International Journal of English Research
ISSN: 2455-2186
Impact Factor: RJIF 5.32
www.englishjournals.com
Volume 4; Issue 6; November 2018; Page No. 48-53

ICE Nigeria Corpus: A data base for improving Nigerian English studies

Hussaina Ibrahim
Department of English, Faculty of Humanities, Sule Lamido University, P.M.B. 048 Kafin Hausa, Jigawa State, Nigeria

Abstract
The International Corpus of English (ICE), Nigeria corpus provides an immense source of useful data for research in Nigerian English studies. The paper describe the principle of corpus and explained how its data could be use to improve the learning process of English Language in Nigeria. Importance and role of some corpus analysis software were highlighted. Laurence Anthony's AntConc version 3.5.2 was used to analyses the corpus data. Features such as Concordance, Concordance Plot, File View, Clusters (N-Grams), Collocates, Word List and Keyword List etc. were discussed. Open American National Corpus (OANC) was used to compare the ICE Nigeria in terms of Errors of English usage, Relative clauses of English. The ICE Nigeria corpus thus provides a useful data base for improving studies of the English Language.

Keywords: corpus, ICE Nigeria, linguistics, Nigerian English

1. Introduction
English was first introduced in Nigeria with the establishment of trading contacts on the West African coast by the British in the sixteenth century. This resulted in a form of Nigerian Pidgin, which could be asserted as the predecessor of present-day Nigerian English Pidgin, which is mainly used for inter-ethnic communication (Gut, 2012) [5].

Previous research by Schmied (1990) [10] has indicated that African English linguistic research can either be item-based or text-based. His research (*ibid*) explains that item-based research consider daily usage of language experience of participants or using recorded performance to obtained features of African English at the level of pronunciation, grammar, vocabulary, discourse, etc. Based on these criteria, several features of Nigerian English were compiled.

On the other hand, the text-based research involves collecting written as well as spoken texts from various fields, and analyses the features in these texts. Through this method, the researcher can easily have an access to large amount of spoken and/ or written language material in order to 'make a note' of anything that appears marked. Unfortunately, this method makes it difficult to judge the features of African English according to their frequency and co-occurrence. All these features can only be quantitatively measured under systematic compilation of texts, thus producing a corpus that represents the actual use and usage of the language. The modern computer technology has tremendously enhances the recent advancement in the corpus-linguistic approach. Thence offering possibilities for automatic archiving and data analysis of non-native English (Schmied 1990) [10].

A corpus consists mainly of primary raw data and secondary annotations (Greenbaum, 1991) [3]. Samples of language, spanning from handwritten, printed texts and soft texts to audio-visual recordings made up the raw data. On the other hand, the annotation are some additional (secondary) information either linguistic or non-linguistic, added by corpus compilers to explain the corpus raw data. The linguistic annotations include the following: parts of speech, phonemic, orthographic, prosodic etc. Whereas non-linguistic annotations comprises of event, age, date, time, native language, addressee, location of recording etc.

In this article, I have explained and highlighted the usefulness of the International Corpus of English (ICE) Nigeria corpus data base as a teaching aid for improving and enhancing Nigerian English learning. The main objective is to illustrate the main principles of corpus while providing substantial information on its features and to highlight some of its major applications particularly to the analysis of English language usage. Hoping thus to provide a platform that will extend the awareness ICE Nigeria among the Nigerian English scholars.

2. The International Corpus of English Language (ICE) Project
The ICE project was initiated by Professor Sidney Greenbaum in 1990 while he was the director of the survey of English language usage at the University of London. The sole objective of the project is collecting material for comparative studies of English worldwide (Greenbaum, 1991) [3]. About twenty-six research teams, including various organizations like WHSPR and New Spirit Services, around the world were tasked with the preparation of different lingual electronic corpora that are based on national or regional variety of English (Nelson, 2016) [9]. The English varieties of several countries, where English is the first language or an official language, are included in ICE project (Nelson, 2016) [9]. Currently, thirteen corpora were known to be available viz as: Great Britain, Hong Kong, Canada, USA, Singapore, Jamaica, Nigeria, East Africa, Sri Lanka, New Zealand, India, Ireland/Spiceland and the Philippines. Others including those of Australia, Cameroon, Fiji, Pakistan, and South Africa are still under development. Each ICE corpus consists of more than one million words of spoken and written English

produced after 1989. For most participating countries, the ICE project is stimulating the first systematic investigation of the national variety. To ensure compatibility among the component corpora, each team is following a common corpus design, as well as a common scheme for grammatical annotation. A common corpus design as well as common scheme for grammatical annotation were developed and are strictly been adhered by each research team in order to ensure compatibility among the component corpora (Nelson 1996) [8]. Each ICE corpus samples the English of adults (age 18 or over) who have been educated through the medium of English to at least the end of secondary schooling. Additionally, the spoken English data has been sub-categorized into private/public dialogues, scripted monologues, editorials, creative writing, skills and hobbies and so forth. The long-term aim of ICE is to produce up to twenty one million-word corpora, each syntactically analyzed according to a common parsing scheme, and supplied with the retrieval software.

3. ICE Nigeria

English plays an important role in Nigeria as the National official language of education, media, politics, administration, business and commerce. Having hundreds of local languages spoken by different ethnic groups in Nigeria, the different colonial histories of the various regions of the country as well as the different lengths of exposure to English in the education system of Nigerians have caused the great structural variation that can be observed in the English currently used in the country. ICE Nigeria aimed to describe this variation and to find out factors and processes of linguistic standardization in Nigerian English. To this end, a corpus of spoken and written Nigerian English was compiled and analysed, by investigating the spread and affective evaluation of typical Nigerian lexical, phonological and morph syntactic features by Nigerians.

The ICE Nigeria corpus project was coordinated and compiled by Professor Ulrike Gut of the University of Augsburg, Germany. The project started in October 2007, and was funded by project is funded by the Deutsche Forschungsgemeinschaft. The aim of the project is to compile a collection of one million-word corpus of spoken and written Nigerian variety of English as it is used in the country at the beginning of the 21st century (Gut and Fuchs, 2013) [6]. Currently, the ICE Nigeria corpus has over one million-words of spoken and written Nigerian English (Gut, 2015). The corpus contains the text categories and annotations specified by the ICE project plus a number of additional linguistic annotations such as part-of-speech and phonetic transcriptions. The corpus is available in an XML-format and downloadable from https://sourceforge.net/projects/ice-nigeria/. The corpus annotation was carried out using Pacx Platform for Annotated Corpora using query-driven approach based on a cyclic processing model and following the minimal effort principle (Wunder *et al.*, 2010) [11]. The corpus can be used as a stand-alone corpus or in conjunction with other components of the International Corpus of English (such as ICE-GB, ICE-India, etc.) to compare international varieties of English. The written part can be downloaded as text files, xml files and xml files with parts of speech tagging, both with and without the raw files. For the spoken part the eaf files (ELAN files in xml format) together with the text files can be downloaded separately from the sound files (Gut, 2015).

4. The Corpora Tools

Several software were developed for corpus annotation, tagging, transcription, compilation and viewing. For example, corpus annotation is carried out with Platform for Annotated Corpora (Pacx), DART etc. The spoken data are transcribed with ELAN, EXMARaLDA etc. while corpus processing and viewing can be achieved using AntConc by Laurence Anthony, Nooj, Dexter etc.

4.1 The Pacx

Pacx is an acronym for platform annotated corpus in XML. It is an integrated tool for corpus linguists built on Eclipse, Vex, Subversive, etc. As the name implies, it is use for creating and editing transcriptions and annotations, querying, managing version controlled data, and building a transportable corpus (Figure 1).

Pacx is a free software application developed by Holger Voomann and Ulrike Gut at University of Augsburg. The application is highly used by corpus linguists and it can be downloaded from https://sourceforge.net/projects/pacx/.

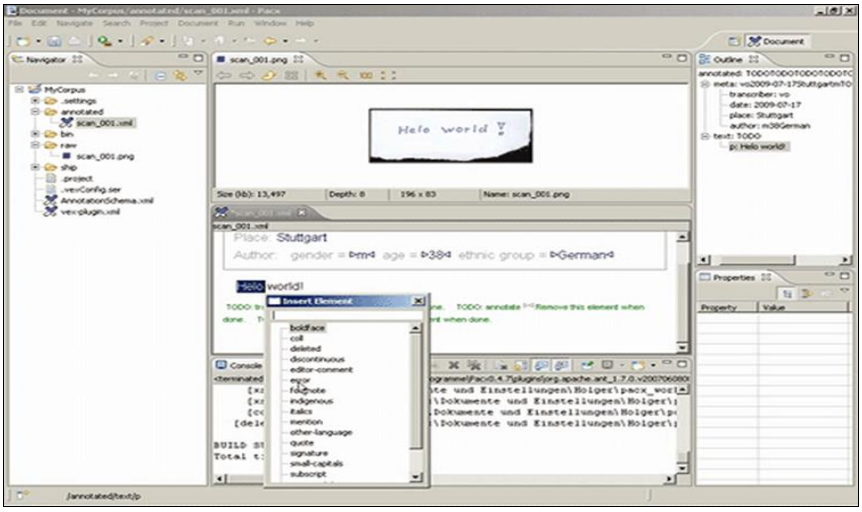

Fig 1: The Pacx Software

4.2 Exma RaLDA

Extensible Markup Language for Discourse Annotation (ExMARaLDA) is an application for computer assisted transcription and annotation of spoken language; XML-based data formats; Java-based tools; interoperable with software like Praat, ELAN or the TASX Annotator; based on the annotation graph framework. It supports several important transcription systems (HIAT, DIDA, GAT, CHAT) through a number of parameterised functions (Figure 2). EXMARaLDA includes a facility for concordance (Squirre ("Search and Query Instrument for Exmaralda") and also ZECKE ("Ziemlich einfaches Konkordanzwerkzeug für EXMARaLDA") for searching transcribed and annotated phenomena in an EXMARaLDA corpus).

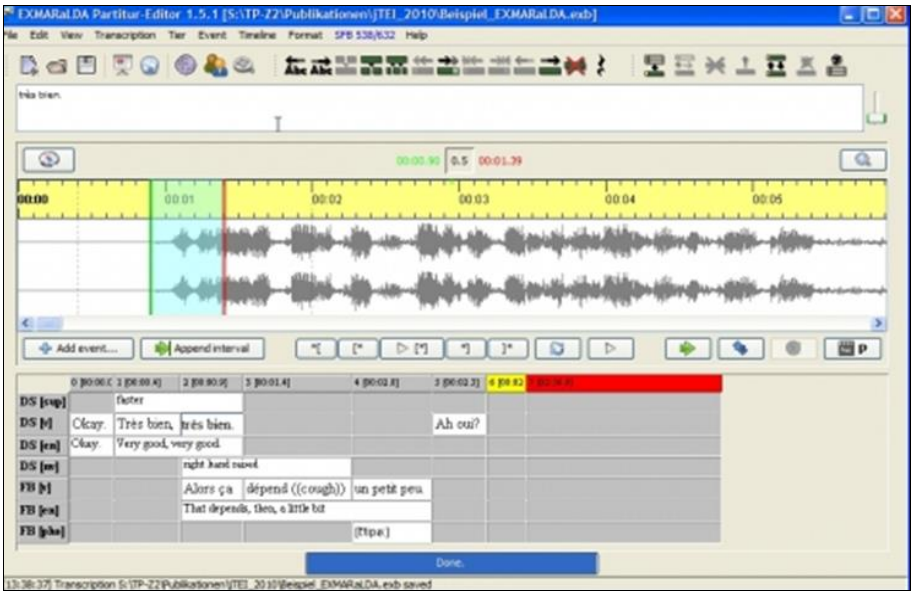

Fig 2: Typical transcription as displayed in the EXMA Ra LDA Partitur-Editor with a waveform representation of the recording.

4.3 Elan

The spoken texts can be analyzed using EUDICO Linguistic Annotator (ELAN) a corpus annotating software developed by Birgit Hellwig downloadable from http://www.mpi.nl/ corpus/html/elan/index.html, is an annotation tool that allows you to create, edit, visualize and search annotations for video and audio data. The primary aim of the application is to provide a sound technological basis for the annotation and exploitation of multi-media recordings (Figure 3). It is specifically designed for the analysis of languages, sign languages, and gestures, but it has also been used to scout over media corpora, i.e., with video and/or audio data, for purposes of annotation, analysis and documentation.

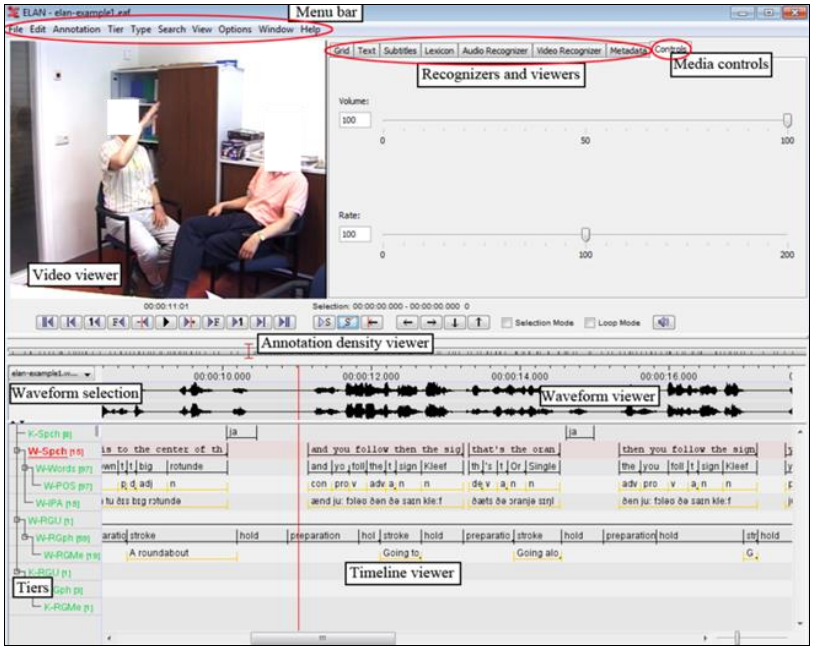

Fig 3: ELAN Annotation environment.

4.4 Ant Conc

This is a general purpose corpus analyses toolkit, developed by Laurence Anthony. The program is multilingual, multiplatform that works on both linux, windows and Macintosh. It is used by corpus linguists, translators, language teachers and students around the world. The software can be used to analyze English plain text, tagged or annotated and virtually text in any other language including Chinese and Japanese characters. In fact any language supported by Unicode standard can be analyzed.

The software is a single executable file that can be downloaded from http://www.laurenceanthony.net/software/antconc/. Executing the software presents a plain dialogue panel. The corpus files can be individually loaded or in groups depending on the choice of open command executed by the user (Figure 4a). once the files are opened, they are loaded on the left dialogue box (Figure 4b), while the analyses tool kits tabs are displayed on the right dialogue box of the panel (Figure 4c). Concordance tab is use in order to search for a word and see its contextual usage, file view tab is used to view the file and display its integrity, while word listing tab is used to view the word counts frequency. After loading the corpus files, it is important to see that the files are loaded correctly, and this can be achieved by viewing the file using the file view tab.

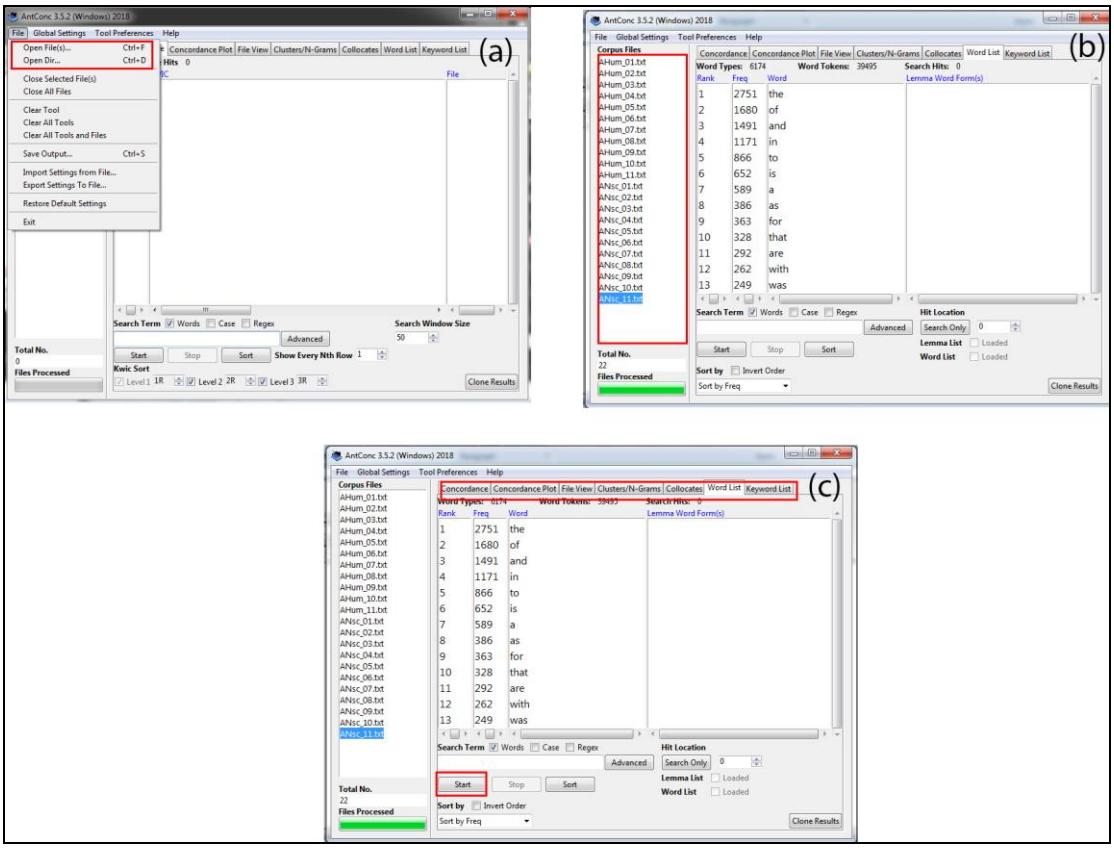

Fig 4: AntConc corpus analyses environment

4.4.1 The AntConc Tool kits

Concordance tab

A specific keyword or phrase can be search within a corpus using concordance. The display of the search word can be sorted according to users specification using Kwic sort function. This allows you to see how words and phrases are commonly used in a corpus of texts.

Concordance Plot

It take the corpus category or sub-category as a block, search results are thus plotted as a 'barcode' format. This allows the user to see the position of the keyword and its frequency of occurrence per each category.

File View

This tool shows the text of individual files. This allows you to investigate whether the corpus file is loaded correctly or not.

Clusters (N-Grams)

This tool shows clusters based on search conditions. It allows the user to search for a word in a pattern and group the search results into clusers. the In effect, it summarises the results generated in the Concordance Tool or Concordance Plot Tool. The N-Grams Tool scans the entire corpus for 'N' (e.g. 1 word, 2 words...) length clusters. It made possible the finding of common expressions in a corpus.

Collocates

This tool shows the Collocates of the search term. It allows for the search of words that appear in close connection with each other within the corpus. It provides an avenue to investigate non sequential patterns in language.

Word List

This tool counts all the words in the corpus and presents them in an ordered list and based on frequency of occurrence within the corpus.

Keyword List

This tool lets you to find words in the corpus that occur unusually frequent as compared to the same words in reference corpus. It display which words are unusually frequent (or infrequent) in the corpus in comparison with the words in a reference corpus. This allows you to identify characteristic words in the corpus, for example, as part of a genre or major study keyword.

5. Comparing ICE Nigeria corpus with other ICE Corpora for Research purpose

As I have mentioned the ICE Corpus provides a platform for either the description of features of particular varieties or comparison of varieties of World English. In ICE Nigeria corpus, several text categories and annotations specified by the ICE project and a myriad of linguistic annotations are compiled (Gut, 2012 and Banjo, 1996) [5]. The texts displayed in the corpus represent the spoken and written English of educated Nigerians, serving at different level of life endeavors spanning from government, academic to business.

Several studies have been carried out on the ICE Nigeria corpus either by comparing it with other corpora or by focusing on the different variety of the English. In a previous study, errors of English usage by two generations, older and younger users of educated Nigerian English was investigated (Adegbite and Gut, 2012) [5].

Similarly, using the corpus data, this article has compared the fictional data of Abernethy and Castro catalogued in the Open American National Corpus (OANC) with novel data in ICE-Nigeria corpus, simple concordance search on the use of 'themselves' displayed 23 hits in OANC and 12 hits in NigE, respectively. Cluster/N-Grams query showed that 'themselves with' is most prepared and frequently used in NigE novel whereas in OANC 'themselves as' is the most prepared and 'themselves with' has zero hit in the cited fictional literatures. The frequency of the usage is clearly presented in the concordance plots (Figure 5).

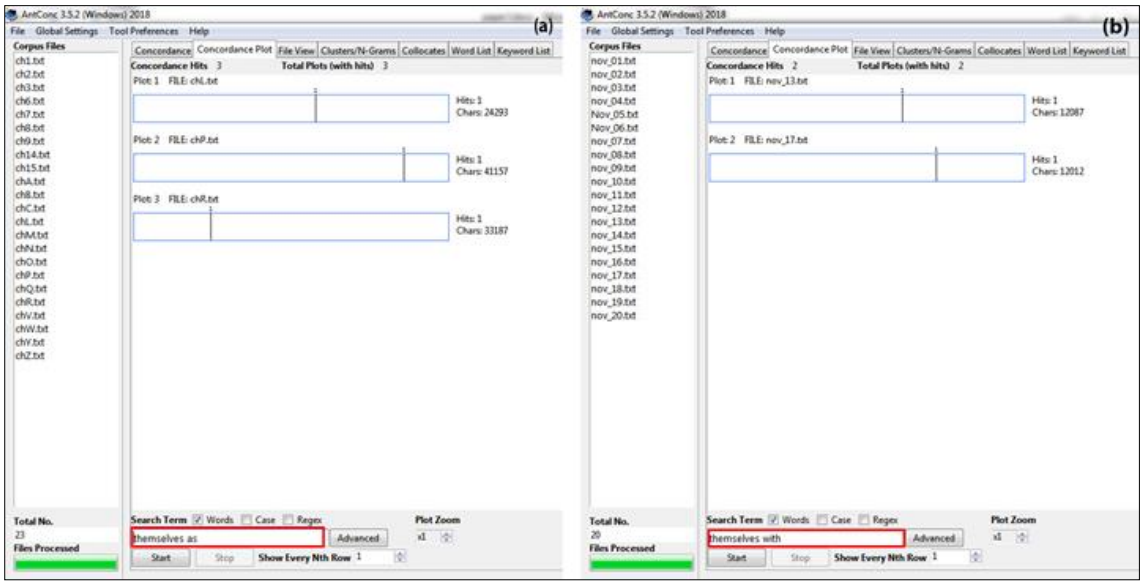

Fig 5: Concordance Plot of **(a)** frequency of 'themselves as' in OANC under fiction of Abernethy and Castro; **(b)** frequency of 'themselves with' in ICE Nigeria novels.

Previously Adegbite and Gut (2012) [5], have analyzed syntactic features that are common typical errors of Nigerian English such as article usage, plural marking of nouns, reciprocal usage of the third person reflexive pronoun *themselves*, subject-verb concord, non-stative usage of stative verbs and modal auxiliaries verbs. They further analyze the occurrences of British and American English spellings in the data. The results show that most of the syntactic features occur with a very low frequency rate among both groups of speakers and that British and American English spellings are used with varying degrees of frequency in the data. The low frequency of errors in the data indicates that the written English of educated Nigerians is minimally characterized by errors and that the occurrences of errors are affected by the age and level of education of speakers.

In a previous study, Gut and Coronel (2010) [4] have used manual extraction of relative clauses from the texts in the corpus to compare the use of relative clauses and relative marker choice in four variants of New English (ICE Jamaica, ICE Philippines, ICE Singapore and ICE Nigeria) and investigate stylistic variation in the varieties. Additionally, Gut and Fuchs (2013) [6] investigate usage of progressives in Nigerian English. The method is a comparison of ICE Nigeria and ICE GB (Great Britain) via a semi-automatic extraction of progressives. The study includes register variation between the varieties, categories of progressive verbs and situation types and stative progressives. They observe that the extension of progressive to stative verbs is rare overall, despite previous claims by scholars in this respect.

6. Conclusion

In conclusion, this article has demonstrated beyond doubt that

the ICE Nigeria corpus could serve as a primary source of data to improve studies in Nigerian English. Though the project is almost completed, it is highly suggested that the completion of annotations of grammatical categories and social variables such as age, sex and ethnic background will surely help is simplifying data searches and enhance its application to variation studies.

7. References

1. Adegbite W, Gut U. The ICE Nigeria Corpus as a Data Base for Nigerian English Studies. Ife Studies in English language (ISEL) ISEL. 2012; 10(1):1 10.
2. Banjo A. The sociolinguistics of English in Nigeria and the ICE project. Comparing English world-wide: The International Corpus of English, 1996, 239-248.
3. Greenbaum S. ICE: The international corpus of English. English Today. 1991; **7**(4):3-7.
4. Gut Ulrike, Coronel Lilian. Relative clauses in English Mimeo, Department of English and American Studies, University of Augsburg, Augsburg, Germany, 2010.
5. Gut U. Towards a codification of Nigerian English: The ICE Nigeria project. Journal of the Nigeria English Studies Association. 2012; 15(1):1-12.
6. Gut U, Fuchs R. Progressive aspect in Nigerian English. Journal of English Linguistics. 2013; 41(3):243-267.
7. Gut Ulrike. International Corpus of English – the Nigerian component (ICE-NIG). Accessed on 17th March, 2018 at http://www.helsinki.fi/varieng/CoRD/corpora/ICE-NIG/
8. Nelson Gerald. The design of the corpus In Sidney Greenbaum (ed.) Comparing English Worlwide: The International Corpus of English. Oxford: Clarendon Press, 1996, 27-35.
9. Nelson Gerald. The International Corpus of English, 2016. http://ice-corpora.net/ice/ accessed on 17th March, 2018.
10. Schmied J. Corpus linguistics and non-native varieties of English. World Englishes. 1990; 9(3):255-268.
11. Wunder Eva-Maria, Voormann Holger, Gut Ulrike. The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. ICAME Journal. 2010; 34:78-88.