CORPUS-ASSISTED DISCOURSE STUDIES: OPPORTUNITIES AND LIMITATIONS FOR THE ANALYSIS OF NEWS DISCOURSE

Hanane ZIH  
hanane.zih@usmba.ac.ma  
Maha EL BIADI  
maha.elbiadi@usmba.ac.ma  
University Sidi Mohamed Ben Abdallah

Abstract
Advances of modern computer technology have recently proved essential in a variety of disciplines. Linguistics in general, and Critical Discourse Analysis (CDA) in particular are no exception, as practitioners working within these areas believe that drawing upon certain computerized methods and using Natural Language Processing (NLP) programs do not only assist language analysts to process larger data, but they also enable them to obtain accurate results from which valid generalizations can be drawn. The present paper aims to discuss the opportunities that Corpus Linguistics (CL) offer when used in critical discourse analysis, Corpus-Assisted Discourse Studies (CADS). It first starts with introducing Corpus Linguistics and outlining the main distinctions between corpus-based and corpus-driven approaches. The paper also provides a brief overview of critical discourse analysis, and the objectives critical discourse analysts seek to attain by qualitatively analyzing a particular discourse. In order to show the possibilities and the limitations of such a methodological synergy, the paper offers a literature review of the studies that employed the CADS framework to analyze news discourse along with drawing upon a recently published pilot study that evaluates the effectiveness of two corpus software in identifying specific linguistic variables compared to manual analysis. It suggests that while the CADS framework is characterized by flexibility and the possibility of processing a larger corpus thanks to its various analytical tools, the extent to which the CADS approach can really elude researcher’s bias as far as the human interpretation is concerned is still a debatable question.

Key words: Corpus linguistics, Critical Discourse Analysis, CADS, Methodological synergy

1. INTRODUCTION
In the last few decades, a great deal of attention has been addressed towards the opportunities offered by technologies in advancing different fields of research, and critical discourse analysis (henceforth CDA) is no exception. As an increasingly significant area of study within applied linguistics, critical discourse analysts have often been criticized for being subjective and biased when dealing with different linguistic forms and the way they are used in discourse, for they mostly rely on the qualitative and not the quantitative method to analyze their data, and the offered interpretations are not based on systematic language description. (Widdowson 1998; Baker 2012). McEnery argues that critical discourse analysts often have a large data to choose from, but since they have to take so many aspects into account, they tend to ‘cherry pick’ their data so that ‘a really detailed in-depth qualitative analysis’ can be carried out (Koller and Mautner, 2004). This raises the issue that the selected text is often the result of the critical discourse analyst own biases. For instance, if the linguists using CDA think the press reports issues related to Muslims negatively, they are more likely to select a small number of texts that prove that. Thus, the findings might confirm that those ‘cherry picked’ texts are really ‘Islamophobic’, but the question that McEnery raises is: ‘How about the 6000 texts which
might represent Muslims and Islam positively?” (McEnery, n.d). Therefore, the main objective of the present paper is to question the extent to which CL combined with CDA can put an end to the criticism directed against critical discourse analysts. It also aims to shed light on the advantages offered by CADS methodology namely in identifying the linguistic features that critical discourse analysts are interested in, allowing a more manageable and objective analysis to be undertaken. Therefore, to counter the criticism that critical discourse analysts tend to prove a ‘preconceived point’, the findings of a corpus-based pilot study will be drawn upon.

Using natural language processing programs can assist the researcher to process large amounts of linguistic data, making it possible for them to avoid being subjective to a certain degree while handling different types of texts. This is achieved by enabling them to come up with quantitative analyses that can point to some discernible patterns, which they can then interpret and explain by linking them to their broader social context of use.

2. CORPUS LINGUISTICS

2.1. Defining Corpus Linguistics and Corpora

Corpus Linguistics first originated in the early 1980s (Leech, 1992) while corpus-based language study dates back to the pre-Chomskyan period when it was used by linguists from the Structuralist school including Boas and Sapir, among others (McEnery et al., 2006:3). Thanks to the advances of modern computer technologies and the many possibilities offered by machine-readable texts including compiling and storing, CL witnessed a revival in the 1980s. Fields such as lexicography, descriptive grammar, lexical and grammatical studies, language variation studies, contrastive and diachronic studies, language teaching and learning, and critical discourse analysis have benefited from the developments of CL.

Kennedy (1998) states that CL is typically concerned “not only with what words, structures or uses are possible in a language, but also with what is probable – what is likely to occur in language use’ (1998:8). Accordingly, corpus linguists employ large texts, what is known as corpora, to study instances of naturally occurring language so as to draw some generalizations. The notion ‘corpus’ has been defined as “anybody of texts’ (McEnery and Wilson, 2001:29), yet in the context of modern linguistics, it is often used to imply “a finite body of text, sampled to be maximally representative of a particular variety of a language, and which can be stored and manipulated using a computer’ (McEnery and Wilson, 2001: 73). Hunston (2002) also defines a corpus as “a collection of naturally occurring examples of language […] collected for linguistic study” (2002:2). Therefore, these definitions point out to the main distinction between a corpus and a computerized archive compilation of texts which was highlighted by Kennedy (1998). He explains, “whereas a corpus designed for linguistic analysis is normally a systematic, planned and structured compilation of a text, an archive is a text repository, often huge opportunistically collected, and normally not structured” (p.4). Interestingly, computer software allows for the manipulation of the corpus resulting in different kinds of observations.

2.2. Corpus-Based Approach Vs. Corpus-Driven Approach

Researchers using corpus linguistics software often distinguish between two different types of corpora approaches viz. corpus-driven approach and corpus-based approach. The former is “an inductive process where corpora are investigated from the bottom up and patterns found therein are used to explain linguistic regularities and exceptions of the language variety/genre exemplified by those corpora.” (McEnery, n.d). This approach is based on viewing the data from ‘a naïve perspective’. McEnery points out that though it is not viable to be entirely naïve when analyzing data, the analysts have to start off with that idea in mind and try not to impose any views or ideas or hypothesis. The corpus should drive the analysis based on the obtained
frequencies and patterns. The second approach, *corpus-based approach*, on the other hand, is a more traditional CDA perspective as the critical discourse analysts begin their analysis with hypotheses which have been formulated based on a certain way of representing a group. It is ‘where corpora are used to test performed hypotheses or exemplify existing linguistic theories.’ (Ibid.). With the hypotheses in mind, the researcher develops a corpus to investigate the use of particular linguistic structures so as to see whether there are any emerging patterns.

The elaborated distinction between corpus-based and corpus driven approaches to the analysis of corpus is not of paramount importance. Critical linguists often prefer to describe the approach as ‘corpus-assisted’, acknowledging “its flexibility and the possibility for exploring discursive patterns based on both researcher intuition and evidence in the corpus.” (Costelloe, 2013:117). Similarly, McEnery argues that combining the two approaches is ‘the best kind of analysis’ since, according to him, carrying out a corpus analysis from a very naïve approach is “really difficult [...] we always have ideas about what a certain group have been represented in, because we live in society, and we are aware of how they are being talked about in advance.” (McEnery, n.d). Likewise, in a study conducted by Baker et al. (2008) analyzing discourses of immigration in 140-million-word corpus of British newspaper texts, they argued that the adopted ‘recursive model’ allows ‘moving back and forth between quantitative and qualitative forms of analysis, with each stage informing the subsequent stage’ (p.248). Hence, combining CDA with corpus linguistics (Krishnamurthy (1996), Hoey (1996), Baker, Gabrielatos, & McEnery (2012), Baker & McEnery (2019) has proved to yield more robust and valid set of findings since the researchers’ interpretation is thus ‘grounded in systematic language description’ (Widdowson, 1998:148). Before delving into the advantages of this ‘methodological synergy’ of CDA and CL- Corpus Assisted Discourse Studies (CADS), it is important to first define CDA as an increasingly growing field of research and its key objectives.

3. CRITICAL DISCOURSE ANALYSIS

CDA is an area of applied linguistics, which has diversely been taken to be a paradigm, a method, and an analytical technique, was originally known as *Critical Language Studies* (Billig, 2003). Though a number of definitions have been attributed to CDA, there is a consensus amongst critical discourse analysts that it is an approach that “takes a particular interest in the relationship between language and power […].” Teun A. Van Dijk contends that Critical Discourse Analysis (CDA) is a type of discourse analytical research that primarily studies the way social power abuse, dominance, and inequality are enacted, reproduced, and resisted by text and talk in the social and political context. With such dissident research, critical discourse analysts take explicit position, and thus want to understand, expose, and ultimately resist social inequality. (Van Dijk, 2001:352).

It is evident that CDA does not view language as innocent or neutral; on the contrary, it aims at revealing the hidden agendas of power and dominance that are implicitly injected in various types of discourse. By doing CDA, critical discourse analysts are doing a sort of ‘mental Gymnastics’ (Hodge and Kress, 1993) so that they can provide the explanations and discern the motivations that have led discourse producers to make certain selections from the available linguistic options. In brief, there is a consensus that CDA has made discourse analysis relevant by linking it to a ‘moral cause and an ideological purpose’ (Widdowson, 2004).

In his book, *Discourse and Social Change*, Fairclough (1992) asserts that discourse as a political practice seeks to “establish, sustain, and change power relations, and the collective entities (classes, blocs, communities, groups) between which powers relations obtain”. He goes on to explain the features of discourse as an ideological practice as it “constitutes, naturalizes,
sustains and changes significations of the world from diverse positions in relations” (Fairclough, 1992: 67). It is worth highlighting that critical discourse analysts are aware of their sensitive roles in society as they tend to deconstruct the political and ideological discursive practices of discourse. That is why CDA is seen as “a reaction against the dominant formal (often ‘asocial’ or ‘uncritical’) paradigms” critical discourse analysts are certain that the social structures and social interactions greatly influence, produce, and reproduce the scholarly discourses (Fairclough, 1992).

The work of critical discourse analysts starts from the premise that the grammatical structure is not devoid of meaning. It aims, therefore, to uncover the hidden meanings underlying the linguistic structure in a given piece of discourse. (Fairclough 2003; Fairclough, Mulderrig and Wodak 2011; Fowler 1991; Van Dijk 1997, 2012; Weiss and Wodak 2002).

CDA is relatively complex as it functions at different levels, which makes it quite challenging, and complicated to sometimes do. Critical discourse analysts do consider issues of production and reception as it seeks to investigate the text and power relations in, between, and behind the constructions aiming by this to offer an in-depth analysis of texts. Having to describe the texts and discursive patterns and relationships and links them to social situations and contexts, critical discourse analysts often choose to analyze a limited number of texts to be able to offer a thorough analysis. Richardson (2004) points out that CDA is a bridge that connects both society and its sociopolitical issues with the critical analysis of language. Therefore, we argue that corpus linguistics can assist critical discourse analysts not only analyze large corpora, but it is believed to diminish the potential ambiguities and make it possible to critical discourse analysts to confidently draw conclusions and generalizations which are grounded on systemic language description.

4. CORPUS-ASSISTED DISCOURSE STUDIES (CADS)

4.1. Opportunities of the CADS Approach for News Discourse Analysis

In the mid-1990s, researchers such as Mautner (1995,2009), Koller and Mautner (2004), Baker (2006), and Baker and McEnery (2005), have attempted to integrate general methodological approaches of corpus linguistics with CDA for the ample benefits this synergy offers. First of all, CADS help “reduce researcher bias” (Baker et al., 2012:8). Compared with manual text analysis, computer-assisted text analysis can be considered potentially ‘more objective’. Murphy points out that text analysis software is free from ‘the presuppositions knowingly or unknowingly imposed by the researcher’ (Murphy, 2001:284) because the text analyst does not have to abide by any procedures, such pre-reading, coding, and pre-specifying categories or concepts to be identified in the text. Being independent from the researcher’s presuppositions, the software generated findings are more objective and valid, as a result. In the same vein, Tankard (2001) describes the researcher criteria, while approaching a text, as ‘vague’ and ‘subjective’ suggesting that this shortcoming can be addressed through ‘a systematic identification of linguistic elements and structural dimensions’ which can be made possible thanks to computer-assisted analysis (Tankard, 2001). When a researcher decides to adopt a corpus linguistic approach, the used procedures are not biased as computers are not biased when analyzing texts. McEnery (n.d) asserts that “[a computer] does not pick out certain things because it thinks they are interesting, or they confirm its own suspicions. […] they kind of direct us to things that we probably wouldn’t have thought of ourselves. And that is really good, because it stops us from being biased.” (McEnery, n.d). Therefore, CL software permits analysts to objectively confirm the extent to which the revealed linguistic structures during the qualitative analysis of the compiled texts can be generalized.
Another significant advantage of CADS is research reliability. As long as there is a fixed algorithm in the computer software, multiple researchers can replicate the process with the same texts resulting in the same findings. The reliability of digital text analysis is 100% (Shapiro, 1997). For example, Zih et al., (2021) and El Biadi and Zih (2022) examined manually the use of passive and nominal constructions in the news discourse of *The Times* online newspaper and compared their findings with the ones yielded by LancsBox and AntConc – corpus linguistics software- and found out roughly the exact findings with few differences that will be discussed in the limitations section. That is, if another researcher is interested in using a particular text analysis software to examine the frequencies of passives and actives in the same texts deployed in El Biadi’s and Zih’s study, the obtained percentages should be exactly the same.

Furthermore, Partington (2003) highlights the benefits of CL methodologies for discourse analysis:

At the simplest level, corpus technology helps find other examples of a phenomenon one has already noted. At the other extreme, it reveals patterns of use previously unthought of. In between, it can reinforce, refute, or revise a researcher’s intuition and show them why and how much of their suspicions were grounded. (Partington, 2003:12)

CADS approach offers the opportunity of not only observing the effect of language use by bringing to light several supporting evidence of the discursive features under question, but also presenting examples of counter-discourse to the predominant discourses which may not be exposed in a particular text. Therefore, the CADS analysis is a method of ‘triangulation’ as it combines the qualitative and the quantitative approaches to linguistic analysis (Baker, 2006).

Mautner (2009) advocates baker’s view asserting that

Corpus Linguistics software offers quantitative and qualitative perspectives on text data, computing frequencies and measures of statistical significance, as well as presenting data extracts in such a way that the researcher can access individual occurrences of search words, qualitatively examine their collocational environments, describe salient semantic patterns and identify discourse functions. (Mautner, 2009: 123).

The combination of CDA and CL allows for a systematic identification, quantification, and examination of various grammatical features, such as nominalization, passivization, pronouns, and metaphors etc. The fact that CL enables critical discourse analysts to analyze significantly larger data which could have been laborious and time-consuming if were done manually. O’Keefe (2012) conceives of CADS methodology for examining newspaper discourse as offering a ‘sharp analytical tool’, asserting that the digital availability of news texts combined with sophisticated processing features of CL software aids critical discourse analysts in their thorough analysis of an infinite number of discourses; CL has positively influenced and hugely contributed to the field of CDA, as a result. In their ‘RASIM’ project at the Lancaster University which examines discourses on refugees, asylum seekers and migrants in the UK press from 1996-2006, Baker et al. (2008) couple the discourse historical approach to CDA with corpus linguistics and reached the same conclusion quoted at length bellow:

CL, in general, and concordance analysis, in particular, can be positively influenced by exposure and familiarity with CDA analytical techniques, and the theoretical notions and categories of DHA can inform the qualitative CL analysis. Also, CL needs to be supplemented by the close analysis of selected texts using CDA theory and methodology. CDA, in turn, can benefit from incorporating more objective, quantitative CL approaches, as quantification can reveal the degree of generality of, or confidence in, the study findings and conclusions, thus guarding against over- or under-interpretation. (Baker et al., 2008: 297).
In addition to Baker et al.’s (2008) study that deploys this methodological synergy, a number of contemporary studies have also deployed a CADS approach to answer their research questions. These include but are not limited to Page’s (2003) examination of patterns of naming Cherie Booth/Blair – wife of Tony Blair – in British online and printed news reports, Baker and McEnery’s (2005) corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts, Caldas-Coulthard’s (2007) study of cross-cultural representations of ‘otherness’ in media discourse, Marchi and Taylo’s (2009) analysis of participants in UK and US news reporting of Iraq war, Taylor’s (2009) study of the representation of immigrants in the Italian press, and Freake et al.’s (2011) investigation of the construction of nationhood and belonging in Quebec popular discourse.

The aforementioned studies regardless of whether some of them have explicitly combined the critical discourse analytical framework with corpus linguistics to answer their research questions (e.g., Baker et al., 2008; Baker and McEnery, 2005; Caldas-Coulthard, 2007; Freake et al., 2011) or relied on a combination of a traditional discourse analytical approach with CL (e.g. Marchi and Taylor, 2009; Page, 2003; Taylor, 2009), they have all acknowledged the quantitative strengths of CL approach with the qualitative thorough analysis of discourse.

To conclude, the possibilities of combining critical analysis of discourse with corpus linguistics make of the CADS approach an appropriate supporting methodology for a number of studies, for it allows for identifying not only main patterns of discourse in a large corpus of texts, but also points out to counter-discourses. This methodological synergy helps facilitate the statistical quantification of the linguistic variables under study. While the CADS approach is relied upon to reduce the researcher subjectivity and the potential for bias along with assisting the processing of large amounts of linguistic data, it is not without its limitations which will be addressed in the following paragraphs.

4.2. Limitations of the CADS Approach for News Discourse Analysis

A number of criticisms have been levelled at CL namely in the areas of context, representativeness, and interpretation. To start with, Baldry (2000) and Widdowson (2000,2004) argue that CL does not account for context as corpora consider mostly language data, being them written or transcribed spoken data. By so doing, important elements, such as intonation, body language and other paralinguistic contextual data are ruled out, as a result. Corpus linguists have been criticized for ‘abstracting text from its context’ (Baldry, 2000:36) and undermining the societal conditions of production and consumption provided in a given corpus (Baker, 2006:18). Nevertheless, Stubbs (2001a) defends these claims asserting that “to accuse Corpus Linguistics of ignoring context strange, since it is essentially a theory of context: the essential tool is the concordance, where words are always studied in their contexts” (2001a, 156). In the same vein, Thornbury (2010) highlights the possibility of rigorous analysis of contextual information by means of using a smaller more localized corpus. While the use of corpus linguistics software, namely LancsBox and AntConc prove to be useful in identifying instances of passives and nominal constructions used in the news reports about Muslim and non-Muslim perpetrators, the software could not specify the exact instances where the passive and nominal constructions were used to report acts perpetuated by the social actors- Muslims and non-Muslims- making the manual analysis of the researcher and the qualitative critical discourse analysis even indispensable to carefully identify and interpret the linguistic features of interest. (Zih et al., 2021; El Biadi and Zih, 2022).

Furthermore, representativeness is another limitation of CL. On the one hand, McEnery et al., (2006) argue that a corpus only shows its contents providing no negative evidence. According to him, “this means that they cannot tell us what is possible or not possible. Everything
included in a corpus is what language users have actually produced. A corpus however large or balanced, cannot be exhaustive except in a very limited range of cases.” (2006:121). Widdowson (2000) shares the same view arguing that a corpus does not account for what is “the encoded possible, nor the contextually appropriate”, it rather documents what he labels ‘textually attested’ language (2000: 7). On the other hand, Stubbs (2001a) refutes this assertion arguing that corpus linguists do not claim that corpora provide evidence of every possible use of language in a particular context, but rather are concerned with a ‘much deeper notion: What frequently and typically occurs.’ (p.151). Therefore, researchers adopting CL should be wary when interpreting the findings of a particular corpus. Baker (2006) asserts that

It is important to bear in mind that the processes of production and reception of any particular article are complex and multiple. Newspaper discourse […] should not therefore be [always viewed as mainstream or hegemonic just because they have occurred within a newspaper. (Baker, 2006: 73).

Consequently, drawing conclusions based on corpora should not be dealt with incautiously, for “a statement about evidence in a corpus is a statement about that corpus, not about the language or register of which the corpus is a sample” (Hunston, 2002:23). Corpora in itself is not indicative of the various discursive functionalities of a particular discourse, it should be rather viewed as a ‘theory of the typical’ (Stubbs, 2001a:151) for making inferences about language use in a given context.

Finally, corpora have been accused of lacking interpretation which makes the researcher’s own explanation and interpretation critical. That is mainly because “corpora can yield findings but barely provide explanations for what is observed” (McEnery et al., 2006: 121). The conclusions resulting from corpus-based analysis require the human interference without which inferences and interpretations would not be possible, as a result. Yet, “researchers may choose to interpret a corpus-based analysis of language in different ways, depending on their position” (Baker, 2006:18). Such multifaceted interpretation might pose some difficulties at the level of lexical and grammatical features. Widdowson (2004) argues that ‘you cannot read off significance from text as if it were a simple projection of textual features’ (p.120). He also rightly comments that CL allows for a description of text as product, not discourse as process. (2000). On the contrary, Stubbs (2001a) counters back this criticism maintaining that potential biased interpretation cannot be read off many empirical disciplines, suggesting that “recognizing the problem obviously does not solve it, but it shows that Corpus Linguistics is trying to develop observational, empirical methods of studying meaning, which are open to the same tests as are applied in other disciplines”. Some of the many analytical tools offered by CL for extracting and identifying particular linguistic patterns are (collocation, keywords, frequency lists, clusters, dispersion plots etc.), and Baker (2006) insists on the significance of setting clear boundaries when deciding on a particular analytical technique to use:

For example, at what point do we decide that a word in a corpus occurs enough times for it to be “significant” and worth investigating? Or if we want to look for co-occurrences of sets of words […] how far apart are we going to allow these words to be? Do we discount cases where the words appear six words apart? or four words? (Baker, 2006: 21).

Clearly, researchers relying on CL should be wary of the precise methodological steps to follow and set clear-cut boundaries as far as the application of analytical tools offered by the corpora are concerned so as to the potential for excessive interpretative bias might be consciously avoided.
5. CONCLUSION
To conclude, Corpus-Assisted Discourse Studies (CADS) approach, which is a combination of a Critical Discourse Analytical approach with the techniques of Corpus Linguistics, is often viewed as a supporting methodological approach. The ample advantages that the CADS approach offer make of it a suitable framework to adopt so as to better address a particular study objective. This methodological synergy is believed to help identify and thoroughly analyze salient linguistic patterns in a given body of texts. CADS framework has been widely adopted by critical discourse analysts for its flexibility and the facilities it offers to minimize if not to elude the researcher’s bias and subjective interpretations along with the possibility of processing a larger corpus of news reports. Regardless of the criticisms which have been levelled against it, critical discourse analysts are mindful that no single methodological approach is without pitfalls. This means that a careful use of the analytical tools offered by the new information technology along with a thorough qualitative analysis of the linguistic variables of interest will yield more well-grounded and reliable conclusions.

References


