The ICE Nigeria Corpus as a Data Base for Nigerian English Studies

Wale Adegbite

Department of English, Obafemi Awolowo University, Ile-Ife, Nigeria

Abstract

The author of this study describes the International Corpus of English (ICE), Nigeria corpus as a useful source of data for research in Nigerian English studies. Explaining the corpus as a part of the world ICE project, he presents the principles and describes the functions of some of the tools and their application to analyses of data from both Nigerian English and World Englishes. The AntConc software tools presented include the Concordance, Concordance Plot, File View, Clusters (N-Grams), Collocates, Word List and Keyword List; and some menu options such as File, Global settings and Tool Preferences. Preliminary studies on the ICE Nigeria include "Errors of English usage by two generations, older and younger users of educated Nigerian English", "Relative clauses in four varieties of English (ICE Jamaica, ICE Philippines, ICE Singapore and ICE Nigeria)" and "Usage of progressives in ICE Nigeria and ICE GB". The ICE Nigeria corpus thus provides a useful data base for studies on English usage by educated Nigerian speakers.

Key words: Nigerian English, corpus linguistics, ICE Nigeria

1. Introduction

Describing research approaches to a description of English in Nigeria, Schmied (1995:338) asserts that in the broadest sense, linguistic research on the formal side of English in Africa can be either item-based or text-based. According to him (ibid.), item-based research records features of African English at the level of pronunciation, grammar, vocabulary, discourse, etc., from the daily language experience of participants or from recorded performance. A number of feature lists of Nigerian English were compiled on the basis of this methodology. In contrast, text-based research collects written and/ or spoken texts from various fields, domains or situations and analyses features in these texts. The advantage of the former procedure is that it allows the researcher to go through a large amount of spoken and/ or written language material and to 'make a note' of anything that appears marked. But the method makes it difficult, however, to judge the African features compiled according to their frequency and co-occurrence. The latter procedure not only provides the raw data but also some guidelines, as far as the frequency and combinations of features are concerned. All these quantitative measures are valid, however, only when the compilation of texts is done on a systematic basis, so that the resulting corpus is roughly representative of actual language use and usage.

The corpus-linguistic approach is a further development of the traditional text-based approach. With the help of modern computer technology, it offers additional possibilities for automatic data analysis of non-native English (Schmied 1990). Many scholars sometimes use the term 'corpus' to refer to a body of texts. But, among linguists, it is better defined as a body of written or spoken material upon which a linguistic analysis is based. Thus, it cannot be seen as just a collection of texts, but is a collection assumed to be representative of a given

language, dialect, or other subset of a language, that is to be used for linguistic analysis (Francis 1982, Esimaje 2011). McEnery, Xiao and Tono (2006) define a corpus as a finite-size body of machine readable authentic texts that is sampled to be maximally representative of the language variety under consideration. According to Voorman and Gut (2008), a corpus consists of raw (primary) data and annotations. The raw data consists of samples of language, ranging from handwritten or printed texts and electronic texts to audio and video recordings, and the annotation refers to additional (secondary) information about the corpus raw data that is added by the corpus compilers. It can be divided into linguistic (orthographic, phonemic, prosodic, parts of speech, etc.) and non-linguistic (age, native language, addressee, event, time, date and location of recording, etc.) information. A corpus is not an end in itself but, rather, a source of improving descriptions of the structure and use of language (Chafe 1991, Kennedy 1998, Svartvik 2007, Esimaje 2011). A corpus linguist, thus, is one who tries to understand language and, beyond language, the mind by carefully observing extensive natural samples with insight and imagination in other to construct plausible understandings that encompass and explain those observations.

The aim of this study is to describe the International Corpus of English (ICE), Nigeria corpus as a viable data base for Nigerian English (NigE) studies. The objectives are to give background information about the corpus, to identify and illustrate some of its principles and features and to highlight some instances of its application to the analysis of English language usage. The goal is to create awareness about the corpus to Nigerian English scholars, many of whom are unaware of its existence, talk less of utilising it to diversify, illuminate and boost research on NigE usage.

2. The ICE Project and ICE Nigeria

The ICE project is a set of corpora representing varieties of English from around the world. Over twenty countries or groups of countries, where English is the first language or an official language, are included (Fuchs 2010, *Wikipedia*). The project, initiated by Sidney Greenbaum, began in 1990 with the primary aim of collecting material for comparative studies of English worldwide. Twenty-three research teams around the world are preparing electronic corpora of their own national or regional variety of English. Each ICE team is compiling or has already compiled a one million-word corpus of spoken and written English (600,000 and 400,000 words respectively). 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 (Nelson 1996). 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. Further sub-categorisation of spoken English data includes private/public dialogues, (un-) 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 analysed according to a common parsing scheme, and supplied with the retrieval software. Ten corpora are currently available: UK, Canada, East Africa, Hong Kong, Singapore, India, Ireland, Jamaica, New Zealand and Philippines. Among those in preparation are Australia, Cameroon, Fiji, US, Pakistan, South Africa and Nigeria.

The compilation of ICE Nigeria corpus, which is coordinated by Ulrike Gut of the University of Augsburg, Germany, started in October 2007. It is the aim of the project to collect a one million-word corpus of spoken and written Nigerian English as it is used in Nigeria at the beginning of the 21st century. The written data of over 400,000 words is already completed and work is in progress on the transcription of the spoken data of 600,000 words. The corpus

is available in an XML-format. Corpus annotation is carried out with Platform for Annotated Corpora (Pacx) and the spoken data transcribed with ELAN (cf. Wunder 2010). The corpus creation process is agile, which means query-driven, based on a cyclic processing model and following the minimal effort principle (see Voormann & Gut 2008).

Fig. 1: The Pacx Software (www.pacx.sf.net)

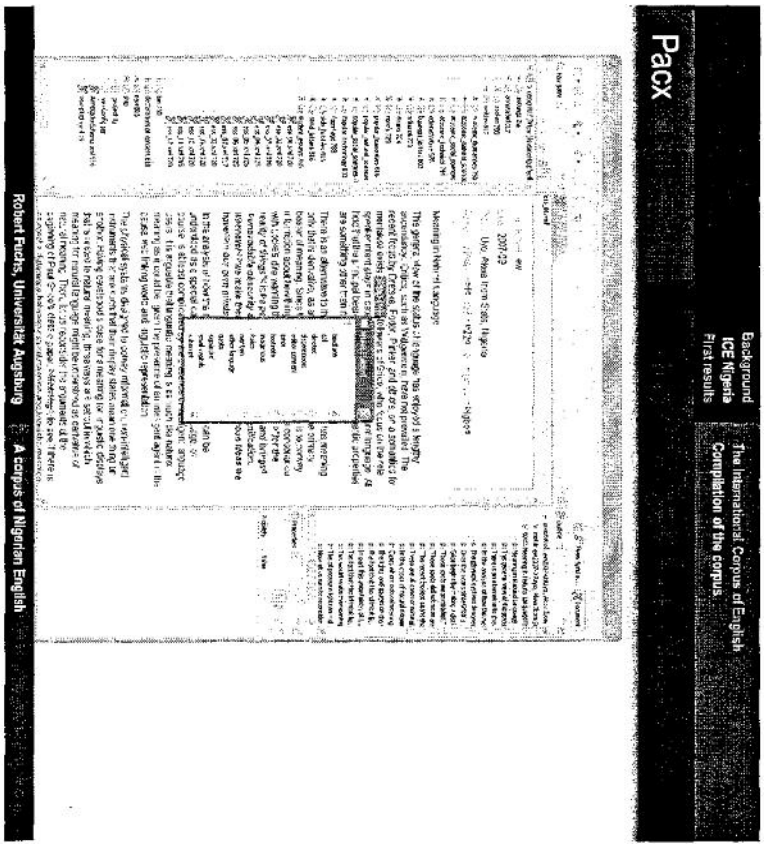
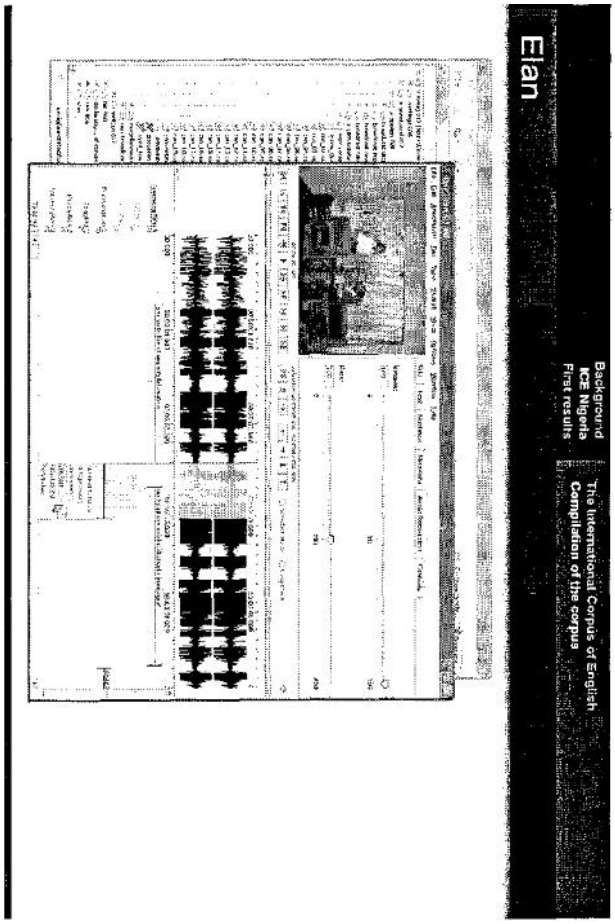


Fig. 2: The ELAN Software (www.lat-mpi.eu/tools/elan)



The corpus contains the text categories and annotations specified by the ICE project plus a high number of additional linguistic annotations such as part-of-speech and phonetic transcriptions. All texts in the corpus represent the spoken and written English of educated Nigerians, undergraduates and graduates and the whole texts or selections of them can be utilised for studies based on this variety. The written texts derive from different genres and sub-genres (see Table 1) below. Manual searches of items from the texts and their frequencies can be done with the aid of some available software and tools such as 'the AntConc' for written texts and 'ELAN' for spoken texts. Furthermore, the annotations of information about ethnic group, age and sex of speakers/writers may guide the selection of texts according to different variables.

Table 1: Categories of ICE Written texts

Categories	**Collected**	**Required**
Academic humanities	20014	20,000
Academic natural sciences	20,025	20,000
Academic social sciences	19,998	20,000
Academic technical	20,006	20,000
Administrative/instructive	20,001	20,000
Business letters	30,004	30,000
Editorials	20,014	20,000
Exams	20,009	20,000
Novels	40,003	40,000
Popular humanities	20,015	20,000
Popular natural sciences	20,035	20,000
Popular social sciences	20,023	20,000
Popular technology	20,074	20,000
Press reports	40,085	40,000
Skills, hobbies/instructive	20,008	20,000
Social letters	30,062	30,000
Essays	20,005	20,000
Total	**400,381**	**400,000**

3. The AntConc Software and Tools

The AntConc software (see http://www.antlab.sci.waseda.ac.jp/software/README-antconc3.2.4.w) has some tools which may be utilised for data processing and analyses. The tools are presented briefly below. For details on instructions, see http://www.antlab.sci.waseda.ac.jp/software/README-antconc3.2.4.pdf.

Concordance

The Concordance tool shows search results in a KeyWord in Context (KWIC) format. This allows you to see how words and phrases are commonly used in a corpus of texts.

Concordance Plot
This tool shows search results plotted as a 'barcode' format. This allows you to see the position where search results appear in target texts.

File View
This tool shows the text of individual files. This allows you to investigate in more detail the results generated in other tools of AntConc.

Clusters (N-Grams)
This tool shows clusters based on search conditions. 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. This allows you to find common expressions in a corpus.

Collocates
This tool shows the collocates of the search term. This allows you 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. This allows you to quickly find which words are the most frequent in a corpus.

Keyword List
This tool shows 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 ESP study.

MENU OPTIONS
Three main categories are presented briefly here: File, Global settings and Tool Preferences.

FILE
Options here relate to reading files into AntConc and writing files to the hard disk containing data of various types. There are also options to export all current settings to a file, and import user settings from a file. If a user settings file becomes corrupted for any reason, simply restart the program or use the "Restore Default Settings" option to return the program to its original state.

GLOBAL SETTINGS
A number of categories under this setting will have an effect on multiple tools in AntConc.

File Settings: the user can choose to display the full path of a file or just the name. The user can also choose to show or hide any tags in the file. The tag boundaries can be specified.

Tag Settings: the user can choose to display or hide any tags that are contained in the corpus files. If tags are to be hidden, the opening and closing tag markers must be specified. The default is <>.

Wildcard Settings: users can edit the default wildcard characters so that they do not clash with a search entry. For example, the "or" wildcard default character (a 'pipe' character |) can be changed to a backslash / here.

Token (Word) Definition: the user can choose which characters, numbers and so on will define a "word".

TOOL PREFERENCES

Each tool (with the exception of Concordance Plot and File View) has a preferences category, where settings can be fine tuned. All tool preference categories 'allow the user to show or hide the different frames in which the results are displayed. For example, the use can choose to hide the frame showing file names in the Concordance tool display window. Also, all tools have the option to treat all data as lowercase and use case when sorting. If results are displayed case sensitively, words including capital letters will appear higher up in the lists.

Concordance Preferences: in addition to the above the following settings can be made: Instead of arranging results by words to the left or right of the search term, it is possible to arrange the results by LETTERS to the left or right of the first letter of the search term. This makes it possible to search for spelling differences. The search term can also be chosen to be hidden in the KWIC lines, allowing instructors to quiz students on possible words to fit the gap. Note: at any time, if the <x> key is pressed while the results window has the focus, the search term can be hidden or shown.

Clusters Preferences: for this tool there are no additional settings that can be set other than those described above.

Collocates Preferences: in addition to the above the settings, the choice of statistical measure can be chosen here. Currently, two statistical measure can be used: Mutual Information (MI) and T-Score. See above.

Word List Preferences: in addition to the above the following settings can be made:

A 'lemma list' can be loaded from a file, which can then be used to generate a lemma list instead of a word list. When the lemma list function is used, the 'lemma word form(s)' column will show the words in the corpus associated with each lemma.

A lemma list can be created by specifying the 'lemma entry' follow by '->' followed by one or more 'words' that should be assigned to the lemma separated by one or more non-tokens. Examples are: be->is, are / play->play, plays, playing, played

Note that in the example above, commas and spaces are assumed to be NOT defined as tokens. For this reason, if the lemma list available on the AntConc webpage is used, a 'dash' needs to be added to the token (word) definition for the lemma list to be processed correctly as the hyphenated words are used to the right of the lemma definition.

The word list can be generated using all words, or a specific set of words, or ignoring a certain set of words (a stop list). This is termed the "Word list Range". The range of words to be used (or ignored) can be entered directly by the user, or can be stored in files which are then read by AntConc by pressing the 'Open' button. A combination of words in a file and words directly entered by the user can also be used.

Keyword List Preferences: in addition to the above the following settings can be made: as described in the section on the Keyword List tool, to generate a keyword list, the user needs to specify a reference corpus, and a statistical measure of 'keyness'. Although, the default options for the 'keyness' measure and threshold values are recommended, changes can be made in this menu. By choosing the "Show Negative Keywords" option, words that are unusually INFREQUENT in the target corpus compared with the reference corpus will be displayed. Also, here you can swap the main and reference corpora.

SAVING RESULTS
Results can be either saved to the clipboard, saved to a text file (.. txt) or saved to a new window using keyboard commands, the appropriate option in the 'File Menu', or by clicking on the "Save Window" button in each tool, respectively. Also, it is possible to launch multiple clones of AntConc by double clicking on the .exe file.

4. Application of ICE Nigeria and other ICE Corpuses to Research

As has been mentioned earlier, the ICE Corpus provides a platform for either the description of features of particular varieties or comparison of varieties of World Englishes. The features can be described in terms of occurrences versus non-occurrences, relative frequencies or contexts of occurrences. Preliminary studies have been carried out on the ICE Nigeria corpus either by focusing on data from the NigE variety or comparing the data with corpuses from other varieties. Some of these are highlighted here.

Adegbite and Gut (2010) investigate errors of English usage by two generations, older and younger users of educated Nigerian English. They utilise data from the ICE Nigeria corpus: a written language corpus of 64,129 words, written by 60 Nigerians and containing four different text types (academic writing, informal letters, formal letters and novels). They analyse the following syntactic features that have been described as typical errors of NigE in previous studies: 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 analyse 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 characterised by errors and that the occurrences of errors are affected by the age and level of education of speakers.

Gut and Coronel (2010) describe relative clauses in New Englishes. They aim to compare relative clauses and relative marker choice in four varieties of English (ICE Jamaica, ICE Philippines, ICE Singapore and ICE Nigeria) and investigate stylistic variation in the varieties. The procedure is via manual extraction of relative clauses from the texts in the corpus.

Furthermore, Fuchs and Gut (2011) 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.

In an earlier report of the Cameroon English Corpus, Tiomanjou (1995) mentions some areas in which data from the Cameroon Corpus of English may be used to further confirm or negate previous findings. He (ibid. pp. 356-365) mentions some topics for future research as follows:

Lexis: Lexical innovation processes, collocations and meanings, modal verbs and meanings

Syntax: prepositions and prepositional phrases, tenses, shift from post modification to premodification of noun head, transitivity versus intransitivity of some verbs

Discourse: intensifiers (amplifiers and downtoners), markers for letter-writing (e.g. vocative beginning letters) and 'If clauses'.

The suggestion above may serve as a guide for future studies on NigE.

5. Conclusion

In conclusion, the relevance of the ICE Nigeria corpus as a veritable source of data for studies in Nigerian English has been demonstrated. Nonetheless, it is expected that the completion of annotations of grammatical categories and social variables such as age, sex and ethnic background will further ease data searches and enhance its application to variation studies. The implication of corpus analysis is that the traditional description of syntactic features of NigE needs to be refined beyond mere attestations and listings of features. Whether a structure occurs in 1% or in 50% of all cases seems to constitute a major difference and should, thus, be marked clearly in the description of a variety of English. It is equally important to begin a discussion about a threshold for rate of occurrence below which a structure should or should not be described as a systematic error of a grammar.

References

Chafe, Wallace (1991) "The importance of corpus linguistics to understanding the nature of language" In Jan Svartvick (ed.) *Directions in corpus linguistics. Proceedings of Nobel Symposium* 82, Stockholm, August 4-8. Berlin: Mouton, pp. 79-97.

Esimaje, Alexandra (2011) "Codification of Nigerian English: Issues in Nigerian English database and empirical linguistics" Paper presented at the 27th Conference of the Nigeria English Studies Association (NESA) held at the Covenant University, Ota, Ogun State, November 2-5, 2010.

Fuchs, Robert (2010) "A Corpus of Nigerian English: The new ICENigeria" Mimeo, Faculty of Philosophy and History, University of Augsburg, Augsburg, Germany. http://www.philhist.uni-augsburg.de/de/lehrstuehle/anglistik/applied

_____ . and Gut, Ulrike (2011) "Progressive Aspect in Nigerian English" Mimeo, Department of English and American Studies University of Augsburg, Augsburg, Germany.

Francis, Nelson W. (1982) "Problems of assembling and computerizing large corpora. Stig Johansson (ed.) *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities, pp.7-24.

Gut, Ulrike and Coronel, Lilian (2010) "Relative clauses in English" Mimeo, Department of English and American Studies, University of Augsburg, Augsburg, Germany.

Kennedy, Graeme (1998) *An Introduction to Corpus Linguistics*. Harlow:Addison Wesley.

McEnery, Tony, Richard Xiao and Yukio, Tono (2006) *Corpus-based Language Studies: An Advanced Resource Book*. New York: Routledge.

Nelson, Gerald (1996) "The design of the corpus" In Sidney Greenbaum (ed.) *Comparing English Worlwide: The International Corpus of English.* Oxford: Clarendon Press, pp. 27-35.

Schmied, Josef (1990) "Corpus linguistics and non-native varieties of English" In J. Schmied (ed.) *Linguistics in the Service of Africa*. Bayreuth: Bayreuth African Studies Series 18, pp. 67-88.

_____ . (1995) "National Standards and the International Corpus of English" In Ayo Bamgbose, Ayo Banjo and Andrew Thomas (eds) *New Englishes: A West African Perspective.* Ibadan: Mosuro, pp. 337-348.

Svartvik, Jan (2007) "Corpus linguistics 25+ years on" In Roberta Facchinetti (ed.) *Language and Computers: Studies in Practical Linguistics.* Amsterdam, New York: Rodopi, pp. 11-25.

Tiomajou, David (1995) "The Cameroon Corpus Project" In Ayo Bamgbose, Ayo Banjo and Andrew Thomas (eds) *New Englishes: A West African Perspective.* Ibadan: Mosuro, pp.349-366.

Voorman, Holger and Gut, Ulrike (2008) "Agile corpus creation" *Corpus Linguistics and Linguistic Theory* 4 (2): 235-251.

Wikipedia, the free encyclopedia (en.Wikipedia.org/wiki/International_Copus_of_English)

Wunder, Eva-Maria, Voormann, Holger and Gut, Ulrike. 2010. "The ICE Nigeria corpus project: Creating an open, rich and accurate corpus". *ICAME Journal* 34, 78-88.