Asian Journal of Humanities and Social Studies (ISSN: 2321 – 2799)
Volume 02 – Issue 04, August 2014

Corpus Linguistics and Language Development in Ghana

Richmond Sadick Ngula

University of Cape Coast
Cape Coast, Ghana
Email: rngula {at} ucc.edu.gh

ABSTRACT— ***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 incredibly popular. As it is now well established, empirical linguistic investigations that do not employ corpus approaches suffer many setbacks, key among them being that interesting lexical, phraseological, semantic and discourse insights derived via corpus techniques would be missed in a manual analysis. Yet unfortunately, not much work on language studies in Ghana is based on corpora and corpus techniques. This paper suggests that a crucial first step towards the development of languages in Ghana lies in the initiation of large-scale electronic corpus projects. Not only would corpora enrich linguistic descriptions of Ghanaian languages (including Ghanaian English), they also have the potential to provide deeper insights into the socio-cultural and religious values of the Ghanaian people through a discourse analysis that relies on corpus methods. The arguments advanced in this paper also have implications for how language teaching at the various levels in Ghana should proceed.***

Keywords— corpora, corpus linguistics, Ghanaian languages, language development

1. INTRODUCTION

In the 1950s, severe 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 was not to be long for all of 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; may be 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: 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: 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.” McEnery and Hardie (2012) further remark that, now, even theoretical linguists see ways corpora can complement their research, which has previously mainly relied on introspective data.

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 (see e.g., Wilson, Rayson and 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 (and more broadly in Africa). 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 I know, a single machine-readable corpus of any type publicly 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 (see e.g., Crystal 2003: 451). As Greenbaum and Nelson (1996: 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 (e.g., ICE-Great Britain, ICE-New Zealand, ICE-Australia, ICE-India and ICE–Singapore), the Ghanaian component, unfortunately, has not been compiled yet. I observe that currently, 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 working on the spoken part (Huber, personal communication).

The aim of this paper is to argue that the first step towards the development of languages used in Ghana (especially English, the local languages taught in schools including Akan, Ewe, Ga, Dagbani, and other foreign languages like French) lies in the initiation of large-scale corpus projects for these languages.The rest of this paper is thus structured as follows: section 2 discusses 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; section 3 focuses on various applications of corpora in English studies, especially in the UK and the USA, as these represent the two most important centres where the most strides have been made in terms of corpus work on the English language and its impact on the development of English. Based on the insights derived from the experiences with major studies on English in these native contexts, section 4 presents what languages in Ghana need in terms of corpora and their applications, and how these languages can benefit from this fascinating approach to the study of language.

2. CORPUS LINGUISTICS

2.1 Its beginnings

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: 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. Any analysis of language that relies on empirical textual data can loosely be regarded as corpus-based, 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: 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: 99), it led to the “explosion of interest in corpora”

2.2 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: 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: 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".
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".

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 has 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 Sinclair (1991), Louw (1993) and Stubbs (2001) has been accomplished because of the availability of computer corpus tools. In the next sub-section, I consider what a corpus is, and also highlight the key issues for consideration in the construction of a corpus.

2.3 What is a corpus?

To engage in a corpus-based study presupposes that there is a corpus upon which the study will be based. Notable corpus linguistics practitioners 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. I 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: 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: 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: 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: 5).

First, it follows from all four definitions that to create a corpus, one has to collect texts. Texts typically come in the form of spoken or written, 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). By 'naturally occurring' means 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: 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 core 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: 158-9), "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". Sinclair (2005) therefore assumes that a corpus, no matter how large and carefully designed, can never have exactly the same characteristics as the language itself. In this regard, a corpus might never be fully representative as it can only, at best, aim to be maximally representative (Reppen and Simpson 2002; Sinclair 2005). To achieve this kind of maximal representation requires that corpus compilers ensure that adequate samples of the relevant text types (genres) and/or authors are included in the corpus. The underlying principle regarding corpus size is that the larger the corpus, the better as this might increase the instances of whatever feature being investigated in the corpus.

Additionally, the definitions given in Leech (1992) and McEnery *et al.* (2006) above highlight the important point of the corpus being machine-readable. I have already shown in section 2.2 how advances in computer technology have made this possible. Here, I 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 and accessible in electronic format. As Baker (2006: 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. 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 usinga scanner with Optical Character Recognition (OCR) software, processes that are time consuming and error- prone, especially with keyboarding.

Another point worthy of mention relates to how the texts are to be stored (file format). The computer allows users several saving options from which to choose, and these can be seen upon clicking the drop-down 'Save as' menu. It is usually preferable to save a corpus text using the file format *Plain text*. This is because, as Reppen (2010) has noted, most corpus analysis tools at present work best with this format, although the *Rich text* and *XML* formats are other workable 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: 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 I 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 *WordSmith* (Scott 2013), *MonoConc Pro* (2000), *Sketch Engine* (Kilgarriff *et al.* 2004), *AntConc*, (Anthony 2005) and *Wmatrix* (Rayson 2009), various kinds of analyses can be carried out on a corpus. As noted by Römer (2006: 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: 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, among others. Overall, then, searching a corpus allows one to see what patterns are associated with lexical, grammatical and discourse features, "patterns that we might not be able to describe purely on an intuitive basis" (Adolphs 2006: 7).

2.4 Main stages in doing (modern) corpus linguistics

Having discussed some of the important issues of consideration in building and using a corpus, I move on now to shed some light on the main stages a researcher might have to work through in contemporary corpus work. The first major stage is probably to be clearly decided as to what linguistic research questions you wish to investigate. The research questions help to determine whether a suitable corpus already exists to be used, or there would be the need to compile a new one which can effectively answer one's questions.

If a suitable corpus already exists, an additional concern might be accessibility: can it be accessed for a study? Some corpora are available free of charge for linguists and researchers to use; for others, however, a researcher may have to pay a fee and become a subscribed user in order to access them. So for example, as Lee (2010) reports, while the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), a five-million-word corpus of spoken British English, is fully restricted and not accessible to the public, the Corpus of Contemporary American English (COCA), a mega corpus of English (over 385 million words) reflecting contemporary usage in the US, is freely searchable online and thus very accessible to the general public. Other examples include the BNC which has been increasingly accessible as there are a number of web interfaces to it, and the Bank of English which allows only about 15% of it (i.e., fifty-six-million word subset) to be freely searched.The point about accessibility is that if a corpus is built with private funding, it is likely to be restricted while its accessibility is determined by the funders.

There are also established international corpus distribution agencies for already existing corpora. Lee (2010), for instance, makes references to the following agencies: the International Corpus of Modern and Medieval English (ICAME), the Linguistic Data Consortium (LDA), the Oxford Text Archive (OTA) and Open Language Archives Community. These sites could always prove useful, as they are able to help a researcher get a sense of what corpora exist and are suitable for particular research projects.

On the other hand, if no ready-made corpus contains the relevant data required for a study, then a new corpus may have to be built by the researcher(s). McEnery *et al.* (2006: 71) have called this a DIY ('do-it-yourself') corpus. Building a DIY corpus requires involves collecting the relevant authentic texts, and then processing them as electronic data using the design procedures I have highlighted in section 2.3. The compiler would have to decide whether to work with the plain or raw corpus texts or move a step further to annotate the texts. Corpus annotation is a means of explicitly adding some kind of linguistic information to an electronic corpus using tags. In the view of Leech (1997: 2) it is a way of enriching or adding value to the corpus, and it might be useful depending on one's research goal.

The most basic type of corpus annotation is the one that assigns part of speech categories to every word in the corpus. Doing this manually in relatively large corpora can be extremely tedious and time-consuming. Luckily, however, there are automatic taggers that can do the job, achieving a near 100 per cent accuracy and saving a lot of time. A well-established automatic part-of-speech tagger, for instance, is the Constituent Likelihood Automatic Word-tagging System (CLAWS) developed at Lancaster University (Garside *et al.* 1987). CLAWS consistently achieves a 97 per cent or more accuracy (the precise accuracy depends on the type of text). So while there could be tagging errors, they usually are very minimal and a post-edit of the annotated corpus might effectively help deal with such errors. Other types of lexical level tagging are lemmatization and semantic fields. Tagging is also possible at the levels of morphology (e.g., prefixes, suffixes, stems), syntax (e.g., parsing), phonology (e.g., intonation, pitch, loudness) and even discourse/pragmatics (e.g., turn-taking, anaphoric relations, speech acts).

Once a suitable corpus becomes available, the final stage is for the researcher to carry out the linguistic analyses using a preferred corpus toolkit, a few of which I have mentioned in section 2.3. An analyst can carry out all kinds of quantitative and qualitative analyses of the naturally occurring texts represented in the corpus. In the next section of this paper, I consider how practical applications of corpora and corpus methods in English language studies have contributed considerably towards advancing descriptions into the linguistics of English.

3. APPLICATIONS OF CORPORA IN ENGLISH STUDIES

Incredibly, the application of corpus methods to the study of English linguistics in the last 30 years or so has covered almost every sub-field of the discipline (see McEnery et al., 2006). This has led to a remarkable advancement in our appreciation of how English works in many different fields and contexts within society, thereby providing immense quantitative dimensions that previously eluded us. As the review of corpus-based studies in McEnery et al. (2006) show, corpora have been insightful in a range of linguistic sub fields including lexical studies, grammatical studies, discourse analysis, register/genre analysis, language change, translation studies, semantics, pragmatics, sociolinguistics, stylistics, literary studies and language teaching. I simply do not have the space in this paper to show how English studies in each one of these areas have benefitted from corpora. Thus, I wish to exemplify the application of corpus resources to English studies in the areas of lexical studies, grammatical studies and discourse analysis, and then refer readers to Hunston (2002) and McEnery *et al.* (2006), where lively accounts of case studies for most of these areas of linguistics are presented. I focus on these three areas because these are the areas best known to me. The discussion here is intended to show how insightful corpus-based studies have been, and how they justify the need for corpora to be incorporated in the study of languages in Ghana.

3.1 Lexical studies

Lexical studies may conveniently be discussed under two headings: lexicography and lexical semantics, and these perhaps are the greatest beneficiaries of corpora in the study of English. Lexicography concerns dictionary making whereas lexical semantics is related more with the study of lexis and its associated meaning characteristics. I shall address the role of corpora in both in turns.

English dictionaries now rely heavily on evidence from corpora and lexicographers spend an awful lot of their time running concordances of words for dictionary entries. A notable corpus resource for English dictionaries is the COBUILD (the Collins-Birmingham University International Lexical Database) corpus, also known as The Bank of English.This is a database of over 500 million words and still expanding, as it is a monitor corpus. The famous Collins COBUILD dictionaries are based on this corpus. Beyond single word dictionaries, the corpus has also considerably enriched studies into multi-word items such as the *Collins COBUILD phrasal verbs dictionary*, and the *Collins COBUILD idioms dictionary*. Other dictionaries such as the *Cambridge Advanced Learner's Dictionary* and the *Oxford Advanced Learner's Dictionary* are all corpus-based. The fact remains that it is now nearly unheard of for dictionaries published from the 1990s not to claim to be based on corpus data. As Leech (1997: 14) notes, corpus-based dictionaries have several advantages, as the corpus data:

- can be searched quickly and exhaustively,
- can provide useful frequency information,
- can be easily processed to produce updated lists of words,
- can provide authentic typical examples of usage for citation,
- can readily be used by lexicographical teams for updating and verifying other levels of descriptions such as dictionary definitions.

On lexical semantics, the power of corpus-based techniques has led to new knowledge and insights about English words and their associated meanings. This has in turn significantly enriched studies on the phenomenon known as phraseology: the study of word combinations. According to Sinclair (2008: xvi), interest in phraseology is mainly because of "present-day use of text corpora as the principal data-source for language analysis". Let me exemplify this aspect of the study of words using the notion of *semantic prosody*, originally outlined by Louw (1993). In a later work, Louw (2000: 57) defines semantic prosody as "a form of meaning which is established through the proximity of a consistent series of collocates, often characterisable as positive or negative, and whose primary function is the expression of the attitude of its speaker or writer towards some pragmatic situation". This kind of meaning is 'prosody' in the sense that it stretches over and beyond the search word (node) to its collocates. This has given rise to a related concept known as *semantic preference*. Stubbs (2001: 225) explains that "There are always semantic relations between nodes and collocates, and among the collocates themselves". Thus collocational meaning arising from the semantic relations between a node and its collocates is semantic prosody, whereas the collocational meaning arising from the semantic relations of a node is semantic preference.

Comprehensive studies into these two notions have been carried out by scholars, notably Stubbs (1995, 2001). For example relying on corpora, Stubbs examines the words *cause* and *provide* in terms of their semantic prosodies and preferences. He reports that the typical collocates of the verb *cause* are 'damage', 'problems', 'pain', 'disease', 'trouble', 'concern', 'degradation', 'harm', 'pollution', 'suffering' 'anxiety', 'death', 'fear', 'stress'. These examples of 'bad company' collocate suggest that the use of *cause* typically carries a negative affective meaning. Semantically, it prefers to co-occur with nouns indicating a negative evaluation. On the other hand, *provide* as a verb usually collocates with words like 'facility', 'information', 'services', 'aid', 'assistance', 'help', 'support', 'care', 'food', 'money', nourishment', 'protection', 'security'. The items here give *provide* a positive affective meaning. The collocates suggest a semantic preference of 'life-enhancing' for the word *provide*. These are the kinds of lexical insights derived from searching corpora and examining concordances and collocational patterns of words.

3.2 Grammatical studies

Grammatical studies of English that utilise corpus techniques have also been carried out quite substantially. There has been increasing consensus that non-corpus-based grammars often contain intuitive, non-evidenced based and biased descriptions, whereas corpora can help to improve grammar description (McEnery and Xiao 2005). Corpus-based studies have appeared in various forms, either as research papers in journals (e.g., Jones and Coates (1999) on indefinite pronouns such as *someone/somebody*; McEnery and Xiao (2005) on the verb forms*help/help to*, Mair (2002) on *verb complementation*), as research papers in edited volumes (e.g., Aarts and Meyer 1995;Aijmer and Altenberg 1991), or as book-length treatment of specific topics (e.g., Coates (1983) on *modality*; Granger (1983) on *passives*; Mair (1990) on *infinitival complement clauses*; Meyer (1992) on *apposition*; Tottie (1991) on *negation*, andde Han (1989) on *nominal clauses*. But perhaps the greatest milestones of the use of corpora in English grammar lie in 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 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.

The grammar of spoken English has also received a great deal of attention by corpus linguists, and in this area, scholars at the University of Nottingham have done a lot of work with the CANCODE. Scholars like Ronald Cater and Michael McCarthy have carried out a series of studies on English grammar from the perspective of discourse, using the CANCODE as data (e.g., Carter and McCarthy 1995; McCarthy and Carter 2001; McCarthy 1998). Clearly, corpora have, in a great measure, improved grammatical descriptions of English over the last 40 years, and these observations can only further entrench corpus methods in future descriptions.

3.3 Discourse studies

Traditional discourse analysis as a method for the study of language has often relied on the analysis of a few texts by hand and eye alone. Corpora have come to add a significant boost to the study of discourse, especially proving immensely revealing in critical discourse studies. The term *discourse* presents a definitional challenge, as it has been applied to mean quite a number of different things. For instance, it often refers to texts beyond the level of the sentence. Another definition of it derives from a functional perspective, leading to such categorisations as news discourse, academic discourse, workplace discourse, religious discourse, to mention only a few.

But corpus linguists have tended to situate corpus-based discourse studies more on a definition that derives from ideas of the French thinker, Michel Foucault, in his (1972) work *The Archeology of Knowledge*, where he defines discourse as "practices which systematically form the objective of which they speak" (cited in Baker 2006: 4). It is this view by Foucault that has resulted in the plural for discourse: *discourses*. Following Foucault, Burr (1995: 48) defines discourse as:

> a set of meanings, metaphors, representations, images, stories, statements and so on that in some way together produce a particular version of events. … Surrounding any one object, event or person etc., there may be a variety of different discourses, each with a different story to tell about the world, a different way of representing it to the world.

Baker (2006: 4) sums up this view by saying that "So around any given object or concept there are likely to be multiple ways of constructing it, reflecting the fact that humans are diverse creatures; we tend to perceive aspects of the world in different ways depending on a range of factors". It is this understanding of discourse that seems to have appealed enormously to analysts using corpora to study discourse, and this is not surprising given that corpus techniques such as concordances and collocation patterns prove very suitable in identifying different types of representations within various discourses.

Substantial amount of work has been done on English discourse analysis at Lancaster University, notably by Paul Baker (e.g., Baker 2005, 2006, 2014; Baker and McEnery 2005). Baker's (2005) study, for example, examined the representations of homosexuality in British newspaper articles. Baker had built his corpus from two British tabloid newspapers, *The Daily Mail* and *The Mirror* and examined concordances and collocation patterns of the words *gay(s)* and *homosexual(s)* for differences in the way gays and homosexuals were constructed in the two newspapers. Baker found that *The Daily Mail* contained more negative representations, connecting gay men with discourses of crime and violence, promiscuity and political militancy, and also seeing homosexuality as a sexual practice rather than an identity. So, again, in this area of study we see how real social phenomena are perceived in society when we analyse language from a corpus perspective.

4. CORPORA FOR LANGUAGES AND THE STUDY OF LANGUAGE IN GHANA

In the previous sections (2) and (3), I have discussed the major issues of concern in corpus linguistics and ways in which the method has been applied particularly to studies in English. I now pose the question: what does Ghana need for the development of her languages? And my answer to this question is simple: corpora. It would be useful to start addressing the role of corpora on a note of the linguistic situation in Ghana at the present time.

4.1 The linguistic situation in Ghana

The linguistic situation in Ghana is rich and diverse. However, the greater majority of communicative activities are carried out in English, which is the official and somewhat 'national' language of the country. Akan seems to be the most commonly used indigenous language in Ghana. It is estimated that there are over fifty (50) indigenous languages (Kropp Dakubu 1996) dotted around the country and this diversity largely corresponds to the ethnic groupings in Ghana. Most of these languages also have several dialects. For example, the Akan language has dialects such as Ashanti Twi, Akuapem Twi and Fante. Among the Konkomba people of Northern Ghana, I am reliably informed that there are at least as many dialects of the language *likpakpan* as there are communities and towns. The diversity of language in Ghana seems to reflect the kind of diversity encountered on the entire African continent. As Pereltsvaig (2012) has suggested, Africa is the most linguistically diverse part of the world.

The majority of these indigenous languages are not written, they are only spoken, and for those that have writing systems only a few, namely Akan (Fante and Twi), Nzema, Ga, Ga-Adangbe, Ewe, Gonja, Dagbani, Kasem and Dagaare are taught in Ghanaian schools. Although there have been calls to promote the indigenous languages, it seems that right from the country's independence in 1957, English has increasingly displaced them and lessened their domestic significance. In higher educational institutions such as the universities there have been, for a long time, departments of Ghanaian languages that teach, promote and research these languages, yet the impact of their work has not been felt much. Linguistic data on many Ghanaian languages hardly exist and descriptions of languages by researchers are still not as visible as expected. This has left even the more widely used ones (Akan, Ga, Ewe and Dagbani) not so well developed linguistically, and interest in their study and use also seems to be declining. English continues to attract the attention of everyone. Boadi (1971) has noted that English remained a colonial legacy after independence, used in a wide range of activities. It is now an institutionalised variety in the Kachruvian sense, and has virtually become the linguistic mainstay of the country. It is the language of most of our internal and external communication and "its pre-eminent position as the language anybody must know if he is seeking a job in the civil and public sectors of the economy" is not in doubt (Sackey 1997: 136).

The importance of English has led to a growing interest in Ghanaian English (GhE), what is supposed to be recognised as the standard Ghanaian variety (Kachru 1986, 1992) different from other varieties of English. But there is still no consensus about the future of GhE, as British Standard English is currently the established model for language teaching and learning in Ghana. While many people believe that GhE is real and have carried out some studies to prove this (e.g., Dako 2001, 2002; Huber and Dako 1998; Ngula 2010, 2011; Owusu-Ansah 1992, 1994), linguistic descriptions of GhE have not been comprehensive and wide enough to give an instructive picture of its features, leading to its codification.

4.2 What do languages in Ghana need?

Given their wide range of linguistic applications, corpora are precisely what languages in Ghana need at the moment. The existence of professionally constructed corpora for GhE, the local languages (at least those used and taught in schools) and other modern foreign languages taught in Ghana (e.g., French) can help in many ways.

We must begin to build corpora for these languages to enrich linguistic descriptions. The crucial starting point should be to develop capacity on how to build corpora, and manipulate them with corpus search tools. Such corpus literacy can be acquired via training, seminars and workshops. Sadly, we still do not have expertise in corpus linguistics; hence university language departments ought to be more interested in investing in the training of their staff to acquire the skills needed in this area. This will not only boost linguistic descriptions of the languages involved through research, but also provide quantitative insights that would be missed in non-corpus-based studies.

On Ghanaian English, a general corpus representing the variety should massively improve research into its linguistic features, providing a good opportunity to effectively codify the variety. It is commendable that Magnus Huber and Kari Dako have taken up the task of building the Ghana component of ICE. However, being a million words in size, ICE-Ghana would be relatively small. We could undertake much bigger corpus projects for GhE. For instance, a general corpus of a 100 million words for GhE (one that can match the British National Corpus, for instance) should be a huge milestone. Of course, this would require a collaborative team to succeed, but the existence of such a general corpus of GhE can improve research on the variety significantly. It can also ultimately inform language policy better and allow for a more fruitful dialogue on whether or not Ghanaians are willing to adopt a standard local variety of English for teaching and learning.

Specialised corpus databases are also a rich source for lexico-grammatical studies and studies into critical discourses. The former should help improve on the knowledge of register/genre specific linguistic features for English and other languages in the Ghanaian context, which are useful for syllabus design and pedagogy at various levels of language teaching. Studies into critical discourses can broaden our knowledge on how certain social, cultural, political and economic issues are constructed or represented in Ghana and talked about my Ghanaians.

Furthermore, corpora and corpus tools can serve as useful technological resources for the teaching and learning of language in Ghana. Classroom applications of corpus technology since the late eighties and early nineties (Flowerdew 2009) in many advanced societies, for instance, have inspired a learner-centred approach to language teaching, making the teacher-centred approach old-fashioned and less effective. In Ghana, however, the situation is still very much a teacher-centred approach, especially in junior and senior high schools where teachers are often seen as sages and students as receptacles.

Given that there are currently reports on the falling standards of the use of English in Ghana for instance (an obvious place to verify this claim is chief examiners' reports on English), an introduction of corpus-based approaches to the teaching of English can improve the situation immensely. Using this technology makes students more interested in investigating for themselves how language works. According to Johns (1997: 101), the application of corpus techniques to language teaching has made students to act as "language detectives", constantly searching corpora to solve problems posed by their teachers and by themselves. In Ghana, this can work effectively but would, again, first require us to equip both teachers and students with the most basic corpus techniques to enable them to run simple concordances of words and phrases and observe the output for possible patterns.

Finally, funding is a crucial element if such corpus projects are to take off successfully in Ghana, and this requires commitment and support from various stakeholders such as the universities, language research institutions, the Ghana Education Service (GES), the Ministry of Education, and all persons interested in seeing a real boost in the linguistic development and sustainability of languages in Ghana.

5. CONCLUSION

As Römer (2006: 81) has observed, it is for very good reasons that "linguists all over the world draw on corpora in language analysis and description". If that is so, why have corpora still not established a firm footing in a rich linguistic context such as we have in Ghana? In this paper, I have sought to argue that building corpora and incorporating corpus-based techniques to the study of Ghanaian languages are an important first step towards language development in Ghana. I have traced the historical development of corpus linguistics, examined the methodological principles underlying its modern practice, and shown how corpora have particularly revolutionarised English, all in an effort to discuss ways languages in Ghana can take advantage of corpora for the development of her own languages. It is quite clear that the earlier we embrace corpora the better for English, the indigenous languages, and even modern foreign languages such as

French in our Ghanaian context. Corpora can help establish a solid foundation for the linguistic development of these languages; they can enhance language technology and education; they can enhance our self-awareness in terms of how the Ghanaian social-cultural reality is constructed through the use of language; and above all else, corpora can offer new opportunities for the linguistics of Ghanaian languages to thrive in a revolutionary fashion.

6. REFERENCES

Adolphs, S. 2006. *Introducing Electronic Text Analysis*. London: Routledge.

Aijmer, K. and Altenberg, B. (eds.) 1991. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman.

Anthony, L. 2005. 'AntConc: a learner and classroom friendly, multi-platform corpus analysis toolkit', in *Proceedings of IWLeL 2004: An Interactive Workshop on Languagee-Learning.* pp.7–13, Tokyo: Waseda University.

Anthony, L. 2013. 'Developing AntConc for a new generation of corpus linguists', *2013 Corpus Linguistics Conference*. Lancaster: Lancaster University.

Arts, B. and Meyer, C. F. (eds.) 1995.*The Verb in Contemporary English: Theory and Description*. Cambridge: Cambridge University Press.

Baker, P. 2005. *Public Discourses of Gay Men*. London: Routledge.

Baker, P. 2006.*Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P. 2014.*Using Corpora to Analyze Gender*. London: Bloomsbury.

Baker, P. and McEnery, T. 2005. 'A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts'.*Journal of Language and Politics*, 4(2): 197–226.

Barlow, M. 2000.*MonoConc Pro*. Houston: Athelstan.

Biber, D., Johansson, S., Leech, G., Conrad, S. and 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, London: Longman.

Burr, V. (1995).*An Introduction to Social Constructionism*. London: Routledge.

Carter, R. and McCarthy, M. 1995.'Grammar and the spoken language', *Applied Linguistics*. 16(2): 58–141.

Coates, J. 1983.*The Semantics of Modal Auxiliaries*. London: Croom, Helm.

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.

de Han, P. 1989. *Postmodifying Clauses in the English Noun Phrase: A Corpus-based Study*. Amsterdam: Rodopi.

Flowerdew, J. 2009. 'Corpora in language teaching', in M. H. Long and C. J. Doughty (eds.) *The Handbook of Language Teaching*. pp. 327–350, Oxford: Blackwell.

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.
Garside, R., Leech, G. and Sampson, G. 1987.*The Computational Analysis of English*. London: Longman.

Granger, S. 1983.*The Be + Past Participle Construction in Spoken English with Special Emphasis on the Passive*. Amsterdam: North-Holland.

Greenbaum, S. and Nelson, G. 1996. ‘The international corpus of English (ICE) project’, *World Englishes*. 3–15.

Huber, M. and 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.

Hunston, S. 2002.*Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

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.

Johns, T. 1997. ‘Contexts: the background, development and trialling of concordance-based CALL program’, in A. Wichmann*et al.* (eds.), pp. 100–115.

Jones, L. and Coates, J. 1999.‘Someone or somebody? A corpus-based investigation into compound pronouns in contemporary English’, *Roehampton Institute Working Papers in Linguistics*. 1: 154–180.

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*(2nd edn). Urbana: University of Illinois Press, pp.48-74.

Kilgarriff, A., Rychly, P. Smrz, P. and Tugwell, D. 2004. *The Sketch Engine*, Proceedings of Euralex. Lorient, France, July: 105–116.

Kropp Dakubu, M. E. 1996. *Language and Community*. Accra: Ghana Universities Press.

Lee, D. Y. W. 2010. ‘What corpora are available’, in A. O’keeffe and M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. pp. 107–121, London: Routledge.

Leech, G. N. 1991. ‘The state of the art in corpus linguistics’, in K. Aijmer and B. Altenberg (eds.) *English Corpus Linguistics: Studies in Honour 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. 1997. ‘Inroducing corpus annotation’, in R. Garside, G. Leech and A. McEnery (eds.) *Corpus Annotation*. pp. 1–18, London: Longman.

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.

Louw, W. E. 1993. ‘Irony in the text or insincerity in the water? The diagnostic potential of semantic prosodies’, in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*. pp. 157–176, Amsterdam: John Benjamins.

Louw, W. E. 2000. ‘Contextual prosodic theory: bringing semantic prosodies to life’, in C. Heffer and S. Hunston (eds.) *Words in Context: A Tribute to John Sinclair on his Retirement*.pp. 48–94, University of Birmingham.

Mair, C. 1990. *Infinitive Complement Clauses in English: A Study of Grammar in Discourse*. Cambridge: Cambridge University Press.

Mair, C. 2002. ‘Three changing patterns of verb complementation in Late Modern English: a real-time study based on matching text corpora’, *English Language and Linguistics*. 6(1): 105–131.

McCarthy, M. and Carter, R. 2001. ‘Ten criteria for a spoken grammar’, in E. Hinkel and S. Fotos (eds.) *New Perspectives on Grammar Teaching in Second Language Classrooms*. pp. 51–75, Mahwah NJ: Lawrence Erlbaum.

McCarthy, M. 1998. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

McEnery, T. and Hardie, A. 2012.*Corpus Linguistics: Methods, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, T. and Xiao, R. 2005. ‘Help or help to: what do corpora have to say?’ *English Studies*.86(2): 161–187.

McEnery, T., Xiao, R. and Tonio, Y. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

McEnery, A. and Wilson, A. 2001.*Corpus Linguistics: An Introduction* (2nd edn). Edinburgh: Edinburg University Press.

Meyer, C. F. 1992. *Apposition in Contemporary English*. Cambridge: Cambridge University Press.

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

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.

Owusu-Ansah, L.K. 1992. ‘So what is new? An initial statement on signalling 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 Englishes*. 13 (3), 341-349.

Pereltsvaig, A. 2012. *Languages of the World: An Introductio*n. Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G. and 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. 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.

Reppen, R. and Simpson, R. 2002. 'Corpus linguistics', in N. Schmitt (ed.) *An Introduction to Applied Linguistics*. pp. 92–111, London: Arnold.

Römer, U. 2006. 'Where the computer meets language, literature and pedagogy: corpus analysis in English studies', in A. Gerbig and A. Müller-Wood (eds.) *How Globalization Affects the Teaching of English: Studying Culture Through Texts*. pp. 81–109, Lewiston: The Edwin Mellen Press.

Sackey, J. A. 1997. 'The English language in Ghana: a historical perspective', in M. E. Kropp Dakubu (ed.) *English in Ghana*. pp. 126–139, Accra: Black Mask Publishers.

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

Scott, M. 2013. *WordSmith Tools.* (Version 6.0), Oxford: Oxford University press.

Silverman, D. 2005. *Doing Qualitative Research* (2nd edn). London: Sage Publications.

Sinclair, J. 1991.*Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. 2005. 'How to build a corpus', in M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. pp. 79–83, Oxford: Oxbow Books.

Sinclair, J. 2008. 'Preface', in S. Granger and F. Meunier (eds.) *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins.

Stubbs, M. 1995. 'Collocations and semantic profiles: on the cause of the trouble with Quantitative methods',*Function of Language*. 2(1): 1–33.

Stubbs, M. 2001.*Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Tottie, G. 1991.*Negation in English Speech and Writing: A Study in Variation*. San Diago: Academic Press.

Wilson, A., Rayson, P. and McEnery, T. (eds.) 2003. *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*. Muenchen: LINCOM GmbH.