Machine Translation of Noun Phrases from English to Igala using the Rule-Based Approach.

Sani Felix Ayegba[1], **Osuagwu O.E.** [2], Njoku Dominic Okechukwu[3]

[1]Department of Computer Science, Federal Polytechnic Idah, Kogi State, Nigeria
felixsani@yahoo.com
[2] Department of Computer Science, Imo State University, Owerri
profoliverosuagwu@gmail.com
[3]Department of Electrical/Electronics Engineering, Imo State Polytechnic, Umuag

Abstract

We live in a multilingual society where large volumes of documents are produced in different languages. Translation is the means by which information generated in one language can be accessed by someone in a different language. Igala is one of the languages spoken in Nigeria. Igala is the ninth largest ethnic group in Nigeria and the language is spoken by about 2.5 million people. The main objective of this research is to model a language processor that can accept as input Noun Phrases in English language and translate same to Igala language. The two core technologies, corpus based and rule based technologies for building machine translation systems were carefully studied. Due to the structural differences between English and Igala, noun phrases coupled with the non- availability of large amount of parallel aligned corpus for English and Igala language, the rule based technology was adopted to develop the model. The model was implemented using VB.net programming language as front end and Microsoft Access as back end. The application was tested on 120 randomly selected English noun phrases using the Bilingual Evaluation Understudy (BLEU) method for evaluating Machine Translation systems. An accuracy of 90.9% was obtained.

Key words: *Translation, Igala language, Language processor, corpus based technology, rule based technology, Bilingual Evaluation Understudy.*

Introduction

Language is the medium of communication. Human language is purposively to communicate ideas, emotions, feelings, desires, to co-operate among social groups, to exhibit habits etc which can be translated along a variety of channels [2]. There are over 6,800 living languages in the world which reflects the scope of linguistic and cultural diversity. Access to information written in another language is of great interest and the means of sharing information across languages is translation, therefore creating tools for translating from one language to another is very crucial contribution to human development. Without translation, there can be no communication, except among those who share a common language and many voices will not be heard without this critical function.

Translation is critical for addressing information inequalities. A study conducted by ***Common Sense Advisory*** on behalf of ***Translators without Borders*** finds that translation is critical for the public health, political stability, and social wellbeing of

African nations [11]. [16] Showed that due to differences in culture and the multilingual environment in India, inter-language translation was necessary for the transfer of information and sharing of ideas. The need for translation is also very glaring in the business community. It has been observed that language barriers between companies and their global customers are stifling economic growth and in fact, forty-nine percent of executives say a language barrier has stood in the way of a major international business deal, nearly two-thirds (64 percent) of those same executives said language barriers are making it difficult to gain a foothold in international markets, whether inside or outside your company, your global audiences prefer to read in their native languages[18], it speeds efficiency, increases receptivity, and allows for easier processing of concepts This clearly shows that language translation is a matter of absolute necessity in the globally united and yet linguistically and culturally separated world in which we live.

The work of translation was originally carried out by human translators. At a point the supply of translation services could no longer keep pace with the demand for translated content, moreover human translation is costly, time consuming and inadequate for addressing the real-time needs of businesses to serve multilingual prospects, partners and customers. The inherent limitations of human translation made the search for an alternative means of translation paramount. The search led to the discovery of what is known today as machine translation or computer assisted translation. Machine Translation is the use of computers to automate some or all of the process of translating from one language to another [1]. This need has prompted research organizations and government agencies to develop tools for automatic translation of text in an attempt to achieve wider outreach and bridge the gap of language diversity [17].

Igala is the language of the ethnic group located at the eastern flank of the confluence of rivers Niger and Benue. They are the ninth largest linguistics group in Nigeria [12]. Geo-politically, they are described as belonging to the middle belt or north-central of Nigeria. They are bordered on the north by Benue and Nassarawa States, on the West by River Niger, on the East by Enugu State and on the South by Anambra State [5]. Igala land is 120 Kilometres wide and 160 Kilometres long. It is located approximately between latitudes 60 30" and 80 North and longitudes 6030" and 7040" East and covers an area of about 13,665 square kilometers. The population of the Igala people is estimated at two-million in the late 1990s [5]. Historically, they are said to be linked to the Yoruba, the Jukuns and the Binis (Edo) and the northern Ibos. Owing to their central location, they have mutually interacted and lived with the Idomas, Bassa-Nkomo, Nupe, Igbirra and Hausa people. The Igala ethnic group is densely populated in their settlements around the major towns such as Idah, Ankpa and Anyigba. They are also found in Edo, Delta, Anambra, Enugu, Nassarawa, Adamawa and Benue States. However, the bulk of them are indisputably found in Idah, Ankpa, Dekina, Omala, Olamaboro, Ofu, Igalamela/Odolu, Ibaji, Bassa (and even Lokoja and Ajaokuta) Local Government Areas of Kogi State [5].

The aim of this research is to develop a system for translating Noun Phrases from English to Igala. The specific objective is to carry out a computational analysis of English to Igala noun phrases translation processes and to model a language processor that will have the capacity to accept as input noun phrase in English

language and translate it into Igala language.

Information Communications Technology (ICT) has not made any significant inroad in empowering Africans towards development because 90 percent of existing content and applications are in the English language [3]. Igala is not left out of this. The impact of ICT among the Igala people has not reached a level where it can be said to be significant. This is due to the fact that existing contents and applications are in a foreign language.

The outcome of the research will result in greater access to information in Igala language. It will also give Igala language a public profile in the information technology world and provide a platform for people to really appreciate the beauty of their indigenous language and also help to develop Igala language and elevate it to the level of languages of developed nations. This will lead to the preservation of the Igala culture and values and also serve as the springboard to the much needed development of Igala society and by extension, the Nigerian nation, African Continent and perhaps the entire globe.

Machine Translation Technologies

Machine Translation systems (MT) can be classified according to their core methodology. Under this classification, two main paradigms can be found: the rule-based approach and the corpus-based approach. Within the rule based paradigm three approaches can be distinguished: Direct, Transfer and Interlingua [17] [13]. Rule-based systems are based on linguistically-informed foundations requiring extensive morphological, syntactic and semantic knowledge. The input is transferred to the target using a large set of sophisticated linguistic translation rules. Translation rules are created manually, demanding significant multilingual and linguistic expertise. Therefore, rule-based systems require large initial investment and maintenance for every language pair [7]. Also within the corpus-based paradigm, three other approaches can be further distinguished: example-based and statistical-based and context based. Under the corpus-based approach the knowledge is automatically extracted by analyzing translation examples from a parallel corpus built by human experts. The advantage is that, once the required techniques have been developed for a given language pair, machine translation systems should (theoretically) be quickly developed for new language pairs using provided training data.

Although the rule based system requires significant amount of linguistic knowledge, the knowledge acquired for one natural language processing system may be reused to build knowledge required for a similar task in another system. [9] posited that rule-based approach is better than its counterpart corpus-based approach for two main reasons:

1: less-resourced languages, for which large corpora, possibly parallel or bilingual, with representative structures and entities are neither available nor easily affordable, and

2: for morphologically rich languages, which even with the availability of corpora suffer from data sparseness. Both the rule based and corpus based technologies have the strength and weaknesses. Due to the inherent weaknesses of these technologies none has been able to singly achieve satisfactory level of accuracy and quality in translation. This development led to the search for a better option which is the hybrid approach. The hybrid approach is a machine translation technology that integrates various machine translation technologies [6] [19]. The technologies compliment each other to produce a more satisfactory result. Some popular machine

translation systems which employ the hybrid technology are PROMT, SYSTRAN ands Asia Online. App Tek delivered its first hybrid machine translation in 2009.

English and Igala Language

Igala and English language differ significantly in morphology, syntax and semantics. Igala is a register tone language, it is isolating which means there is a one-to-none correspondence between words and morphemes, and it also has agglutinative features. Igala is a fixed word order SVO (subject, verb, Object) like English but the arrangement of words in noun phrase and adjective phrase are not the same. English places modifiers before nouns in noun phrases, Igala does the reverse, and nouns are placed before modifiers.

Words are grouped into categories called parts of speech. There are eight parts of speech in English language. They are Nouns, Verb, Adjectives, Adverb, Conjunctions, preposition and determiners. In Igala there are only two major parts of speech - Nouns and verbs, the others are seen as derivatives of these two main types. In Igala language tone plays a significant part in realizing parts of speech [10].

Noun Phrase

A noun Phrase (NP) is a phrase in which a noun or pronoun is the governor or head word, optionally accompanied by a modifier set. NP can be pre-modified or post-modified. If the modifier is placed before the noun, the NP is pre-modified. If the modifier is placed after the noun then the NP is post-modified. English allows both forms of modification. Igala allows only post-modification, modifiers are placed after nouns. A noun phrase consists of three parts, the head which is the principal part and other two optionally occurring parts. Possible modifiers of NP are:

Definite Articles (the)
Indefinite Articles (a, an)
Demonstratives (this, that)
Quantifiers (few, every, several…)
Cardinal numbers (one, two, three …)
Ordinal numbers (first, second, third…)
Possessive Pronouns (my, your, their…)
Pre- determiners (all, both, half..)

Igala and English differ in the syntax of noun phrases. The placement of modifiers is not the same in the two languages. We denoted Definite Articles by DA, Indefinite Articles by IDA, Demonstratives by Dem, Cardinal Numbers by Cdn, quantifiers by Qtn, Ordinal Numbers by Odn, Possessive pronouns by PPN and Pre-determiners by PDT, and generated the table below that shows the rules that governs the arrangement of lexical units in noun phrases for English and Igala languages.

.

Table 1: Noun Phrase Transformational Rules

Proposed Systems Components

Rules	English: NP=	Igala: NP=	Examples	
R1	DA + N	N + DA	E	The house
			I	Unyi lẹ
R2	IDA + N	N	E	A house
			I	Unyi
R3	DA + ADJ + N	N + ADJ + DA	E	The beautiful house
			I	Unyi alifia lẹ
R4	IDA +ADJ + N	N + ADJ	E	A red cap
			I	ọtajiya kpikpa
R5	DA + ADJ + ADJ + N	N + ADJ + ADJ + DA	E	The beautiful white dress
			I	Ukpo alifia fifu lẹ
R6	IDA + ADJ + ADJ + N	N + ADJ + ADJ	E	A beautiful white dress
			I	Ukpo alifia fifu
R7	PPN + N	N + PPN	E	My dog
			I	Abia mi
R8	PPN + ADJ + N	N + PPN + ADJ	E	My red cap
			I	ọtajiya mi kpikpa
R9	PPN + ADJ + ADJ + N	N + PPN + ADJ + ADJ	E	My beautiful white shirt
			I	Afẹ mi alifia fufu
R10	DEM + N	N + DEM	E	That house
			I	Unyi lẹ
R11	PDT + N	N + PDT	E	Few books
			I	ọtakada re e
R12	PDT + PPN + N	N + PPN + PDT	E	All their houses
			I	Amunyi ma chaka a
R13	CDN + N	N + "m" + CDN	E	Two doors
			I	ọna meji
R14	ODN + N	N + "ẹk" + ODN	E	Second house
			I	Unyi ekẹji
R15	DA + CDN + N	N + "m" + CDN + DA	E	The three doors
			I	amọna mẹta lẹ
R16	DA + ODN + N	N + "ẹk" + ODN + DA	E	The third building
			I	Unyi ẹkẹta lẹ
R17	DEM + CDN + N	N + "ẹk"+CDN + DEM	E	That fourth house
			I	Unyi ẹkẹlẹ lẹ
R18	PPN + ODN + N	N + PPN +"ẹk" + ODN	E	My second car
			I	Moto mi ẹkẹji
R19	QTN + N	N + QTN	E	Many rooms
			I	Ejefu wewe
R20	QTN + ADJ + N	N + ADJ + QTN	E	Few beautiful houses
			I	Amunyi alifia re e

The English-Igala Dictionary

Dictionaries are the largest components of a machine translation system in terms of the amount of information they hold. If the system is expected to perform well the dictionary should be more than simple word list. The size and quality of the dictionary limits the scope and coverage of the system and the quality of the translation that can be expected.

A full form bilingual lexicon of English and Igala is developed. The dictionary contains Igala equivalents of all English words. English words together with their derivatives, part-of-speech tag and Igala equivalents are listed in the dictionary. The dictionary is used for the translation of words of English to Igala. The fields in the dictionary are: word_id, English word, igala_equivalent, part_of_speech.

Parts - of – Speech Tagging (POS Tagging)

The process of assigning part-of-speech to each word in a sentence is called part-of- speech tagging (POS tagging). It is the annotation of each word in a sentence with a part-of- speech marker. POS tagging is used in machine translation for word sense disambiguation, syntax analysis (parsing) and reordering of words to obtain correct sentence structure in the target language. There is no available dictionary for Igala language either in hard or softcopy and no parallel corpora of any size; the full-form bilingual dictionary is filled manually. During the entry of words in the full-form bilingual dictionary, the POS of each word is identified and appropriate tag is assigned to it from our tag set.

Morphological Analysis

Morphological analysis is the identification of a word-stem from a full word-form and sometimes also the identification of the syntactic category of the stem. According to [13] there are two techniques for dealing with morphological analysis in machine translation:

i. ***Full – form lexicon***: in this technique all the inflected variants of the word are listed in the lexicon.

ii. ***Morphological analysis Component***: this is a rule based module that analyzes a word and relate it to its root or base form. There is no existing morphological analyzer for Igala language. The full form lexicon technique which lists all words and their derivatives in the bilingual dictionary is used in this study. The reason for this choice is that English is not a highly inflecting or morphologically rich language [8] [1] [15] and Igala is an isolating language which means that is it also poor in morphology.

The proposed System Architecture and Modules

Figure 1 is architecture diagram of the overall design of the proposed system. It has four main components, a ***parser***, an ***analyzer***, a ***transformer*** and a ***generation*** component.

Parser is an algorithm which produces a syntactic structure for a given input. The parser is the first component of the rule based machine translation system and it is used on the source (English) side. The Parser built in Natural Language Processing Toolkit was downloaded and used. The parser is used to verify the grammatical correctness of the English input.

The ***analysis component*** consists of three modules: Preprocessor, Tokenizer and Postprocessor.

The ***preprocessor*** counts the number of words in the English noun phrase input and declares three arrays of the size of the

number of words for use by the other modules.

A basic text processing operation is **tokenization** which is the breaking up of raw text or sentence into words. This function is performed by the ***tokenizer***. The input sentence is broken up at this point into words. It recognizes a word whenever a space is encountered which signifies the end of the word. It then puts each of the tokens (words) into one of the arrays created by the preprocessor.

The Postprocessor opens the full-form bilingual dictionary for each of the tokens in the array, retrieves its part of speech (pos tag) and Igala equivalent. It stores the retrieved pos tags and Igala equivalents in the remaining two arrays respectively. Thus arrays of English word (the tokens), pos tags and Igala equivalents are generated by the postprocessor. Tokens not found in the full–form bilingual dictionary are listed and displayed by the postprocessor.

The ***transformer*** consists of a set of transformation rules. These rules are used to build the Igala language equivalent of the input English sentence.

The generation component contains a module called synthesizer which formats the Igala language and displays it as output.

System Implementation

The full form bilingual lexicon which contains the English words and Igala equivalents together with the parts of speech was developed in Microsoft Access database platform. The ***rule engine*** which applies a collection of lexical and syntactic transfer rules to generate the Igala noun phrase was developed using ***vb.net platform***. The translation interface is shown in figure 2.

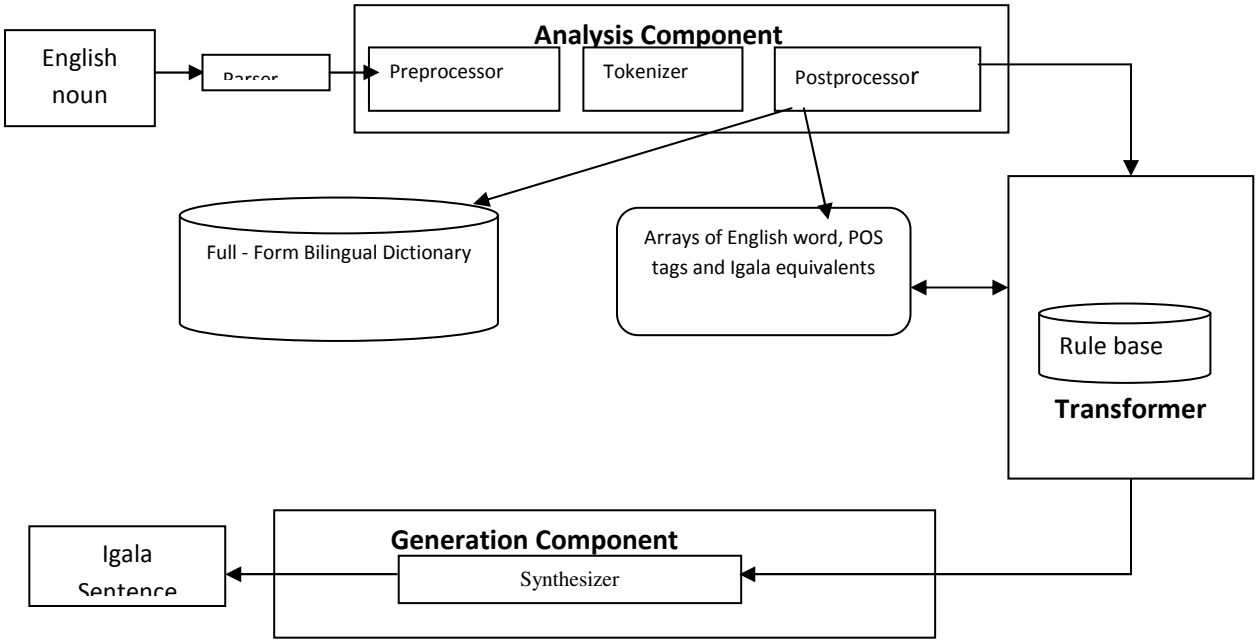


Figure 1 Proposed system architecture

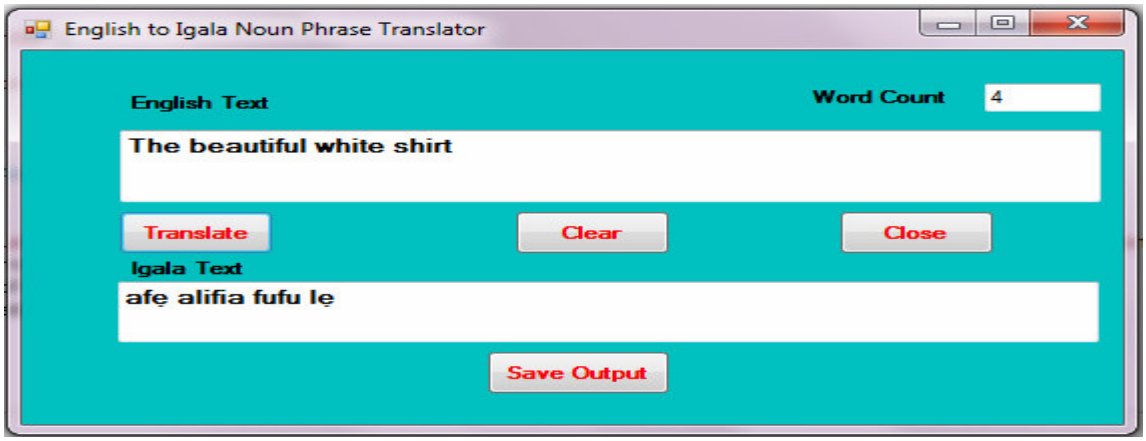


Figure 2. Translation interface

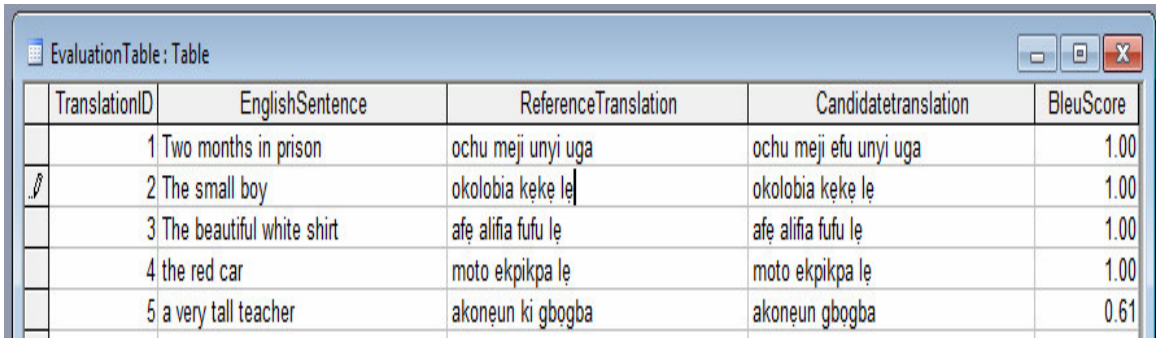
After entry of the English noun phrase the user clicks the Translate button. The Igala equivalent of the noun phrase is generated and displayed in the Igala text control. Both the input and output can be saved in a table by clicking the save output button.

Testing and Evaluation

The accuracy of any machine translation system is always evaluated by comparing the results to human judgment. The most popular automatic evaluation method which is Bilingual Evaluation Understudy (BLEU) proposed by IBM [14] was adopted for evaluating the output of the developed machine translation system. The method is language independent and highly correlated with human evaluation. BLEU is based on the idea that the closer the output of a machine translation is to a reference (professional human) translation, the better. BLEU scores range from 0 to 1. [4] stated that BLEU scores above 0.30 generally reflect understandable translations and BLEU scores above 0.50 reflect good and fluent translations.

A corpus of 120 English noun phrases was created and given to professional English Expert for translation from English to Igala. This was used as the reference translation. The reference translation was stored in a table in the database called Evaluation Table which has the following fields: Translation ID, English Sentence, Reference Translation, Candidate translation and BleuScore. The same set of English noun phrases were given to the developed system for translation. When a noun phrase is entered into the control on the template, it is translated into Igala, its ID is retrieved from the table, the translated text is stored in the Candidate translation field based on the Translation ID. This Operation was repeated until all the noun phrases were translated and stored. A module called BLEU evaluation module is then executed to compute the BLEU score for each of the translated phrase.

The sample output is shown in figure 3.


EvaluationTable : Table

TranslationID	EnglishSentence	ReferenceTranslation	Candidatetranslation	BleuScore
1	Two months in prison	ochu meji unyi uga	ochu meji efu unyi uga	1.00
2	The small boy	okolobia kẹkẹ lẹ	okolobia kẹkẹ lẹ	1.00
3	The beautiful white shirt	afẹ alifia fufu lẹ	afẹ alifia fufu lẹ	1.00
4	the red car	moto ekpikpa lẹ	moto ekpikpa lẹ	1.00
5	a very tall teacher	akoṇẹun ki gbọgba	akoṇẹun gbọgba	0.61

Figure 3. Sample output

The result of the 120 test noun phrases was analyzed. The table blow shows the analysis.

BLEU Score range	No. of phrases	Percentage value
>=0.8 and <=1.0	86	71.7
>=0.5 and <=0.7	23	19.2
<=0.4	11	9.2

Table 2. Analysis of Results

According to [4] BLEU scores above 0.50 reflect good and fluent translations. From the table total percentage score above 0.50 is 71.7 + 19.2 = 90.9. Therefore accuracy of 90.9% was achieved.

Conclusion and Future work

The rule based English to Igala noun phrase machine translation system was successfully implemented. Accuracy of 90.9% was achieved. The accuracy can be improved by improving and extending the full form bilingual lexicon. This work only handles the translation of noun phrases which is part of a complete sentence. The system will be developed further to produce acceptable translation of complete sentences and hosted on the internet for public use. The system will be of immense benefit to the Igala people as it will help to further develop the language and elevate it to the level of languages of developed nations.

References

[1] Arnold, D, etal (1994) Machine Translation: An Introductory Guide, NCC \ Blackwell, London, ISBN: 1855542-17x.

[2] Banjo, A.E, Jibowo, A.V *The use of principles and Theories of Translation in Languages:* A case study of Yoruba. Journal of Communication and Culture; International Perspective Vol. 2 No 3, Dec, 2011.

[3] . Chetty, M (2004) Information and Communications Technologies (ICTs) for Africa's development. Published by African Forum on Science and Technology for Development, NEPAD, www,nepadst.org. available on line at www.eldis.org/go/home&id=16694&type=Document, retrieved on 3rd march, 2014.

[4] .Denkowski, M., Lavie, A. (2010b). Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks. In Proceedings of the Association for Machine Translation in the Americas AMTA.

[5] Egbunu, Fidelis Eleojo, *Education and Re-orientation of Igala Cultural Values,* African Journal of Culture, Religious, Educational and Environmental Sustainability (AJCREES), Vol. 1, No. 2. Pp. 66 – 82. Dec., 2013.

[6] Fahime Mohammadpour, Abbas Ali Ahanger, Nader Jahangiri, *Building a Hybrid \ Translation System for Translating from English into Persian.* English Linguistics research, Vol. 1, No. 2, 2012. ISSN 1927-6028, e-ISSN 1927-6036.

[7] Hieu, H, (2011) *Improving Statistical Machine Translation with Linguistic Information,* PhD Thesis, Institute for Communicating and Collaborative Systems School of Informatics University of Edinburgh.

[8] Hutchins, W.J, Somers, H.L. (1992). *An Introduction to Machine Translation.* Academic Press, London, ISBN: 0-12-362830-x

[9] Khaled Shaalan, Rule-*based Approach in Arabic Natural Language* Processing International Journal on Information and Communication Technologies, Vol. 3, No. 3, June 2010.

[10] Mbah,E and Ayegba M. *Tone in Igala Language: An Autosegmental Analysis*. Journal of Igbo Language and Lingustics, No. 4 ISSN 05987518, pp 67-75, 2012

[11] Nataly, Kelly, Donald A. DePalma and Vijayalaxmi *Hedge, The Need for Translation In Africa,* May 2012. Available at http://www.commonsenseadvisory.com/Portals/0/downloads/Africa.pdf. Retrieved on 20th December 2013.

[12] Omachonu, G.S (2011) *Igala Language Studies*. LAP LAMBERT Academic Publishing GmBH & Co. KG, Saarbrucken, Germany, ISBN: 978-3-8465-5822-5.

[13] Omar Shirko, Nazlia Omar, Haslina Arshad and Mohammed Albared: *Machine Translation of Noun Phrases from Arabic to English Using Transfer-Based Approach,* Journal of Computer Science 6 (3): pp 350-356, 2010, ISSN 1549-3636.

[14] Papineni, K., Roukos, S., Ward, T., Zhu, W. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311—318, 2002.

[15] Reuta Tsarfaty, Djame Seddah, Yoav Goldberg, Sandra Kubler, Marie Candifo, Jennifer Foster, Yannick Versely, Ines Rehbin, Lama Tounsi. *Statistical Parsing of Morphologically–Rich Languages: What, How and Whither*. In Proceedings of NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically- Rich languages , Los Angeles, Califiornia, pp 1-12, June 2010.

[16] Sitender, Seema Bawa, *Survey of Indian machine Translation System*, International Journal of Computer Science and Technology, Vol 3, Issue 1, Jan-March, 2012, pp 286-290.

[17] Sneha, Tripathi, Juran, Krishna Sarkhel, *Approaches to machine Translation* Anals of Library and information Studies, Vol. 57, December 2010, pp 388-393.

[18] Top 5 Big Language Business Problems Solved by Machine Translation. Downloaded from http://www.sdl.com/campaign/lt/enterprise/general/wp-top-5-big-language-problems.html on 23rd February, 2014.

[19] YAM AB ANA Kiyoshi, KAMEI Shin-ichiro, MURAKI Kazunori, DOI Shinchi, TAMURA Shinko, SATOH Kenji. *A hybrid approach to interactive Machine Translation integrating rule-based, corpus based, and example- based method*, NEC Research Institute, U.S.A. available online at http://ijcai.org/Past%20Proceedings/IJCAI-97- VOL2/PDF/026A.pdf, retrieved on 10th April, 2014.