LANGUAGE IN THE 21ST CENTURY: RISING ISSUES IN THE DEVELOPMENT OF NIGERIAN ACADEMIC ENGLISH CORPUS (NAEC)

Yewande Mulikat Olabinjo

Abstract

Over the years, linguists like Fries (1952) and Quirk et al (1985) have employed the use of language corpus in the study of natural language use. Language corpus like The British Academic Written English (BAWE) corpus, Lancaster Corpus of Mandarin Chinese (LCMC), American English corpus, have set the pace for language study. They have contributed immensely to the development of sustainable language development. In order for languages spoken in Africa to benefit from this global trend, it needs to explore the technological advancement being applied to language study in other languages. In this work, we created a Nigerian Academic English Corpus (NAEC). It is a collection of texts published by academics in Nigeria. Since works with large body of texts are written by scholars in the humanities, the corpus contains more texts from scholars in this field. Acceptable works selected are in English language, but may contain language use examples from local languages. This work charts the experience in data collection, highlights problems encountered during data collection and the approach towards finding a solution. A preliminary research was done using our data to show the endless possibilities a quantitative analysis of data has. We chose two words from the NAEC and calculated the MI score and the Log-Dice of their collocates. It also reveals the temporary academic finding so far with view of facilitating and encouraging future development of other corpus text in African languages.

Keywords: Corpus Linguistics, Nigerian English, Collocation, MI-score, Log Dice;

Introduction

English originated in England and it is the dominant language of the United States, the United Kingdom, Canada, Australia, Ireland, New Zealand, and various island nations in the Caribbean Sea and the Pacific Ocean. It is also the official language of a great number of countries, including India, Ghana, Singapore, and Nigeria.

Kachru, Braj Behari (1962) studied the 'Indianness' in Indian English and categorized the speakers of the English language into three concentric circles: the inner circle, the outer circle, and the expanding circle. According to him, the inner circle represents the traditional base of English. This circle consists of traditional English-speaking nations like the United Kingdom, the United States, Australia, New Zealand, Ireland, etc.

The outer circle consists of places where English is not the native tongue but is important for historical reasons and plays a part in the nation's institutions, either as an official language or otherwise. The *Outer Circle* is said to have produced the second diaspora of English, which spread the language through imperial expansion by Great Britain in Asia and Africa. In these regions, English is not the native tongue but serves as a useful lingua franca between ethnic and language groups. The legislature, judiciary, higher education, and national commerce are all carried out predominantly in English. The countries in this circle include India, Nigeria, the Philippines, Bangladesh, Pakistan, Malaysia, Tanzania, Kenya, etc.

The expanding circle includes those countries where English plays no historical or governmental role, but English is nevertheless widely used as a foreign language or lingua franca. This includes much of the rest of the world's population, including China, Russia, Japan, most of Europe, etc.

Nigerian English, also known as Nigerian Standard English, is a dialect of English spoken in Nigeria. Nigerian Standard English is used in politics, formal education, the media, and other official purposes. Nigerian English has become a nativized language that functions uniquely within its own cultural context. Guthrie (1964) is of the opinion that English languages spoken overseas have a feature that operates in a cultural void, resulting in local varieties arising with their own canon of correctness. In Nigeria, however, English cannot be considered to be anyone's mother tongue. This implies that every time a Nigerian speaks English, there is a contact between English and the mother tongue, thereby influencing the languages involved. Weinreich (1953) stated that it is the language user who provides the point of contact between two languages, which in turn gives room for language interference. This interference he described as those instances of deviation from norms of either language that occur in the speech of bilinguals arising from their familiarity with more than one language, i.e., as a result of language contact.

The interference could vary from pronunciation to phonology, lexical, or even cultural interference. This form of English language, with inferences from Nigerian culture and mother tongue, resulted in an English variation called 'Nigerian English'. A handful of scholars have done research on this variation of English being spoken in Nigeria. One of such scholars is Bambose (1996)[1], in his work, he identified the use of words in Nigerian English. Farooq A. Kperog (2012)[2], wrote on prepositional and collocational abuse in Nigerian English. Danica

[1] Ayo Bamgbose, "Identifying Nigerian Uses in Nigerian English." *English: History, Diversity, and Change*, ed. by David Graddol, Dick Leith, and Joan Swann. Routledge, 1996).

[2] Farooq A. Kperog, "Nigeria: Prepositional and Collocational Abuse in Nigerian English." *Sunday Trust*[Nigeria], July 15, 2012).

Salazar (2020)[3] highlighted the common clipping and word formation in Nigerian English. What all of these articles lack is a quantitative analysis of the Nigerian English being analysed. The form of English regarded in this work goes beyond the regular colloquial English; it is the finest form of English being spoken and written in the Nigerian academic environment.

Theoretical frame work

The theory used in the preliminary analysis of the data in NAEC is the information theory. Information theory is an applied mathematics theory that deals with the quantification, storage and communication of information. The theory was purported by Claude Shannon (1948) in his seminar work, "A mathematical theory of communication". He established the fundamental concepts of information theory, including entropy, mutual information, and the source-channel coding theorem.

Information theory plays a crucial role in corpus linguistics, especially in the analysis of large collections of text (corpora). It provides a mathematical and statistical framework for analysing linguistic data in a quantitative manner, making it an invaluable tool for researchers in corpus linguistics. It helps in the analysis of the distribution of word frequencies in a corpus. Information theory enables metrics like entropy and measures of diversity that help characterize the richness and variability of a vocabulary, while identifying the collocations, which are frequent word combinations that occur more often than would be expected by chance.

Using Mutual information, a concept from information theory, the measurement of the strength of association i.e. MI-score and

[3] Danica Salazar, an OED world editor, wrote in an article titled "Nigerian English in the OED January 2020 update"

Log-Dice, between randomly selected words in the NAEC were analysed.

Use of Corpus

Corpora like the PERC corpus of professional English which consists of academic journals, textbooks, webpages, and textbooks; the TOEFL 2000 spoken and written academic language corpus (Biber et al., 2002); International Corpus of Learner English (see Granger *et al.*, 2002) and the Louvain Corpus of Native English Essays (see Granger and Tyson, 1996); all have enormous roles in language research. Corpora like these are capable of developing academic literacy skills and have been used to develop literary works in English. They enable researchers to analyse and describe the linguistic properties of the language documented.

According to Nesi, H., et al. (2002), corpus informed the design of virtually all recent major English dictionaries and grammar reference books while also being helpful in the description and accuracy of linguistic texts.

About the Nigerian Academic English Corpus (NAEC)

The Nigerian Academic English corpus is populated with formal academic writing by Nigerian academic environment. They are papers that have already been published in reputable journals within and outside Nigeria. Selected works are works readily available online on Google Scholar. NAEC contains the works of scholars in 10 universities across Nigeria; they are mainly works of scholars in the humanities, being that scholars in this field presumably write in formal English spoken and written in academia. The universities selected for the purpose of this research include: Ambrose Alli University, Ekpoma, Edo State, Nigeria; McPherson University, Seriki-Sotayo, Abeokuta, Ogun State, Nigeria; University of Calabar, Calabar, Cross River State, Nigeria; University of Abuja, Abuja, Nigeria; Federal University of

Technology, Akure, Nigeria; Federal University, Oye-Ekiti, Nigeria; Afe Babalola University, Ado-Ekiti (ABUAD), Nigeria; Tai Solarin University of Education, Ijagun, Ijebu-Ode, Ogun State, Nigeria; University of Ibadan, Ibadan, Oyo State, Nigeria; University of Lagos, Lagos State.

NAEC contains over 70 published works by scholars from the listed universities. It consists of over 23000 lemmas and 400,000 tokens. As earlier mentioned, the works collated in the corpus have been assessed by named journals and subsequently published. This implies that the corpus consists mainly of expertly written, peer-reviewed, and edited texts. They are works readily available in the public domain and accessible on the World Wide Web. By creating a corpus with varying academic writing, word frequency and their co-occurrence in the language can be studied, amongst other things. NEAC is a pilot program capable of illustrating the infinite possibilities of a larger corpus.

The purpose of the corpus project is to collect as many samples of prolific academic writing as possible over a period of six months, from January to July. This corpus is currently not available on the World Wide Web but can be available on request by researchers. The data used to populate the corpus are mainly from academics in the humanities and social sciences, before the decision was made. On these demography of academics, data samples from academics in other fields of study like the sciences and environmental science were also taken. After careful consideration of the data, it was realized these that groups of academics have lots of calculations and scientific equations in their works rather than the use of pure Nigerian English. Hence, the works of scholars in the humanities and social sciences are statistically valid for our corpus.

Issues

At the initial stage, the intention was to build a corpus search engine for the purpose of this research. With adequate funding, the corpus would be able to collaborate with the computer science department to build a corpus search engine capable of analysing the English language and other local languages like Yoruba and Igbo. But, due to limited funds and a tight time schedule, the out-sourcing of an already developed corpus search engine was the best option. The Lancaster University corpus toolbox (Lancbox), was the selected corpus engine of choice.

LancsBox[4] is a new-generation software package for the analysis of language data and corpora. It was developed by scholars at Lancaster University. It enables users to work with their own data to create a corpus; its function includes annotation of data into parts of speech, enables comparative analysis. It is also downloadable as a compact pack, which suits our purpose adequately. After the resolution of the issue regarding a search engine, the next hurdle, was data collection. As easy as it sounds, it was not the easiest in practice.

The initial plan was for data to be collected physically from individual researchers. A collection would be done for a period of time for the population the corpus. After populating the engine with the collected data then ultimately commence research. However, in regards to data collection, it was exhausting to have to explain to every researcher the reason for the collection of their published works. While some authors were worried about plagiarism on the one hand, others had concerns about physically searching for the papers and then compiling due to their busy schedules. The plagiarism concerns were

[4] Brezina, V., Weill-Tessier, P., & McEnery, A. (2021). #LancsBox v. 6.x. [software package]

quickly placated with the assurance that the research is only interested in already published works in journals and the public domain. The explanation was well understood, but there was still some form of reluctance. Consequently, a decision was made to do the collation using the World Wide Web.

The works featured in the corpus have to meet some certain specific contextual criteria for a research paper to be selected. It was viable to have multiple papers written by a singular author.

Preliminary Research

Collocates

There are two distinct approaches fondly used in the notion of formulaicity and classification of collocations: the phraseological approach and the frequency-based or distributional approach. The phraseological approach focuses on the semantic relationship between two or more words and the degree of non-compositionality of their meaning. The frequency-based approach, on the other hand, draws on quantitative evidence to identify the co-occurrence of words in corpora (Paquot & Granger, 2012), with three subtypes as a tool for identification of co-occurrence: surface, textual, and syntactic. Another method of identifying collocates is the distance and proximity between co-occurring words.

Collocation Using statistical definitions can distinguish two major criteria (Ellis et al. 2015)[5] , namely, absolute frequency and word combinations' strength of association. Absolute frequency counts the co-occurrence in word form, while the latter provides information about frequencies and other properties that can be expressed

[5] Ellis, N. C., Simpson-Vlach, R., Ro¨mer, U., Brook O'Donnell, M., & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition. In S. Granger, G. Gilquin, & F. Meunier (Eds.), The Cambridge handbook of learner corpus research (pp. 357–378). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139649414.016.

mathematically (McEnery and Hardie, 2011)[6], they are also called the Associate Measures (AM). AMs are termed the measure of the strength of word combinations. In measuring the AM of frequency, three formulaic dimensions are to be considered: dispersion, exclusivity, and directionality[7]. Dispersion expresses how evenly the pattern of co-occurrence of words appears in a corpus. Exclusivity looks at the extent to which two words occur predominantly in each other's company. The mutual information score is a strong indicator of this property in a corpus. Directionality postulates the attraction strength of collocates; that is, the probability that two words will co-occur in a pair. There are three kinds of AMs commonly used in corpus-based language study: t-score, MI-score, and log-dice.

When it comes to t-score, there are two opinions that reign supreme: "it is the measure of certainty of collocation" (Hunston, 2002); "it is the measure of the strength of co-occurrence" (Wolter & Gyllstad, 2011). However, the t-score has been established to have short comings[8][4]. The T-score is calculated based on raw frequency data from which random co-occurrences are eliminated; this is then divided by the square root of the raw frequency. The T-score depends on the size of the corpus and thus operates on a different scale based on corpus size. As observed by Durrant & Schmitt (2009), it highlights the different frequency combinations of words; thus, making it close to raw frequency ranking. The T-score is generally used in comparing two or

[6] McEnery, T., & Hardie, A. (2011). Corpus linguistics: Method, theory and practice. Cambridge, UK: Cambridge University Press.

[7] For more on formulaicity dimensions see: Dana Gablasova, Vaclav Brezina, and Tony mcenery (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. A Journal of Research in Language studies.

[8] See Evert (2005, pp. 82–83)

more corpus using the standard deviation of co-occurrence counts. NAEC is currently not big enough to calculate the t-score.

MI-Score

MI-score is a measure of strength in relation to the tightness, coherence, and appropriateness of word combinations. The MI-score uses a logarithmic scale to express the ratio between the frequency of the collocation and the frequency of random co-occurrence of the two words in a combination (Gulasova et al., 2017). Using the formula below, we can calculate the MI-score of two concurring words. The formula as used by Cover et al(2006)[9] and Shannon et al (1948) is shown below.

$$MI(X;Y)=\sum x\in X\sum y\in YP(x,y)\log\left(\frac{p(x)p(y)}{P(x,y)}\right)$$

MI score compares the probability of observing x and y together (the joint probability) with the probabilities of observing x and y independently (chance)."[10] Church et al (1990). We will calculate the Mutual Information (MI) score between the words "national" and "development" that often co-occur together NAEC, they are two words nouns randomly selected. Most of the other words with higher collocates, are either propositions, nouns and propositions

In fig. 1.0, *P(national)* is the probability of the word "national" occurring. The frequency of "national" is then divided by the total number of tokens in the corpus. *P(development)* is the probability of the

[9] Cover, T. M., & Thomas, J. A. (2006). Elements of information theory (2nd ed.). Wiley-Interscience.

[10] **Church, K. W., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, 16(1), 22-29.**

word "development" occurring. Similarly, it is calculated as the frequency of "development" divided by the total number of tokens in NAEC. With p(national development) same is done for the phrase occurring. After which we then apply the Mutual Information Formula. The resulting MI score using the log., is approximately -0.00186, in the negative. If the MI scores is higher, it would indicate a stronger relationship between the words. In this case, the MI score is very close to zero, suggesting that the occurrence of "national development" is not strongly associated compared to what would be expected if "national" and "development" were independent.

Fig. 1.0: MI-Score of "National Development"

$$\boldsymbol{P(national)} = \frac{780}{349762} \approx 0.00223$$

$$P(development) = \frac{863}{349762} \approx 0.00247$$

$$P(\text{national development}) = \frac{77}{349762} \approx 0.00022$$

Mutual Information formula:

$$MI(\text{national; development})=\text{P(national development)}\cdot$$

$$\text{Log} \quad \left(\frac{\text{P(national development)}}{\text{P(national)}\cdot\text{P(development)}}\right)$$

$$MI(\text{national; development})= \frac{77}{349762} \,.\, log\left(\frac{\frac{77}{349762}}{\frac{780}{349762}\cdot\frac{863}{349672}}\right)$$

$$\approx \qquad 0.00022.\,log\left(\frac{0.00022}{(0.00223).\,(0.00247)}\right)$$

$$MI(\text{national; development}) \approx \frac{77}{349762} \,.\, log\left(\frac{77}{671940}\right) \approx$$

$$0.00022.\,\text{log}(0.04)$$

$$\approx \frac{77}{349762} \,.\,(-0.848)$$

$$MI\ score\ \text{(national; development)} \approx -0.00186$$

Log Dice

Log Dice[11] is a lexicographer –friendly association score (Rychlÿ, 2008). It is used in extracting terms. It can be used in analysing and highlighting the exclusivity in collocation between words without low frequency bias. It is a measure of coefficient association between terms using a formula that quantifies the strength of association between words A and B based on their co-occurrence patterns. The Log Dice measure was introduced by Evert (2005) in the paper titled "The Statistics of Word Co-occurrences: Word Pairs and Collocations". Evert's work focused on statistical measures for word co-occurrences and collocations in corpus linguistics. The Log Dice measure is one of the metrics he proposed to assess the strength of association between two words in a text corpus.

In analysing the result of a Log Dice, a higher Log Dice coefficient indicates a stronger association between words A and B. A lower value indicates a weaker association. Using the formula:

$$\text{Log Dice } (A,B)\text{=log}\left(\frac{2nAB}{nA+nB}\right)$$

With this, we can calculate the co-efficiency of the words “National development” through Log Dice, using the same co-occurrence and

[11] Evert, S. (2005). The Statistics of Word Cooccurrences: Word Pairs and Collocations. In S. Kepser & M. Reis (Eds.), Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives (pp. 139-158). Mouton de Gruyter.

individual frequency counts.

nAB(the co-occurrence of "national development") = 77

nA(the frequency of “National”) =780

nB(Frequency of “development”) = 863

To calculate the Log Dice coefficient:

Fig. 2.0

Log Dice("national”, “development")=$\log\left(\frac{2\times77}{780+863}\right)$

$$\approx\log\left(\frac{154}{\text{dx}1643}\right)$$

$$\approx\log(0.09367)$$

Log Dice coefficient approximately ≈-2.97

Colloquial English and Nigerian English

According to the *OED* (2020), words in Nigerian English include “*agric, adj. & n.*” , “*qualitative, adj.*”, “*severally, adv*”, “*zoning, n*” “*zone, v.*” these words are reported to be peculiar to Nigerian English. However, with a preliminary research in NAEC, the word “agriculture” occurred 101 times in the corpus; not once was the word “*agric*” used either as a noun or an adjective. The word “*several*” has an occurrence of 170, but not once was "severally" used. “*Qualitative*” occurred 11 times, “*zone*” had 18 hits, and “*zoning*” had no hits. While the corpus is still rather small to make an assertive assumption, it is rather peculiar that a quantitative research can be done on standard Nigerian English without prejudice. *"Agric,"* as often used, is more of a colloquial use in Nigeria than a Nigerian English standard; the same can also be said about the word “*severally*”. "Qualitative" can be said to be a part of Nigerian English, while *"zoning" will* have conditional or special use. With a corpus like this, a distinction can be made to differentiate between

Nigerian colloquial English and standard Nigerian English, which can help in the standardization of this genre of English language.

Conclusion

This paper began with an introduction to the nativized English language spoken in Nigeria, called Nigerian English, discussed the process of the development of the Nigerian Academic English Corpus (NAEC), the criteria used in the collation process, and the various participating universities in the NAEC. Ultimately, the work iterates the various issues encountered in the process of documenting and populating the corpus, the initial plan of the research team for the corpus, and the eventualities that led to the proffered solutions.

LancBox was then employed for the preliminary analysis of the data in the corpus. With the preliminary analysis done with the data in the study, it shows an endless possibility of quantitative analysis. Two words were chosen from the NAEC and their MI scores and Log-Dice of the collocates were calculated. The MI score was-0.00186 while the Log Dice was -2.97, showing a weak association between the words. The possibility of standardizing Nigerian English using the NAEC was also explored showing a prospect of being able to draw a precise and conclusive decision between Nigerian colloquial English and standard Nigerian English.

Referencse

Adegbija, Efurosebina. (1989) "Lexico-semantic variation in Nigerian English", *World Englishes*, 8(2), 165–177.

Biber, D., S. Conrad, R. Reppen, P. Byrd and M. Helt. 2002. 'Speaking and writing in the university: a multidimensional comparison', *TESOL Quarterly* 36, pp. 9–48.

Dana Gablasova, Vaclav Brezina, and Tony McEnery (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *A Journal of research in Language studies.* 67: S1, pp. 155-179.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL-International Review of Applied Linguistics in Language Teaching*, 47, 157–177. doi:10.1515/iral.2009.007

Ellis, N. C., Simpson-Vlach, R., Ro¨mer, U., Brook O'Donnell, M., & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 357–378). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139649414.016.

Evert, S. (2005). *The statistics of word co-occurrences: Word pairs and collocations*. (Doctoral dissertation, Institut fu ̈r maschinelle Sprachverarbeitung, University of Stuttgart.)

Evert, S. (2005). The Statistics of Word Cooccurrences: Word Pairs and Collocations. In S. Kepser & M. Reis (Eds.), *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives* (pp. 139-158). Mouton de Gruyter.

Granger, S., E. Dagneaux and F. Meunier. (2002). *The International Corpus of Learner English/Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S. and S. Tyson. 1996. 'Connector usage in the English essay writing of native and non-native EFL speakers of English', *World Englishes* 15, pp. 19–29.

Guthrie, M. (1964). Multilingualism and cultural factors. Symposium on multiligualism. London: CCTA/CSA, 107-8.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139524773.

Kachru, B. B. (1985). *Standards, Codification and Sociolinguistic Realism: The English Language in the Outer Circle*. Oxford University Press.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge, UK: Cambridge University Press.

Nesi, H., Sharpling, G. and Ganobcsik-Williams, L. (2002) Student papers across the curriculum: designing and developing a corpus of British student writing. Computers and Composition, volume 21 (4): 439-450.

Oxford English Dictionary. (2020, January). Addition of Nigerian English words. https://www.oed.com/news/post/nigerian-english

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149. doi:10.1017/S0267190512000098.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal,* 27(3), 379-423.

Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, *32*, 430–449.

Weinreich.U.(1953*). Language in contact.* New York: publications of the linguistic circle of New York.