Language Corpora: The Case for Ghanaian English

RICHMOND SADICK NGULA
Department of English
University of Cape Coast, Ghana
Lancaster University

MARK NARTEY
Department of Language and Literature
Norwegian University of Science and Technology
Trondheim, Norway
narteynartey60@gmail.com

ABSTRACT

In the last two decades, the compilation of corpora and the analysis of linguistic phenomena via corpus data have become a fascinating linguistic practice around the world and by this, corpus linguistics is now firmly established as a credible approach for the study of language. Linguists and other researchers of varied persuasions are appreciating more and more the interesting dimensions corpora are introducing to language studies. Yet, not much corpus-based work goes on in Ghana. In this paper, we suggest that a vital first step towards the development of Ghanaian English (GhE) lies in the initiation of large-scale electronic corpus projects. The paper argues that corpora can go a long way to enhance the linguistic descriptions of GhE, making the features of the variety more visible and providing a good opportunity for its codification. The results of building corpora for the study of GhE will not only highlight its rich features, but also help Ghanaians and policy makers to determine its proper status in the country.

Keywords: codification; corpora; corpus linguistics; Ghanaian English; linguistic description

INTRODUCTION

In the 1950s, strident attacks from generative linguist Noam Chomsky (and his followers) did not only make corpus linguistic research unpopular, but also influenced many linguists to begin to think of this approach to linguistics as not worthy of any serious intellectual attention. But it did not take long for this to change. In the last few decades, the building of corpora and the uses to which they are put have regained popularity among linguists of varied persuasions in a manner that is unimaginable; maybe not so unimaginable, given the remarkable contributions corpora have made (and continue to make) in the description of language, and in the construction of linguistic theory. As Meyer (2002, p.1) has observed, "Linguists of all persuasions are now far more open to the idea of using linguistic corpora for descriptive and theoretical studies of language."

Clearly, corpora have positively affected research in linguistics which explains why, in the words of Leech (1991, pp.13-14), "corpus linguistics need no longer feel timid about its theoretical credentials, nor does the earlier Chomskyan rejection of corpus data carry such force." Hence for very good reasons, researchers of language nearly everywhere now draw on corpora in language analysis and description.

It is probably true to say that the English language has been the leading beneficiary in terms of corpus building, annotation and analysis, as this is evident in the many existing corpora on English. For example, the first notable computerised corpus of English, the Brown

Corpus, was developed by Nelson Francis and Henry Kučera at Brown University in the 1960s (Meyer 2002). Thereafter, many other general corpora on English (e.g., the American National Corpus (ANC), the British National Corpus (BNC), the Lancaster-Oslo/Bergen Corpus (LOB), the Australian Corpus of English, etc.) as well as more specialised ones (like the Corpus of English Conversation, the Zurich Corpus of English Newspapers (ZEN), the International Corpus of Learners' English (ICLE), the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) Corpus, etc.) now exist, and are being used to describe the linguistics of English. But corpus linguistic research has extended beyond English to quite a number of other languages around the world including Chinese, Danish, Dutch, German, Maltese, Russian, Slovene and Spanish (Wilson, Rayson & McEnery 2006).

Unfortunately, however, this global trend of building computerised corpora and using them for linguistic analysis has not yet received any serious attention by researchers and scholars in Ghana. Schmeid (1991) suggests that very few countries in Africa are currently engaged in corpus linguistic projects. At the seventh Corpus Linguistics Conference (2013) held at Lancaster University, developer of the *AntConc*corpus search tool, Laurence Anthony, gave a graphical view of places around the world where researchers were downloading the free software for corpus analysis. The map did not show any downloads of the search tool on the African continent, further suggesting that not much corpus linguistics work is going on in Africa.

In Ghana, there is not yet, as far as we can determine, a single machine-readable corpus of any type available for the analysis of the use of English. The International Corpus of English (ICE) project, which evolved from a proposal by Sidney Greenbaum of blessed memory in the early 1990s, had Ghana as one of the original 20 research countries to embark on the building of corpora for varieties of English (Crystal 2003, p.451). As Greenbaum and Nelson (1996, p.3) explain, "ICE was initiated to provide the resources for comparative studies of the Englishes used in countries where it is either a majority first language or an official additional language." At the moment while a number of countries have completed their corpora and have released them for use (ICE-Great Britain, ICE-New Zealand, ICE-Australia, ICE-India and ICE–Singapore), the Ghanaian component, unfortunately, has not been compiled yet. We observe that it is only now that Professor Magnus Huber at the University of Giessen in Germany is collaborating with Professor Kari Dako of the University of Ghana to develop the Ghanaian ICE component. They have now completed the written component and are now completing work on the spoken part of the corpus.

At present in Ghana, there seems to be no unilateral agreement on the validity of Ghanaian English (henceforth, GhE), as people still have doubts about whether distinctive features of the variety are real innovations or are simply markers of deviation from a standard native norm (Adika 2012). Thus one of the major obstacles of GhE is its lack of proper recognition and acceptance in the country. The aim of this paper is to stress the importance of corpora for language description, and argue that large-scale corpus projects can serve a good starting point towards enhanced linguistic descriptions into GhE for its proper recognition as a variety of English within Kachru's (1986) *Outer Circle* of World Englishes.

The rest of this paper is structured as follows: we first discuss the historical background to corpus linguistics, drawing attention to how this approach to the study of language has developed over the years to its modern practice. Following this, we explore the linguistic situation in Ghana, with particular focus on the role and functions of English. Here, we try to show how corpus applications could help enhance linguistic descriptions of GhE, and provide a solid basis for classifying and codifying the distinctive linguistic features of GhE. We argue that incorporating corpus methods and applications in the studies of GhE

would not only enrich linguistic studies into GhE, but could ultimately help establish the legitimacy and recognition of the variety for its proper recognition and acceptance in the Ghanaian society. We conclude with a summary of the main concerns of this paper.

BEGINNINGS OF CORPUS LINGUISTICS

Although corpus-based studies of language have a substantial history, the term *corpus linguistics* itself was first introduced in the early 1980s (Leech 1992, McEnery*et al.* 2006). The history predates the advent of the computer which became an important facility in contemporary corpus work. As Reppen and Simpson (2002, p.92) observe, "Before the advent of computers … many empirical linguists who were interested in function and use did essentially what we now call corpus linguistics"

An empirical approach to the analysis of language is one that relies on naturally occurring spoken or written texts and which stands in opposition to an approach that gives priority to introspection. This kind of language study can loosely be regarded as a corpus-based approach to the study of language, and indeed such was the work of quite a number of linguists prior to the emergence of the use of computers in corpus linguistics. For example, as Hyland (2011) notes, the English grammars of Otto Jespersen, Franz Boas' studies of poorly documented languages, and the grammatical descriptions of structuralists like Zellig Harris and Carpenter Fries were all based on real, authentic examples of usage and could be classified as corpus-based. Hyland (2011, p.99) goes on to say that most of these early language analysts "believed that linguists were virtually obliged to study authentically occurring texts to gain any understanding of the ways language worked".

This notion thus informed much of the work of these researchers and even though they neither used computers nor the sophisticated tools and methods associated with contemporary corpus linguistics, the simple processing methods they used produced basic frequency counts, syntactic patterning, word associations and the meanings of words in different contexts. These methods could essentially be regarded as corpus-based and their practice was thus to serve as the spring board for modern corpus linguistics to take off. In the words of Hyland (2011, p.99), these practices led to the "explosion of interest in corpora".

MODERN CORPUS LINGUISTICS

Modern corpus linguistics is closely connected with the use of computers and today no corpus linguist would imagine anyone doing corpus linguistics simply by relying only on a manual analysis of a few texts in printed format. The computer has become so important to corpus linguistics to the extent that it is reflected in the definition of a corpus. For example, Leech (1992, p.106) writes that a corpus is "a *helluva* lot of text, stored on a computer". In fact, Leech has further suggested that a more appropriate term for the discipline would be *computer corpus linguistics*, owing to the role computers play in the work of practitioners.

So why have computers become indispensable in the work of modern corpus linguists? McEnery *et al.* (2006, p.6) address this question in line with the 'machine-readability' attribute of a modern corpus and outline four major advantages that the use of electronic corpora in language study has over their paper-based equivalents as follows:

i. The most obvious advantage relates to the speed computers offer in the processing of electronic corpora and the ease with which a researcher can manipulate a corpus using such techniques as searching, selecting, sorting and

formatting. It takes only a few seconds for a search query to display results even if the corpus one is working with is pretty huge (a million word and over).

ii. Computers are able to process machine-readable data with such accuracy and consistency that cannot be achieved without them.

iii. the use of computers in the analysis of corpus data "can avoid human bias in an analysis, thus making the results more reliable", and

iv. The use of computers to store a corpus has made it possible for "further automatic processing to be performed on the corpus so that corpus texts can be enriched with various metadata and linguistic analyses".

Thus, McEnery *et al.* (2006) hold the view that computers and computer programs are the tools that have given analyses of corpus texts a tremendous boost, such that corpus-based studies carried out in the last 20 years would not have been possible without these tools. A typical case in point is the renewed interest in lexical studies that have come about as a result of electronic corpus analysis. Recent work on the semantic association of words (collocation, semantic prosody, and semantic preference), such as Hunston (2008), Römer & Schulze (2010) and Sook Beng & Chee Keong (2014), has been accomplished because of the availability of computer corpus tools. In the next sub-section, we consider what a corpus is, and also highlight the key issues for consideration in the construction of a corpus.

WHAT IS A CORPUS?

To engage in a corpus-based study presupposes that there is a corpus upon which the study will be based. Leading practitioners in the discipline of corpus linguistics have given their own versions of what a modern corpus (plural corpora) is, and while each of the various definitions has its unique readable style, the different perspectives, considered together, capture the salient methodological issues that one has to be mindful of when designing and constructing a corpus. We wish to provide just four examples of the definition of a corpus, and on the basis of these definitions discuss some of the important issues considered when collecting texts for the construction of a corpus.

> A corpus is a collection of naturally-occurring language texts, chosen to characterise a state or variety of a language (Sinclair 1991, p. 171).
> A corpus is a collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis (Francis 1982, p. 7).
>
> A corpus is a *helluva* lot of text, stored on a computer…computer corpora are rarely haphazard collections of textual material: they are generally assembled to be (informally speaking) *representative* of some language or text type (Leech 1992, pp. 106 - 116).
>
> A corpus is a collection of (1) *machine-readable* (2) *authentic texts* (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of a particular language or language variety (McEnery *et al.* 2006, p. 5).

First, it follows from all four definitions that to create a corpus, one has to collect texts. Texts typically come in spoken or written forms, and depending on the corpus compiler's purpose, the corpus may include either or both of these forms. It is generally agreed that it is more time-consuming and tedious to create a spoken corpus than a written

one because of the additional processes of recording and transcribing speech. This also partly explains why for some existing corpora that incorporate both spoken and written texts, the disparity between the two modes is considerable. An example that immediately comes to mind is the British National Corpus (BNC), a 100 million word corpus whose spoken component is only 10 million as against a 90 million written component.

Secondly, the texts collected for a corpus are ideally naturally-occurring (authentic) texts, as highlighted in the definitions in Sinclair (1991) and McEnery *et al.* (2006). The term 'naturally occurring' suggests that the texts to enter a corpus are produced as language within specific communicative events, and are without the intervention or inducement of the corpus compiler. Relying on naturally occurring texts therefore affords the opportunity of basing one's linguistic analysis on instances of language use in real-life situations rather than on language derived from induced data-gathering techniques such as interviews, questionnaires and administration of tests.

A major weakness these latter techniques have over corpus data is that they often involve "setting up particular 'artificial' research environments" (Silverman 2005, p.119), a procedure which may end up reducing the authenticity of the data (texts) collected. So for example, if we were interested in studying error patterns in the writing of senior high school students, the corpus researcher would prefer to build a suitable learner corpus (if one does not already exist) which would rely on, say, previous argumentative or expository essays written by the students rather than asking the students to write similar essays in an 'artificial' set-up for the study. Using existing essays, thus, assures the researcher of the naturalness and authenticity of the data.

Another very important issue in corpus design and compilation relates to the idea of representativeness, and this is explicitly mentioned in all our definitions above, except in Sinclair's where it is alluded to as well. This concerns the 'corpus' and the 'language' that it represents, and can be likened to the relationship between a 'sample' and a 'population' used in most social science research. But in corpus building it is difficult to refer to a language (or language variety) as 'the population' from which a sample is to be drawn, since the texts available for the language may be unlimited. The unlimited nature of language poses a problem in constructing a representative corpus.

According to Leech (2011, pp.158-159), "we see the difficulty of determining whether what is found to be true of a corpus can be extrapolated to the language as a whole". It seems correct, then, to assume that no corpus, no matter how large and carefully designed, can have exactly the same characteristics as the language itself. Thus, a corpus might never be fully representative; it can only at best aim to be maximally representative (Reppen & Simpson 2002). In other words, representativeness in corpus compilation is an ideal to strive for but not necessarily to be achieved.

So far, the way corpus practitioners have handled the issue of representativeness has been to carefully design the corpus, considering the range of text types to be included, the range of authors (or speakers), and the number of text samples and the selection of particular texts. All these require some sampling decisions to be made by the compilers and/or designers. The overriding principle is that, in the end, the corpus is seen to provide a reasonably realistic picture of the language or subset of language it represents. This means a more realistic approach is to select sufficiently large and relevant samples of the different genres (text types) to be included in the corpus, and then try to establish ways in which a fairly representative corpus may be reached (McEnery & Wilson 2001, Hunston 2002). After all, a corpus-based study is only as good as the corpus is sufficiently large and relevant to the research questions for which answers are being sought.

Additionally, the definitions given in Leech (1992) and McEnery *et al.* (2006) above highlight the important point of the corpus being machine-readable or stored electronically. We have already shown how advances in computer technology have made this possible. Here, we wish to draw attention to one or two practical issues regarding the computerisation of corpus texts. The internet has now made it a lot easier to collect many texts that are already available in electronic format. As Baker (2006, p. 31) has noted, "due to the proliferation of internet use, many texts which originally began life in written form can be found on websites or internet archives". So for example, building an electronic corpus of editorials from newspapers in Ghana now would not be as arduous a task as it would have been many years ago, the reason being that there are now websites for the major newspapers in Ghana (e.g. *The Daily Graphic*, *The Ghanaian Times, The Daily Guide*) where editorial archives for these newspapers are available. The editorials can thus be downloaded easily. In fact, using digital versions of newspaper articles to build corpora for language study is already in practice even in languages other than English. A typical example is Abdul Razak's (2014) study, where he compiles digital versions of Arabic newspapers (totalling 87,000 tokens) into an electronic corpus for a study of word variations. Without such internet accessibility, the compiler would have to start the collection from the written hard form by either entering (keyboarding) the editorial texts directly onto a computer or scanning them using a scanner with Optical Character Recognition (OCR) software, processes that are time-consuming and error-prone, especially with keyboarding.

Another point relates to how the texts are to be stored (file format). To save a text on a computer, you are always given an option as to how you want it saved. The options are seen from the drop-down 'Save as' menu. It is usually preferable to save a corpus text using the file format *Plain text* because most corpus analysis tools at present work best with this format, although the *Rich text* and *XML* formats are other options. It is important to note that most corpus analysis tools will not read texts in *word* or *pdf* format, and so when texts in these formats are downloaded from the internet, they would have to be further converted to and saved as *Plain text*. However, as Reppen (2010) suggests, file naming conventions need to be established before saving a text. According to Reppen (2010, p.33), file names should "clearly relate to the content of the file to allow users to sort and group files into sub-categories or to create sub corpora more easily".

A final point we wish to make in this section is explicitly stated in Francis' (1982) definition: a corpus, once it is built to completion, is to be used for linguistic analysis and description. With the help of search tool packages such as *Wmatrix* (Rayson 2009) and *WordSmith* (Scott 2013), various kinds of analyses can be carried out on a corpus. As noted by Römer (2006, pp.84-90), these tools allow you to do such things as word listing and counting (tearing the text apart), tracing repeated occurrences of an item in a text (examining dispersion plots), compiling a concordance (putting words back into context), sorting the context in a concordance (uncovering patterns) and examining the context of a word (looking for collocations).

McEnery and Hardie (2012, p.2) further note that "concordances and frequency data exemplify respectively the two forms of analysis, namely qualitative and quantitative, that are equally important to corpus linguistics". It is these corpus-handling techniques that further enable a researcher to comprehensively study word meaning in context, frequency distribution patterns, collocation patterns, use and function of grammatical parts (morphology and syntax), aspects of discourse and many more. Overall, searching a corpus allows one to see what patterns are associated with lexical, grammatical and discourse features, patterns that can easily elude an analysis that relies entirely on human introspection.

CORPORA FOR GHANAIAN ENGLISH

THE LINGUISTIC SITUATION IN GHANA

Ghana is a multilingual country. Quite clearly, then, the linguistic landscape in Ghana is rich and diverse, resulting in a high degree of linguistic heterogeneity in the country. Essentially, English plays a very cardinal role in Ghana's linguistic ecology as it exists alongside over fifty indigenous languages as the official and somewhat 'national' language. Despite this co-existence with indigenous Ghanaian languages, English attracts the attention of everyone, given its wide range of applications in the domestic affairs of Ghana (used for example in governance, education, media, law, business, etc.) as well as in international communication. In this regard, Morris (1998, p.15) avers that "fundamentally, in contemporary Ghanaian society, English is used for official purposes: governance, education and diplomacy".

As far as the indigenous languages are concerned, it is worth mentioning that there is no consensus of opinion on the total number of languages spoken in Ghana. Hence, Kropp Dakubu (1988, p.10) submits that "we could say there are between forty-five and fifty languages in Ghana". Owing to the high degree of linguistic diversity in the country, languages perform different functions in different communicative contexts, but crucially the linguistic situation has led to the emergence of English, Akan and Hausa as the most important lingua francas (Obeng 1997, p.63). Of these three francas, there is no gainsaying the point that English is the commonest, most pervasive and most central as rightly noted by Morris above.

The centrality of English in the public domains of life in the country as acknowledged by Morris is further tacitly reinforced by Sackey (cited in Morris) when in response to why English was retained as the official language even after independence he explained that "… the country had no single indigenous language of government, law, education and social intercourse at all levels …." Sarfo (2011, p.460) summarises the quintessence of English in Ghana when he opines that "English language has come to stay as a communicative tool for social, political and economic development".

Owing to its importance in both national and international affairs of the nation coupled with the emergence of the New Englishes, there has been growing interest – in the last two and half decades –in Ghanaian English. The term Ghanaian English actually poses a definitional challenge, not just because of the conflicting views on its status, which we will address shortly, but also because in Ghana, different professionals with different levels of education and exposure to English use different varieties of English (Boadi 1997). In this paper, we define Ghanaian English (GhE) in terms of what we may qualify as a local Ghanaian standard variety of English. Thus for us, it refers to the English produced by educated Ghanaians who have been brought up and schooled up to the university level in Ghana, and who are using English for major communicative purposes.

We must hasten to add that there is not only educated GhE, but also a continuum of more or less successful approximations to the standard (Sey 1973, Boadi 1997) based on whether or not English was acquired in school coupled with the educational attainment of the speaker. In this regard, Boadi's work discusses and shows three main varieties of English in Ghana roughly corresponding to the sociolinguistic varieties of basilect, mesolect and acrolect.

The first variety or group of speakers could be considered as belonging to the incipient bilingual stage. According to Boadi, this group of speakers either do not have any form of education –and so might have acquired English through observation or within their

environment –or have attended only elementary school. This group of speakers are uneducated or less educated, and so can be regarded as speaking the basilectal variety of English. The second group of speakers, roughly corresponding to the mesolectal variety, have a fairly reasonable amount of education. Usually, these speakers have progressed to the upper forms of secondary school. The third group of speakers Boadi identifies are highly educated, invariably having a university (or an equivalent) level of education. Quite clearly, this group would correspond to the acrolectal variety –the variety our definition of GhE above relates to.

Given the cline of proficiency of GhE presented above, it is our view that if Ghanaians are to promote a local standard variety in Ghana, the acrolectal variety would naturally and obviously be the variety to turn to. It is this acrolectal variety of English in Ghana that could possibly replace native standard varieties in the Ghanaian context of English use –the institutionalized variety recognised in Kachru's *Outer Circle* of World Englishes (Kachru 1986, 1992). Although the existence of a Ghanaian variety of English has long been recognised with several articles written on different aspects of it (Gyasi 1990, Owusu-Ansah 1994, Sackey 1997, etc.), studies on its linguistic characterisation and/or typical linguistic features have not been intensive and comprehensive. Hence, there still isn't an instructive picture about what the variety entails to give any hopes of it becoming an authoritative local model in Ghana, thereby strengthening counter arguments against its legitimacy and validity. For instance, while researchers like Ngula (2011), Adika (2012) and Owusu-Ansah (2012) have maintained that GhE is an emerging Ghanaian standard variety of English, other scholars, including (Sey) 1973 and (Ahulu) 1994 insist that these supposed distinct features of the so-called GhE are, in fact, markers of deficiency in the use of English, rather than legitimate innovations.

EFFORTS SO FAR TOWARDS DESCRIBING GHANAIAN ENGLISH

Since Sey's (1973) seminal work on Ghanaian English, there have been several attempts by other researchers –Ghanaians and non-Ghanaians alike –to describe the innovative features of GhE, and more importantly, to establish its legitimacy as a standard nativised English. Most of these studies, however, were based on small amounts of data, thus revealing only superficial tendencies without any conclusive findings. Indeed, the vast majority of these studies can safely be labelled as preliminary investigations or even pilot studies, and so (perhaps) would not suffice to be used as conclusive evidence for the linguistic distinctiveness of GhE.

Commendably, these studies – albeit with minimal data – have not been delimited to merely a couple of grammatical levels, but have spanned or touched on quite a number of these levels, including vocabulary, phonology, syntax and semantics. This has culminated into a good number of studies on GhE, especially in the last two and half decades. Notable studies that have attempted to describe the phonological peculiarities of GhE include Simo Bobda (2000), Adjaye (2005), Huber (2008) and Ngula (2011). With respect to unique grammatical features of GhE, the studies of Owusu-Ansah (1994), Huber and Dako (2008) and Ngula (2010) are worth mentioning. Still, studies carried out by Bamiro (1997) and Dako (2001, 2002) try to give a representation of the innovative forms of vocabulary or lexis in GhE, while Owusu-Ansah (1992) touches on GhE, exploring some of its discourse features. Further, some other studies, *viz.* Platt *et al.* (1984) and Mesthrie and Bhatt (2008) have examined the linguistic characterisation of GhE by drawing on two or more of the afore-mentioned levels.

Following from the above, there is no gainsaying the point that some significant strides have been made towards the description of the typical lexico-grammatical properties

and features of GhE. Good as these efforts might be, it does remain a fact that many of these previous studies have failed to capture interesting and nuanced quantitative findings of the features they explored owing to the rather small sizes of their data. Owusu-Ansah (1994), for instance, while examining some of the distinct grammatical features of GhE looked at a mere twenty-two personal letters of Ghanaian students. Similarly, in his study on the semantics of modal auxiliary verbs in GhE, Ngula (2010) used a total sample size of only 52, 000 words.

Owing to the use of small data sets which have characterised most of the existing scholarship on GhE, we are presented with the somewhat difficult situation where we are unable to explicitly and confidently tell whether the few instances of occurrences in the rather small data qualify as innovative features, especially so when the pervasive and wide spread use is crucial in determining such features (Bamgbose 1998). Consequently, some of these inherent setbacks in previous studies of the innovative features of GhE have somewhat given credence to arguments put forward by purists against GhE, saying that what are often described as innovative features are inter-language difficulties or deviant language usage, deficiencies and fossilized errors. They have argued therefore that these so-called innovative features are not desirable (Sey 1973, Ahulu 1994). In view of this, the characterisation of GhE norms is seriously being challenged at the present time.

WHAT DOES GHANAIAN ENGLISH NEED?

Significantly, we argue that the arguments about the existence of a standard or an emerging standard variety of Ghanaian English would be laid to rest, once and for all, if extensive and comprehensive corpus-based approaches are brought to bear on GhE. While the debate regarding the validity of GhE as a 'new' English as well as the extent to which the linguistic features associated with it can be considered innovative continue to rage on, it is our view that applying corpus-based approaches to the various linguistic aspects of GhE would provide credible evidence and hard proof of the distinctiveness of the identified features and, subsequently, the legitimacy of GhE.

We therefore suggest that a vital first step towards the development and global acceptance of Ghanaian English lies in the initiation of large-scale electronic corpus projects. Phrasing it slightly differently, we advance that there has to be a general corpus of GhE that is quite substantial. While the ICE project on GhE being currently undertaken by Huber and Dako is good, we maintain that it is quite small since it is made up of just over a million words and therefore can be duly regarded as not sufficient enough to address most effectively certain research questions, especially in the areas of lexis and lexicography where huge data sets are often required. Baker (2006, p.28), for instance, has noted that for lexical studies, "a million words is unlikely to be enough …"

Not discrediting, discouraging and faulting Huber and Dako's efforts – because we think it is a laudable one – we strongly suggest that their effort should not be the beginning and the end of large scale electronic corpus studies on GhE. Instead, subsequent corpora should follow this worthy first effort. A corpus of GhE of about 100 million words (containing a good balance between written and spoken genres and/or registers), for instance, should be a huge milestone. Inasmuch as we concede that such a landmark feat would require a collaborative team and a huge financial injection, it would be worth the while since the existence of such a corpus of GhE would improve research on GhE significantly.

Such a general corpus of GhE being proposed by this study would in no small measure enhance the linguistic descriptions of GhE, making the features of the variety more visible and providing the opportunity for its codification. By highlighting its rich and typical

innovative features, the results of building a general corpus for the study of GhE would put an end to the issue of what findings should be tolerated and what should not (Owusu-Ansah 1997). This is because such a corpus will be carefully constructed to reflect only educated uses of GhE.

From the foregoing paragraphs, it is quite discernible that the role of a general corpus of GhE as far as the determination of the unique linguistic characterisation of GhE, and indeed the overall legitimacy of GhE as'new' English are concerned neither can be underestimated nor overemphasised. First and foremost, such a general corpus of GhE would enhance vigorous research into the various linguistic aspects of English. For instance in the areas of lexicography and lexical studies, concordance lines and collocates could help a great deal in illuminating patterns of meaning associated with words in the Ghanaian context that would probably escape an analysis based on hand and eye alone. This kind of analysis should therefore serve as a useful framework to help identify peculiar uses of words that characterise GhE. Besides, other corpus tools such as clusters/n-grams, word lists and keyword lists could provide invaluable insights into English usage in the Ghanaian context, thereby enhancing our ability to effectively classify markers of GhE in the areas of grammar, discourse analysis, pragmatics, and even phonology. Eventually, as studies of GhE corpora accumulate, the innovative features of GhE with respect to these areas would be clearly manifested for all to see and attest to.

Second, the general corpus of GhE would provide very nuanced and interesting quantitative findings of the lexico-grammatical and discourse properties of GhE, which virtually all the previous studies on GhE have failed to capture owing to their rather small data sets. While qualitative differences between varieties of English (differences relating to kinds of linguistic feature used) are much easier to spot and discuss with some level of confidence, quantitative differences (differences relating to degrees of use of a feature) can often be suspect and somewhat misleading, especially when data size is way too small and therefore contains only a few instances of the feature(s) being explored. Such quantitative deficits can be considerably improved when working with relatively large electronic corpora. Conrad (2011, p.54) points out that because "identifying [linguistic] patterns relies on quantitative analysis…much of the strength of Corpus Linguistics comes from the role of quantitative analysis".

Third, a large general corpus of GhE would provide a sound basis for which GhE can be compared with other varieties of English (native, non-native or both) such as American, British, Indian and Singapore Englishes. After all, the different varieties of English around the world are so characterised largely by virtue of the linguistic variations/differences that distinguish one from another. Such comparative studies are more effectively carried out on varieties for which electronic corpora are available and readily accessible. There are good examples of corpus-based comparative studies carried out between varieties of English in recent years. Much work has already been done on the ICE project such as the edited volume by Hundt, Marianne and Ulrike Gut published in 2012.

Apart from ICE, other corpora have aided descriptions into varieties of English. Biber (1987), for instance, relying on Brown and LOB compared American and British English writing features. With the existence of more recent and much bigger corpora like the BNC, Bank of English and COCA for the leading standard varieties (British and American), the GhE corpus we are proposing should greatly enrich comparisons between GhE and other varieties, especially with these two leading varieties. The resulting studies would not only establish the point(s) where GhE significantly diverges from these other varieties, both qualitatively and quantitatively, but also highlight in a credible manner the distinctive

patterns of English usage in Ghana, and by that further provide a firm basis for classifying and codifying the typical linguistic features of GhE.

Issues relating to the codification of the features of a variety of language (in this case English) are crucial, and must therefore be given the deserved attention if the variety is to be recognised and accepted. Codifying a language (or variety of a language) involves setting up official rules and norms as a process of standardising the language to guide general usage and pedagogy. The availability of a general corpus of GhE would ultimately facilitate the codification of GhE at the various linguistic levels: orthography, vocabulary, grammar and pronunciation, leading to the production of dictionaries, published grammar books, lexical guides, phraseological information manuals and other similar guideline materials. Such documentation is extremely important as it would provide a credible source of reference for the variety.

The codification of GhE would, no doubt, make the features of the variety more visible than they currently are. We argue that the more visible the features of the variety, the more it injects confidence in the people (here, Ghanaians), and make them begin to see what they can really identify as their own. This would result in a positive change in attitude towards GhE and provide the impetus for the nationals to own the language. Again, the visibility of the innovative features of the variety is likely to gradually promote the international acceptance of GhE, as the tangible evidence of the linguistic distinctiveness of the variety would be palpably reflected in the documentation for all to see and attest to.

Ultimately, the availability and accessibility of a large-scale electronic corpus of GhE (which would result in an extensive codification of the innovative features of the variety) would settle, once and for all, the validity concerns that purists have consistently raised about the variety. Besides, the presence of the corpus would also go a long way to nullify the negative designation that some scholars have ascribed to the characterisation of GhE norms of usage.

Finally, we would like to stress that the corpus linguistic approach being mooted by the present study, as a clear-cut means of establishing GhE as a credible and legitimate non-native variety of English, is realistic. Indeed, our conviction is borne out of the fact that there are a good number of examples of other varieties of English whose linguistic descriptions have been greatly enhanced due to the exploitation and/or deployment of corpus data. Because corpora in English have proved their usefulness in empirical language analysis as well as in the building of linguistic theory, many countries where English is used either as a first language (Britain, North America, Australia), additional second language (e.g., India, Singapore, Kenya), or foreign language (Spain, China, Japan) have successfully embraced the corpus approach to the study of English. In the area of the grammar of English in native contexts, for instance, so much corpus-based research has been documented, the greatest milestones in this regard being, perhaps, the reference works by Quirk *et al.* (1985), *A Comprehensive Grammar of the English Language* and Biber *et al.* (1999), *Longman Grammar of Spoken and Written English*. The former presents refreshing insights on grammatical differences between British English and American English whereas the latter reports interesting differences along four major registers (conversation, academic prose, news, fiction), showing how different social contexts of language are reflected in grammar. If corpora and corpus techniques can be used to achieve such great feat in the description of language elsewhere, then we in Ghana should be able to utilise its great potential to help enrich linguistic descriptions into GhE.

CONCLUSION

In this paper, we have shown that corpus linguistic methods, if vigorously applied to the study of English in Ghana, could richly enhance the prospects of GhE by proving beyond reasonable doubt its credibility and, legitimacy as a non-native variety of English within Kachru's *Outer Circle* of Englishes. But as the present situation suggests, the status of GhE is uncertain. While there are still arguments against its legitimacy and validity, studies that have sought to highlight its innovative features, thus far, have been inadequate owing to the limited data sets upon which these studies are based, and therefore their findings have often been inconclusive.

We argue that the shortfalls that confront studies into GhE (especially regarding data size and modes of analysis) can be overcome by the application of corpora and corpus tools. It is not for nothing that corpus linguistics is increasingly gaining popularity in many parts of the world and among linguists of different persuasions, with even theoretical linguists, who in the past would have had nothing to do with 'performance data', now incorporating corpus approaches in their work (McEnery & Hardie 2012). This scientific method to language analysis is now utilised in nearly all subfields of linguistics, virtually changing the direction of the study of linguistics in the 21st century. We envisage the methodology being the panacea needed to aid the codification and ultimate acceptance and recognition of the local Ghanaian standard variety of English (GhE). The starting point is the building of large electronic corpora of GhE. Once such corpora become available, effective linguistic descriptions (both qualitative and quantitative) into GhE can be intensified in order to classify and codify genuine distinctive markers of GhE. Until all of this is done, many of us would continue to talk about GhE largely in abstraction, having very little in terms of its features and norms of usage to show for it, and doubts surrounding its legitimacy will continue to linger in the minds of many.

REFERENCES

Abdul Razak, Z. R. (2014). Word usage variations in Arabic newspapers: A corpus investigation. *GEMA Online Journal of Language Studies 14*(2): 19–45.

Adika, G.S.K. (2012). English in Ghana: Growth, tensions and trends. *IJLTIC* 1(1): 151-166.

Adjaye, S. (2005).*Ghanaian English pronunciation*. London: Edwin Mellen.

Ahulu, S. (1994). How Ghanaian is Ghanaian English? *English Today 38*: 25-29.

Baker, P. (2006).*Using corpora in discourse analysis*. London: Continuum.

Bamgbose, A. (1991). *Language and the nation: The language situation in Sub-Saharan Africa*. Edinburgh: Edinburgh University Press.

Bamgbose, A. (1998). Torn between the norms: Innovations in world Englishes. *World Englishes 24* (1): 85-93.

Bamiro, E. O. (1997). Lexical innovation in Ghanaian English: some examples from recent fiction. *American Speech* 72(1): 105–112.

Biber, D. (1987). A textual comparison of British and American writing. *American speech*. *62*(2): 99–119.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Boadi, L.A. (1971). Education and the role of English in Ghana. In: J. Spencer (Ed.) *The English Language in West Africa*, pp. 49-65 (English Language Series). London: Longman Group.

Conrad, S. (2011). Variation in corpora and its pedagogical implications: interview with Susan Conrad. In: V. Viana, S. Zyngier and G. Barnbrook (Eds.) *Perspectives on corpus linguistics*, pp. 47-62, Amsterdam: John Benjamins.

Crystal, D. (2003). *The Cambridge encyclopedia of the English Language* (2nd edn). Cambridge: Cambridge University Press.

Dako, K. (2001). Ghanaianisms: Towards a semantic and formal classification. *English World Wide*. *22*(2): 23–53.

Dako, K. (2002). Code-switching and lexical borrowing: Which is what in Ghanaian English? *English Today 18*: 48–54.

Francis, W. N. (1982). Problems of assembling and computerizing large corpora. In S. Johansson (ed.) *Computer corpora in English Language research*. pp. 7–24, Bergen: Norwegian Computing Centre for the Humanities.

Greenbaum, S. & Nelson, G. (1996). The international corpus of English (ICE) project. *World Englishes* 3-15.

Gyasi, I.K.C. (1990). The state of English in Ghana. *English Today 23*: 24-26.

Huber, M. (2008). Ghanaian English phonology. In R. Mesthrie (Ed.) *Varieties of English: Africa, South and Southeast Asia*, (pp. 67 – 92). Berlin: Mouton de Gruyter.

Huber, M. & Dako, K. (2008). Ghanaian English: Morphology and syntax. In R. Mesthrie (Ed.) *Varieties of English: Africa, South and Southeast Asia*, (pp. 368–380). Berlin: Mouton de Gruyter.

Hundt, M. & Gut, U. (Eds.). (2012). *Mapping unity and diversity world-wide: Corpus-based studies of New Englishes*. Amsterdam: John Benjamins.

Hunston, S. (2008). Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics* 13(3): 271–295.

Hyland, K. (2011). Looking through corpora into writing practices: Interview with Ken Hyland. In: V. Viana, S. Zyngier and G. Barnbrook (Eds.) *Perspectives on Corpus Linguistics, (*pp. 99–113). Amsterdam: John Benjamins.

Kachru, B.B. (1986). *The Alchemy of English: The spread, function and models of non-native Englishes*. Urbana: University of Illinois Press.

Kachru, B.B. (1992). Models for non-native Englishes. In B.B. Kachru (Ed.) *The other tongue: English across cultures* (pp.48-74). Urbana: University of Illinois Press.

Kropp Dakubu, M. E. (1988). *The languages of Ghana*. London: Keagan Paul International for the International African Institute.

Leech, G. N. (1991). The state of the art in corpus linguistics. In K. Aijmer and B. Altenberg (Eds.) *English Corpus Linguistics: Studies in Honor of Jan Svartvik*, pp. 8-29, London: Longman.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (ed.) *Directions in Corpus Linguistics*, (pp. 105-122). Berlin: Mouton de Gruyter.

Leech, G. (2011). Principles and applications of corpus linguistics: Interview with Geoffrey Leech. In V. Viana, S. Zyngier and G. Barnbrook (Eds.). *Perspectives on corpus linguistics (*pp.155–170). Amsterdam: John Benjamins.

McEnery, T., Xiao, R. & Tonio, Y. (2006). *Corpus-based language studies: An Advanced Resource Book*. London: Routledge.

McEnery, A. & Wilson, A. (2001). *Corpus linguistics: An introduction* (2nd ed). Edinburgh: Edinburg University Press.

McEnery, T. & Hardie, A. (2012). *Corpus linguistics: Methods, theory and practice*. Cambridge: Cambridge University Press.

Mesthrie, R. & Bhatt, R. M. (2008). *World Englishes: The study of new linguistic varieties*. Cambridge: Cambridge University Press.

Meyer, C. F. (2002). *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Morris, L. (1998). The function of English in contemporary Ghanaian society. .*African Diaspora ISPs*.Paper *52*.http: //digitalcollections.sit.edu/African_diaspora_isp/52.

Ngula, R. S. (2010). Variation in the semantics of modal verbs in Ghanaian English. *Drumspeak: International Journal of Research in the Humanities*. 2: 1–27.

Ngula, R. S. (2011). Ghanaian English: Spelling pronunciation in focus. *Language in India*. *11*: 22–36.

Obeng, S.O. (1997). An analysis of the linguistic situation in Ghana. .*African Languages and Cultures 10*(1): 63-81.

Owusu-Ansah, L. K. (1992). So what is new? An initial statement on signaling new information in non-native spoken English. *RevistaCanaria de EstudiosIngleses* (Univeridad de la Laguna) *25*: 83-94.

Owusu-Ansah, L. K. (1994). Modality in Ghanaian and American personal letters. *World Englishes13* (3):341 349.

Owusu-Ansah, L. K. (1997). Nativisation and the maintenance in non-native varieties of English. In M. E. Kropp Dakubu (Ed.) *English in Ghana*, (pp. 23-33). Accra: Black Mask Publishers.

Platt, J., Weber, H. & Ho, M. L. (1984). *The new Englishes*. London: Routledge & Kegan Paul.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English Language*. London: Longman.

Rayson, P. (2009). *Wmatrix: A web-based corpus processing environment*. Computing Department, Lancaster University.

Reppen, R. & Simpson, R. (2002). Corpus linguistics. In N. Schmitt (Ed.) *An introduction to applied linguistics*, (pp. 92–111). London: Arnold.

Reppen, R. (2010). Building a corpus: what are the key considerations? In A. O'Keeffe and M.McCarthy (Eds.) *The Routledge handbook of corpus linguistics*, (pp. 31–37). London: Routledge.

Römer, U. (2006). Where the computer meets language, literature and pedagogy: Corpus analysis in English studies. In A. Gerbig and A. Müller-Wood (Eeds.) *How globalisations affect the teaching of English: Studying culture through texts*, (pp. 81–109). Lewiston: The Edwin Mellen Press.

Römer, U & Schulze, R. (Eds.) (2010). *Patterns, meaningful units and specialized discourses*. Amsterdam: John Benjamins.

Sackey, J. (1997). The English language in Ghana: A historical perspective. In M. E. Kropp Dakubu (Ed.) *English in Ghana*, (pp. 126-139). Accra: GESA.

Sarfo, E. (2011).English language and sustainable development in Ghana. .*Language in India* 2011(11): 460-469.

Schmied, J. (1991). *English in Africa*. London: Longman.

Scott, M. (2013). *WordSmith tools.* (Version 6.0), Oxford: Oxford University Press.

Sey, K. A. (1973). *Ghanaian English*. London: Macmillan.

Silverman, D. (2005). *Doing qualitative research* (2nded). London: Sage Publications.

Simo, B, A. (2000). The uniqueness of Ghanaian English pronunciation in West Africa. *Studies in the Linguistic Sciences 36*(26): 185–198.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

SookBeng, C. O & Chee Keong, Y. (2014). A corpus study of structural types of lexical bundles in MUET reading. 3L*: The Southeast Asian Journal of English Language Studies 20*(2): 127–140.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Wilson, A., Rayson, P. & McEnery, T. (Eds.). (2003). *A Rainbow of corpora: Corpus linguistics and the languages of the world*. Muenchen: LINCOM GmbH.