English – Igala Parallel Corpora for Natural Language Processing Applications

Sani Felix Ayegba
Department of Computer Science, Federal polytechnic Idah, Kogi State, Nigeria

Abu Onoja
Department of Mathematics/Statistics
Federal Polytechnic Idah
KogiState, Nigeria

Musa Ugbedeojo
Department of Computer Science, Federal polytechnic Idah, Kogi State, Nigeria

ABSTRACT

Parallel text is a fundamental requirement for the development of corpus based or data driven machine translation systems and other Natural Language Processing applications. The unavailability of this valuable linguistic resource has greatly hampered the development of NLP applications in English and Igala language. This study is aimed at creating English – Igala parallel text. The result of the study in addition to providing linguistic resource will enhance language learning. Various algorithms for automatic construction of parallel text such as STRAND, PTMiner, PTI, WPDE, BITS were studied to determine their appropriateness in creating English –Igala parallel text. Wikipedia and the Bible which are excellent sources of parallel or comparable corpora were also gleaned. Existing algorithms and other sources of Parallel text were found to be unsuitable for the construction of English – Igala parallel text due to the unavailability of contents rendered in Igala language on the web and in electronic form. A combination of manual and machine assisted translation was used to generate the parallel text. English – Igala parallel corpora comprising of 50,000 aligned sentences was obtained.

General Terms

Artificial Intelligence, Natural Language Processing, Computational Lingusitics

Keywords

Parallel text, Natural Language Processing, Machine Translation, comparable corpora, corpus based or data driven machine translation systems, linguistic resource

1. INTRODUCTION

Translation is the task of transforming a text written in one human language called source language into an equivalent text in another human language referred to as the target language. Translation which is automatically performed from a human language into another by using computer software and hardware technology is called automatic translation, or machine translation (MT). Machine translation evolved as a result of the scarcity and slow speed of human translators. Machine translation systems provide people with greater access to information making them more informed, widen their horizon and improve their national understanding.

The development of language technology applications such as machine translation, information extraction, and multilingual text analysis systems requires the availability of basic linguistic resources such as large parallel corpora (texts accompanied by their translation in another language, also known as bi-texts or text paired with its translation into a second language). Parallel resources offer the same functionality across languages, using the same categories and the same input and output format.

Igala history dates back to Egypt between 36 and 841 AD. The invasion of Egypt by Arabs from Southern Europe led to their migration from Egypt to their present abode in Kogi State [18]. 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].

2. USEFULNESS OF PARALLE TEXT

Large scale parallel corpus is an indispensable language resource for the creation of models for corpus based machine translation. English – French Parallel corpus obtained from proceedings of the Canadian parliament debates was used for the initial work on statistical machine translation. Parallel corpora such as those made up of United Nations publications were used to train machine translation models that translates Chinese and Arabic to English [6]. EuroParl corpus developed by Koehn has been used in the creation of SMT systems for up to 110 language pairs [7]. JRC – Acquis parallel corpus enabled the creation of SMT systems for 462 European language pairs [8]. Studies by [3] [2] indicated that MT results can be improved using parallel corpora involving more than two languages through exploiting triangulation. According to [14] Parallel corpus have the following additional uses:

- Producing multilingual lexical and semantic resources such as dictionaries and ontologies

- Training and testing information extraction software
- Annotation projection across languages for Named Entity Recognition
- Improving monolingual text analysis by exploiting patterns in various other languages
- Automatic creation of parallel tree banks
- Cross-lingual textual entailment
- Cross-lingual plagiarism detection
- Cross-lingual word sense disambiguation
- Checking translation consistency automatically
- Multilingual and cross-lingual clustering and classification
- Creation of multilingual semantic space
- Translation studies and comparative language studies

Other uses include cross-language information retrieval and deriving multilingual and monolingual text processing tools [13] projected learning of linguistic structure, annotation projection across languages for Named Entity Recognition [4].

It is clear from the above mentioned uses of parallel corpus that the creation of English-Igala parallel corpus will result in the availability of a valuable language resource which will facilitate the development of language technology software for Igala language.

3. MOTIVATION

The advancement of any group of people to a large extent is dependent on the extent to which their language is developed for creative and productive thinking as well as self-mobilization and mass communication. Language development is so important that [12] posited that the survival of any group of people in the world is tied to what they can make of their language.

The study is aimed at providing linguistic resources for the 'resource-poor' Igala language. "Resource poor" with respect to language technology refers to lack of a large corpus in electronic form or lack of native speakers trained in computational linguistics [17]. The availability of this linguistic resource which facilitates development of language technology applications can result in the following benefits: (i) development of more business potentials, (ii) maintenance of linguistic diversity and preservation of rich cultural diversity, (iii) Igala language is in danger of digital extinction because the available digital support for the language is very weak. This situation will be averted by the availability of this resource. (iv) Aid research in computational linguistics (v) Increase globally visibility of Igala language.

4. APPROACHES AND TOOLS FOR PARALLEL CORPUS CONSTRUCTION

Parallel corpora can be created manually or automatically. Manually constructing large parallel corpora is a tedious task that consumes both time and resources (Mohammed etal. 2015). Due to the cumbersomeness of creating parallel corpora manually several algorithms and systems for mining parallel corpora from the web were developed. These include:

STRAND [16] is a popular application for extracting parallel text from the web by identifying pairs of pages that are mutual translations.

PTMiner was developed by [11] to construct large parallel corpora from the web. It used search engines to identify web sites that are likely to contain parallel pages, and then used the URLs collected as seeds to further crawl each web site for more URLs.

BITS [9] was developed and used to generate English-German parallel text by computing content-based similarity in English-German bilingual dictionary.

Parallel Text Identification System (PTI) [1] facilitates the construction of parallel corpora through the alignment of pairs of parallel documents from collections of multilingual document. The system crawl the web to extract multilingual web documents using a web spider.

Web Parallel Data Extraction (WPDE) [20] is an automatic system for large scale mining of parallel text from existing English-Chinese bilingual web pages in a variety of domains.

[19] proposed two approaches for creating large scale parallel corpora. The first approach is to extract translated sentences from source – target translation text while the second is to obtain new translations from supporting volunteer translators.

4.1 Challenges of creating English-Igala Parallel Corpus

Survey of the web and existing literatures clearly shows that parallel data are nonexistent for English – Igala. It is also evident that there are no parallel data for Igala and any other language. Parallel data here refers to any of the following: (i) A Parallel corpus is a sentence-aligned corpus containing bilingual translations of the same text. (ii) A noisy parallel corpus contains non-aligned sentences, but is mostly parallel. (iii) A comparable corpus contains non-sentence-aligned, non-translated bilingual documents that are topic aligned. (iv) A very-non-parallel corpus contains disparate, non-parallel bilingual documents that can either be on the same topic or not. Sufficient quantity of Igala text does not exist in electronic form either on electronic media such as CDs and other storage devices and the web. There are no collections of bilingual articles that although are not exact translations, but contain similar information for the language pairs. Although [16] identified the Web as the largest and most accessible source of comparable texts, there are no comparable text for English and Igala language. For these reasons the algorithms discussed above (STRAND, PTMiner, WPDE, BITS, PTI) which work where there are documents on the web that are mutual translations for the language pairs are not suitable for the construction of English – Igala parallel text. Both methods proposed by [19] cannot be used in the creation of English – Igala parallel corpora due to the unavailability of documents that have contents rendered in English and Igala. Moreover supporting volunteer translators are not handy.

Wikipedia is a web-based free multilingual encyclopedia where articles are generated through public collaboration. Currently it contains over 40 million articles in 293 languages [21]. Wikipedia has features that make it a rich source of parallel data. Two articles in different languages that describe the same concept as comparable documents can be viewed. Many Wikipedia articles link directly to the counterpart articles describing the same concept in other languages. Though a rich source of comparable document, Wikipedia contains no article in English and any other language that has a counterpart in Igala.

Studies by [15] [4] showed that the Bible is used extensively in the creation of parallel corpora. Using the Bible in the construction of parallel corpora is feasible where the Bible exist in electronic forms for the language pair. English Bible exists in electronic form but for Igala there is none. Attempt to obtain Igala Bible in electronic form by scanning the Pages of Igala Bible and converting to editable text failed due to encoding mismatch. Text resulting from the scan appeared garbled or gibberish.

It is clear from the foregoing that although the web and the Bible are rich sources from which parallel text can be mined but the almost complete absence of contents in English-Igala language on the web and unavailability of Igala Bible in electronic form greatly limits their use in the construction of English –Igala parallel corpora. We therefore had to look away from the automatic method of parallel text creation to achieve our objective.

5. CONSTRUCTION OF ENGLISH-IGALA PARALLEL TEXT

English – Igala parallel text was constructed using the following methods and sources.

5.1 Transcription and Translation

Christian messages in English on audio and video were transcribed. 20,000 sentences were generated from the transcription. 5000 of the sentences were manually translated to Igala language by professional translators. The remaining 15000 sentences were translated using the ruled based English to Igala Machine translation system [17].The output of the post edited by professional translators. 20,000 parallel texts were generated from transcription, manually and machine assisted translation. Table 1 is an extract from the generated parallel text.

Table 2: Sample English – French Parallel text (Night Watch)

	English	**Igala**
1	Let us be on our feet as we welcome the man of God.	Akwanẹ dago oji ẹrẹ wa alu ka gwa Adu ỌJỌ ki wọla lẹ
2	Praise the Lord	A jẹnyu ỌJỌ
3	I bring you greetings from the throne of grace	U nẹ ugwa nwu mẹ kwo ọgbede ufẹdọ ẹnyọ ỌJỌ
4	This morning by the grace of God	Odudu yi lefu ejumomi ỌJỌ
5	We shall be considering the topic	Anya dibe kado ọla koji nwu kakini
6	The power of your foundation	Ukpahiu ki defu uchanẹ wẹ
7	Evil things happen to believers and they wonder why	Ẹnwu biẹnẹ nache kẹrẹbu amakẹdọno, ma la na kakini ẹnwu chi.

Table 1: Sample English – Igala PT generated through transcription and translation

5.2 Conversion of exiting French – English P.T to English – Igala P.T

Lonweb.org translates Short stories in English into many languages in a convenient parallel text format to facilitated language learning. The story (Night Watch) translated from English to French in Parallel text format was copied from the web site [http://www.lonweb.org/daisy/ds-french-nightwatch.htm]. Table 2 is a sample.

SURVEILLANCE DE NUIT	**NIGHT WATCH**
Daisy s'était levée tôt ce matin de printemps car elle travaillait sur une affaire dans la ville voisine.	Daisy had got up early that spring morning because she was working on a case in the nearby town.
Elle arriva à son bureau à huit heures moins le quart avec à la main un sac en papier contenant des petits pains au lait, et elle mourrait d'envie d'une tasse de café.	She arrived at her office with a paper bag in her hand containing fresh cream buns at a quarter to eight and was dying for a cup of coffee.
Comme elle mettait la clé dans la serrure, une voix de femme appela : "C'est ouvert, Daisy".	As she put the key in the lock, a woman’s voice called out, "It’s open, Daisy."
C'était Pam, la femme de ménage.	It was Pam, the cleaner.

The French column of the corpora was deleted leaving only the English column. The English sentences were manually translated into Igala by Igala native speakers and validated. Table 3 is an extract from the parallel text obtained.

Table 3: English – Igala Parallel text obtained from translation of Night watch from English to Igala

	English	Igala
1	Daisy had got up early that spring morning because she was working on a case in the nearby town.	Desi takpa gwanẹ odudu lẹ todu k’ nẹ ọla efu ewo ọwọ ka.
2	She arrived at her office with a paper bag in her hand containing fresh cream buns at a quarter to eight and was dying for a cup of coffee.	I gwudu t’ unyi ukọlọ nwu kpai ikpa are efu ọwọ nwu k’ amekpo j’efu nwu adiko agogo mẹjọ bọ miniti mẹgwẹlu, ebi iti emọ la na kpọ nananana.
3	As she put the key in the lock, a woman’s voice called out, "It’s open, Daisy."	Alu ki na du ọma igede tunyu ọna, igbọ ukomu onobulẹ ki kakini ichẹ bi.
4	It was Pam, the cleaner.	Pam alianẹ dẹ
5	"How about some	Pam, ukpuchọ ha?, Desi

	breakfast, Pam?" said Daisy with a smile and then noticed Pam had obviously been crying.	nene kpai anyi taki li ki Pam tete arakwu.
6	"Pam, whatever has happened?	Pam, ẹnwu che?
7	Come on sit down and have some breakfast with me.	Lia gwanẹ kẹ jẹ ukpuchọ kpai omi.
8	Please tell me what's bothering you."	Ka mi ẹnwu k' adihianyi wẹ
9	Pam was a hard-working woman with two children to bring up.	Pam che onobulẹ ki kpẹdọ ru ukọlọ. I nẹ amọma meji ki anẹ
10	She did the cleaning for the whole building which meant seven offices.	I li ọgbọ unyi lẹ chakaa, uña unyi ukọlọ mebie.

The process was repeated for "The Search for Lorna", "The Surprise", "A Nice Little trip", "Imogen". A total of 350 English – Igala parallel sentences was obtained.

5.3 5.3 News Articles

Nigeria Television Authority (NTA) has responsibility for broadcasting. The branch of NTA located at Idah which is the principal town of Igala people broadcast news in English and Igala language. Versions of English and Igala news exist in hardcopies. Hardcopies of News and Igala language that are mutual translations of each other were collected and converted to electronic form. The electronic copies of the English and Igala documents were carefully studied. No noise was observed. The translations were faithful.

A php application that uses MYSQL database as backend called **EIPTExtractor** was developed to extract English – Igala parallel sentences from the electronic copies. The conceptual architecture of the php application is shown in figure 1.

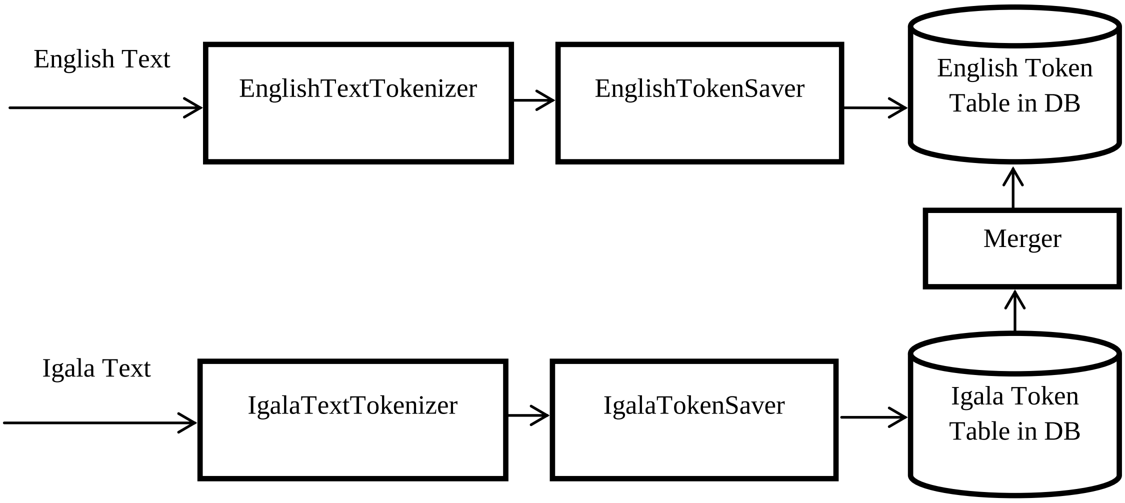

Figure 1: Conceptual Architecture of EIPTExtractor

The structure of **EIPTExtractor** above breaks down into the following phases:

Phase (1) - English text: The first input to the system consists of text in English language.

Phase (2) – EnglishTextTokenizer: The input in English text could be a paragraph which is made up of a number of sentences. In this phase the system splits the text into sentences. It recognizes a sentence whenever a full stop is encountered which signifies the end of the sentence.

Phase (3) – EnglishTokenSaver: This module stores the resulting English sentences in a database table called tbl_english_tokens.

Phase (4) - Igala text: The second input to the system consists of text in Igala language.

Phase (5) – IgalaTextTokenizer: The input in Igala text could be a paragraph which is made up of a number of sentences. In this phase the system splits the text into sentences. It recognizes a sentence whenever a full stop is encountered which signifies the end of the sentence.

Phase (6) – IgalaTokenSaver: This module stores the resulting Igala sentences in a database table called tbl_igala_tokens.

Phase (7) – Merger: This module merges the records in tbl_english_tokens with the records in tbl_igala_tokens based on a common field called id, so that each Igala sentence is aligned with the corresponding English sentence.

A total of 20,000 parallel English – Igala parallel sentences were generated. Table 4 shows the sample.

Table 4: Sample Parallel text obtained from EIPTExtractor

id	English Sentence	Igala Sentence
1	Ayah community in ibaji Local government area has called on Kogi State government to provide them with social amenities	Abo ojanẹ Ayah efu gọmeti ọwanẹ Ibaji rodu bọ gọmeti kuma ki du amẹnwn kekwu nwu akwu ma nwu ma todu ki bu ọbata ma lọ

	to alleviate their suffering	
2	Chairman Ayah Development Association Mr Felix Ejima Ugbena made the call at a thanksgiving reception organized for political appointees appointed by the Kogi State government from the area	Ẹnẹ ki k abo Ojanẹ Ayah nyẹrẹ Onẹnyi Fẹlix Ejima Ugbẹna kọla ekidẹi ẹgba ki che icholo ugwa kpai ẹjẹ kpai emọ dama ki du tẹ nwu abo ki ne uña efu ijabe ki gọmeti chi ma dago kwo Ojane lẹ
3	Ajibola Christopher who was there for NTA News made us to understand that Ayah community in Ibaji local government area witness a large turnout of people at a reception ceremony in honour of their political appointees among who are special adviser to Kogi State Deputy governor Honorable Arome, ibaji local government chairman Thomas Offor and PA to Kogi State deputy governor Newton Ukwu	Agba inabali Ajibọla Kristofa ki gba inabali dufu nwu wa che jẹ nwu wa ma kakini amonẹ wewe che danẹ efu icholo ugwa kpai ẹjẹ kpai emọ dama oji abo kuma chi kuma che ọda koji abo Ayah efu ijabe le dabu Adibe nwu arọne gọbina Kogi, Enojima Arọmẹ, Chamani Ojane Ibaji Tomochi OFor kpai ene kuma lere nwu li gobina Newton Ukwu
4	Chairman, Ayah Development Association Mr Felix Ejima, who expressed gratuitude to God for making thee ceremony a reality, says the appointment of their illustrious sons in government is the first in the history of the land	Chamani Ayah Ọnẹnyi Fẹlix Ejima la che dugwa gw' ỌJỌ todu kuma fu ma chi kuma chọda koji Ojanẹ Ayah todu eyi ch' ọmama echi amonẹ kw' Ojanẹ ma
5	He commended Kogi State Governor Alhaji Yahay Bello and his Deputy chief Simon Achuba for making them proud	I la dugwa kẹrẹbọ gọbina Ojanẹ Kogi Alaji Yahaya Bello todu ukọlọ ki dọmọ ache lẹ kpai arọne nwu Simọn Achuba kuma du ojima le na amonẹ oji

6. CONCLUSIONS

This paper has concentrated on the construction of parallel corpora for English and Igala language to deal with the bottleneck faced in developing Natural Language Processing applications for the language pair. The unavailability of documents which are mutual or comparable translations of each other for the language pair on the web informed the decision to build the corpora from scratch instead of using existing algorithms or their variants for constructing the corpora automatically. Our method though herculean in nature was undertaken due to the necessity of the parallel corpora for the language pair. Our objective of creating the valuable linguistic resources was achieved since we now have a reasonable quantity of sentence aligned parallel text for English and Igala language. This will fast track the development and deployment of NLP systems such as Cross language information Retrieval Systems, Corpus based Machine translation systems etc. for the language pair.

7. REFERENCES

[1] Chen, J., Chau, R., Yeh, C.H. 2004 Discovering parallel text from the World Wide Web**.** In Proceedings of the 2nd Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation, Dunedin, New Zealand, Australian Computer Society. 157-161

[2] Chen Yu, Martin Kay and Andreas Eisele. 2009. Intersecting multilingual data for faster and better statistical translations. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 128-136. Boulder, Colorado.

[3] Cohn Trevor and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,pp. 728-735. Prague, Czech Republic.

[4] Christodouloupoulos C, Steedman M. 2015. A massively parallel corpus: the Bible in 100 languages. Lang Resources & Evaluation (2015) 49:375–395 DOI 10.1007/s10579-014-9287-y.

[5] Egbunu, F. E. 2013. 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] Eisele A. & Yu C. 2010. MultiUN: A Multilingual Corpus from United Nation Documents. Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010), pp. 2868-2872. Valletta, Malta.

[7] Koehn P. 2005. EuroParl: A Parallel Corpus for Statistical Machine Translation. Proceedings of the Machine Translation Summit, pp. 79-86, Phuket, Thailand.

[8] Koehn P., Alexandra B., & Ralf S. 2009. 462 Machine Translation Systems for Europe. In Proceedings of the Twelfth Machine Translation Summit (MT-Summit XII), pages 65-72. Ottawa, Canada, (August 2009).

[9] Ma, X. and M. Liberman. 1999. BITS: A Method for Bilingual Text Search over the Web. In Proceedings of Machine Translation Summit VII.

[10] Muhammed M. S., Mohammed M. K., Ali M. N A. 2015. Automated Construction of Arabic-English Parallel Corpus.

[11] Nie, J. Y., Isabelle, M. S. P., and Durand R. 1999. Cross-language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development.

[12] Omachonu G.S. 2012. Igala Language Studies and Development: Progress, Issues and Challenges, Text of a paper presented at the 12th Igala Education Summit held at Kogi State University, Anyigba- Kogi State, Nigeria. (Dec. 2012).

[13] Pianta E., Bentivogli L. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus Natural Language Engineering 11 (3): 247–261. 2005 Cambridge University Press.

[14] Ralf S. et al. 2014. An overview of the European Union's highly multilingual parallel corpora. EUROPEAN COMMISSION (EC) & EUROPEAN PARLIAMENT (EP) & EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL (ECDC).

[15] Resnik, P., Olsen, M., and Diab, M. 1999. The Bible as a parallel corpus: Annotating the ''Book of 2000 Tongues''. Computers and the Humanities, 33, 129–153.

[16] Resnik, P. and N. A. Smith. 2003. The Web as a Parallel Corpus. Computational Linguistics, 29(3)

[17] Sani F. A. 2016. English to Igala Machine Translation System. PhD Dissertation, Universidad Azteca, Mexico.

[18] Sani Rita I. 2013. Adaptation of the Staff of Office of Attah Igala into Textile Design Forms and Products. Master's Thesis, University of Nigeria, Nnsukka.

[19] Utiyama Masao. 2012. Efficient Technologies for Creating Parallel Corpora. Journal of the National Institute of Information and Communications Technology Vol. 59. Pp 41 – 47.

[20] Ying Zhang, etal. 2006. Automatic acquisition of Chinese–English parallel corpus from the web. ECIR'06 Proceedings of the 28th European conference on Advances in Information Retrieval. London, UK — April 10 - 12, 2006, pp 420-431.

[21] www.wikipedia.org