Baker, P. & McGlashan, M. (2020) 'Critical Discourse Analysis'. In: Adolphs, S. & Knight, D. (Eds.) The Routledge Handbook of English Language and the Digital Humanities. London: Routledge. pp. 220-241.

Critical Discourse Analysis

Paul Baker and Mark McGlashan

INTRODUCTION

Critical Discourse Analysis (CDA) is a methodological approach to the analysis of language in order to examine social problems, with a focus on power, particularly issues around abuses of power including discrimination and disempowerment. Analysis considers the language in texts but also considers the texts in relationship to the wider social context in which they are produced and received. Many critical discourse analysts carry out close readings of a small number of texts, focussing on how linguistic phenomena like metaphor, agency, backgrounding, hedging and evaluation can help to represent various phenomena or present arguments for different positions. This chapter focusses on how automatic forms of computational analysis (in this case, corpus linguistics) could be used in combination with the kind of 'close reading' qualitative analyses associated with CDA, particularly in cases where analysts are working with large databases containing hundreds or thousands of texts.

The type of database that this chapter focusses on is online news. News is particularly relevant for studies in CDA as it tends to involve reports of social issues and the way that news is reported can have important consequences for societies. Van Dijk (1991) argues that newspapers influence public opinion, whereas Gerbner et al. (1986) contend that the media's effect on an audience compounds over time due to the repetition of images and concepts. Searchable online databases of news articles like Lexis Nexis and Factiva have enabled researchers within the digital humanities to quickly gain access to vast datasets of stories around particular topics. Users can conduct targeted searches for articles from certain newspapers or time periods, as well as specify words that articles must contain. Such databases have been used effectively in critical-oriented research, e.g. on gay men (Baker 2005), refugees and asylum seekers (Baker et al 2008), and Muslims (Baker et al 2013). By employing software to identify linguistic frequencies in corpora (essentially large representative collections of electronically encoded texts), it is possible to make sense of data comprising millions of words. This enables generalisations to be made of the whole, as well as directing researchers to typical (or atypical) cases so that more qualitative, close readings can be carried out. The three studies mentioned above just focussed on the articles themselves, although more recently, McEnery et al (2015) examined both news articles and discussion on the social media platform Twitter relating to the ideologically inspired murder of the British soldier Lee Rigby.

In the last two decades the way that people engage with print news has changed; as (access to) technology has changed, so to have practices. People are now able to access information (including news) more readily on mobile and internet-connected devices, which has ‘ shifted the roles traditionally played by newspapers, television stations, [etc.]’ (Westlund and Färdigh 2015), and online news now offers an alternative to print (Thurman 2018).With these new technologies, breaking news can be almost instantaneously reported and, apart from a small number of news outlets which charge subscriptions for access to their content, can be accessed largely free of charge. Newspaper editors are able to gather large amounts of data about their readers in terms of which links they click on and how long they remain on a web page (as well as other forms of ‘customer analytics’). Such information can then be used in order to refine news content to prioritise the sorts of stories that readers are most likely to read. However, perhaps the most marked change in news production and consumption involves the fact that readers can publish their own reactions to stories, as well as interact with others, creating dialogue between the news producer and other readers. As most user comments appear after the article itself, the phrase ‘below the line’ is sometimes used as a shorthand to refer to this user-generated discussion.

In line with work by Chovanc (2014), we argue that reader comments have the potential to alter the way that news articles are engaged with and understood. They offer readers the opportunity to engage (passively or actively) within a community of practice (cf. Wenger 1998) which involves other readers, and to learn about different types of responses to news stories. It cannot be argued that reader comments represent the most typical views of the public or even that they are typical of those who read the article as people who take the time to comment on articles may have stronger opinions than most, or may have other motives (such as trolling other readers). However, we argue that the comments have the potential to influence opinion, particularly if a large number of comments put forward the same argument across multiple articles on the same topic.

Therefore, an aim of this study is to acknowledge that digital engagement with news involves an additional dimension, one that is important to consider when carrying out analyses of the news articles themselves, and one which the digital humanities is well-placed to examine – user generated comments.

In this chapter we first briefly describe critical discourse analysis, its aims and methods and how it can be utilised on large collections of digital texts or corpora. We then describe the focus of our research, the analysis of newspaper articles and reader comments taken from the British newspaper the *Daily Express* which refer to Romanians. After outlining why this topic is of relevance to critical discourse analysts and providing information relating to the social context of the topic, we list our research questions and discuss how the corpus was collected.

This is followed by an illustrative analysis of the corpus data, applying the keywords technique to identify salient words in the articles and comments, and then using a second technique, collocation to gain an idea about how particular keywords are used in context. To illustrate how our method works we focus on a small number of keywords (*Romanians*, *them* and *us*), identifying collocates of these words which help to give an impression of how they are used in context and contribute towards ideological positions. We then show how the tool ProtAnt can be used in order to rank the news articles in order of prototypicality, helping us to focus on a 'down-sampled' set of articles which are the most representative of the entire set. We are thus able to carry out a close qualitative analysis of the language in a single prototypical article and then compare its content against the reactions of commenters. The chapter concludes with a discussion of potential issues surrounding taking a critical approach within the digital humanities, and consideration of future directions for this field.

BACKGROUND

Critical discourse analysis (CDA) is an 'interdisciplinary research movement' interested in examining 'the semiotic dimensions of power, injustice, abuse, and political-economic or cultural change in society' (Fairclough et al., 2011: 357). Wodak and Meyer (2009: 10) define CDA as aiming 'to investigate critically social inequality as it is expressed, constituted, legitimized, and so on, by language use (or in discourse)'. Developing from the 'East Anglia School' of critical linguistics (Fowler et al 1979), which involved analysis of texts in order to identify how phenomena are linguistically represented, and was strongly influenced by the Frankfurt School, particularly the critical work of Habermas (1988), CDA sees discourse as semiotic and social practice (Fairclough 1992, 1995, 2003), constituting both modes of social action and representation. Where CDA differs from critical linguistics is that analysis goes beyond the text to consider various levels of context, including the processes of production and reception around the text, the social, political, economic and historic context in which the text occurred, and the extent to which texts contain traces of other texts through direct quotation, parody or allusion (intertextuality, cf. Kristeva 1986) or involve interdiscursivity (e.g. the presence of discourse associated with advertising might be found in an educational text, cf. Fairclough 1995).

A number of different 'schools' or approaches to CDA exist, each with different (and sometimes overlapping) theoretical and analytical foci and tool-sets e.g. Fairclough's (1992, 1995, 2003) dialectical-relational approach, Reisgl and Wodak's (2001) Discourse-Historical Analysis, Van Dijk's (2006) socio-cognitive approach and van Leeuwen's (1996, 1997) focus on social actor representation. Within these schools there is no single step-by-step guide to analysis, but research questions and data-sets can influence the analytical procedures which are selected and the order in which they are carried out.

A relatively new form of CDA – corpus-assisted critical discourse analysis – was conceived as a way of handling large collections of texts. This approach, which draws directly on theory and methods from corpus linguistics (CL), can be seen as being directly related to the field of digital humanities. Jensen (2014) argues that, 'corpora are *per se* digital [and that] building a corpus is inherently a DH [digital humanities] project' but that 'CL is on the fringes of contemporary DH, which is itself currently on the fringes of the humanities'. The exploration of this method here, therefore, attempts to position corpus assisted critical discourse analysis as a relevant (although potentially fringe) approach to DH scholarship. Moreover, corpus-assisted critical discourse analysis enables analysts to draw their observations from a wide range of texts rather than 'cherry-picking' a few texts which prove a preconceived point, thus attempts to mitigate researcher bias (Widdowson 2004). Important to this approach are corpus techniques based on frequency, statistical tests and automatic annotation (e.g. of grammatical categories), which enable frequent linguistic patterns to be identified within texts while concordance tools allow a qualitative reading to be carried out across multiple occurrences of a particular pattern.

The model proposed by Baker, et al. (2008) advocates moving forwards and backwards between corpus techniques and close reading in order to form and test new hypotheses. The analysis is supplemented with background reading and the sorts of contextual interrogation typical of other forms of CDA. In dealing with large sets (which we define as any set which is beyond the capabilities of the researcher or research team to carry out a close reading of every text) of electronically-encoded texts, this approach is particularly germane to the digital humanities and in the analysis which follows we take a corpus-assisted approach to both identify large-scale patterns in corpora of newspaper articles and readers' comments, and to enable us to focus on a smaller number of prototypical articles for a more detailed close reading. In the following section we outline the topic under investigation and explain why it is relevant to take a CDA approach.

Background and Data Collection

On 23 June 2016, a national referendum was held across the UK to decide whether the country should 'Remain' in the political and economic union of member states that are located mainly in Europe, known as the European Union, popularly referred to as Brexit (British Exit). Over 33 million votes were cast, with the 'Leave' vote winning by a narrow margin (51.89% to 48.11%). While the split decision indicated a general feeling of disunity, the reasons given for people wanting to leave suggested more specific social problems. For example, in *The Guardian*[i], Mariana Mazzcato argued that people felt aggrieved by the government's programme of austerity to reduce borrowing since the global economic crash in 2008. On the other hand, the *Daily Mail*[ii] reported that Brexit was blamed on the German chancellor Angel Merkel's 'open-door approach' to immigration. A poll of over 12,000 people carried out by Lord Ashcroft (a former deputy Chairman of the Conservative Party who now works as a major independent public pollster of British public opinion), after the vote gave the top three reasons for voting Leave as ‘decisions about the UK should be taken in the UK’, ‘the best chance for the UK to gain control over immigration and its borders’ and ‘remaining meant little or no choice about how the EU expanded its membership or powers’.[iii]

Clearly, concerns about immigration were an important factor in the decision, and in this indicative study we investigate how attitudes towards immigration may have been influenced by the ways that the media constructed immigration in the years leading up to the Brexit vote. Prior to this research, a large corpus-assisted CDA research project focussed on the words *refugee*, *asylum seeker*, *immigration* and *migrant* (and their plurals) in the national British press between the years 1996 and 2005 (see Baker et al 2008, Gabrielatos and Baker 2008). Our study takes a different approach by examining the representation of a single national identity group, Romanians. This group was chosen because Romania was a relatively recent entrant to the European Union, joining in 2007. Romania’s entrance into the EU was not straightforward, indicating that there were concerns about its membership in some quarters. For example, the EU had imposed monitoring of reforms based on the rule of law, judicial reform and the fight against corruption, as well as putting in place a transitional cap on migration meaning that Romanians were not able to become resident in the UK until the beginning of 2014 (unless they were self-employed or worked in seasonal jobs). The change to the law resulted in headlines such as ‘Sold Out! Flights and buses full as Romanians head for the UK’[iv], a story which was later found to be untrue and amended.[v] These existing indicators of concern around Romanian’s entry to the EU make it a particularly relevant country for examination of media discourse. However, unlike most previous corpus-based work on press discourse, this study also takes into account readers’ comments on the online versions of articles. The online comments act as type of proxy for reader reception and thus fit well with CDA’s remit for considering different forms of context around texts.

Our research questions were:

1) How were Romanians typically linguistically represented by the *Daily Express* in the period up to Brexit?
2) What were the differences/similarities in linguistic representation between articles and reader comments?

To answer these questions, we carried out a corpus-assisted critical discourse analysis of the corpora of articles and comments, as well as conducting a qualitative analysis of the articles and comments identified as most representative of the whole.

As this is a relatively small-scale study aimed more at demonstrating principles and techniques of analysis we have chosen to analyse just one newspaper, *The Daily Express*, a right-leaning or conservative tabloid.[vi] We collected data from approximately a decade-long period (24th July 2006 until 23rd June 2016 which was the day of the Brexit vote). The articles and their corresponding comments that form the data were identified using the search functionality of *The Daily Express* website (http://www.express.co.uk/search) which allows users to search all online pages that contains a particular search term and gives several methods for refining and ordering searches by date, time, section, and headline. The search used here returned a list of 1,945 articles published online during 'All Time' and containing the search term 'romanian'. This list displays a headline, sub-headline, image, paper section and time/date of publication referring to each page, which, when clicked, uses a hyperlink to navigate to the appropriate page. It would be possible to manually click through the entire list and collate, by copying and pasting, all of the articles and comments sections for each page. However, this process would be arduous and open to human error. Instead, computational methods were implemented to harvest data automatically from these pages – an activity typically referred to as 'web scraping' – using the Python[vii] programming language and the

Python libraries Beautiful Soup[viii] and Selenium.[ix]. Beautiful Soup enables users to parse and scrape data from the HTML code that from which webpages are made and Selenium gives the ability to simulate user behaviour (e.g. scrolling and clicking links) and automate activity in web browsers such as Mozilla Firefox[x] and Google Chrome[xi]. Concerning ethics, this approach simply emulates behaviours that could be carried out manually and only targeted data published for free, public, online consumption. No data collected were shared or redistributed in any way and no personally identifiable information (e.g. comment/article author names/handles) were targeted for collection. The web scraping programme developed for this data collection activity first identified the hyperlink for each of the 1,945 pages returned, assigned a number to each page and then navigated to each page in turn. The programme then extracted the HTML code for the page that it navigated to and identified and saved the headline, sub-headline, and the body text for the article on that page into a text file. If an article had any comments, they were saved into another text file. The files for articles and comments were assigned the same number as the page from which they were extracted and saved into separate 'articles corpus' and 'comments corpus' folders. The process is outlined graphically in Figure 1.

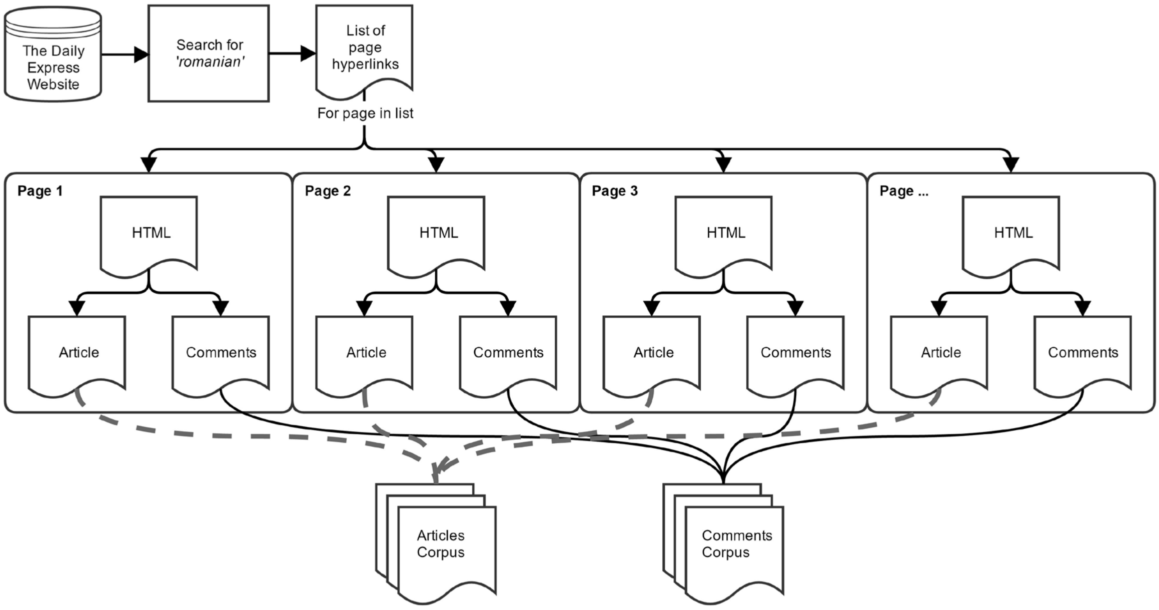

Figure 1: process for extracting articles and comments corpora

As data was collected during July 2016 articles published after the sampling cut-off date of 23rd June 2016 were also returned. 1,929 articles remained in the sample following the removal of those articles published after the cut-off date from which an articles corpus (771,878 words) and comments corpus (2,166,148 words) remained.[xii]

Programming languages like Python and web scraping technologies provide important automatable methods for researchers in the digital humanities. They enable huge amounts of data such as the corpora described above to be collected relatively quickly. However, use of these methods requires an acknowledgement of several issues. *Vis-à-vis* time, learning how to use programming languages (even for *ad hoc* purposes) is a lengthy process and the data collection itself is not instantaneous – it can take several hours to collect corpora such as those used in this study. Time is also a factor when considering data 'cleaning'. The data collected for this study, for example, required manual investigation to check that the program collected what was required. In the case of the user comments there was an issue where the first few words in a comment were repeated and followed by "\r\r" (e.g. 'Strange how the police will act when the prob\r\rStrange how the police will act when the problem affect the rich and powerful!'). This error was likely due to a redesign of the way in which comments were displayed on *The Express* website but the repetition of words would have adversely impacted on how word frequencies were calculated which is fundamental to corpus linguistic studies. An additional script had to be developed to automatically identify and delete these instances of repetition.

Another issue concerned annotation. As the contents of files in the articles and comments corpora contained different kinds of data, different annotation schemes had to be developed to delineate different parts of a document. For example, in the articles corpus, headlines were placed between tags similar to those found in XML documents, e.g. "<Headline> EU workers rushing to get jobs in Britain ahead of potential Brexit </Headline>"; sub headlines and the body text of an article in these files were annotated similarly. In the comments corpus, a similar annotation scheme was used to simply distinguish one comment from another in the order in which comments were posted on the article from most to least recent. However, some comments were nested as replies to other comments, creating a conversation hierarchy. One issue with the annotation scheme used was that this hierarchy was not preserved, thus potentially decontextualizing some comments from their original conversational structure.

Analysis

First, the tool AntConc version 3.4.4 (Anthony 2016) was used in order to obtain frequency lists of the datasets and obtain keywords. This tool has been chosen for analysis because it is free and relatively easy for beginners to learn how to use. A potential disadvantage is that it does not offer as wide a range of processes or statistical tests as other tools like WordSmith

Tools[xiii], #LancsBox[xiv], CQPweb[xv] or SketchEngine[xvi] but for the purposes of introducing new analysts to the field, it is useful. Keywords are words which appear in one corpus statistically significantly more often than another, and are useful in identifying repeated concepts which may be overlooked by human analysts. They can help to focus analysis on a small set of salient words which are then subjected to a more detailed qualitative interrogation. Using the keywords facility of AntConc, comparing a wordlist of the articles against the comments (and *vice versa*) produces words which are more frequent in one corpus as opposed to another, which is useful for identifying the main lexical differences between two corpora, but will overlook similarities. To focus on both difference and similarity, both corpora were separately compared to a third dataset, a reference corpus of general written published British English (the 1 million word BE06, see Baker 2009). The BE06 corpus offers a reasonably good reference for the way that written British English was used around the period that the news articles and comments were written. A potential issue is that it does not include informal Computer Mediated Communication (CMC) so comparing it against comments could result in keywords appearing that are more concerned with the register of CMC (such as the acronym *lol*) as opposed to the topic of immigration. However, this concern proved to be unfounded and none of the keywords obtained from the comments corpus were due to the aspects of the CMC register.

In AntConc keywords are calculated by performing log-likelihood tests, taking into account the relative frequencies of all the words in the two corpora being compared as well as the sizes of each corpus.[xvii] Table 1 shows the 50 keywords which have the highest log-likelihood scores (in other words, these are the most 'key' words) for both the articles and comments.[xviii]

Categories	Articles	Comments	Both
Politics	David	government, Nigel, politicians, vote	UKIP, Cameron, Labour, Farage
Nationality	Eastern, European, Bulgaria, Bulgarian, Bulgarians, Roma, Bucharest, country	country	EU, Romania, Romanian, Romanians, Britain, British, UK, Europe, countries
Movement of people	influx, migration, migrant		migrants, immigration, immigrants
Football	champions, Chelsea, Manchester, game, win, league		
Grammatical words	after, who, against	they, we, our, all, you, this, them, us, here, why, what, if, no, not	
Other adjectives	false, last	just	
Other verbs	said, has	come, want, get, should, have, do, are, is, stop, don’t	will

Other nouns	workers, police, express, year, restrictions, million	people	benefits, racist

Table 1. Keywords in articles and comments

For the top 50 keywords in each corpus, 19 are shared, while each corpus has 31 unique keywords. While it would have been possible to categorise the keywords according to say, an existing automatic semantic tagging system like USAS, this would have focussed on dictionary definition meanings of words. A more nuanced categorisation scheme, based on the context in which the words were used, was more helpful in terms of identifying the themes that the articles accessed (see Baker 2004). For example, the keyword *David* could have been classified as a proper noun or a male name using an automatic categorisation scheme. However, for the purposes of our analysis it is more useful to know that *David* almost always refers to the Prime Minister David Cameron and thus indicates discussion of Romanians in a political context. Therefore, the set of categories was created by subjecting the keywords to concordance analyses (involving reading tables containing citations of every keyword – see Table 2 for an example of a sample of some of the mentions of the word *influx* of the total that were examined) and then categorised into groups based on theme. This is an admittedly subjective approach, and in such circumstances it is useful to have more than one analyst (which is the case here) to work on the scheme.

Highlighting the impact of such an	influx	on the NHS, Vote Leave forecast A&
the whole truth about the annual	influx	of newcomers. Alp Mehmet, vice chair
find it harder to cope with an	influx	of children who do not speak English.
when their neighbourhood sees a large	influx	of migrants. While that might seem

people whose neighbourhoods see an	influx	of migrants tend to be made unhappy
workers into the town. The huge	influx	began in 2004 when Britain opened its
came into the country during the recent	influx	from Syria of being the culprits. Town
benefit pledge and reduce the foreign	influx	. Denying that the Prime Minister has
"Others voiced concerns that the huge	influx	of people would put additional strains
desperate bid to control huge migrant	influx	. TANKS and riot police have been

Table 2. Concordance table of 10 citations of the keyword *influx* in the Articles corpus.

This resulted in four main themes: Politics, Nationality, Movement and Football. In most cases categorisation into these themes was easy as words were used unambiguously (as Table 2 suggests, *influx* normally referred to movement of people). One word which was slightly more difficult to classify was *Manchester*. Concordance analysis revealed that 95% of cases referred to football teams (Manchester City and Manchester United). However, a small number of cases referred to the city itself (e.g. Car cleaner Adrian Oprea began selling the Big Issue when he came to Manchester from Romania in 2009 with his wife and son.) As such cases were a clear minority, the word *Manchester* was categorised as referring to the theme of football.

Not all words could be clearly categorised into themes, so a category of closed-class ‘grammatical words’, consisting of pronouns, determiners, prepositions, wh-questions and negators was also created. This left a smaller set of adjectives, verbs and nouns which appear at the bottom of the Table 1. Some of these words are suggestive of themes too (e.g. *police* could be put in a new category of ‘Crime’, or *workers* and *benefits* in ‘Economics’) but this is a preliminary stage of analysis and was carried out in order to understand the themes underlying reference to Romanians which are indicated by multiple words referencing the same concept. The categorisation also helps us to focus on similarities and differences between the two corpora. For example, it is notable that the words about Football are only keywords in the news articles, not the comments, indicating that commentators do not seem to be as interested in responding to articles which refer to Romanians in the context of football.

When *The Express* does refer to Romanians in the context of football it is usually to mention a team or player e.g.:

> Romanian side CFR Cluj-Napoca, in Group F along with Manchester United, notched a 2-0 win away to Braga (Article 1191)

The articles about football do not discuss Romanians in the context of immigration and for that reason will not play a large role in our analysis. However, it is worthwhile discovering their existence because they indicate that *The Express* did not always discuss Romania in the context of immigration, and also that these articles do not tend to provoke much reader comment.

A full keyword analysis would involve taking the words in Table 1 in turn and subjecting each to a more detailed analysis of context. This means examining a word's collocates (words which frequently occur near or next to one another, usually more than would be expected by chance), as well as concordance lines, in order to obtain an impression of what a word is being used to achieve, particularly in relation to the research questions asked. For illustrative purposes, we have selected a smaller number of words from Table 1 for this kind of analysis. These words are *Romanians*, *us* and *them*.

Romanians

It is unsurprising that *Romanians* is a keyword - it actually appears as the focus of Research Question 1 and was also one of the search words used in creating the corpus. Due to space restrictions, we focus our analysis on this word, although note that it would be useful to consider related word forms like *Romania* and *Romanian*. As a plural noun, *Romanians* is an example of collectivisation, a type representation that involves assimilation of individuals into a group (see van Leeuwen 1996). Ideologically, collectivisation can be relevant to focus on as it can result in differences between individuals being obscured and generalisations being made. The word *Romanians* occurs 890 times in the articles and 1,433 times in the comments. It is notable that in the comments *Romanians* is more frequent than *Romania* and *Romanian*, while in the articles it is less common than those words, indicating a potential difference between journalists and readers – the readers appear more likely to engage in this form of collectivisation. As a way of obtaining an impression of how *Romanians* is used in context, we can examine its collocates in both corpora. We calculated collocates using the Mutual Information (MI) statistic (which shows the strength of a relationship between two words). We considered all words 5 words either side of *Romanians* which had an MI score of 6 or above (following Durrant and Doherty (2010: 145), who had carried out experiments on human subjects and indicated that this was a cut-off where collocates became 'psychologically real'). As the MI measure can produce collocates that are of low frequency (and are thus not very representative of the corpus as a whole), we stipulated that a pair must occur at least 5 times together before we could consider it as a collocate.

Table 3 shows the collocates of *Romanians*.

Category	Articles	Comments	Both
Quantification	percent, estimates, estimate, eight, number, tens, nearly	hundreds	
Movement	arriving, settle, gain, arrived	enter, moved, arrive	
Nationalities and groups	Lithuanians, group	Latvians, Indians, Indian, Bulgarian,	Bulgarians, Poles

		Roma, gipsies, gypsies, gypsy, Africans	
Law and Order	visa, loophole, attacks, gang, curbs, restrictions, controls	legally, steal, stealing, honest, decent, metropolitan	arrested
Homelessness	rough, homeless	begging	sleeping
Time	night, minutes, yesterday, temporary		January
Grammatical words		between	
Images		picture	
Reporting Verbs	revealed, showed, warned	showing	
Misc Verbs	lifted, expected, seeking, applied, imposed, gain	meant, met	
Others	marble, arch, league, abroad	errr, um, neighbours, difference, educated, guys	

Table 3. Collocates of *Romanians*.

As with the keywords, collocates of *Romanians* have been grouped into thematic categories. While not an analysis in itself, this helps the analyst to spot repeated associations and also differences and similarities between the two corpora. For example, it is notable that both articles and comments tend to quantify Romanians and focus on their movement e.g.

> The most reliable estimates have always come from the Migration Watch think tank. It **estimates** that about 50,000 **Romanians** and Bulgarians will arrive on our shores every year for the first five years when we lift border restrictions. If they are correct that will mean 250,000 Romanians and Bulgarians arriving. (Article 981)

> We are just over 4 months away from "Entry Day" when potentially tens of thousands, sorry I meant **hundreds** of thousands of **Romanians** & Bulgarians can legally enter Britain and the three main party leaders haven't a clue what to do. (Comments, 854)

The comment above was noteworthy because it appeared eight times in the comments corpus, all in the same comment thread. It appears that this comment had been made repeatedly by the same person. This was an unexpected feature of the comments data, where we found that some collocates were the result of repeated comments. Not all cases involved repetition to the same thread. For example, in Table 3 the collocates *errr* and *um* with *Romanians* were the result of the following comment:

> Interview goes along these lines....Hello "ANY" MPs. Are these people Romanians or Roma gypsy's?..... errr well um.....errr we welcome Romanians & Bulgarians & i'm off to the airport to have a photo shoot.

This comment appeared six times across four different threads, made by the same commenter, suggesting that it had been cut and then pasted multiple times. Analysis of clusters (e.g. fixed sequences of words occurring multiple times) revealed a small number (18) cases across the corpus where the same comment appeared at least 5 times. Such cases ultimately do not detract from the overall findings of the analysis, and where collocates could be attributed to such cases, they should be noted, but they do indicate a way that individuals can take advantage of the affordances of digital data by ensuring that a particular message they want to convey reaches a wider audience. Obviously, if we encounter such cases, we may want to indicate that a repeated pattern is not necessarily widely shared among a discourse community but more due to a single commenter with a particular axe to grind.

Some collocates suggest a different sort of focus between the news reports and comments. Consider *steal* and *stealing*, which are collocates of *Romanians* in the comments corpus. Collectively, they occur 15 times with *Romanians* (15/1,433; a relative frequency of 0.01). A detailed analysis of concordance lines revealed 5 cases that refuted the idea that Romanians steal e.g. 'There are many honest Romanians who work and do not steal' while the other 10 constructed Romanians as stealing. Even within this distinction, there were differences, with Romanians accused of stealing jobs, EU funds, UK benefits and in one case, a lawnmower from someone's garden.

One of the comments which challenges the idea of stealing Romanians purports to be from someone who identifies as Romanian:

> And by the way you got so much youth unemployment because you`re lazy and have so big requirements that nobody wants to work with you. Show me a brit girl who would wipe your elders **** when they can`t do it by themselves. Show me a brit mopping the streets of your cities, show me a brit doing all these kinds of low esteem jobs and then come tell me about **ROMANIANS stealing** your **** jobs. (747)

The analysis indicates that a collocational pair like *Romanians* + *stealing* does not indicate a straightforward 'discourse position' but instead reveals a range of potential stances. This is important to note, although we should also indicate that the more popular stance is one of Romanians as thieves, and this is something that commenters have to orient towards. On the other hand, the articles do not directly refer to Romanians stealing by using these collocates (there is not even one case where *steal* or *stealing* appears near the word *Romanians* in the articles corpus). With that said, the words *steal* and *stealing* do occur within the news articles (in positions other than near the word *Romanians*), a total of 122 times (an average of 0.06 times per article), and they do typically refer to people from Romania involved in crime.

Additionally, in the articles, other law and order collocates exist like *gang*, *curbs*, *restrictions* and *controls*. The latter three words involve citations of *The Express*'s stance that there should be restrictions on Romanians arriving or working in the UK. This often involves quoting from others who are in favour of such restrictions.

> However, critics fear a fresh wave of migration next year when visa **restrictions** on millions of **Romanians** and Bulgarians are lifted. (Article 1037)
>
> The efficiency with which the police destroyed huts and tents – using a digger to crush caravans – led some observers to question why British authorities can't be as ruthless. Britain, like France, has transitional **controls** on **Romanians** seeking to settle in the UK. Until next year, only Romanian migrants who have a job, or can support themselves, are allowed to stay in Britain. (Article 1202)
>
> Warning over 'flood' of immigrants… Temporary **curbs** were imposed on **Romanians** and Bulgarians in 2005 to protect the British labour market. (Article 1109)

Gang clearly contributes towards the discourse of Romanians as criminals with 7 of 9 instances referring to crime e.g. 'A gang of Romanians has been jailed for a total of 24 years after a string of crimes across the UK' (Article 1181). The other 2 cases refer to *gang* negatively so could more indirectly contribute towards an association of Romanians with crime. Article 136 has the headline 'POLICE raiding a suspected slavery gang found 40 Romanians crammed inside a three-bedroom house'. Reading the article, the Romanians are constructed as victims of the gang, whose nationality is not specified. The ninth instance involves a case of a couple whose home was taken over by squatters. It contains the line ‘When Samantha asked the police if the gang were Romanians on benefits she was ticked off for being “racist”.’ The article refers to Romanian squatters elsewhere and is sympathetic towards Samantha, putting *racist* in distancing quotes and implying that the police were not doing their jobs properly.

Not all the Law and Order collocates of *Romanians* are negative. Consider *honest* and *decent* which are collocates in the comments corpus. These cases involve descriptions of Romanians as *honest* (11 cases) and *decent* (12 cases). However, closer reading indicates that these cases tend to make a distinction between Romanians and gypsies. The following example (which also appears to be written by someone who identifies as Romanian), is a response to news article 386 which describes a television programme about a car thief called Mikki. The article does not refer to Mikki as Romanian, although it does appear in the comments section. One commenter writes

> He is not a Romanian! He is a GYPSY - from Romania. Gypsys came from INDIA few hundred years ago and we cannot get rid of them. They are doing the same job in Romania, they steal from **Romanians** also. We dont want them but we are forced to live with them. They cannot be integrated or sent to school. They are raising 6-7 a family and send them to beg. PLEASE CONTACT THE POLICE AND SEND HIM TO PRISON. THEY ARE A DISGRACE FOR THIS COUNTRY. 90-95 % OF THEM. Real **romanians** have **honest** jobs and they work very very hard to earn a living,but this scumbags know only to steel,cheat, rob, bare knuckle fights. Dont judge Romania for this GYPSYS. Ask a romanian about this GYPSYS - they will tell you the same thing as me. And then ask brits about a native Romanian - you will hear good things.

Indeed, this comment contains other words which were identified as collocates of *Romanians* like *gypsys*, Furthermore, a closer look at Table 3 reveals collocates like *Roma* and *Indians* which also appear in comments which attempt to separate ‘honest, decent’ Romanians from the more negatively constructed gypsies. E.g.:

> It is important to be informed before forming an opinion. Bulgarians and **Romanians** work, they don't sleep on the streets. Only *gypsies* do as they don't have dignity and don't care where they live, as long as there is something to steal because this is what they call "work". (Comment 517)

This is a distinction which is made repeatedly in the comments section but does not appear to be so clear in the newspaper articles. The term *Roma* does occur in the articles, along with *gypsies*, although these terms are not frequent enough to be key. An illustrative case is article 45 which has the headline ‘Romanian gang 'use fireworks to spark TERROR on Paris Metro - then pickpocket the victims'’. The body of this article does not use the word *Romanian(s)* again but instead refers to the gang as ‘a group of Roma women’ and ‘Roma pickpocket gangs’. From a critical perspective a number of issues arise here – the difference between Romanians and Roma gypsies, which appears to be highlighted in comments as more important than some newspaper articles appear to be acknowledging, and whether the more negative constructions of Roma gypsies are actually fair. However, at this point we will turn to look at some other keywords.

Pronoun use

One difference between the two corpora was that the comments corpus contained several personal pronouns as keywords: *they*, *we*, *our*, *you*, *them* and *us*. To an extent this is to be expected as the commenting register is more colloquial and interactive than the news reporting register. However, pronouns can also have ideological functions; for example, they can help to create a shared sense of identity between a writer and reader, or they can construct in and out groups. For this reason, we have decided to examine two key pronouns, *them* (9873 occurrences in the comments corpus) and *us* (5944 occurrences in the comments corpus). We are particularly interested in the sorts of actions that are described as being carried out on these pronouns, so using the same settings for collocation as described above Table 4 lists the verb collocates of each.

Keyword	Verb collocates
us	foisted, enslave, prevents, defraud, offload, forbids, outnumber, costing, betray, enlighten, warn, telling, inflicted, sponge, fleece, leech, deceive, betraying, dictate, treating, ripping, tells, denying, forgiven, enrich, store, bleed, hates, rob, dragging, bust, threaten, swamp, imposing, betrayed, sold, drag, tell, forcing, invade, imposed, forgive
them	carving, clothe, packing, shoving, bribe, send, rounding, offends, punish, escort, forgive, chuck, deport, turf, ship, educate, smuggle, shove, prosecute, educating, deporting, persuade, invite, shelter, shipping, sending, waving, encourages, handing, kick, lock

Table 4. Verb collocates of *us* and *them* in the comments corpus.

The verb collocates of *us* suggest a number of discourse prosodies (Stubbs 2001). The concept of discourse prosody relates to the ways that words can repeatedly co-occur in contexts which suggest a negative or positive evaluation. For example, Sinclair (1991: 112) has observed that ‘the verb *happen* is associated with unpleasant things – accidents and the like.’ While the word *happen* does not appear to imply anything negative in itself, its prosody would indicate that people who encounter the word in text are likely to be primed to interpret what follows as likely to be negative.

For the pronoun *us* in the comments corpus, the words *foisted*, *inflicted*, *imposing*, *forcing* and *imposed* suggest that some form of unwanted action has been carried out on 'us'. These verbs refer to immigrants or EU-related laws which relate to immigration:

> Don't forget that we have 'gained' millions of unwanted and in many cases, worthless immigrants who were **foisted onto us** by the EU. (123)

> The dire effects of this forced multiculturalism **inflicted on us** by the cultural Marxists has already changed this country beyond recognition (747)

> Britons are seeing what has bee **imposed on us** from London and Brussels. ENOUGH IS ENOUGH, , NO MORE MIGRANTS.

Another (negative) discourse prosody involves burden, particularly relating to Romanians or immigrants in general, and involving verbs which operate as metaphors: *costing*, *sponge*, *fleece*, *leech*, *bleed*, *ripping*, *rob*:

> 80% of immigrants are just here to **fleece us** and moan in the process (37)

> we as a country are fighting for survival against this rabble who want to come here and **sponge of us** (231)

> Another mongrel immigrant to rob, rape, groom and **leech off of us** (651)

> They are coming here to **bleed us** dry because we are dull enough to give away FREE everything! (203)

A third (negative) discourse prosody involves a sense of being overwhelmed or attacked: *outnumber*, *invade*, *swamp*, *enslave*, *threaten*:

> Two faced Germans nothing changes they could not **enslave us** so they are buying us instead (231)

> The whole idea is to **swamp us** so that we lose our identity and they can rule us for ever. (478)

And there is a sense of being coerced or lied to, not by Romanians but by those who are seen as elite decision makers: *deceive*, *betray*, *betraying*, *betrayed*, *defraud*, *dictate*, *forcing*:

> And isn't it strange how our news channels never actually feature the depraved and disgusting behaviour of some of these immigrants? Presumably so our treacherous politicians can continue to lie, **deceive and betray us**, and keep pretending they are a great asset to our economy. (617)

> The EU serves to **dictate** to **us** what we can and can't do, whether we like it or not. (14)

How about the verb collocates of *them*? The negative discourse prosody containing the most collocates involves some form of assisted movement: *shoving*, *send*, *rounding*, *escort*, *chuck*, *deport*, *turf*, *ship*, *smuggle*, *shove*, *deporting*, *invite*, *shipping*, *sending*, *waving* and *kick*. It is notable that these collocates do not appear to refer to Romanians specifically but more generally involve immigrants coming to the UK, particularly asylum seekers and undocumented migrants (this was also the case for the subject of the verb collocates above like *sponge*, *bleed* and *fleece*).

> **Turf them** out, **send them** back we are sick to death of these scroungers. (193)

> **Ship them** back to turkey regardless of where they claim to be from (22)

The analysis of pronouns indicates that discussion of Romanians in the comments tends to take place within a wider context of immigration to the UK. This is borne out by reference to some of the other keywords in Table 1 e.g. *migrants*, *immigration* and *immigrants* are keywords in both corpora. The analysis so far reveals a fairly negative picture of how commenters view immigration generally and Romanians more specifically, as well as indicating how commenters create a shared sense of grievance, from a European political class who is seen to care more about immigrants than British nationals.

Identifying prototypical texts

This part of analysis takes a different track by using a tool called ProtAnt which ranks a set of texts based on their lexical prototypicality. ProtAnt calculates the keywords in a corpus of texts, using a reference corpus and then presents the list of texts in order, based on the number of keywords in each. Those at the top of the list are thus viewed as most typical. Experiments with ProtAnt e.g. Anthony and Baker (2015, 2016) showed that the tool was able to successfully identify the most typical newspaper articles and novels from corpora, as well as identifying atypical files e.g. those which came from a different register to all the others in a corpus. Using ProtAnt with the corpus of 1,929 *Express* articles of Romanians (and the BE06 corpus again used as the reference corpus) resulted in the articles in Table 5 being ranked as being most prototypical.

	Headline	File number
1	Number of Romanians coming to Britain TRIPLES as EU workers boost migrant totals	152
2	We are losing control of our borders. Number of EU workers on the up	478
3	UKIP's Nigel Farage: 'Three-month benefit ban for migrants is not enough'	570
4	Number of Romanians and Bulgarians working in UK soars by nearly 10 PER CENT	381
5	Tory MPs warning over new tide of immigrants	1075

6	Romanians and Bulgarians were already working in Britain before ban ended	522
7	Up to 350,000 Romanians and Bulgarians coming to UK	970
8	Young Romanians want to work in the UK	907
9	New immigration BOMBSHELL as number of Romanians and Bulgarians in Britain TREBLE	515
10	Even Romanian MPs warn us that swarms of migrants are coming to live in UK	604

Table 5. The most prototypical articles in the corpus.

The headlines in Table 5 all point to articles about immigration to the UK, with seven of them directly referring to Romanians. From the perspective of an analyst carrying out CDA, it is also notable how metaphors are employed in some of the headlines. Article 1075 uses a water metaphor in referring to a *tide of immigrants* whereas article 604 mentions *swarms of migrants*, figuratively constructing migrants as a kind of flying pest or insect.

One way that ProtAnt could be used is to carry out down-sampling of a larger data set so that a close analysis could be carried out on a manageable number of typical files. For example, one of the files ranked in the 10 most typical (file 515) is shown in full below.

> New immigration BOMBSHELL as number of Romanians and Bulgarians in Britain TREBLE
>
> BRITAIN was today hit with a new immigration bombshell as official figures showed the number of Romanians and Bulgarians arriving in the UK trebled in 2013.

Latest figures show net migration to Britain is rising again. The Office for National Statistics said this morning that 24,000 citizens from the two countries arrived in the year to September 2013, nearly three times the 9,000 who arrived in the previous 12 months. The ONS said this was ‘statistically significant’ and that around 70 per cent came to work, while 30 per cent came to study. Their figures also showed the Government was slipping behind its target for reducing overall net migration – the difference between migrants leaving and arriving in the UK. David Cameron wants the net number reduced to the tens of thousands but the figure soared to 212,000 in the period from 154,000 the previous year. Figures are not yet available for the numbers of Romanians and Bulgarians who have arrived in the UK since January 1 when labour market restrictions were lifted to comply with EU rules. UKIP leader Nigel Farage said: ‘The immigration targets were nothing more than spin to appease in the short term the public who are concerned about uncontrolled immigration But how can you have any targets when you can't control who comes to live, settle and work in this country? It's as sensible as hefting butterflies.’ Despite the sharp rise, Immigration and Security Minister James Brokenshire said a dip in the number of migrants arriving from outside the EU showed the Government was getting to grips with the issue. He said: ‘Our reforms have cut non-EU migration to its lowest level since 1998 and there are now 82,000 fewer people arriving annually from outside the EU than when this government came to power And overall figures are also well down from when we first came to government in 2010 — with nearly 70,000 fewer migrants coming to the UK. Numbers are down across the board in areas where we can control immigration, but arrivals from the EU have doubled in the last year.’ Mr Brokenshire admitted the limitations of the Government's ability to restrict immigration from the EU. He said: ‘We cannot impose formal immigration controls on EU migrants, so we are focusing on cutting out the

> abuse of free movement between EU member states and seeking to address the factors that drive European immigration to Britain.’ A Home Office spokesman said reducing net migration to the tens of thousands ‘remains the government’s objective’.

A close qualitative analysis would focus on aspects of the language of this article (such as lexical choice, evaluation, metaphor and patterns around agency), as well as its text structure (for example, what pieces of information are foregrounded or appear at the end and are thus implied to be less important, who is quoted and how much space are they given, and how does the main narrative voice of the article align itself with any quotes). For example, it is notable how in the headline the information about Romanians and Bulgarians is dramatically described as a *bombshell*. In the body of the article we are first given numerical information about numbers of immigrants, described as *official figures*, helping to legitimise the way in which the information is given. This is followed by a metaphorical description of how the government is *slipping behind* Prime Minister David Cameron’s targets for immigration, which instead is said to have *soared*. Then the article notes that figures are not available since January 1st, an important date in this corpus as it was the date when Romanians and Bulgarians were allowed to travel to the UK to work. The lack of such information could implicitly invite readers to speculate on the amount of EU immigration since this date. After giving this information, there is a quote by the leader of the UK Independence Party (UKIP), Nigel Farage, who describes the targets as ‘spin’. The article then describes the rise in immigration as *sharp* and then gives a longer quote from a member of the government (James Brokenshire) who is directly responsible for immigration and attempts to downplay the amount of immigration to the UK. However, the minister is also described as admitting that the government are limited in being able to restrict immigration from the EU, and he refers to the existence of abuse of free movement between EU member states. The article then ends with a short statement from the Home Office about the objective of reducing net migration.

Although the language in the article is reasonably restrained in comparison to the language used in the comments, it is clear that immigration is taken for granted as problematic, with Britain described metaphorically as being *hit with an immigration bombshell*. The opinion of Nigel Farage is foregrounded with his quote appearing before those of government officials. Additionally, Farage is represented as speaking on behalf of the whole country when he describes the public as being *concerned about uncontrolled immigration*. While the article appears to give a balanced view in that it presents opinions from members of the government which appear to be more reassuring about numbers of immigrants, the minister's speech appears to agree with Farage in terms of blaming membership of the EU for Britain's inability to restrict immigration. It is notable that nowhere in the article are any reasons given for why immigration to the UK is considered to be problematic. The public are described as 'concerned about uncontrolled immigration', but the article does not describe actual negative consequences of such immigration. This is left for readers to infer.

It is interesting to consider how the comments respond to the information in the article. The article produced 149 comments, of which over 95% express concern and disapproval about immigration to the UK. For example, commenters variously denounce immigrants as 'human shyte', 'scrounging load of rubbish', 'deluge of immigrants' and 'foreign spongers'. They are also implied to be criminal:

> If we had a goverment in control of our country, not the puppets with their strings being pulled by the EU, we could deport all the foreign homeless, unemployed, beggars, pickpockets, rapists and murderers

The government is described as clueless, soppy, pretend, shambolic, powerless and useless. David Cameron is also viewed critically e.g. 'BALL—LESS apoligy for man', 'USELESS', 'A SICKENING SIGHT' and 'A WEAK SELF SERVING TRAITOR'. He is also described as ‘clinging like a child to the skirt of Angela Merkel dictator of the EU’. On the other hand, UKIP is constructed as the only viable option, with 36 references to UKIP across the 149 comments, and 18 cases of the phrase *vote/voting UKIP*. A typical comment is ‘The best for THIS Country is to have the UKIP in Power THEN we WIll get a decient Britain like we used to have’.

Not all of the comments express anti-immigration and anti-EU views, with 5% offering an alternative perspective or are critical of the content of the article e.g.:

> Yesterday this paper reported that Romanians and Bulgarians returning to their countries because they were 'disillusioned with low pay and benefits' today the are apparently all coming back again- this has to be the most confused, dismal rag ever...Start reporting sensibly or I for one wont be buying this paper. TYou seem to be more confused than the govt is and that's saying something.

While such comments indicate that not everyone who reads the article is as accepting of the newspaper’s stance, they are in the minority and the alternative voices feel drowned out by the higher numbers of comments which not only agree with the article, but go further in terms of negatively constructing EU and British policy around immigration using much more offensive language. It is important to consider that the articles and comments therefore work in concert, presenting different but complementary perspectives. In a sense, the article initiates a discussion about Britain’s role in Europe and the impact of immigration, putting forward a negative perspective but in a relatively ‘restrained’ tone. The comments continue this dialogue, using more inflammatory, emotive and colloquial language. They engage in ‘decoding’ work of the initial article, spelling out its implications, negatively characterising immigrants and the political establishment while advocating voting for UKIP, but in a minority of cases there is critical engagement giving an alternative perspective. It is by considering both the articles and the comments *in tandem* that we can begin to understand the role that newspaper like *The Express* and its commentators played in the decision of the British people to leave the EU.

Future Directions

A potential criticism of corpus linguistic approaches involves the application of cut-off points. For example, analysts need to determine a cut-off point for what counts as a keyword and there are no clear guidelines on what this should be. While the log-likelihood test is a statistical and thus associated p values can be produced, the application of such tests on language data produces somewhat different results from the sort of hypothesis-testing experiments that are normally conducted in the social sciences. Thus, applying standard p value cut-offs of 0.05 or 0.01 will often result in hundreds of keywords to examine. Also, the larger the corpus, the more keywords are likely to be produced. The notion of statistical significance is thus problematic and corpus analysts either impose very high p values like 0.0000001 or may simply select the strongest 10, 20 or 100 keywords, depending on how much analysis they are able to do. A good principle to bear in mind is transparency, so even if not all keywords are analysed, it is worth pointing out how many *could* have been analysed. For example, 1,541 keywords were produced from the news articles which had a log-likelihood score of above 15.13 (corresponding to a p value of < 0.0001), while 1,359 keywords were produced from the comments corpus using the same cut-off.[xix] Of the 50 which were examined from each corpus, they had a LL score of between 359 and 6,470 in the articles and 728 and 7,508 in the comments. Further research needs to be carried out on the application of keyword cut-offs but see Wilson (2012) for a discussion of how Zip’s Law might be gainfully used to this effect.

The consideration of user-generated data raises issues for corpus linguists. Readers will have probably noticed that the comments quoted in this chapter contain numerous typos. Such typos will impact on word frequency counts and it is worth carrying out some preliminary research on the corpus to ascertain the extent to which this will impact on results. Tools such as VAARD (Archer, et al. 2015) can be used to regularise inconsistent spelling in a dataset although this too would require an initial survey of texts to outline the problem areas.

While corpus linguistics methods are effective at handling large amounts of electronic text, they are currently not particularly good at dealing with other types of data (such as images, audio or video data). In order to analyse the relationship between image and text in a corpus of children's picture books, McGlashan (2016) developed a method called collustration which involved pairing concordance lines with images that occurred in the vicinity of the text under consideration. His analysis found how the images supplemented the text in interesting ways e.g. the phrase *love each other* tended to appear in the vicinity of images of people engaged in physical contact. However, this analysis needed to be carried out painstakingly by hand rather than automatically and would be difficult to implement on a large scale. Corpora could be encoded for images which are cross-linked to the parts of text that they appear next to and a future analysis tool could then show concordance lines alongside the corresponding images. Similarly, parts of the spoken British National Corpus have been cross-linked to audio files so that concordance lines can be clicked on to play audio clips of the speech. As new tools are developed, more sophisticated forms of analysis will become available.

In this chapter we have shown how corpus techniques can help to direct Digital Humanities analysts to salient or frequent linguistic patterns in large databases so that methods of analysis associated with the kinds of close readings carried out by CDA can be utilised. Although we used a newspaper corpus (along with reader comments) as a case study, such techniques could be used on other electronic datasets of texts including court proceedings, works of fiction, legal debates, diaries or business reports. The keywords procedure offered a 'way in' to the corpora we analysed, whereas the analysis of collocates enabled us to more quickly see repeated patterns of association with those words.

Finally, we note that corpus linguistics techniques have a great deal to offer the Digital Humanities, although we advocate that they are used in conjunction with qualitative forms of inquiry which take various forms of context into account and involve close readings of texts. Otherwise there is a danger in placing too much reliance on automatic forms of analysis which may produce at best a shallow interpretation and at worst a completely inaccurate one.

Further Reading

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

An essential text outlining practical approaches to corpus-based discourse analysis.

Baker, P., Gabrielatos, C. & McEnery, T. (2013) *Discourse Analysis and Media Attitudes: the representation of Islam in the British Press*. Cambridge: Cambridge University Press.

This book brings together corpus lingusitics and CDA to examine representations of RASIM (refugees, asylum seekers, immigrations and migrants) in a large collection of UK news articles and provides a valuable example of how methods from CDA may be applied to very large corpora.

Baker, P. & McEnery, T. (Eds.) (2015) *Corpora and Discourse Studies: integrating discourse and corpora.* Basingstoke: Palgrave Macmillan.

A recent addition to the literature on and about the growing field of corpus-based discourse studies, this title presents a number of diverse approaches that might be taken when conducting corpus-based approaches studies of discourse in large datasets.

Wodak, R. & Meyer, M. (Eds.) (2015) *Methods of Critical Discourse Studies* (3rd ed.). Los Angeles: SAGE. (https://capitadiscovery.co.uk/bcu/items/1248946)

A now touchstone text for students of CDA. This accessibly written edited collection covers fundamental theory in CDA and presents several chapters from experts in the field on a range of theoretical and methodological approaches, including corpus linguistics.

REFERENCES

Anthony, L. (2016). AntConc (Version 3.4.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Anthony, L. and Baker, P. (2015). *ProtAnt*: A tool for analysing the prototypicality of texts. *International Journal of Corpus Linguistics*, 20(3), pp. 273-292.

Anthony, L. and Baker, P. (2016). Evaluating the Effectiveness of Prototypical Text Detection in Teaching and Research: New Developments and Applications of *ProtAnt*. Paper given at TaLC 2016 Gissen, Germany, July 20-23.

Archer, D., Kytö, M., Baron, A. and Rayson, P. (2015). Guidelines for normalising Early Modern English corpora: Decisions and justifications. *ICAME Journal*. 39(1), pp. 5-24.

Baker, P. (2004) 'Querying keywords: questions of difference, frequency and sense in keywords analysis.' *Journal of English Linguistics*. 32(4): 346-359.

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

Baker, P. (2009). 'The BE06 Corpus of British English and recent language change.' *International Journal of Corpus Linguistics*, 14(3), pp. 312-337.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., and 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), pp. 273-306.

Baker, P., Gabrielatos, C. and McEnery, T. (2013). *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press.* Cambridge: Cambridge University Press.

Chovanec, J. (2014) *The representation of Eastern Europeans in British internet debates*. Paper given at the 12th ESSE Conference, Košice, Slovakia. 29 August-2 September 2015.

Durrant, P. and Doherty, A. (2010). 'Are high frequency collocations psychologically real? Investigating the thesis of collocational priming.' *Corpus Linguistics and Linguistic Theory,* 6 (2), pp. 125-155.

Fairclough, N. (1992). *Discourse and Social Change*. Cambridge: Polity Press.

Fairclough, N. (1995). *Critical Discourse Analysis: The Critical Study of Language*. London: Longman.

Fairclough, N. (2003). *Analysing Discourse: Textual Analysis for Social Research*. London: Routledge.

Fairclough, N., Wodak, R., and Mulderrig, J. (2011). Critical discourse analysis. In T. van Dijk (Ed.), *Discourse Studies: A Multidisciplinary Introduction* (2nd ed.). London: Sage. pp. 357-378.

Fowler, R., Hodge, B., Kress, G., and Trew, T. (1979). *Language and Control*. London: Routledge.

Gabrielatos, C., and 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(1), pp. 5-38.Gerbner, G., Gross, L., Morgan, M., and Signorielli, N. 1986. 'Living with television: The dynamics of the cultivation process', in

J. Bryant and D. Zillman (eds.) *Perspectives on Media Effects*. Hilldale, NJ: Lawrence Erlbaum Associates, pp. 17-40.

Habermas, J. (1988). *On the logic of the social sciences*. MIT Press.

Jensen, K. E. (2014). Linguistics and the digital humanities: (computational) corpus linguistics. *Journal of Media and Communication Research*, 30(57), pp. 115-134.

Kristeva, J. (1986). Word, dialogue and novel. In T. Moi (Ed.), *The Kristeva reader*. Oxford: Basil Blackwell. pp. 34-61

McEnery, T., McGlashan, M. and Love, R. (2015). Press and social media reaction to ideologically inspired murder: The case of Lee Rigby. *Discourse and Society*, 9(2), pp. 1-23.

McGlashan, M. (2016). The representation of same-sex parents in children's picturebooks: a corpus-assisted multimodal critical discourse analysis. Lancaster University, UK: PhD Thesis.

Reisigl, M., and Wodak, R. (2001). *Discourse and Discrimination: Rhetorics of Racism and Antisemitism*. London: Routledge.

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

Stubbs, M. (2001) *Words and Phrases*. London, Blackwell.

Thurman, N. (2018) 'Newspaper consumption in the mobile age' *Journalism Studies* 19: 1-21.

van Dijk, T. 1991. *Racism and the Press*. London: Routledge.

van Dijk, T. (2006). Discourse, context and cognition. *Discourse Studies,* 8(1), pp. 159-177.

van Leeuwen, T. (1996). The representation of social actors. In C. R. Caldas-Coulthard & M. Coulthard (Eds.), *Readings in Critical Discourse Analysis*. London: Routledge.

van Leeuwen, T. (1997). 'Representing social action.' *Discourse and Society*, 6(1), pp. 81–106.

Wenger, E. (1998). *Communities of Practice: learning, meaning and identity.* Cambridge: Cambridge University Press.

Westlund, O., and Färdigh, M. A. (2015), ‘Accessing the news in an age of mobile media: Tracing displacing and complementary effects of mobile news on newspapers and online news’ *Mobile Media & Communication*, 3(1), pp. 53-74.

Widdowson, H. G. (2004), *Text, Context, Pretext: Critical Issues in Discourse Analysis.* Oxford: Blackwell.

Wilson, A. (2012) Using corpora in depth psychology: a trigram-based analysis of a corpus of fetish fantasies. *Corpora*. 7, 1, pp. 69-90.

Wodak, R., and Meyer, M. (2009). Critical discourse analysis: History, agenda, theory and methodology. In R. Wodak & M. Meyer (Eds.), *Methods of Critical Discourse Analysis* (2nd ed.). London: Sage. pp. 1-33

[i] https://www.theguardian.com/commentisfree/2016/jun/27/austerity-economic-woes-eu-referendum-brexit

[ii] http://www.dailymail.co.uk/news/article-3666511/Merkel-s-open-door-policy-caused-Britain-leave-EU-German-leader-blamed-Brexit-failure-deal-migrant-crisis-open-arms-immigration-policy.html

[iii] http://lordashcroftpolls.com/2016/06/how-the-united-kingdom-voted-and-why/

[iv] (http://www.dailymail.co.uk/news/article-2531440/Sold-Flights-buses-Romanians-Bulgarians-head-UK.html

[v] http://www.theguardian.com/media/greenslade/2014/mar/17/dailymail-pcc

[vi] We initially considered a number of possible newspapers although found that each one had a different online commenting facility which would have required a range of bespoke retrieval solutions which this chapter would be unable to do justice to. The *Daily Express* worked well with the web scraping tools we wanted to showcase in this chapter.

[vii] https://www.python.org/

[viii] https://www.crummy.com/software/BeautifulSoup/ - Beautiful Soup is a library)

[ix] http://www.seleniumhq.org/

[x] https://www.mozilla.org/en-GB/firefox/new/

[xi] https://www.google.com/chrome/

[xii] Further regarding ethics, it could be argued that when posting a comment on a freely available website which is not password protected, one’s comments are already in the public domain. However, we felt it was good ethical practice to protect, as much as possible, the identities of commenters, especially as it would have been very difficult to ask permission of the hundreds of commenters to include their comments in the corpus. People who posted comments to the Daily Express website are required to create an anonymous account with a log-in name. With each comment only the log-in name is given with it – e.g. cheeky1952. We were unable to find references to real names in the dataset but to ensure an additional level of anonymity we do not refer to the log-in names in our analysis.

[xiii] http://www.lexically.net/wordsmith/index.html

[xiv] http://corpora.lancs.ac.uk/lancsbox/

[xv] https://cqpweb.lancs.ac.uk

[xvi] https://www.sketchengine.co.uk

[xvii] AntConc offers two ways of deriving keywords, chi-square tests and log-likelihood tests. There are currently at least 8 ways of deriving keywords, some of which have strong champions and detractors, see http://ucrel.lancs.ac.uk/llwizard.html. It is beyond the remit of this chapter to do justice to the ongoing debate over which measure gives the 'best' set of keywords, but we would instead note that the log-likelihood test, which was used for this analysis, measures the confidence with which we can say that a given word is actually a keyword (e.g. the difference in frequencies is not the result of chance). However, the test does not give an indication of how strong the difference is.

[xviii] The highest 50 LL scores range from between 330-3888 for the articles and 728-91407 for the comments. Having experimented with different cut-offs, (e.g. the top 10, the top 20, the top 50, the top 100), we felt that the top 50 keywords provided a reasonable number in order to allow keywords to be placed in categories of similarity to identify themes, without resulting in too much analytical repetition (e.g. resulting in lots of keywords which refer to the same theme). This is a matter of trial and error and different cut-offs need to be established for each project, relevant with its aims and taking into account real-world practicalities.

[xix] http://ucrel.lancs.ac.uk/llwizard.html