one that gives the greatest value for Pr (*S*) Pr (*T/S*). Meaning that, if we want to translate Yorùbá sentence *Y* to English sentence *E*, we have to look for the highest value for Pr (*Y*) given Pr (*E/Y*). To do this, Keohn (2010) stipulates that rules such as the chain rule or Bayes rule allow us to reformulate distributions. He explains further that interpolation allows us to compensate for poorly estimated distributions due to sparse data.

Therefore, to have a good SMT, there is need for a very large language corpus for language modeling probably adopting n-gram model for linguistic simple-minded for the language pairs. This enables the computer to identify the linguistic structure of the languages. It is however unfortunate that many of African languages have very little corpus required for an exercise of this nature. For example, Yorùbá language has very few translated material reported by Awofolu (2002) and Odoje; so is Swahili (see Pauw, Wagacha and Schryver, 2011). Hence, there are very view Statistical Machine Translation in African languages and none in Yorùbá language apart from this ongoing project. This is as a result of the fact that African languages are considered as resource-scarce from a language

technological point of view (see Pauw, Wagacha and Schryver, 2011). Meaning that, African languages are languages that have small or economically disadvantaged user base which are typically ignored by the commercial world (Chan and Rosenfeld 2012).

4. Language Toolkit Systems

Language toolkit is a software or application for natural language processing usually to perform sentence detection, tokenization, POS-tagging, text chunking, lemmatisation, coreference analysis and resolution, named-entity detection among others[1]. For SMT, the following are the software available as open sources: Giza ++, Moses, Phrasal, cdec, Joshua, Jane, Pharaoh, etc. Moses is adopted for this research because it has an online discussion group where researcher discuss their challenges and seek help. It also has a website and it manual is readily available.

5. Challenges of Developing SMT

Och (2006) points out three significant challenges of SMT. The collection and the use of huge amount of data is the first challenge explained. He maintains that hundreds of billions of words and translation models trained on hundreds of millions of words will require very large computational resources, a corresponding software infrastructure, and a focus on systems building and engineering. The second challenge he made mention of is the need for new evaluation metric since the large corpus available now made it difficult for BLEU to distinguish MT output and human translation output on some standard data sets. While the third challenge is the SMT independence since currently, the best data-driven MT systems do not employ Natural Language Processing tools such as linguistic parsers, parts-of- speech taggers or explicit word sense disambiguation, and there have been very few success stories in integrating those components. As accurate as Och's observation is, African languages were not put into consideration while reaching this conclusion hence there is need to consider the challenges of SMT in the light of an African language

5.1. Peculiar Challenge to SMT for African Languages

Like Och (2006) observed, there is a need for very large corpus for data driven MT. This is lacking for many African languages, for example, SMT translates better for specialized domains like hotel reservation, flight booking and safety instruction, and weather forecast etc than general domain; may be because of its limited words (see Koehn 2009 and Hutchins and Somers1992). It is observed that translation for these specialized domains are scarce at least for Nigerian languages. Egbokhare

(2011) observe this particularly the need for local airplane to use Nigerian languages in Nigeria. He commented that:

> One has heard it said that there are too many languages in Nigeria; hence, it will be virtually impractical to meet the needs of every group. This argument is specious because with five languages, English, Nigerian Pidgin, Hausa, Yorùbá and Igbo, the linguistics needs of over 90% of Nigerians can be met. Nigeria must insist that airlines address Nigerians in the languages they understand best and planes flying in our airspace must adhere to language requirement as part of airline safety requirement…

However, there are literary materials that have been translated basically for academic and social purposes. To achieve SMT for Yorùbá-English language, the following literary books were used

Title of Book	*Author*	*Translation Title*	*Translation Author*
Ògbójú Ọdẹ nínú Igbó Irúnmọlẹ̀	D. O. Fagunwa	The Forest of a Thousand Demons	Wole Soyinka
Igbó Olódùmarè	D.O. Fagunwa	The Forest of God	Gabriel Ajadi
Igbó Olódùmarè	D.O. Fagunwa	In the Forest of Olódùmarè	Wole Soyinka
Àdìtú Olodùmare	D. O. Fagunwa	The Mystery of God	Olu Ọbafemi
Aké: The Years of Childhood	Wole Soyinka	Aké Láti Ìgbà Èwe	Akinwùmí Ìsọ̀lá

This leads to peculiar challenges of Yorùbá-English SMT

5.1.1 Lack of Online Data for SMT

Initially one thinks that Jehovah Witness website could have translated materials which could be used for this exercise but a visit to their website reveals that most of the Yorùbá materials on the website were not translated materials. Hence it is difficult to get online corpus like European Parliament documents available online at http://www.europarl.europa.eu/ and the available ones like in the table above are hard copies which need to be converted to soft copies for usage. In other words, there is no website where Yoruba–English translated equivalent material could be accessed today.

5.1.2. Orthography

Another challenge for Yorùbá-English SMT is the issue of orthography. Among the five books above, only the Aké's translation or otherwise language retriever as Isọ̀lá calls it was the only book written in the new orthography, the rest were written in old orthography. This will greatly affect tokenization and language modeling if not converted to new orthography which is consequently time consuming and capital intensive and may affect the translation of the machine. Getting a typist with the knowledge of new orthography to convert the hard copy to soft became a very hard nut to crack since none of the available typist has the required knowledge for the conversion.

5.1.3. Diacritization

Yorùbá like many African languages is a tonal language which must be represented in the written form as well as other diacritic representations in the language. Many typists have just been using any symbol of their choice for the diacritics without standardization most especially using Unicode for any computer to recognize. Therefore, Yorùbá text typed by a typist could not be used because of this challenge since it was not typed in Unicode that can be recognized by Linux operating system.

5.2. Operating System

In this part of the world, Window Operating System is very common but our research findings shows that Moses is not usually run on Window. Therefore, Ubuntu which is a Linux operating system was used which took some time to learn before the process of building the system began. More so, the researcher does not have much knowledge of computational science.

5.3. Demarcation of Disciplines

Demarcation of academic disciplines as it is observed in Africa and Nigeria in particular denies to greater extent the collaboration that is needed for this kind of research, after all, nobody can do it all.

5.4. Funding

Just as Adegbola (2009) repots, there is no funding for projects like this in Africa ass is being done in other part of the world. This has not encouraged interested individuals or groups to join in the development of MT in Africa. If funding is available, it will be easy to source for equivalent translated material in their soft copies and get enough date for works of this nature.

6. Conclusion

With the peculiar challenges of developing SMT for Yorùbá and in general African languages identified above, there is need for a reclassification of the challenges of SMT generally which might give an insight to solving these challenges since African languages have features that can be of tremendous improvement to SMT generally if properly harnessed.

Note

1. For the list of language toolkit visit http://en.wikipedia.org/wiki/List_of_natural_language_processing_toolkits

References

Adegbola, T. 2009a. Building Capacities in Human Language Technology for African Languages. EACL 2009 Workshop on Language Technologies for African Languages – AfLaT (pp. 53- 58). Athens: Association for Computational Linguistics.

Adegbola, T. 2009b. *Indigenising Human Language Technology for National Development*. A lecture delivered as First ARCIS Distinguished Guest Lecture, Mach 18.

Awobuluyi, O. 1978. *Essentials of Yoruba Grammar*. Ibadan: Oxford University Press

Awobuluyi, O. 2010. The Role Linguistics in Nation Building: Lecture in Honour of Ayo Bamgbose, delivered at the University of Ibadan.

Awofolu, O. 2002. *The Making of A Yoruba-English Machine Translator*. St Mary's City: St. Mary College of Maryland.

Bamgbose, A. 2005 "Language and Good Governance", Nigerian Academy of Letters (NAL) 2005 Convocation Lecture, Conference Centre, University of Lagos, August 11.

Clarkson, R. - *Adaptation of Statistical Language Models for Automatic Speech Recognition*. A PhD Dissertation Submitted to Engineering Department, University of Cambridge

Egbokhare F, 2011. *The Sound of Meaning*. An Inaugural Lecture delivered at the University of Ibadan on July 14, 2011.

Eludiora, Salawu, Odejobi and Agbeyangi's (2011) Ife: Machine Translation.

Kamssu, J. Siekpe J. S. and Elizy J.A 2004 Shortcomings to Globalization: Using Internet Technology and Electronic Commerce in Developing

Countries. *The Journal of Developing Areas*, Vol. 38, No. 1 pp. 151-169 http://www.jstor.org/stable/20066700 Accessed: 25/01/2012, 06:23.

Koehn, P. 2010, *Statistical Machine Translation*. New York: Cambridge University Press.

Lopez, A, 2008 Statistical machine translation.ACM Comput. Surv., 40, 3,Article 8 (August 2008), 49 pages DOI = 10.1145/1380584.1380586 Nigeria.

Noone, G. 2003. *Machine Translation A Transfer Approach*. www.scss.tcd.ie/undergraduate/bacsll/bacsll_web/nooneg0203.pdf.

Och 2006. Challenges in Machine Translation. TC-STAR Workshop on Speech-to-Speech Translation, June 19-21, 2006 Barcelona, Spain.

Odoje, C. 2010, The Role of Syntax in Rule-Based Machine Translation. An M.A Thesis submitted to the Department of Linguistics and African Languages, University of Ibadan, Nigeria

Pauw, G., Wagacha P., G. Schryver 2011. Towards English-Swahili Machine Translation. MTMRL

Rowland, C. 1969. *Teach Yourself Yoruba: A Complete Course for Beginners.*

Sundat, T. 2008. *Yoruba-English Interlinear Translation Machine (software)*. An Undergraduate Long Essay submitted to the Department of Linguistics and African Languages, University of Ibadan, Ibadan.