Towards a decade of synergizing corpus linguistics and critical discourse analysis: A meta-analysis

Mark Nartey [a] and Isaac N. Mwinlaaru [b]

[a] Department of English, The Hong Kong Polytechnic University; mark.nartey@connect.poly.hk
[b] Department of English, University of Cape Coast, Ghana; isaac.mwinlaaru@ucc.edu.gh

Abstract

The incorporation of corpus linguistics (CL) methods within critical discourse analysis (CDA) has increasingly gathered momentum in the last decade. This paper surveys studies using this triangulated framework, drawing on a database of 121 studies collected from three citation indexes: Social Sciences Citation Index, Arts & Humanities Citation Index and Scopus. It presents a meta-analysis of these studies focusing on four variables, namely their chronological development, the domains of engagement, the issues that have been topicalized and the area/regional coverage of the studies. In particular, the paper accounts for the factors that have contributed to the popularity of corpus-based CDA in the last decade as an approach to discourse analysis, provides insights into the evolution of this eclectic approach, and anticipates the future of the framework by offering suggestions. The paper concludes that corpus-based CDA presents both discourse analysts and corpus linguists with a robust methodology to tackle research questions bordering on discursive reflections of social issues and to identify new sites of public discourse for systematic analysis.

Keywords: *Corpus linguistics; Critical discourse analysis; ideology and power; meta-analysis*

1 Introduction

Since the advent of discourse studies in the 1960s and its subsequent development in various forms, scholars have continued to explore approaches and tools that are more effective for analyzing discourses. One discourse-oriented research paradigm that has come under pressure to refine its methodology of data collection and analysis is critical discourse analysis (CDA). It has been criticized, for instance, for a biased selection of texts and categories for analysis (Stubbs, 1997; Orpin, 2005). In response to this criticism, several scholars (e.g. Stubbs & Gerbig, 1993; Hardt-Mautner, 1995; Stubbs, 1996; Baker, 2006) proposed a synergy between approaches to discourse analysis (including CDA) with corpus linguistics (CL) methods. Baker et al. (2008) subsequently applied this approach to the study of media representation of refugees in the UK, putting the approach firmly on the research agenda of the applied language sciences. This approach has since been replicated and extended in CDA research. In addition, the emergence of 'big data' stemming from the digitization of traditional media and the proliferation of new media makes it important for CDA studies to use innovative, including CL methods, in managing and analyzing data.

The present study investigates the dynamics involved in the triangulation of CDA and CL in order to inform future research that will use this integrated approach and in order to identify methodological implications for the analysis of public discourses in general. Specifically, the study carries out a meta-analysis on corpus-based CDA studies, focusing on four main variables: the chronological development of the application of CL in CDA, the institutional domains

covered by these studies, the specific issues investigated and the geographical distribution of the studies. Thus, by systematically identifying and profiling studies in the extant literature that have incorporated tools within CL and CDA in the last decade, the present study aims to show the research topics that have been covered, limitations and challenges of the approach, and to point out gaps for future research. One project related to our study is Gabrielatos' (2017) comprehensive bibliography of corpus-based discourse studies as well as discourse studies deploying corpus linguistics techniques (including PhD theses). Our study is, however, a more focused investigation of corpus-oriented CDA studies to identify the nuances in the scholarship in this area and reveal research gaps to drive the field forward.

The rest of the paper is organized as follows. Section 2 briefly discusses the characteristics of CL and CDA while Section 3 discusses those limitations of CDA and CL that necessitate a triangulation of the two frameworks. These discussions are meant to create a conceptual context for the meta-analysis. Section 4 describes the data source and methods of the study. Section 5 examines the findings of the study in relation to the objectives outlined above and Section 6 concludes the study and draws implications for further research.

2 Conceptual background

2.1 Corpus linguistics

CL is an empirical approach to the study and description of 'real life' language use which relies heavily on corpora.[1] It investigates various linguistic phenomena from the corpus, identifying "probabilities, trends, patterns, co-occurrences of elements, features or groupings of features" (Teubert & Krishnamurthy, 2007: 6) from which generalizations about language can be made. One issue that has generated considerable debate in CL is the corpus-based vs. corpus-driven points of view first argued by Tognini-Bonelli (2001). The former refers to "a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions" (p 65) "in order to validate, refute or refine" them (McEnery & Hardie, 2012: 6). The latter "claims instead that the corpus itself should be the sole source of our hypothesis about language" (McEnery & Hardie, 2012: 6), i.e. the corpus 'drives' the research in the sense that the analyst observes what is salient to explore in the corpus and theory is derived from the corpus.

Some scholars reject "the binary distinction between corpus-based and corpus-driven linguistics" and opine that "all CL can justly be described as corpus-based" (McEnery & Hardie, 2012: 6, 147-153). We agree with this position and assert that the two approaches are complementary and not mutually exclusive, and can, thus, be useful in diverse ways. Biber et al. (1998: 4) identify four characteristics of corpus-based research, namely:

1. It is empirical, analyzing the actual patterns of use in natural texts;
2. It utilizes a large and principled collection of natural texts, known as a corpus, as the basis for analysis;
3. It makes extensive use of computers for analysis, using both automatic and interactive techniques;

[1] The presentation on corpus linguistics here is a very brief one. The following works offer a comprehensive introduction to corpus linguistics as a research methodology in language studies, including its historical development, issues of its theory, methods and procedures, and practice: McEnery & Wilson (2001), McEnery, Xiao & Tono (2006), and McEnery & Hardie (2012).

4. It depends on both quantitative and qualitative analytical techniques.

The last point above – which borders on triangulation – is particularly significant because it enables quantitative analysis of frequent linguistic and repetitive patterns and also includes "qualitative, functional interpretations of quantitative patterns" (Biber et al., 1998: 5). CL is, however, generally quantitative and focuses attention on the local context of situation such as in concordance analysis, which is "a collection of the occurrences of a word-form, each in its own textual environment" (Sinclair, 1991: 32). Methods in CL have been applied to the analysis of discourse-level phenomena such as "characteristics associated with the use of a language feature", "realizations of a particular function", "characterizing a variety of language" and "mapping the occurrences of a feature through entire texts" (Conrad, 2002: 75). And in the last decade, its application in CDA research, especially to media and political discourse, has gained momentum (Baker et al., 2008; Mautner, 2015; Lee, 2016; see Section 5 below for details).

2.2 Critical discourse analysis

Emerging from critical linguistics, CDA views language as embedded in its sociolinguistic context and, therefore, examines how grammatical or lexical choices are used to express social processes and social phenomena (Fairclough, 2010).[2] The concern is not to analyze language as a static linguistic entity but to investigate how text reflects social practice. CDA "is problem- or issue-oriented, rather than paradigm-oriented"; and it highlights "the *underlying ideologies* (original emphasis) that play a role in the reproduction of or resistance against dominance or inequality" (van Dijk, 1995: 17-18). Thus, by examining various dimensions of discourse and semiotic resources, "CDA deals with the discursively enacted or legitimated structures and strategies of dominance and resistance in social relationships" (van Dijk, 1995: 18). It is informed by three main stages of analysis: description of text, interpretation of the relationship between text and interaction, and explanation of the relationship between interaction and social context (Fairclough, 2001: 21-22).

CDA perceives language as 'social practice' (Fairclough, 2001: 21) and considers the 'context' within which language is instantiated to be crucial. By considering 'discourse as social practice', CDA relates meaning in text to social and ideological phenomena and aims to make more visible the opaque aspects of discourse as social practice (Fairclough, 2010). Although CDA has made a claim to investigating the issues of power (abuse) in discourse and has focused particularly on demystifying the power asymmetry in society in different contexts, it is also essentially concerned with exploring the co-constitutive relationship between discursive practices and their institutional, cultural, social and political contexts. In sum, CDA is "fundamentally interested in analyzing opaque as well as transparent structural relationships of dominance, discrimination, power and control as manifested in language" (Wodak & Meyer, 2009: 10) and can be seen as a means for deconstructing the implicit ideologies embedded in text and "identifying and defining social, economic and historical power relations between dominant and subordinate groups" (Henry & Tator, 2002: 72). Since its inception in the late 1980s, CDA has been fruitfully applied to language use in different domains, including media, law, politics,

[2] CDA encapsulates different theoretical approaches. Notable among them are a combination of poststructuralist theory and systemic functional theory (led by Norman Fairclough); a socio-cognitive approach (led by Teun van Dijk) and the discourse -historical approach (led by Ruth Wodak). For want of space, we only synthesize the general tenets of these approaches to characterize CDA.

education, etc., often exposing the non-neutrality and taken-for-granted assumptions underlying language use. Also, these studies have revealed that discourse shapes/constitutes and is shaped/constituted by various social structures, competing ideologies and power relations.

2.3 The need for a synergy: Limitations of CDA and CL

The synergy between CL and CDA responds to limitations of both fields, even though it appears that the limitations of CDA are often highlighted the more in the literature. The criticisms against CDA have focused on its methodological weakness, mainly due to its qualitative approach to linguistic analysis. Flowerdew (1997), for instance, indicates that corpora in CDA studies usually serve as a repository of examples, as opposed to an analysis that adheres to the "principles of total accountability" (Leech, 1992: 112; see also Fowler 1996 on critical linguistics). Furthermore, the data size in CDA research has been criticized for usually being small (e.g. 25, 000 words; Clark, 2007). Such small data may lack some of the features under consideration, "or contain them in too small frequencies for results to be reliable, particularly when issues of statistical significance are not addressed" (Baker et al., 2008: 275). In addition, others like Widdowson (2000) have criticized CDA for its lack of academic and analytical rigor, arguing that the data analysis is highly informed by the analyst's subjective preconceptions and assumptions (see also Orpin, 2005: 38).

Finally, Stubbs' (1997) concern is with what he considers a lack of representativeness of the often small and arbitrarily selected texts. Such 'cherry-picking' as referred to by other critics, in extreme cases, may result in a biased database (Magalhaes, 2007) and, thus, the results proceeding from such data may be lacking in generalizability, replicability and reliability. Stubbs (1997) adds that the linguistic features in CDA research are seldom compared with norms in the actual language (see also Orpin, 2005: 38). This means that CDA research may miss the opportunity to reflect the "finer shades of meaning and nuances in representation" (Duguid, 2010: 215). Following these criticisms, Stubbs (1997) suggests the bolstering of CDA by large corpora or by corpus linguistics methodology, "specifically through using random sampling, analyzing a large collection of text and comparing the textual features under study with language norms captured in a corpus in order to make reliable generalizations about typical language use" (Cheng, 2013: 1-2). In the incorporation of CL into CDA, however, Mautner (2015: 157) sounds a note of caution that there must be "checks and balances" since the idea is for CL to contribute to CDA rather doing CDA in its own right.

The main weakness for which CL is often criticized is that it neglects the socio-cultural context of discourse owing to the large size of data and since the texts that constitute a corpus are invariably decontextualized examples of language use (Baker, 2006: 25). A pure corpus analysis is, therefore, often said to lack explanatory and interpretive power since in-depth exploration of context is usually not feasible with a large corpus. As Mautner (2007: 65) explains, "what large-scale data are not well suited for … is making direct text-by-text links between the linguistic evidence and the contextual framework it is embedded in". This point in reinforced in Bednarek (2009: 20), who asserts that "beyond its consideration of syntagms, large scale corpus linguistics usually has less to say about context, and the unfolding of meaning in texts (e.g., intratexual patterning)". Moreover, the focus of CL on frequent usages and repeated patterns may be done at the risk of losing out on the insights of outstanding singular texts that might have an impact on phylogenesis and ontogenesis and whose relevance transcend 'ordinary' singular texts that achieve their impact through repetition (Bednarek, 2009: 21). The over-sensitivity to frequency can also result in semiotic impoverishment (Mautner, 2009) whereby the results miss out on

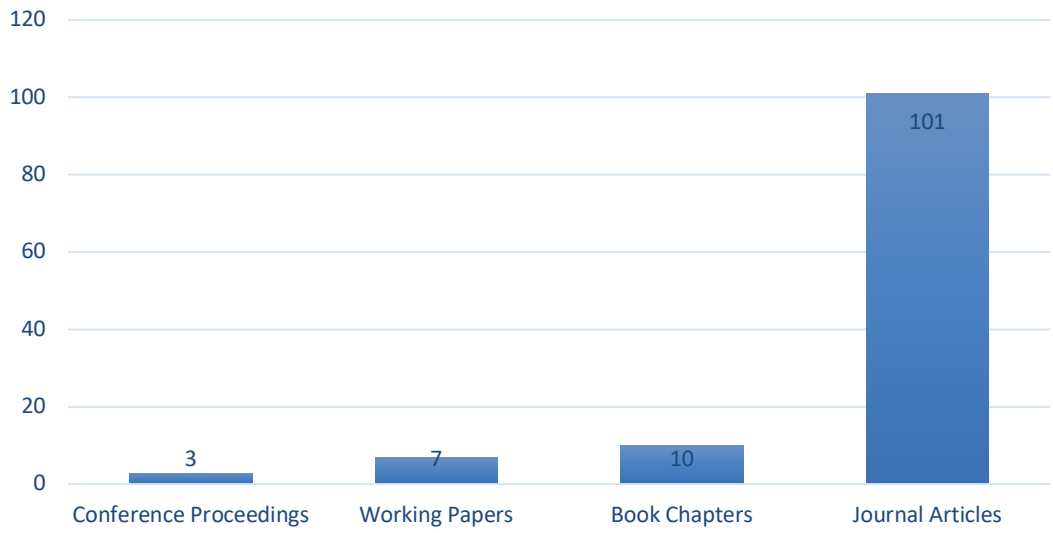
insights that absences or outliers in a corpus (i.e. what could have been said but was not) might reveal.

It is these limitations that mainly provided a motivation for synergizing CDA and CL since the synergy helps to cancel out or make up for the constraints in the respective approaches, thereby making their combination a stronger methodological framework.

4 The database of studies: sources and selection criteria

Following the conceptual review above, we proceed to survey how this triangulated methodology involving CDA and CL has been implemented in the literature. The survey is based on a database of 121 studies, published between 1995 and 2016. The compilation of the database follows the methodology developed by Norris & Ortega (2000, 2006) for research synthesis and meta-analysis. The guiding principle of this methodology is to retrieve the relevant literature for the survey in a replicable and systematic manner as much as possible. The studies were purposively sampled from two main online databases. The first is Thomson Reuters' flagship citation indexes, the Arts & Humanities Citation Index (A&HCI) and the Social Science Citation Index (SSCI) and the second is Scopus (Elsevier). These citation indexes allowed access to prestigious sources in the field and at the same time ensured the retrieval of a wide range of studies as much as possible. These consist of journal articles, book chapters, conference proceedings and working papers. Figure 1 presents the frequency distribution of the studies across these research genres, with the research article being the avenue for most of the published work in the area (101, 83.5%). Although many relevant studies may have been published in sources not included in the databases consulted, we believe that the studies retrieved are representative of the highly relevant studies in this area since the citation indexes we consulted include many of the internationally recognised CDA and CL publication sources. One limitation, however, is that relevant studies published in languages other than English are not captured in the study and readers should interpret the findings considering this limitation. Only one study was found for 2017 (Kim, 2017) and was excluded from the dataset for the sake of representativeness across the years since the data collection was completed in March 2017.

Fig. 1 Frequency distribution of studies across publication mode
(total number of studies = 121)



The criteria for including or excluding a study from the dataset are outlined in Table 1 along four dimensions: accessibility of studies, the nature of the data used in the study, analytical techniques deployed and the theoretical orientation of the study. The general guiding principle was to include studies that are publicly accessible, which is defined here as published materials that are available in online databases. A comment also needs to be made on the 'theoretical orientation' dimension. As indicated in Table 1, the studies included are limited to those with a 'critical' focus, which is operationalized in this study as research that highlights how language use reflects the inequalities, ideologies, and power dynamics embedded in social and political contexts. Admittedly, the boundary between what is CDA and what is DA is sometimes fuzzy and the two orientations of discourse analysis are better conceived of as a continuum with 'critical' analysis at one end of the continuum and purely DA analysis on the other end. Our study is essentially limited to studies in which authors explicitly declared to use some CL methods/tools for some form of critical discourse analysis or self-categorize as 'critical' in their titles. Consequently, studies that lie along the fuzzy boundary between CDA and DA (e.g. Baker, 2006; Partington, 2010; Partington et al. 2013) are not included in our database. Our findings should therefore be interpreted against this limitation. The search terms used included "critical", "discourse", "corpus-based", "corpus-driven", "corpus-assisted", "corpus-informed", "corpus-supported", "corpus analysis", "corpus approach", "corpus study", "corpus/corpora", and "critical discourse analysis". And studies that did not align with our inclusion criteria were excluded after reading the abstract, examining the key words, and, when in doubt, reading the full text. We avoided using more specific keywords such as "power and domination", "inequality", "ideology and "immigration and racism" in our search in order to prevent skewing the data in favour of particular topics. We assumed that studies on these issues would also include terms like "critical", "discourse" or "critical discourse analysis" if CDA is central to their approach.

Table 1 Inclusion and exclusion criteria for compiling corpus-based CDA studies

Dimension	Include	Exclude
Accessibility of studies	Published with online access to scholarly audience: journal articles, book chapters, papers in conference proceedings, working papers	Unpublished conference presentations and local publications not available online
Data	Corpus-based: either self-compiled large data sets or use of existing corpora or a combination of both	CDA studies using small samples of texts (’classical CDA’)
Analysis	Use of corpus software e.g. *WordSmith*, *ConGram*, *Wmatrix*, *UAM Tool*, etc.	Manual analysis of large data sets or analysis of large data sets aided by spreadsheets
Theoretical orientation	Must be an empirical study (at least one case study) with an explicit ‘critical’ focus (e.g. ideology, power, gender and masculinity, etc.)	Corpus-based analysis of ideologically-oriented or gendered discourses without a ‘critical’ focus (e.g. genre analysis of media or corporate discourses) and purely theoretical studies.

As mentioned earlier, the analysis of our data focuses on the temporal and geographical spread of the studies, institutional domains and specific issues and topics investigated by corpus-based CDA research. The modes of dissemination of the studies are also considered. The analysis was aided by an Excel spread sheet. Except for the publication year and mode of research variables, it was difficult classifying a few studies into single categories since they examine multiple values of the same variable. Such studies are counted two or more times, as the case may be. This is indicated in the relevant sections in the discussion below.

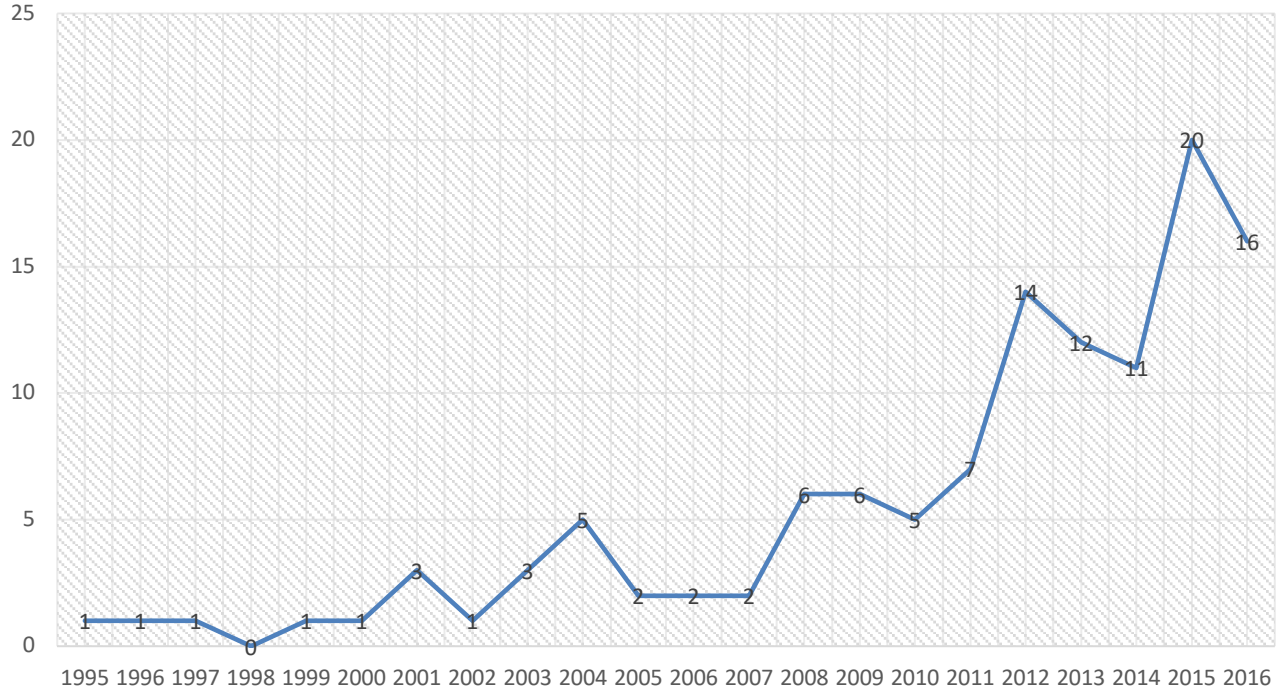
5 Meta-analysis of studies

This section proceeds to present the findings of the meta-analysis. First, we discuss the chronological progression of the studies together with the media through which researchers have disseminated the findings of their studies. Next, we examine the domain or discourse sites which have been the focus of attention and the issues which have been addressed. Finally, the regional distribution of the studies is discussed. The qualitative analysis and discussion is complemented by frequency counts and percentage distribution.

5.1 Chronological development of corpus-based CDA studies

Figure 1 presents the frequency distribution of publications from 1995-2016. It can be observed that although corpus-based CDA methodology has been employed since the mid-1990s (see Stubbs & Gerbig, 1993; Hardt-Mautner, 1995; Stubbs, 1996), it was not until 2008 that the approach began to gain momentum, reflecting Baker et al.'s (2008: 295-297) influential emphasis on the usefulness of the methodological synergy. Before 2008 (from 1995 to 2007), only 23 (19.2%) of the 121 studies in our dataset were recorded, with many years recording only one or two studies. The number of studies began to rise from 2008 and reached its peak in 2015, accounting for 81 (66.9%) of the 121 studies in the database. There was a drop in 2016, albeit, comparatively, a high number of 16 studies was still recorded. These findings show that whereas Hardt-Mautner's (1995) study is seminal with respect to corpus-based CDA and is, therefore, a keystone paper, it was Baker et al. (2008) that popularized the methodology which has now become commonplace.

Fig. 1 Number of publications from 1995 - 2016 (total number = 121)



The fact that there is almost no study before Hardt-Mautner (1995) points to the historical relevance of her work.[3] The non-existence of corpus-based CDA before the 1990s is not alarming because both CDA and CL started gaining grounds around the same period (early

[3] We thank one anonymous reviewer for drawing our attention to an earlier corpus-based CDA study, Stubbs and Gerbig (1993). We have made references to this study, but it is not included in the quantitative counts in the present study.

1990s). During this period, researchers in each field of research, as it were, were pre-occupied with establishing the tenets and principles of the respective frameworks, making a synthesis between them an unlikely occurrence. Indeed, Hardt-Mautner (1995) attributes the decision to incorporate corpus tools into CDA to a methodological need required by the nature of her data:

> Originally, the project was to draw solely on the theoretical foundations and descriptive resources of the framework known as critical discourse analysis, or CDA for short. However, the mainly qualitative methodology used in CDA proved ill-suited to handling the sizeable corpus that formed the basis of the study. It was this mismatch between the chosen framework and the nature of the data that led to the development of an alternative analytical procedure, combining the use of concordance programs with CDA's traditional qualitative analysis (Hardt-Mautner, 1995: 1).

Apart from Hardt-Mautner (1995), other earlier studies that set the stage for the integration of CL and CDA include Morrison and Love (1996), Flowerdew (1997), Sotillo and Starace-Nastasi (1999), Simon-Vandenbergen (2000) and Downs (2002). These studies mainly focused on corpus methods/tools such as bare analysis of frequencies and key-word-in-context (KWIC) concordances. Rarely were collocations examined in this first period and when they were, this was done qualitatively via sorted concordance lines and not through statistical/frequency calculation (e.g. Simon-Vandenbergen, 2000). Also, these earlier studies were largely conducted using relatively small corpora such as 123 letters to the editor (Sotillo & Starace-Nastasi, 1999), 167 published letters (Morrison & Love, 1996), 200 instances of *I think* in radio political interviews and casual conversations and 40 speeches (Van de Mieroop, 2007). Some of the corpus analysis software employed in these studies include *Micro-Concord*, *WordSmith* and *Longman Mini Concordancer*. Some studies (e.g. Sotillo & Starace-Nastasi, 1999; Simon-Vandenbergen, 2000; Cotterill, 2001) were silent on the software package used while others like Flowerdew (1997: 465) used constructions like "computer-generated word frequency lists indicate …" in lieu of explicitly stating the name of the corpus analysis software. The incorporation of CL in CDA in this early period was, therefore, rudimentary in terms of the use of small corpora by today's standards, frequent use of wordlists and, for some studies, less explicit use of corpus techniques and software.

The publication of Baker et al. (2008) marked a turn in the development of studies synthesizing CL and CDA, what we would like to call the 'canonizing' phase. In this study and in another by two of the authors in the 2008 study (Baker & Gabrielatos, 2008), Baker and his colleagues at Lancaster argue for the usefulness of synergizing corpus approaches and CDA, drawing on a large corpus of 140 million words, a database of the discourse of refugees, asylum seekers, immigrants and migrants (RASIM) in the UK press over a ten-year period (1996-2005). With the publication of Baker et al. (2008), there was a boost in CDA's engagement with CL from the initial studies of the 1990s and early 2000s. Baker et al. (2008: 295) also advanced a nine-step procedure of analysis as the possible stages in corpus-based CDA. In addition to using corpus tools and techniques from the earlier studies which mainly focused on wordlists and concordance analyses, Baker et al. (2008) used other methods, including the analysis of colligation and collocational profiles (and c-collocates), keywords/key cluster analysis, semantic preference, semantic prosody and/or discourse prosody, normed vs. raw frequencies, the use of dispersion plot as well as lemmatization and standardization. In terms of the analysis of collocates, Baker and his colleagues foregrounded the notion of statistical collocation, including issues of statistical significance of collocates, collocation span, frequency thresholds, and strong non-adjacent collocates. They also highlighted a distinction between corpus-based and corpus-

driven CDA research. From a history-of-science point of view, it is difficult to say why an uptake of the idea of a corpus-based CDA in the mid-nineties (cf. Stubbs & Gerbig, 1993; Hardt-Mautner, 1995) was less widespread, and why more than ten years later things really got off the ground. We can probably attribute this situation to the influential position of Lancaster University in the field of corpus linguistics and the fact that the authors of Baker et al. (2008) strategically included leading and influential scholars in both CDA and CL.[4]

Following 2008, the corpus-based CDA studies conducted are akin to Baker et al. (2008). Describing their methodologies, Taylor (2009: 6), for instance, states that her work is "a para-replication of work carried out at Lancaster University" Freake et al. (2011: 27) write that "We followed Baker et al.'s (2008) 'eclectic approach'", and Subtirelu (2013: 42) asserts that the approach he adopts is "a combination of critical discourse analysis (CDA) and corpus linguistics (CL), an approach described extensively in Baker et al. (2008)". Studies such as KhosraviNik (2008), Mulderrig (2008), O' Halloran (2008), Prentice & Hardie (2009), Don et al. (2010) and Augoustinos et al. (2010) all emphasize the analysis of keywords and key clusters, following the tradition of Baker et al. (2008), but with an addition of how automated semantic tagging or the analysis of key semantic categories (e.g. Prentice, 2010) can enrich corpus-based CDA. Interestingly, several of these studies were conducted in the UK and papers like Prentice (2010), KhosraviNik (2008), and Baker & Gabrielatos (2008) were directly from the Lancaster School. These findings reveal the impact of Baker et al.'s (2008) paper on the studies that followed theirs. Don et al.'s (2010) study also stands out as a notable study to have come from Asia (Malaysia) quite early in the development of corpus-based CDA studies. By this second phase, the construction of large (and often specialized) corpora was coming up strongly, with many studies using a corpus size of at least a million tokens.

The period between 2011 and 2016 represents the third phase in the development of corpus-based CDA studies, the broadening phase. Not surprisingly, this period witnessed a massive surge in the number of studies published, a phenomenon which could be attributed to the view that by the third year of Baker et al.'s (2008) publication, corpus-based CDA had become popular with many discourse analysts. Here too, many of the studies were still based directly on Baker et al. (2008) with little or no modification (e.g. Salama, 2011; Mulderrig, 2012; Baker, 2012; Sealey, 2012; Jeffries & Walker, 2012; Boyd, 2013; Aull & Brown, 2013; Brindle, 2015; Hardaker & McGlashan, 2015). However, other studies departed from Baker et al. (2008) by either adopting a bilingual and/or comparative corpus (e.g. Freake et al., 2011; Jaworska & Krishnamurthy, 2012) or examining key function/grammatical words (e.g. Pearce, 2014), unlike Baker et al. (2008) and several other studies that focused exclusively on key lexical/content words with the explanation that such words usually help in the identification of discourses.

An important development in this broadening phase is the extension of the framework to include other analytical models and/or theories. That is, in addition to CL and CDA, other studies directly or indirectly incorporate other frameworks into their analysis, including systemic

[4] We acknowledge one anonymous reviewer for suggesting a reflection along the lines of a history-of-science point of view. Although number of citations was not part of the variables we investigated, we note the frequencies of citations to seminal publications by studies in our database as follows: Baker et al. (2008) cited by 45, 37% (GS = 822); Baker (2006) cited by 29, 24% (GS = 1, 310); Hardt-Mautner 1995 cited by 7, 6% (GS =186); Stubbs & Gerbig (1993) cited by 2, 2% (GS = 36). Corresponding Google Scholar (GS) citations are in parentheses. If we limit the calculation of the counts of citations to Baker et al. (2008) to the period after its publication, we record 50% citations.

functional theory (El-Falaky, 2015; Lee, 2016),[5] translation theory (Murphy, 2013; Pan & Zhang, 2016), social theory (Mulderrig, 2008; 2011), critical stylistics (Jeffries & Evans, 2013), contrastive analysis (Schroter & Storjohann, 2015), sociolinguistics (Chiluwa, 2012), critical literacy pedagogy (Abid & Manan, 2015), multimodality (Edwards & Milani, 2014), topic modeling (Tornberg & Tornberg, 2016), pragmatics (Triebl, 2015), genre theory (Skalicky, 2013), theory of governmentality (MacDonald & Hunter, 2013) and theory of argumentation (Lippi & Torroni, 2016). Of all these other theories, systemic functional linguistics tends to be dominant. Given the centrality of SFL to the development of CDA, it is rather surprising that the explicit application of SFL descriptive categories (e.g. mood, theme, and transitivity analyses) to corpus-based CDA is only found in this third phase. This brings us to the techniques and approaches of CDA often used.

The evidence from the database suggests that, generally, studies do not explicitly mention the CDA theory or traditional methods of CDA which inform their study. It seems, therefore, that the main aim of many studies in combining corpus methods and CDA is to arrive at more accurate, insightful, objective, and generalizable findings rather than contributing to a specific discourse-oriented theory. In the few studies where the CDA techniques or theory are demonstrated, the discourse-historical approach is prominent, and it is used in many of the studies by Paul Baker and his colleagues and in other works such as Almeida (2011) and Salama (2011). In addition to the discourse-historical approach, other CDA t echniques found in our database include critical metaphor analysis (Davidson, 2013), categorization analysis (Caldas-Coulthard, 2010), social actor theory (Subtirelu, 2013), the dialectical approach to CDA (Pan & Zhang, 2016), and the notion of ideological square (Prentice et al., 2009).

Further, this third phase brought to the fore several corpus analysis software that can be utilized for corpus-based CDA, with *WordSmith* and *AntConc* being the common ones. Other tools that emerged include *Sketch Engine*, *ConcGram*, *Wmatrix*, *Graphcoll*, *UAM Corpus Tool*, *Standford NLP Tools*, *MonoConc Pro* and *WordPilot*. In this broadening phase also, the extent to which corpus methods have been integrated into CDA research has been varied. Studies from the Lancaster School and others like Aull & Brown (2013), Potts et al. (2015) and Bednarek & Caple (2014) tend to be more balanced, incorporating corpus methods at all stages of the research. Conversely, some studies (e.g. Hou, 2016, Leung, 2016) minimally used corpus methods, especially in the analysis and discussion, as a way of boosting the empirical credence of analyses. The measure of integration achieved in corpus-based CDA, thus, remains a challenge of the framework (cf. Baker et al., 2008: 295-297). Like the first two phases, concordance analysis still remains the common corpus method in the third phase, giving an indication of its centrality to corpus-based CDA. In the third phase, however, concordance analysis is usually combined with the analysis of collocates (including semantic preference and semantic/discourse prosody) and n-gram and concgram analyses. Based on this observation, it can be said that the range of corpus techniques and methods, as should be expected, is expanding. That said, the maintenance of concordance analysis as the common corpus technique in all the three periods is not altogether surprising since this can be attributed to its similarity with traditional methods in CDA in terms of allowing analysts to do more contextual analysis. Further, the statistical notion of collocation observed in the second phase continues to feature prominently in this third phase.

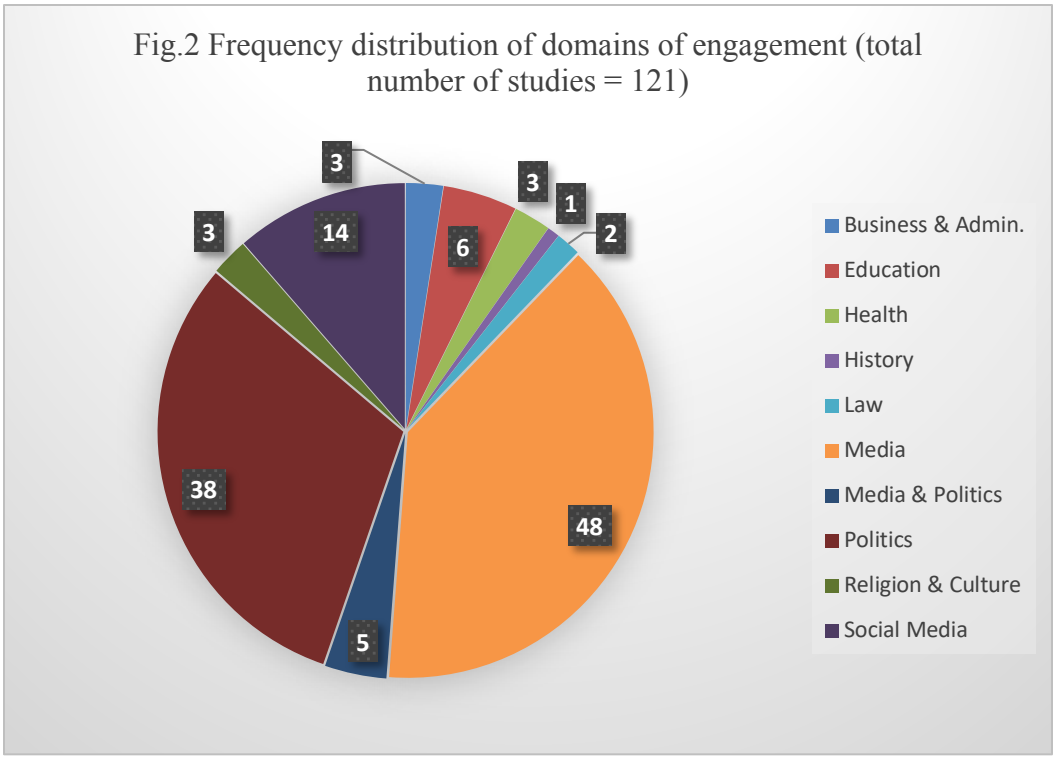
[5] One anonymous reviewer questioned our mention of SFL here as an additional approach since SFL has been so central to CDA. We note that the reference to SFL here is with regard to studies that explicitly set out to systematically combine CDA agenda with CL methods and SFL descriptive categories and that state this eclectic approach as an objective of the research.

In addition to Baker et al.'s (2008) influential study, other contextual factors that may have contributed to the prevalence of corpus-based CDA post-2008 include the emergence of 'big data', the digitization of traditional modes of discourse, and the proliferation of new media. Moreover, the notion of interdisciplinary discourse analysis which characterized the 'critical and discursive' turn in discourse analysis in the early 2000s (cf. Bhatia et al., 2008) gives an indication of why corpus-based CDA began to gain momentum around the same period since the idea of interdisciplinarity makes CDA very open to other frameworks. This paper maintains that the development of a theoretical or an analytical model is invariably linked to the need to better explain and understand as well as challenge and extend existing knowledge within the limits of critical bounding assumptions. This position together with the contextual factors identified above lends credibility to the view that in the 1990s, corpus-based CDA could not have been dominant since both CDA and CL, at the time, were now gaining prevalence as analytical models in language studies.

5.2 Domains of engagement

This section examines the domains or discourse sites within which the studies have been conducted. The domains of engagement of corpus-based CDA studies are extremely important. First, they indicate the preoccupation of (critical) discourse analysts and corpus linguists in the application of the methodology and where research energy has been concentrated. Both CL and CDA underscore descriptions that not only contribute to theoretical and intellectual findings in language but, more importantly, contribute to practical applications in critical contexts of language use. Additionally, the domains of engagement give an indication of the elasticity of the framework in terms of its applicability to different contexts of language use. Before the discussion, however, it is worth stating that the domains of engagement were identified through a content analysis approach, i.e. based on a close reading of the abstracts, an examination of the key words, and sometimes the reading of the entire text (see Hsieh & Shannon (2005) on content analysis). Each of the authors separately identified the domains after which the results were compared and collectively discussed to arrive at the final decision.

Ten domains are identified in the database, namely, business and administration, education, healthcare, history, law (defined broadly as the legal and justice system), media, media and politics, politics, religion and culture, and social media. It can be stated that the methodology has been applied to a wide range of texts in various contexts. These discourse sites impact directly on human activity and relationship, suggesting the practical value of the methodology. As Figure 2 shows, the top three domains of corpus-based CDA research are media (48, 39.7%), politics (38, 31. 4%) and social media (14, 11.6%). Three studies (Boyd, 2013; McEnery et al., 2015; Pan & Zhang, 2016) are counted twice in Figure 2 as each of them can be classified into two categories. Given the orientation towards ideological critique that characterized the inception of CDA, the emergence of these three domains as the dominant areas of engagement for corpus-based CDA research is unsurprising. One observation that can be gleaned from this empirical information is that while, on the one hand, it seems that diverse domains have received scholarly attention (at least ten areas), on the other hand, the focus of attention has been rather narrow – that is, politics and (social) media. We argue that the picture presented here in terms of the distribution of studies across domains is not limited to corpus-based CDA per se but to CDA research in general.



Further to the point that the top three domains of engagement are relevant to CDA itself, the focus on politics and media have indeed being a wider trend in applied linguistics. The pioneering work of Hardt-Mautner (1995) and Baker et al.'s (2008) influential study both focused on media discourse. Hardt-Mautner (1995) examined European discourse in the UK press whilst Baker et al. (2008) analyzed RASIM in the UK press, both studies straddling media and politics. Subsequently, many studies (82.6 % of 121) have taken a similar line. Further, the accessibility of data may yet be another reason for the much focus on the three domains. Generally, the issues discussed in traditional media, on social media or in politics are considered issues/discourses of public domain. The same cannot, however, be said of domains like health, law (i.e. the legal and justice system), business and administration and even education, where issues of privacy, confidentiality, anonymity and sensitivity of information make it challenging to access data for CDA research.[6] It takes more effort or may be completely impossible for a researcher to negotiate one's entry into these research sites. Another observation from the survey on the domains of engagement is that with the exception of Salicky's (2013) study on *Amazon.com*, Chiluwa's (2012) work on Biafra online tweets and Thelwall's (2008) paper on swear words on *MySpace* pages, all the other studies on social media took place between 2014 and 2016 (e.g. Potts et al., 2014; Subtirelu, 2015; Tornberg & Tornberg, 2016). Thus, we see the impact of social media, in particular, and new media technology, in general, in the last five years on corpus-based research.

[6] With our reference to the restrictions in these domains, we are here concerned about those private contexts that will be of most interest to the CDA agenda. These include doctor-patient interaction, police interrogation, courtroom trial and teacher-student interaction.

It can be deduced from the surveyed data that in the application of corpus-based CDA to various aspects of language instantiation, the framework has been in accord with the core ideals of CDA. However, it should prove useful for future studies to venture into other areas, including health, education, environment (especially, climate change/global warming), business and economy; and religion and culture. Admittedly, engaging with these domains will be challenging but it will be equally rewarding, owing to the usefulness the findings identified in these critical contexts will have on society. For example, in the last decade, terrorism and its nexus with religion and culture have gained global attention. The concentration of corpus-based CDA studies in this vein could help in illuminating the issues, creating critical awareness, setting the agenda and shaping public discourse. And from a social standpoint, the domain of sports with accompanying discourses such as locker room talk, pressers, commentary, etc. are areas where the synergy can make a contribution.

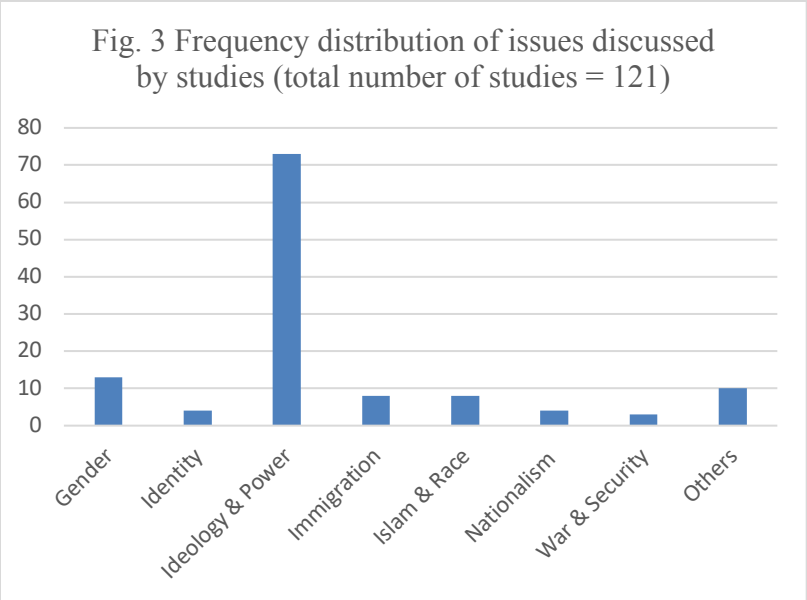
5.3 Topics/issues

One important consideration in any field of research is the issues that receive attention by researchers, as this together with the domains of engagement construe the epistemological base of the field. In CDA, the issue of which topics merit researchers' attention is even more important since CDA is, generally, "a socio-politically conscious and oppositional way of investigating language, discourse and communication" (Van Dijk, 1995: 17). As already mentioned, CDA does not analyze text for the sake of text analysis and thus regards language not only in terms of an abstract system but, more importantly, as a form of social practice.

From the database, seven key issues have garnered scholarly attention: gender (13, 10.7%), identity (4, 3.3%), ideology and power (73, 60.3%), immigration (8, 6.6%), Islam and race (8, 6.6%), nationalism (4, 3.3%) and war and security (3, 2.5%). As with the domains discussed in Section 5.2 above, these categories were derived by conventional content analysis, which Hsieh and Shannon (2005) define as identifying coding categories directly from the text data as opposed to using preconceived categories. In other words, we developed the coding scheme as we systematically skim the literature in our database until we reach saturation point, i.e. when new categories were not emerging anymore. At the end, some topics that were similar in focus such as 'war' and 'security' were merged together into a broader category. As Figure 3 shows, ideology and power (73, 60.3%) is the dominant topic of coverage, reinforcing the principal thesis underlying CDA and its focus on critically analyzing what people do with their words. (Three studies, Baker, 2004; Al-Hejin, 2014; Hardaker & McGlashan, 2015, are counted twice as each of them investigates gender and identity or Islam). Arguably, all the other categories, as is characteristic of CDA, assume some kind of ideology; however, the 'Ideology and Power' category consists of only studies that make this issue their central focus. The view that the chief objective of CDA research over the years has been on issues of power and ideology embedded within texts is further buttressed by the fact that ideology and power is explicitly foregrounded throughout the period of coverage of the present study (e.g. Flowerdew, 1997; Baker, 2003; Mulderrig, 2008; Alhejin, 2015).

Such studies have discussed power and ideology as it is manifested in xenophobia (Downs, 2002), food politics (Cook et al., 2009), disillusionment (Morrison & Love, 1996), (dis)-empowerment (Prentice & Hardie, 2009), discrimination, stereotypes and prejudice (Hardaker & McGlashan, 2016), solidarity (El-Falaky, 2015) and in the construction or reproduction of and resistance against dominance, inequality and power abuse (Chiluwa, 2012; Sotillo & Starace-Nastasi, 1999). Altogether, these studies show how various social groups,

depending on their position and role, use discursive strategies to, among other things, preserve face, create and promote a myth or an illusion, discredit others, reproduce or resist a stereotype (or categorization), invoke solidarity, unification or enemification, evoke fear or threaten others into submission or action as well as manipulate others. The topicalization of ideology and power in the corpus-based CDA literature is prevalent in Western Europe (the UK, especially), North America (notably the US) and Southeast Asia (notably China). Consequently, some studies (e.g. Chen, 2014; Hou, 2015; Einsiedel et al., 2015) compared the textualization of ideology and power in these regions.



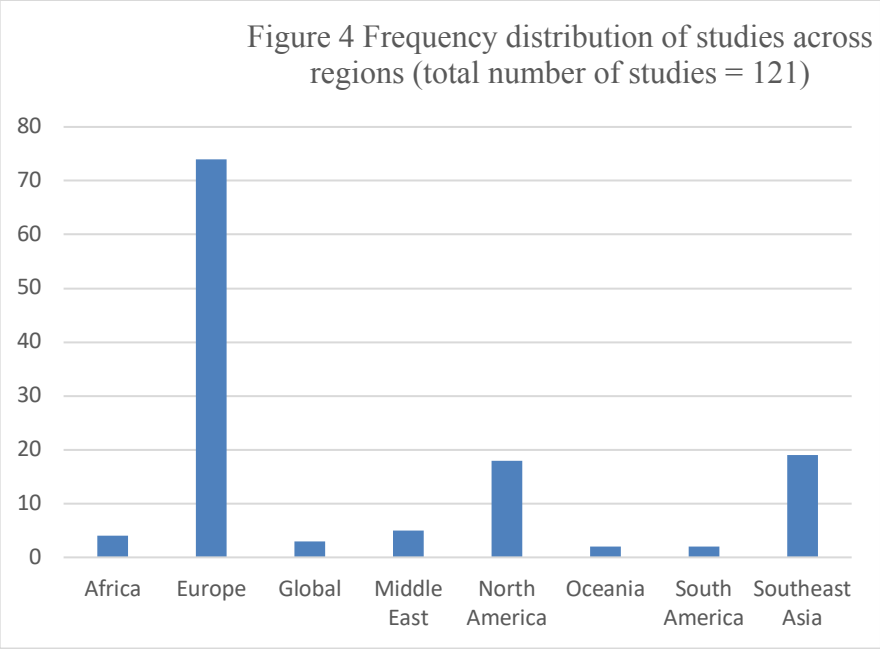
Following ideology and power is the issue of gender (13, 10.7%) – with its focus on sexualization (e.g. Caldas-Coulthard & Moon, 2010; Edwards & Milanis, 2014), homosexuality (e.g. Baker, 2003; Mongie, 2015) and identity (Baker, 2004) – Islam and race (8, 6.6%) (e.g. Salama, 2011) and immigration (8, 6.6%) (e.g. Salahshour, 2016). Interestingly, all the studies on Islam and immigration come after Baker et al. (2008), suggesting the possible influence that their study (which also focused on Islam and immigration) has had on corpus-based CDA research. The bulk of the studies on Islam and immigration are from Western Europe (Spain and the UK – e.g. the studies of Baker and his colleagues) and, occasionally, from the Middle East (e.g. Hakam's (2009) study on Arab newspaper discourse and the 'Prophet Mohammed cartoons controversy') or Oceania (e.g. Salashour's (2016) paper on liquid metaphor as positive evaluations of migrants in New Zealand). This finding is also not alarming given the kind of news that has captured world attention in recent years.

Although war and security (3, 2.5%) is the topic to have received the least attention in the database, these studies are worth mentioning because of the importance that issues of security have on world affairs, especially the increased debate on terrorism, disarmament and nuclear weapons across the globe. From the database, the studies on war and security have focused on (regional) conflict (Almeida, 2011), counter-terrorism (MacDonald & Hunter, 2013) and nuclear proliferation (MacDonald et al., 2015). Also, they featured prominently in the UK and US, rather unsurprisingly, and rarely in the Middle East. More studies are certainly needed in this

area. It is indeed interesting that issues on war and security are not the focus of scholarly critique in the Middle East, given the sustained hostilities in this region for decades. The label 'Others' in Figure 3 refers to topics that appeared only once in the database, including healthcare (McDonald & Woodward-Kron, 2016), tourism (Figueiredo & de Santa Catarina, 2015), elections (Antonakaki et al., 2016), fiscal crisis (Schroter & Storjohann, 2015), justice (Boyd, 2013), climate change (Koteyko, 2012) and multilingualism (Murphy, 2013). It can, therefore, be said that scores of issues have occupied the attention of discourse analysts synthesizing the methodological models provided within CL and CDA.

5.4 Representativeness: areal and regional coverage

The areal coverage of studies that have integrated CL and CDA approaches is worth discussing, as it gives us an idea of the representativeness of the use of the methodology since it was introduced in the 1990s and popularized in 2008. The section also discusses the possible factors accounting for the preponderance or paucity of studies in the various regions. As already noted, the selection criteria for the database included only studies written in English (the academic language the authors are literate in). As this can have implications for the geographical spread of the framework, readers should interpret the discussion in this section with this constraint in mind. Figure 4 lists all the regions where corpus-based CDA has ever taken place with respect to the database and indicates the number of research output in each area – the term 'Global' refers to studies conducted on datasets such as *Amazon.com* compiled from multiple regions. Five studies are concerned with two or three countries/regions, namely China and US (Chen, 2014; Wang & Chen, 2015), UK and US (Thelwall, 2008), Southeast Asia and Europe (Hou, 2015) and Europe, North America and Oceania (Einsiedel et al., 2015). These are counted for each of the regional categories they include (see Figure 4). In all, the broad regions covered by studies include Africa, Europe, the Middle East, North America, Oceania, South America and Southeast Asia. Europe (74, 61. 2%) has the highest research output –the studies here are more than all the other regions put together – followed by Southeast Asia (19, 15. 7%), North America (15, 12.4%) and the Middle East (5, 4. 1%).



In Europe, several studies (e.g. Davidson, 2013; Baker & Levon, 2015; Breeze, 2015; Meng & Yu, 2016) have been conducted in Western Europe, especially in the UK and these studies have primarily focused on issues such as power and ideology, gender and identity, race and Islam and immigration. The studies from North America (e.g. Downs, 2002; Sotillo & Wang-Gempp, 2004; Freake et al., 2011) are normally from the US, with their focus on race, gender and identity, nationalism, and power and ideology. The topical issues that have been the focus of the studies in the UK and US, as stated earlier, show the preoccupation of the media in these countries. For instance, in recent years, immigration and issues on human rights have dominated the media in the UK and US. Moreover, the availability of funding and access to technology (especially, software for corpus analysis) in these countries could be a contributing factor to the prevalence of corpus-based CDA research there. Still, to the extent that Baker et al.'s (2008) influential study was carried out in the UK, it seems logical that the framework will gain ground, first, in the UK before being adopted and/or adapted elsewhere.

Of the countries in Southeast Asia where corpus-based CDA research has been done, Hong Kong is prominently featured, with the work of Flowerdew (1997; 2004), Cheng & Lam (2012), Cheng & Yu (2012), and Zhang & Mihelj (2012) being some notable studies. Generally, these studies have focused on Hong Kong people's identity vis-a-vis China and the world's perception of Hong Kong, given that it was the last British colony, returning to Chinese sovereignty in 1997. Notably, only few studies have come out of Mainland China in addition to our observation that all these studies except for Zhang (2015) and Hou (2016) were comparative studies, with their focus on identity and ideology (but not explicitly on power). Perhaps, the issue of information control and censorship could be the reason for the lack of corpus-based research in China compared to Hong Kong. Another factor possibly accounting for this situation is that many Chinese CDA scholars often write in Chinese and publish in local Chinese journals (Liu Xiangdong, pc.) which are not represented in our database. The studies that have focused on the Middle East discussed conflict (Almeida, 2011) and religion (Islam) (Hakam, 2009; Al-Hejin, 2014) with two of the authors writing from outside the jurisdiction where their studies were

conducted and the other two residing in the Middle East. An interesting observation is that unlike in Europe and North America, issues such as gender (and the accompanying notions of sexuality, masculinity vs. feminity, etc.) and power were hardly broached in Asia and the Middle East. It seems then that the socio-cultural orientation of the regions and, possibly, their political situations tend to inform the extent to which corpus-based CDA research has been applied and the issues and topics addressed by scholars. The regional details provided here are, therefore, important in ascertaining the external factors shaping and constraining CDA research around the world.

The regions with the least number of corpus-based CDA research are Africa (four studies), South America (two studies) and Oceania (two studies). The studies conducted in South America and Oceania (e.g. Figueiredo & de Santa Catarina, 2015; Salahshour, 2016) were all done between 2015 and 2016, suggesting that corpus-based CDA is only now advancing into this region. In Africa, however, as far back as 1996, a corpus-based CDA study had been conducted in Zimbabwe by Morrison and Love. They examined the discourse of disillusionment as manifested in letters sent to the editors of two Zimbabwean magazines ten years after Zimbabwe's independence. It is, therefore, quite startling that from 1996 to 2016, only three other studies have been added to Morrison and Love's. Two reasons may account for this lull: funding and technology. Morrison and Love's study was "an outcome of a British Council-funded Link Scheme between the School of English at the University of Birmingham and the Linguistics Department at the University of Zimbabwe" (p. 70) – this is a very rare case in point. Additionally, the first author was based at the University of Oslo, Norway, as a doctoral student when the research was conducted. Although some corpus analytical software may be expensive, the point must also be made that one of the distinctive qualities of CL is that it has a strong tradition of making software freely available. It seems then that (critical) discourse analysts in several parts of Africa may be unaware of the utility of corpus analysis or they may not be aware of some of the free corpus software (e.g. *AntConc* and *UAM Corpus Tool*) available for download. In terms of technology, until recently, many African universities were not equipped with the tools and expertise needed to carry out corpus-based research and even now, not much corpus-based work goes on in Africa (Ngula & Nartey, 2014). For instance, at the 2013 Corpus Linguistics Conference in Lancaster when Laurence Anthony, developer of *AntConc,* took stock of his software, giving a graphical representation of downloads of the free corpus analysis toolkit in the various continents, it was zero per cent for Africa (Ngula, 2015).

Of the four studies from Africa, three were done in Southern Africa (Zimbabwe and South Africa) (Morrison & Love, 1996; Edwards & Milani, 2014; Mongie, 2015) and one was done in Nigeria, West Africa (Chiluwa, 2012). The South African studies looked at homosexuality and sexualization – these being issues of concern in South Africa in recent years. Chiluwa's (2012) study examined Biafra online tweets as a case of resistance discourse and Morrison and Love's (1996) paper, as already mentioned, analyzed the discourse of disillusionment in the Zimbabwean media. These studies, like many of the corpus-based CDA studies discussed in this paper, focused on issues pertinent to the countries where they were conducted. It is clear from the surveyed data that a lot more research and collaboration across regions such as between Europe and Africa are needed for even a minimal areal representation to be achieved.

6 Concluding remarks

This study surveyed studies that combine CDA and CL. It shows that attempts to deploy corpus methods in CDA studies began in the mid-1990's but this approach was popularised in the late 2000s by Baker et al. (2008). It also shows that over the years, the commonly used corpus techniques in CDA include wordlists, keywords, collocations, clusters, and concordances while the discourse-historical approach, social actor theory and the dialectical-relational approach are the dominant CDA techniques/approaches combined with CL. One underlying assumption of this study is that corpus-based CDA research is important in advancing the social and political commitment of CDA. It enables the analysis of a large quantity of data to ascertain the dynamics and nature of wide-spread ideologies relating to macro-issues such as race and identity, gender and sexuality, and immigration and prejudice. It also helps to narrow down from large corpora, encompassing a wide variety of domains, as defined in this study, to specific domains for a closer investigation and appropriate social and political action. The study further assumes that corpus-based CDA is relevant to descriptive corpus linguists interested in language as a semiotic action. It equips such linguists with the techniques to intensively analyze "the localized construal of social phenomena such as identity in particular contexts, often resulting in a complex, rich, interpretive, dynamic, and flexible analysis of micro-contexts, and capturing the dynamic and negotiatory nature of much language use" (Bednarek, 2009: 22). In other words, CDA makes it possible for CL to answer socially inspired research questions such as power, inequality, identity, and change so that CL is not limited to grammar or lexicography. One contribution of the present study lies in its revealing of some research dynamics that could be applicable to CDA as a broad field of language and communication research. Examples are the regional distribution of studies and domains and issues covered. Further research will be useful in confirming the extent to which these findings apply to the general sub-discipline of CDA beyond corpus-based approaches.

Considering that corpus-based CDA is *en vogue*, it is likely that in future, the approach will become even more widespread, diversified, and more versatile with a greater degree of sophistication of corpus analytic tools, concordancing software, and the expansion of research paradigms within CDA. It is also likely that regions like Africa, South America, and Oceania where corpus-based CDA is still not common will experience an increase in the use of this methodology, especially given the organization of corpus linguistics workshops across the globe.[7] Despite the gains that have been made in the utilization of corpus-based CDA in discourse studies, the issue of doing justice to the two research traditions when combined remains a challenge in several studies. From the survey of the studies in our database, only few studies (e.g. Orpin, 2005; Baker et al., 2008 and other studies from this camp) effectively demonstrate this balance. This situation highlights one of the main lacunae that future corpus-based CDA research must endeavor to fill. Therefore, the need for collaboration between corpus linguists and (critical) discourse analysts cannot be over-emphasized. In terms of the other analytical models and theorems that have been combined with corpus-based CDA, this study reveals that frameworks in the allied field of critical genre analysis and forensic contexts as well those of contrastive/intercultural rhetoric, narrative inquiry/analysis, mediated discourse analysis, and conversion analysis have hardly been engaged. It would be interesting to see how future

[7] One of such notable corpus linguistics training events is organized by The ESRC Center for Corpus Approaches to Social Science, Lancaster University.

corpus-based CDA studies can be combined with these models. Other theoretical approaches to CDA (e.g. French discourse analysis, critical linguistics, and social semiotics) can also be combined with CL. Finally, this study demonstrates that specialized corpora of mainly media and political discourses have been the focus of corpus-based CDA. Future studies could examine other text types, including academic classroom discourse, business media discourse, textbooks and curriculum texts, feedback commentaries in education, and sermonic discourses. Although 'tailor-made' corpora usually serve the purpose of corpus-based CDA studies, it will be useful for future studies to combine such corpora with large reference corpora (but bearing in mind its static nature) given the latter's role in safeguarding against over- and under-interpretation' (O' Halloran & Coffin, 2004). As Mautner (2009: 35) notes, "The discovery of usage patterns in large reference corpora can be seen as a worthwhile pursuit in its own right, making as it does a legitimate contribution to the critical study of language and society".

Acknowledgements

We sincerely thank Malcolm MacDonald, Kathleen Ahrens and Christian Matthiessen for very useful comments and discussions on earlier drafts of this paper. We also thank Paul Baker, the Commissioning Editor of *Corpora* and two anonymous reviewers, for their valuable advice and insightful comments.

References

Abid, R. Z., & Manan, S. A. (2015). Integrating corpus linguistics in critical literacy pedagogy: A case study of Lance Armstrong's transformation from a titleholder to a fraud. *Procedia - Social and Behavioural Sciences*, 208(20): 128-137.

Al-Hejin, B. (2015). Covering Muslim women: Semantic macrostructures in BBC news. *Discourse & Communication*, 9(1): 19-46.

Almeida, E. (2011). Palestinian and Israeli voices in five years of U.S. newspaper discourse. *International Journal of Communication*, 5: 1586-1605.

Antonakaki, D., Spiliotopoulos, D., & Samaras, C. (2016). Investigating the complete corpus of referendum and elections tweets. In: Kumar R., Caverlee J., & Tong, H. (eds.) *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM 2016*, pp. 100-105.

Augoustinos, M., Crabb, S., & Shepherd, R. (2010). Genetically modified food in the news: media representations of the GM debate in the UK. *Public Understanding of Science*, 19(1): 98-114.

Aull, L.L., & Brown, D.W. (2013). Fighting words: a corpus analysis of gender representations in sports reportage. *Corpora,* 8(1): 27-52.

Baker, P. (2003). No effeminates please: a corpus-based analysis of masculinity via personal adverts in Gay News/Times 1973-2000. *The Sociological Review*, 51(1): 243-260.

Baker, P. (2004). 'Unnatural Acts': Discourses of homosexuality within the House of Lords debates on gay male law reform. *Journal of Sociolinguistics*, 8: 88-106.

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

Baker, P. (2012). Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies*, 9(3): 247-256.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and

corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3): 273-306.

Baker, P., Gabrielatos, C., & McEnery, T. (2013). Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim' in the British press 1998–2009. *Applied Linguistics*, 34(3): 255-278.

Baker, P. & Levon, E. (2016). 'That's what I call a man': representations of racialized and classed masculinities in the UK print media. *Gender and Language*, 10(1): 106-139.

Bednarek, M. (2009). Corpora and discourse: A three-pronged approach to analyzing linguistic data. In M. Haugh et al. (eds.) *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, pp. 19-24, Somerville, MA: Cascadilla Proceedings Project.

Bednarek, M. & Caple, H. (2014). Why do news values matter? Towards a new methodological framework for analysing news discourse in critical discourse analysis and beyond. *Discourse & Society*, 25(2) 135-158.

Bhatia, V. K., Flowerdew, J., & Jones, R. H. (2008) (Eds.). Advances in discourse studies. Routledge: London.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Brindle, A. (2015). A corpus analysis of discursive constructions of the Sunflower Student Movement in the English language Taiwanese press. *Discourse & Society*, 2015: 1-17.

Boyd, M.S. (2013). Representation of foreign justice in the media: The Amanda Knox case. *Critical Approaches to Discourse Analysis across Disciplines*, 7(1): 33-50.

Breeze, R. (2015). "Or so the government would have you believe": Uses of "you" in Guardian editorials. *Discourse, Context & Media*, 10: 36-44.

Caldas-Coulthard, C.R. & Moon, R. (2010). 'Curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse & Society*, 21(2): 99-133.

Cheng, W. (2013). Corpus-based linguistic approaches to critical discourse analysis. In: C. A. Chapelle *The Encyclopedia of Applied Linguistics*, pp. 1-8, New Jersey: Blackwell.

Chen, S. (2014). Corpus linguistics in critical discourse analysis: A case study on news reports of the 2011 Libyan civil war. *Stream: Culture/Politics/Technology*, 5(1): 21-28.

Cheng, W., & Lam, P.W.Y. (2012). Western perceptions of Hong Kong ten years on: A corpus-driven critical discourse study. *Applied Linguistics* 2012: 1-13.

Chiluwa, I. (2012). Social media networks and the discourse of resistance: A sociolinguistic CDA of Biafra online discourses. *Discourse & Society*, 23(3): 217-244.

Clark, C. (2007). A War of words: A linguistic analysis of BBC embedded reports during the Iraq conflict. In N. Fairclough, G. Cortese & P. Ardizzone (eds.) *Discourse and Contemporary Social Change*, pp. 119–40. Bern: Peter Lang.

Cotterill, J. (2001). Domestic discord, rocky relationships: semantic prosodies in representations of marital violence in the O.J. Simpson trial. *Discourse & Society*, 12(3): 291-312.

Davidson, P. (2013). The role of 'social exclusion' and other metaphors in contemporary British social policy: a cognitive critical approach. *Journal of International Relations and Development*, 16(2): 206-226.

Don, Z.M., Knowles, G., & Fatt, C.K. (2010). Nationhood and Malaysian identity: a corpus-based approach. *Text & Talk*, 30(3): 267-287.

Downs, (2002). Representing gun owners: Frame identification as social responsibility in news media discourse. *Written Communication*, 19(1): 44-75.

Duguid, A. (2010). Investigating anti and some reflections on modern diachronic corpus-assisted discourse studies (MD-CADS). *Corpora*, *5*(2):191-220.

Edwards, M. & Milani, T.M. (2014). The everyday life of sexual politics: A feminist critical discourse analysis of herbalist pamphlets in Johannesburg. *Southern African Linguistics and Applied Language Studies*, 32(4): 461-481.

Einsiedel, E.F., Remillard, C., & Gomaa, M. (2015). The representation of biofuels in political cartoons: Ironies, contradictions and moral dilemmas. *Environmental Communication: A Journal of Nature and Culture*, 11(1): 41-62.

El-Falaky, M.S. (2015). Vote for me! A corpus linguistic analysis of American presidential debates using functional grammar. *Arts Social Science Journal*, 6(4): 1-13.

Evans, M. & Jeffries, L. (2013). The rise of choice as an absolute 'good': a study of British manifestos (1900-2010). *Journal of Language & Politics*, 14(6): 751-777.

Fairclough, N. (2001). *Language and power* London: Longman.

Fairclough, N. (2010). *Critical discourse analysis: the critical study of language*. London: Routledge.

Figueiredo, D. & Pasquetti, C.A. (2016). The discourse of tourism: an analysis of the online article "Best in Travel 2015: Top 10 cities" in its translation to Brazilian Portuguese. *Ilha do Desterro*, 69(1): 201-212.

Flowerdew, J. (1997). The discourse of colonial withdrawal: A case study in the creation of mythic discourse. *Discourse & Society*, 8(4): 453-477.

Flowerdew, J. (2004). Identity politics and Hong Kong's return to Chinese sovereignty: analyzing the discourse of Hong Kong's first Chief Executive. *Journal of Pragmatics*, 36: 1551-1578.

Fowler, R. (1996). On critical linguistics. In: C. R. Caldas-Coulthard. and M. Coulthard (eds.), *Texts and Practices: Reading in Critical discourse analysis*, London: Routledge.

Freake, R., Gentil, G., & Sheyholislami, J. (2011). A bilingual corpus-assisted discourse study of the construction of nationhood and belonging in Quebec. *Discourse & Society*, 22(1): 21-47.

Gabrielatos, C. (2017). Bibliography: Corpus approaches to discourse studies. Available at: https://www.edgehill.ac.uk/english/dr-costas-gabrielatos/?tab=bibliography-corpus-approaches-to-discourse-studies.

Gabrielatos, C. & Baker, P. (2008). Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK press 1996-2005. *Journal of English Linguistics*, 36(5), 5-38.

Hakam, J. (2009). The 'cartoons controversy': a critical discourse analysis of English language Arab newspaper discourse. *Discourse & Society*, 20(1): 33-57.

Hardaker, C., & McGlashan, M. (2016). "Real men don't hate women": Twitter rape threats and group identity. *Journal of Pragmatics*, 91: 80-93.

Hardt-Mautner, G. (1995). Only connect: Critical discourse analysis and corpus linguistics. *UCREL Technical Paper 6*. Lancaster, UK: University of Lancaster.

Henry, F., & Tator, C. (2002). *Discourse of domination: racial bias in the Canadian English-language press*. Toronto: University of Toronto Press.

Hou, Z. (2015). A critical analysis of media Reports on China's air defense identification zone. In: FuertesOlivera P. A, AlvarezDeLaFuente E., & FernandezFuertes, R. (eds.) *Current Work in Corpus Linguistics: Working with Traditionally Conceived Corpora and Beyond (CICLC 2015)*, pp. 194-201.

Hou, Z. (2016). A corpus-driven analysis of media representations of the Chinese Dream. *International Journal of English Linguistics*, 6(1): 142-149.

Hunston, S. (2011). *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. London: Routledge.

Hsiu-Fang Hsieh Sarah E. Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9): 1277-1288.

Jaworska, S. & Krishnamurthy, R. (2012). On the F word: A corpus-based analysis of the media representation of feminism in British and German press discourse, 1990-2009. *Discourse & Society*, 23(4): 401-431.

Jeffries, L., & Walker, B. (2012). Key words in the press: A critical corpus-driven analysis of ideology in the Blair years (1998- 2007). *English Text Construction*, 5(2): 208-229.

KhosraviNik, M. (2008). British newspapers and the representation of refugees, asylum seekers and immigrants between 1996 and 2006. *Working Papers Series 128*, Department of Linguistics and English Language, Lancaster University. UK.

Kim, K.H. (2014). Examining US news media discourses about North Korea: A corpus-based critical discourse analysis. *Discourse & Society*, 25(2), 221-244.

Kim, K.H. (2017). Newsweek discourses on China and their Korean translations: A corpus-based approach. *Discourse, Context & Media*, 15: 34-44.

Koteyko, N. (2012). Managing carbon emissions: A discursive presentation of 'market-driven sustainability in the British media. *Language & Communication*, 32(1): 24-35.

Lee, C. (2016). A corpus-based approach to transitivity analysis at grammatical and conceptual levels: A case study of South Korean newspaper discourse. *International Journal of Corpus Linguistics*, 21(4): 465-498.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (ed.) *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, 4–8 August 1991*, pp. 105–22. Berlin: Mouton de Gruyter.

Leung, R.C.H. (2016). Representation of gamblers in the Singaporean press since casino legalization: A corpus-driven critical analysis. *International Journal of Applied Linguistics & English Literature*, 5(6): 51-63.

Lippi, M. & Torroni, P. (2016). Argument mining from speech: Detecting claims in political debates. *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, pp. 2979-2985, Phoenix-Arizona.

MacDonald, M.N., Homolar, A., & Rethel, L. (2015). Manufacturing dissent: The discursive formation of nuclear proliferation (2006-2012). *Discourse & Communication*, 9(2): 173-197.

MacDonald, M. & Hunter, D. (2013) Security, population and governmentality: UK counter-terrorism discourse (2007-2011). *Critical Approaches to Discourse Analysis across Disciplines*, 7(1): 123-140.

Mautner, G. (2007). Mining large corpora for social information: The case of elderly. *Language in Society,* 36(1): 51-72.

Mautner, G. (2009). Corpora and critical discourse analysis. In P. Baker (ed.) *Contemporary Corpus Linguistics*, pp. 32-46, London: Continuum.

Mautner, G. (2015). Checks and balances: How corpus linguistics can contribute to CDA. In R. Wodak & M. Meyer (eds.) *Methods of Critical Discourse Studies*, pp. 154-179, London: Sage.

Magalhaes, C.M. (2006). A critical discourse analysis approach to news discourses and social practices on race in Brazil. *DELTA*, 22(2): 275-301.

McDonald, D. & Woodward-Kron, R. (2016). Member roles and identities in online support groups: Perspectives from corpus and systemic functional linguistics. *Discourse & Communication*, 10(2): 157-175.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Meng, C. & Yu, Y. (2012). "We should…" versus "We will…": How do the governments report their work in "One Country Two Systems"? A corpus-driven critical discourse analysis of government work reports in Greater China. *Text & Talk*, 36(2): 199-219.

Mongie, L.D. (2015). The discourse of liberation: Frames used in characterizing the gay liberation movement in two South African newspapers. *Stellenbosch Papers in Linguistics 46*, Department of General Linguistics, Stellenbosch University, South Africa.

Morrison, A., & Love, A. (1996). A discourse of disillusionment: Letters to the editor in two Zimbabwean magazines 10 years after independence. *Discourse & Society*, 7: 39-76.

Mulderrig, J. (2008). Using keywords analysis in CDA: evolving discourses of the knowledge economy in education. In Jessop, B., Fairclough, N. & Wodak, R. (eds.) *Education and the Knowledge-based Economy in Europe. Educational Futures: Rethinking Theory and Practice*, pp. 149-169, Rotterdam: Sense.

Mulderrig, J. (2011). The grammar of governance. *Critical Discourse Studies*, 8(1): 45-68.

Mulderrig, J. (2012). Manufacturing consent: A corpus-based critical discourse analysis of New Labor's educational governance. In D. R. Cole & L. J. Graham (eds.) *The Power In/Of Languag*e, pp. 13-28, Oxford: Blackwell.

Murphy, A. (2013). Corpus analysis of European Union documents. In Chapelle C. A. (ed.) *The Encyclopedia of Applied Linguistics*, pp. 1-6, Oxford: Blackwell.

Ngula, R.S. (2015). *Epistemic modality in social science research articles written by Ghanaian authors: A corpus-based study of disciplinary and native vs. non-native variations*. Unpublished PhD thesis, Lancaster University, United Kingdom.

Ngula, R.S., & Nartey, M. (2014). Language corpora: The case for Ghanaian English. *3L: The Southeast Asian Journal of English Language Studies*, 20(3), 79-92.

Norris, J.M., & Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning*, 50: 417-528.

Norris, J.M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In: J. M. Norris & L. Ortega (eds.) *Synthesizing Research on Language Learning and Teaching*, pp. 3-50, Amsterdam: Benjamins.

O'Halloran, K. A. & Coffin, C. (2004). Checking overinterpretation and underinterpretation: Help from corpora in critical linguistics. In C. Coffin, A. Hewings, & K. A. O'Halloran (eds.) *Applying English Grammar: Functional and Corpus Approaches*, pp. 275-297, London: Hodder Arnold.

O'Halloran, K. (2013). A corpus-based deconstructive strategy for critically engaging with arguments. *Argument & Computation*, 4(2): 128-150.

Orpin, D. (2005). Corpus linguistics and critical discourse analysis: Examining the ideology of sleaze. *International Journal of Corpus Linguistics*, 10(1): 37–61.

Pan, H. & Zhang, M. (2016). Translating for a healthier gaming industry Keywords and translation in Macao's gaming discourse. *Translation Spaces*, 5(2): 163-180.

Partington, A. (2010). Modern diachronic corpus-assisted discourse studies (MD-CADS) on UK newspapers: an overview of the project. *Corpora*, 5(2), 83-108.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meaning in discourse: Theory and practice in corpus-assisted discourse studies*. Amsterdam: John Benjamin.

Pearce, M. (2014). Key function words in a corpus of UK election manifestos. *Linguistik Online,* 65(3): 1-22.

Piper, A. (2000). Some have credit cards and others have giro cheques: 'Individuals' and 'people' as lifelong learners in late modernity. *Discourse & Society*, 11(3): 515-42.

Potts, A., Simm, W., & Whittle, J. (2014). Exploring 'success' in digitally augmented activism: A triangulated approach to analyzing UK activist Twitter use. *Discourse, Context & Media*, 6: 65-76.

Potts, A., Bednarek, M., & Caple, H. (2015). How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina. *Discourse & Communication*, 9(2): 9(2) 149-172.

Prentice, S. & Hardie, A. (2009). Empowerment and disempowerment in the Glencairn Uprising A corpus-based critical analysis of Early Modern English news discourse. *Journal of Historical Pragmatics*, 10(1): 23-55.

Salahshour, N. (2009). Liquid metaphors as positive evaluations: A corpus-assisted discourse analysis of the representation of migrants in a daily New Zealand newspaper. *Discourse, Context & Media*, 13: 73-81.

Salama, A.H.Y. (2011). Ideological collocation and the re-contextualization of Wahhabi-Saudi Islam post-9/11: A synergy of corpus linguistics and critical discourse analysis. *Discourse & Society*, 22(3): 315-342.

Schröter, M. & Storjohann, P. (2015). Patterns of discourse semantics: A corpus-assisted study of financial crisis in British newspaper discourse in 2009. *Pragmatics & Society*, 6 (1): 43-66.

Sealey, A. (2012). 'I just couldn't do it': representations of constraint in an oral history corpus. *Critical Discourse Studies*, 9(3): 195-210.

Simon-Vandenbergen, A. (2000). The functions of *I think* in political discourse. *International Journal of Applied Linguistics*, 10(1): 41-63.

Skalicky, S. (2013). Was this analysis helpful? A genre analysis of the Amazon.com discourse community and its "most helpful" product reviews. *Discourse, Context & Media*, 2(2): 84-93.

Sotillo, S.M., & Wang-Gempp, J. (2004). Using corpus linguistics to investigate class, ideology, and discursive practice in online political discussion. In U. Connor & T. A. Upton (eds.) *Discourse in the Professions: Perspectives from Corpus Linguistics*, pp. 91-122, Amsterdam, Netherlands: John Benjamins.

Sotillo, S.M. & Starace-Nastasi, D. (1999). Political discourse of a working-class town. *Discourse & Society*, 10(3): 411-438.

Stubbs, M. (1997). Whorf's children: Critical comments on critical discourse analysis. In: A. Ryan and A. Ray (eds.) *Evolving Models of Language*, pp. 110-16, Clevedon: BAAL in association with Multilingual Matters.

Stubbs, M. & Gerbig, A. (1993). Human and inhuman geography: on the computer-assisted analysis of long texts. In: M. Hoey (ed.) *Data, Description, Discourse: Papers on the English Language in Honor of John McH Sinclair*, pp. 64-85. London: Harper Collins.

Subterilu, N.C. (2013). 'English… it's part of our blood': Ideologies of language and nation in United States Congressional discourse. *Journal of Sociolinguistics*, 17(1), 37-65.

Teubert, W., & Krishnamurthy, R. (2007). General introduction. In: W. Teubert & R. Krishnamurthy (eds.) *Corpus Linguistics: Critical Concepts in Linguistics*, pp. 1–37, London, England: Routledge.

Thelwall, M. (2008). *Fk yea I swear*: Cursing and gender in a corpus of MySpace pages, *Corpora*, 3(1): 83-107.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: Benjamins.

Tornberg, A. & Tornberg, T. (2016). Muslims in social media discourse: Combining topic modeling and critical discourse analysis. *Discourse, Context & Media*, 13: 132-142.

Triebl, E. (2015). ⋯or not to be. The strategic and non-strategic use of negative identifiers in online forums. In: B. Kettemann (ed.) *AAA - Arbeiten aus Anglistik und Amerikanistik*, pp. 247-270, Tübingen, Germany.

Van Dijk, T.A. (1993). *Elite discourse and racism*. Newbury Park, CA: Sage.

Van Dijk, T.A. (1995). Aims of critical discourse analysis. *Japanese Discourse*, 1, 17-27.

Van De Mieroop, D. (2007). The complementarity of two identities and two approaches: Quantitative and qualitative analysis of institutional and professional identity. *Journal of Pragmatics*, 39(6): 1120-1142.

Widdowson, H.G. (2000). On the limitations of linguistics applied. *Applied Linguistics,* 21(1): 3-25.

Wodak, R. (2001). What CDA is about - a summary of its history, important concepts and its developments. In R. Wodak & M. Meyer (eds.) *Methods of Critical Discourse Analysis*, pp. 1-13, London: SAGE.

Wodak, R., & Meyer, M. (2009). Critical discourse analysis: History, agenda, theory and methodology. In R. Wodak & M. Meyer (eds.) *Methods of critical discourse analysis*, pp. 1-33, London: SAGE.

Zhang, W. (2015). The shifting representation of common people in China's news media. *Journal of Language & Politics*, 14(2): 285-307.

Zhang, M. & Mihelj, S. (2012). Hong Kong identity and the press-politics dynamics: a corpus-assisted discourse study. *Asian Journal of Communication*, 22(5): 506-527.