Corpus approaches to language in the media

Book or Report Section

Accepted Version

Jaworska, S. (2018) Corpus approaches to language in the media. In: Cotter, C. and Perrin, D. (eds.) Routledge Handbook of Language and Media. Routledge. ISBN 9781138014176 Available at http://centaur.reading.ac.uk/40957/

This is a pre-publication version accepted for Routledge Handbook of Language and Media edited by Colleen Cotter and Daniel Perrin, Routledge: London/New York (to appear early 2016). Please refer to the published version of this article if you wish to quote from it.

Corpus Approaches to Language in the Media

Sylvia Jaworska

1. INTRODUCTION

Over the last two decades, the tools and methods of Corpus Linguistics (CL) have been increasingly used to study language in the media. Although CL is commonly associated with quantitative techniques, most researchers adopting corpus methods to study media language combine quantitative with qualitative procedures used widely in the many branches of discourse studies. This has led to fruitful methodological synergies (e.g. Baker et al. 2008; Baker and Levon 2015) and approaches (e.g. Partington et al. 2013), which have revealed much more nuanced patterns of language use and representations than a quantitative or qualitative analysis alone is able to uncover.

The main aim of this chapter is to offer an overview of research that has adopted the methodology of CL to study aspects of language use in the media. The overview begins by introducing the key principles and analytical tools adopted in corpus research. To demonstrate the contribution of corpus approaches to media linguistics, a selection of recent corpus studies is subsequently discussed. The final section summarises the strengths and limitations of corpus approaches and discusses avenues for further research.

2. DISCIPLINARY PERSPECTIVES I: CORPUS PRINCIPLES AND TOOLS

Corpus Linguistics is mainly concerned with examining language use in large corpora, where *corpus* refers specifically to an electronic compilation of naturally-occurring texts (McEnery and Hardie 2012). Corpus Linguistics has developed rapidly over the last three decades owing to the advances and the availability of linguistic software programmes that allow linguists to search through large corpora quickly and reliably. Insights derived from corpus research have largely increased our understanding of language use by providing empirical evidence for the existence of regularities and patterns that are not immediately visible to the naked eye. As John Sinclair pointedly remarked: “The language looks rather different when you look at a lot of it at once” (Sinclair 1991: 100).

The attractiveness of Corpus Linguistics lies particularly in the ease with which large amounts of data can be automatically scanned to reveal distinctive and typical patterns, for

example, in forms of frequency lists and keywords. While the speedy processing has much advantage over a manual analysis, there are also potential pitfalls, of which a researcher embarking on a corpus-based analysis needs to be aware. Thus, before discussing the contribution of corpus approaches to media linguistics, the following section outlines the key principles and analytical procedures commonly used in corpus research. To illustrate the practical uses of corpus tools, examples are drawn from the MinD corpus, which stands for **M**ultilingualism **in** Public **D**iscourse and contains 469 articles (646,847 words) on the topic of bi- and multilingualism published in the major British newspapers from 2000 to 2014.[i] The corpus is part of a larger interdisciplinary project which investigates media representations of multilingualism over time and the extent to which the representations are shared or refuted in views of general public. Issues related to corpus building and corpus annotation have been discussed extensively elsewhere (e.g. Biber et al. 1998; McEnery and Hardie 2012) and for reasons of space are not considered here.

3.1. Principles and analytical tools of corpus research

The first point of entry into a corpus is often frequency. In corpus linguistic terms, frequency refers to the count of items in a corpus, whereby *item* can be a word, a part of speech or a keyword. Most of the widely available concordancers, that is, linguistic software programmes such as AntConc (Anthony 2011), WordSmith Tools (Scott 2008) or Sketch Engine (Kilgarriff et al. 2004) can produce frequency lists within seconds. Table 1 below shows a frequency list retrieved from MinD in Sketch Engine.

Table 1: The 15 most frequent words in MinD

Word	**Freq.**
THE	22,702
TO	12,111
OF	11,556
A	11,027
AND	10,813
IN	10,638
IS	5,696
THAT	4,740
FOR	4,224
IT	3,826
AS	3,081
I	3,079
ARE	2,898
WAS	2,749

LANGUAGE	2,667

As can be seen, the first most frequent words in MinD are functional words such as articles, prepositions and conjunctions. Most corpora of English will have this type of words on top of the frequency list. The first content word, which appears on this list, is 'language', which is not surprising given that the corpus includes articles about bi- and multilingualism. While functional words can point to important grammatical features of the studied data set, corpus researchers studying aspects of language use in the media are often interested in media discourses and tend to focus on content words, as they are more likely to tell us something about discourse (Baker 2006).

Specific content words can be gleaned from a raw frequency list, but this procedure could be time consuming. They can also be retrieved automatically, if the corpus was annotated with parts of speech (POS). Some concordancers such as Sketch Engine (Kilgarriff et al. 2004) have an inbuilt POS-tagger, also for languages other than English. Table 2 shows the 10 most frequent nouns and adjectives in MinD.

Table 2: The 10 most frequent nouns and adjectives in MinD

Noun	**Freq.**
LANGUAGE	2,374
SCHOOL	1,274
TIME	641
YEAR	562
WORLD	466
COUNTRY	444
EDUCATION	419
HOME	350
LIFE	288
BUSINESS	285

Adjective	**Freq.**
BILINGUAL	989
FRENCH	982
ENGLISH	829
FOREIGN	622
FIRST	531
NEW	498
GOOD	371
DIFFERENT	328
WELSH	310
PRIMARY	299

The lists highlight several interesting features of the contemporary press discourse about bi- and multilingualism. They contain a number of words that point to schooling ('school', 'education'), which could suggest that bi- and multilingualism are predominantly discussed in the context of education. We also have here a number of items that point to languages or specific linguistic identities such as 'English', 'French' and 'Welsh'. If we assume that frequency can be a marker of saliency, then the list is quite revelling. It shows that multilingualism is strongly associated with languages or linguistic contexts that, apart from Welsh, are seen as prestigious and useful worldwide. In contrast, languages that are spoken by various communities in the UK appear with much lower frequencies, for example Polish only 105 times, Urdu 49 and Punjabi 36. Such representations seem to reinforce the concept of 'elite' bi- and multilingualism which values some resources (prestigious languages) and excludes others (community languages) (e.g. De Mejía 2002).

Another way of revealing the main themes of a given data set is via *keyword analysis*. In Corpus Linguistics, a keyword is a word which occurs unusually frequent in a given corpus, as compared to another mostly larger reference corpus (Scott 2010). This unusualness is established by using a test of statistical significance, mostly log-likelihood. Keywords retrieved in this way are seen as good indicators of the text's aboutness and register. Table 3 below lists the first 15 keywords retrieved from MinD, using the British National Corpus (BNC) as a comparator.

Table 3: The first 15 keywords in MinD as compared with BNC

	MinD		**British National Corpus**		
word	**Freq**	**Freq/mill**	**Freq**	**Freq/mill**	**Score**
LANGUAGES	1,365	2710.9	2,806	25.0	22.5
BILINGUAL	1,044	2073.4	304	2.7	21.2
LANGUAGE	2,374	4714.7	17,203	153.4	19.0
ENGLISH	1,685	3346.4	22,604	201.5	11.4
FRENCH	1,191	2365.3	16,441	146.6	10.0
SPEAK	783	1555.0	8,711	77.7	9.3
WELSH	563	1118.1	3,834	34.2	9.1
SCHOOL	1,274	2530.1	29,410	262.2	7.3
PUPILS	546	1084.3	7,519	67.0	7.1
FOREIGN	622	1235.3	10,493	93.5	6.9
SPANISH	377	748.7	3,430	30.6	6.5
GAELIC	262	520.3	397	3.5	6.0
LEARNING	475	943.3	8,355	74.5	6.0

	MinD		British National Corpus		
word	**Freq**	**Freq/mill**	**Freq**	**Freq/mill**	**Score**
SCHOOLS	619	1229.3	13,862	123.6	5.9
CHILDREN	1,282	2546.0	42,075	375.1	5.6

While the list above signals the main themes discussed in the context of media representations of bi- and multilingualism, that is education and prestige languages, the list is much longer with nearly 900 keywords, and hence, there may be many other *key* issues worth examining. To capture a wider range of topics, researchers often scan the first 100 keywords (strongest in terms of statistical significance) and group them manually into semantic categories (Ensslin and Johnson 2006; Gabrielatos and Baker 2008; Baker et al. 2013). Table 4 shows an example of such an analysis based on the first 100 keywords in MinD. 10 semantic groups were identified with the most salient being EDUCATION followed by LANGUAGES AND LINGUISTIC VARIETIES. It is interesting to see that beyond education and languages, bi- and multilingualism are predominantly discussed in the context of specific countries and cities (mostly metropolises) and groups of people. The latter category includes social groups such as 'children' and 'parents', but also 'immigrants'.

Table 4: The first 100 keywords categorised into semantic categories

Semantic Category	Examples of keywords
EDUCATION	school, pupils, schools, learning, learn, primary, lessons, teaching, taught, education, GCSE, teach, students, teachers,
LANGUAGES/ LANGUAGE VARIETIES	English, French, Welsh, Gaelic, Spanish, German, Chinese, Catalan, Irish, Mandarin, Russian, Italian, Polish, Latin
PEOPLE	children, speakers, parents, friends, immigrants, minority, professor, foreigners
COUNTRIES/REGIONS	France, Wales, UK, Malta, EU, China, Belgium, Britain, Gaeltacht, Europe
LINGUISTIC TERMS	languages, bilingual, language, multilingual, (mother) tongue, bilingualism, linguistic, monolingual
EVALUATION	native, fluent, foreign, cultural, ethnic, global, international
COMMUNICATIVE SKILLS	speak, says, speaking, spoken, speaks, translation
CITIES	London, Beijing, Paris, city, Brussels, Manchester
MEDICAL/BODILY TERMS	brain, Alzheimer, dementia
OTHER	culture, signs, age, career, Internet, website, online

Interestingly, apart from Polish languages used by the many immigrant communities in the UK do not appear to belong to the 'strongest' keywords in the corpus. It was also unexpected to see a number of medical keywords pointing to diseases such as Alzheimer and dementia.

Keywords and groupings of keywords into semantic categories can be useful tools in signposting main topics and issues discussed in relation to the studied phenomenon. But a keyword analysis has its limitations too. The type of keywords retrieved from the target corpus will greatly depend on the selection of the reference corpus, its size and contents as well as the metrics used. For example, the last category OTHER includes items such as 'Internet' and 'website'. This could imply that online communication is an important theme in the context of public representations of bi- and multilingualism. However, the keyness of these terms does not have much to do with the topic, but rather with the data included in the comparator corpus. The BNC was compiled at the beginning of the 1990s, at a time when web-related terminology was only starting to reach public domains. Hence, references to the Web are not frequent in BNC, but because they appear statistically more frequently in MinD, they were identified as key. Also, examining the top 100 or 200 keywords in-depth seems rather selective, as all keywords are statistically significant and hence, all are 'equal candidates' for analysis. Gabrielatos and Marchi (2012) also argue that the metrics commonly used to retrieve keywords such as log-likelihood are not necessarily appropriate; they may point to statistical significance, but do not signal how large the effect is. Instead, the authors propose frequency difference as a more suitable metric. Furthermore, a list of keywords shows isolated lexical items only. We know from studies in phraseology that the form of a word does not necessarily encompass all the meanings that the word in question has. These often arise from the typical combinations of the word with other lexical items. Keywords and groupings of keywords can help us develop some hypotheses regarding the meanings of lexical items. However, to test these hypotheses, we need to move beyond single words and investigate wider textual patterns. This can be done by examining *collocations* and *concordances.*

In Corpus Linguistics, collocation is understood as the co-occurrence of two or more words within a certain span, for example five items to the left and five to the right (-5 and +5) and a certain cut-off point (e.g. occurring 5 times or more). A distinction is normally made between co-occurrences that are determined on the basis of raw frequency or significance testing (Barnbrook et al. 2013; McEnery and Hardie 2012). When determining collocations on the basis of frequency alone, we cannot be certain whether a co-occurrence is a true reflection of a relationship between two items or whether it emerged by chance, for example, due to the fact that one of the items is a frequent word in the given corpus (Baker 2006). To eliminate this concern, various tests of statistical significance are used, of which the most popular are Mutual Information (MI), Log-Likelihood, T-score and LogDice. It needs to be

borne in mind that each of the test yields different results, because they favour different types of words (Baker 2006). Collocations are not only useful indicators of strong lexico-grammatical associations. They can also point to salient themes and value judgments associated with a studied item (Mautner 2007). To illustrate an example, table 3 shows collocations of the keyword 'bilingual' sorted in accordance with the LogDice score.

Table 3: The 20 most frequent collocations of 'bilingual' in the MinD Corpus within +5 and -5 span

Word	Freq.	LogDice	Word	Freq.	LogDice
1. CHILDREN	163	10.926	11. STREAM	26	9.472
2. EDUCATION	83	10.720	12. TEACHING	30	9.471
3. SCHOOL	155	10.687	13. FIRST	38	9.459
4. SIGNS	52	10.243	14. BECOME	34	9.418
5. PRIMARY	52	10.138	15. ENGLISH	56	9.330
6. STATE	38	9.822	16. PUPILS	34	9.291
7. ROAD	30	9.557	17. SECRETARY	23	9.245
8. FOR	122	9.541	18. PROGRAMME	23	9.153
9. HAVE	150	9.500	19. GOOD	33	9.131
10. PEOPLE	48	9.494	20. FRENCH	30	9.127

The list above suggests that the term 'bilingual' is strongly associated with teaching and schooling, and some prestigious languages (English and French). Interestingly, 'children' and 'pupil' are frequent collocates of 'bilingual', but there are no items pointing to adults. This could suggest that bilingualism is seen as something to be achieved during childhood and not something typical for adults. This, in turn, echoes a common assumption that languages are best learned at the earlier stages of life. While undoubtedly there are many advantages of learning another language before puberty, there is evidence to suggest that adults can become proficient bilingual speakers too (e.g. Bongaerts et al. 1997). However, this message is somewhat silent in the media. Equally, community languages do not appear to be associated with being bilingual and are not mentioned in the most frequent 100 collocates. This again might suggest that bi- and multilingualism are predominately framed as an 'elite product' that foregrounds prestigious international varieties and excludes those spoken by bilingual communities in the UK.

Related to the concept of collocation is the notion of *semantic prosody*. This concept goes back to Sinclair's (1991) observations that some words have a tendency to occur with pleasant situations, while others may be associated with negative events. The term semantic prosody was coined by Louw (1993: 157), who defined it as the "consistent aura of meaning with which a form is imbued by its collocates". The notion was subsequently expanded by

Stubbs (2001: 65) who prefers the term *discourse prosody* to distinguish between lexical choices that express speakers' attitudes (discourse prosody), and associations that are more or less explicit in the semantics of a lexical items (semantic preference). To illustrate an example of a semantic/discourse prosody, we will consider the item 'immigrants', which is one of the keywords in MinD. Table 4 below lists the 20 strongest collocates of the term. When studying the item, a certain profile of immigrants emerges. Accordingly, they appear to be frequently associated with criminality ('illegal') and large numbers ('influx'). They are also 'young' and 'poor'. It is interesting to note that certain geographical attributes are foregrounded such as 'African' and 'Eastern', the latter associated exclusively with Eastern Europe.

Table 4: The 20 most frequent collocations of 'immigrant' in the MinD Corpus within +5 and -5 span

Word	**Freq.**	**LogDice**	**Word**	**Freq.**	**LogDice**
1. ILLEGAL	23	12.364	11. OFFERING	3	9.236
2. INFLUX	4	10.057	12. POOR	3	9.176
3. POLITICALLY	4	10.011	13. CULTURES	3	9.150
4. EASTERN	4	9.881	14. EUROPE	4	8.757
5. NATIONALITY	3	9.715	15. STARTED	3	8.721
6. BLUNKETT	3	9.690	16. BRITAIN	4	8.262
7. POSTERS	3	9.690	17. FROM	15	8.002
8. AFRICAN	3	9.563	18. PARENTS	3	7.885
9. YOUNG	8	9.393	19. FOREIGN	5	7.806
10. ARRIVED	3	9.236	20. PRIMARY	3	7.771

According to OED, the term 'immigrant' describes a person who moved from one country to another. Linking it insistently with illegality and large numbers adds a new evaluative and in this case pejorative dimension, which is not inherent in the form or the prime meaning of the word. By persistently associating immigrants with illegality, the press constructs the newcomers as criminals and reinforces the message that immigration is a 'bad thing'. Such representations and semantic prosodies have been shown in previous research that was specifically interested in the representations of immigration in the media (e.g. Gabrielatos and Baker 2008; Taylor 2014). What is interesting is that such discourses are transposed to the context of bi- and multilingualism. Effectively, the negative prosody surrounding the keyword 'immigrants' in MinD imbues the perceptions of bi- and multilingualism with

negativity, which might reduce the general positive associations with the ability of speaking other languages.

This example illustrates that studying frequent collocations can be useful in revealing persistent discursive associations used to construct social phenomena and social actors in the media. These are interesting, because they are often a matter of writer's choice who selects some lexical items over others to propagate a particular version of reality (van Dijk 1995). Hence, collocations and the semantic and discourse prosodies they create can indicate shared evaluative judgments and highlight ideological uses of language (Baker et al. 2013) that are reinforced even in contexts in which they are less expected, as the above case of 'immigrants' has shown.

While collocations can signal semantic and evaluative associations, there may exist subtle variations regarding their use. *Concordances* can help us reveal further discursive patterns. To illustrate an example, we will now look at the item 'English'. As Table 5 shows, 'English' in MinD collocates strongly with words such as 'speak', 'language', 'French', the conjunction 'and' and the negation 'not'. The co-occurrence of 'English' with 'not' seemed intriguing and was further examined via concordance lines (Figure 1).

Table 5: The 10 most frequent collocations of 'English' in the MinD Corpus within +5 and -5 span

Collocation	**Freq.**	**LogDice**	**Collocation**	**Freq.**	**LogDice**
1. SPEAK	307	11.623	6. NOT	164	10.106
2. LANGUAGE	297	10.751	7. OTHER	86	10.036
3. FRENCH	126	10.479	8. THEIR	110	9.997
4. AND	515	10.404	9. FIRST	62	9.773
5. LEARN	101	10.302	10. THEY	101	9.729

In MinD, there are 164 occurrences of 'not' in the vicinity of 'English', of which more than half points to one particular group – children who do not speak English as their first language (see Figure 1). While this seems neutral, the concordance lines below point to strong evaluative messages conveyed. Striking is the frequent reference to specific numbers ('three-quarters') and other types of quantifiers ('not one pupil', 'not a single', 'twice as many').

Figure 1: Concordance lines of the collocation pair 'English' and 'not'

for children whose first language is *not* **English** ,warned Tory councillor Imtiaz Ameen.

that 9.3 per cent (632,000) of pupils in	**English**	schools do *not* speak English as a first
of pupils in English schools do *not* speak	**English**	as a first language. Education chiefs claim
country where not a single child speaks	**English**	as his or her first language, and the results
others whose first language is *not*	**English**	,the groundbreaking scheme is promoting
Peterborough, where *not* one pupil speaks	**English**	as a first language," thundered Peter Hill
twice as many (11%) as in 1999 (6%) say	**English**	is not their first language. In 1999, only
than three-quarters of pupils do not have	**English**	as a mother tongue has introduced lessons

Out of the 164 concordance lines, 48 include a quantification of some sort. The references to numbers are used to amplify the 'concern' that English is not the first language for a growing number of pupils in schools in the UK. This is immediately visible in the choice of verbs that follow the quantifications such as 'thundered' or 'warned'. The alarmist stance can be further examined by studying the passages in which these statements were made. Extract 1 and 2 below show examples of expanded concordance lines from Figure 1.

Extract 1:

MILLIONS of pounds are being wasted on paying for bilingual classroom assistants in schools which have large numbers of ethnic minority pupils, it was claimed yesterday. The policy results in lowered standards and blighted prospects for children whose first language is not *English* because they are not being made to learn it in class. *The Daily Mail*

Extract 2:

"If you wonder what's gone wrong with Britain look no further than Gladstone Primary School, Peterborough, where not one pupil speaks *English* as a first language," thundered Peter Hill in The Express, without actually explaining why. Is Gladstone Primary a vision of a dystopian future or a triumph of multiculturalism? *The Guardian*

In the first extract, the view of bilingualism is clearly negative. It is linked with immigration and perceived as a burden for schools and communities. It is perhaps not surprising to see such views expressed in the middle-range tabloid *The Daily Mail* which is well known source of scaremongering. The message is reiterated in Extract 2 taken from *The Guardian*, but questioned, at the same time, by pointing to a lack of justification for such concerns. This appears to indicate the existence of a counter discourse. However, if we take into account the circulation of the sources, then we can understand what discourses are more likely to be widely disseminated in the public sphere. *The Daily Mail* with nearly 5 million readers is one of the bestselling newspapers in the UK and hence, the views expressed in this source are likely to be more widely shared. The readership of *The Guardian* is much smaller and oscillates around 1 million readers.[ii]

The brief corpus analysis of media representations of bi- and multilingualism exemplifies how corpus tools and techniques can be used to interrogate and interpret media data and what strengths and limitations they have. To demonstrate the wider applications of corpus tools to study aspects of language in the media, the next section discusses a selection of some of the current corpus studies on the subject.

3.2 Current contributions of corpus research to media linguistics

Corpus research concerned with aspects of language in the media can be broadly divided into two strands: 1) research that investigates structural, pragmatic and rhetorical features within and constitutive of a variety of media, and 2) research that is interested in studying language in relation to representation. Whereas the former draw on descriptive models offered in systematic functional linguistics or register analysis, the latter tend to combine corpus methods with concepts developed within discourse studies, often Critical Discourse Analysis (CDA).

Representative for the first strand of research are studies by Morley (2003), Biber (2003), Partington (2003), Bednarek (2006) and Duguid (2010). Morley (2003) examines strategies of persuasion adopted in a corpus of newspaper editorials. The analysis reveals that in contrast to news stories, persuasion in editorials takes place typically in the last paragraph. The main linguistic features of persuasion in this context include modal verbs, stance adverbials and the structure *it is* + evaluative adjective. By combining register analysis with a corpus approach, Biber (2003) studies the frequencies and forms of noun phrases in a corpus of British newspapers. The findings are compared with three other registers including academic, fiction and conversation. Although noun phrases are used with similar frequencies across all four registers, substantial differences are revealed regarding their forms. For example, more than 60% of all noun phrases in newspaper discourse contain a modifier and many have multiple modifiers. Partington (2003) explores rhetorical strategies in a corpus of 50 press briefings held at the Office of the White House Press Secretary from 1996 to 1999. The study demonstrates differences in the use of rhetorical strategies by the press and the podium. Whereas the press employs a wide range of tactics including belligerence, sarcastic reformulation or devil's advocacy, the podium restores predominantly to the strategy of lexico-grammatical parallelism. Bednarek's (2007) corpus research focuses on evaluation in press discourse. Her novel parameter-based framework combines some former approaches to evaluation, and simultaneously, offers new dimensions such as the distinction between the

core and peripheral evaluative parameters. Based on a corpus of 100 articles from British broadsheets and tabloids, her study offers comprehensive insights into the forms and functions of evaluation in news stories. Following the approach of the Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) (Partington 2004), Duguid (2010) explores changes in discourse informalisation over time by focusing on evaluative keywords in two large corpora of British broadsheets SiBoL 1993 and SiBol 2005[iii]. Her study points to an increase in a conversational and informal style represented, for example, by greater use of hyperbolic keywords.

Whereas research discussed thus far has focused on selected lexico-grammatical and pragmatic devices, the second strand of corpus research on language in the media is concerned with representations. This strand of research goes back to pioneering studies by Hardt-Mautner (1995) and Krishnamurthy (1996). Hardt-Mautner's (1995) work on the representations of Europe in the British press sets the framework for integration of corpus approaches with CDA, while Krishnamurthy's (1996) study on the representations of race and ethnicity also in the British press combines Corpus Linguistics with a lexicographical approach. Both studies instigated greater interest in discourse and social phenomena in Corpus Linguistics and inspired much of corpus-based analyses of language in the media. For example, Baker and McEnery (2005) examine the discourse of refugees and asylum seekers in the British press and in documents produced by the United Nations High Commissioner for Refugees. The analysis demonstrates that the groups were portrayed differently in the two sources. Whereas the press tended to represent refuges and asylum seekers negatively as potential invaders or out-of-control mass, the UN texts focused on global issues and help. However, there was also evidence pointing to some positive representations in the press. Interestingly, traces of some of the negativity surrounding refuges and asylum seekers could also be identified in the UN texts. A subsequent study by Gabrielatos and Baker (2008) extends the previous research by examining a much larger corpus of UK press articles, a wider range of terms including 'immigrant' and 'migrant', and including a wider coverage spanning 10 years. The analysis focuses on collocations and keywords, and on differences between broadsheets and tabloids. The results reveal a much more complex and ambiguous picture. While tabloids tend to convey negative discourses surrounding immigration, often reinforced by the use of pejorative metaphors, the broadsheets appear to use terms that have neutral or positive connotations. Similarly, Taylor (2014) investigates the representations of immigrants in the British and Italian press. Whereas previous work look at immigrants as a homogenous group, Taylor (2014) breaks up the category into different nationalities in order

to investigate the mismatch between the attention given to each national group and the actual demographics. Her study confirms that certain nationalities tend to be foregrounded and viewed negatively, although demographically, they do not necessarily constitute the largest groups. Taylor's research is a good example demonstrating how corpus results can be linked with external data to interpret findings in a more objective and replicable manner.

A substantial body of corpus research on media representations have been inspired by the aforementioned MD-CADS methodology. The term CADS (Corpus-Assisted Discourse Studies) was first coined by Partington (2004) in order to account for the growing discourse research that uses corpus tools and methods alongside qualitative techniques (Partington et al. 2013). The MD abbreviation stands for Modern Diachronic and adds the historical dimension to this research. CADS and its sister approach of MD-CADS are not affiliated with any specific school of discourse analysis and unlike CDA, do not pursue any specific political agenda (Partington et al. 2013: 10).

Using the MD-CADS approach, Marchi (2010) investigates the representations of morality in the British press within a decade in the SiBol corpora. These are interrogated for the use of terms pointing to morality. Her analysis points to an overall decline of moral terms, but this result could be 'seasonal', in that it might reflect a particular socio-political climate. Using the same methodology and data sets, Taylor (2010) investigates the role of science in British newspapers. By examining keywords in articles containing the term *scienc**, Taylor (2010) identifies considerable changes in the ways science has been conveyed within a decade. A rise in references to science has been noted as well as the use of science and scientific matters for sensationalist purposes.

Caldas-Coulthard and Moon (2010) examine representations of gender in the British press and provide empirical evidence for the persistence of gendered stereotypes in media reporting. This study shows a tendency to associate men with roles and status, and women with appearance and sexuality. Similarly, Jaworska and Krishnamurthy (2012) examine the representations of feminism in British and German press discourse. Starting with claims proposed by social and cultural scientists that feminism is marginalised and its representations dominated by negative and sexualised images, the researchers interrogated a corpus of British and German national newspapers from 1990 to 2009 to verify the extent to which such representations are salient. The analysis shows that feminism receives little attention in the press and there is a general negativity surrounding the movement. At the same time, there seems to be less evidence for sexualised images and a greater focus on the academic and intellectual status of feminism. Baker et al. (2013) examine representations of

Islam in the British press. Although newspapers seem to refrain from explicit islamophobic statements, the analysis point to overall negative representations. This negativity is expressed implicitly and in more subtle ways, for example, by persistently linking Islam and Muslims with extremism, terrorism and conflicts.

Alongside the focus on immigration, social movements, gender and religion, corpus tools and methods have also been used to study media representations of shock events including the recent financial crisis and Hurricane Katrina. Koller and Farrelly (2010) use examine keywords and metaphors in articles discussing the economy in British print media including *The Economist* and *The Financial Times* published just before, during and after the financial crisis. The analysis demonstrates interesting shifts in the use of keywords and metaphors; for example, at the beginning of the financial crisis, there were more keywords from the domain of markets and the economy. As the crisis evolved, the focus moved towards politics. Similarly, Storjohann and Schröter (2011) examine discursive representations of financial crisis in a corpus of German newspapers published in 2009. The study reveals that in the German press, the global financial crisis was mostly associated with items indicating consequences and less with lexis of causes and responsibilities. The media representations of Hurricane Katrina are investigated by Potts et al. (2015). Specifically, this research looks at how the disaster has been constructed as newsworthy in a corpus of news articles from 24 major American newspapers published as the storm hit. By examining lemma frequencies, collocations, key parts-of-speech and key semantic domains, the researchers demonstrate how the corpus tools can be effectively used to establish news values and complement the existing discursive approaches to newsworthiness (e.g. Bednarek and Caple 2014).

Within the corpus research on language in the media, there is also a strand interested in the representations of language and linguistic matters. For example, Ensslin and Johnson (2006) examine the discourse surrounding the phrase ‘the English language’ in a corpus of articles from *The Times* and *The Guardian*. The analysis reveals the existence of consistent and conflicting discourses that represent English as either a victim or a superior language. In a subsequent study, Johnson and Ensslin (2007) investigate the intersections of language ideologies with gendered representations by analysing the phrases ‘his language’ and ‘her language’ in the two broadsheets *The Times* and *The Guardian*. Language ideologies are also of concern to work conducted by Vessey (2013). By examining terms denoting English and French language, and their speakers in Canadian newspapers, Vessey (2013) shows how language ideologies permeate press discourse. Her research also offers a comprehensive

discussion about the methodological challenges that cross-linguistic and comparative corpus research needs to address.

3. FUTURE DIRECTIONS

This chapter presents only a limited selection of the corpus studies concerned with aspects of language in the media. The research discussed above was chosen to indicate the main research interests, themes and analytical tools adopted in corpus-based analyses. As this overview shows, there is a broad range of topics and issues that has been covered to date. Similar to CDA, most of the corpus studies start with a problem pertaining to society including politics, economy, religion, gender and science. Most of the studied areas involve relationships of dominance, power and control and researchers are often interested in uncovering the mechanisms of ideological work as expressed and legitimatised through language use. Also, most studies attempt to contextualise the results by carefully scrutinising the historical, political, social, cultural and political contexts surrounding the studied phenomena (e.g. Marchi 2010; Vessey 2013; Taylor 2014). Links are often made with research conducted in fields other than linguistics, including political, cultural and social sciences (e.g. Jaworska and Krishnamurthy 2012). Hence, the earlier criticism of Corpus Linguistics that it does not sufficiently account for the wider context cannot be upheld, certainly for corpus research concerned with aspects of language in the media.

Unlike other approaches to language in the media, the corpus research discussed above is normally based on a larger amount of data which is studied at first quantitatively. This can deliver results that are more reliable and generalisable than those obtained from examining a few texts as commonly done in studies following the qualitative discourse approach. It also fosters a greater distance to the data and increases objectivity of research in that it can help reduce some of the cognitive biases such as the primacy effect or confirmation bias (Baker 2006: 10-12). Moreover, corpus analysis enables us to see significant patterns and associations that are not immediately visible to the naked eye, unexpected or simply run counter to our intuition leaving room for serendipitous effects (Partington 2014). The ability of corpus tools to reveal the existence of repeated patterns also allows the researcher to see the discourses that are systematically and continually disseminated gradually influencing media audiences. As Baker et al. (2013) point out, there are many of ways to write about a topic, but certain ways will be preferred over others and these may indicate underlying ideological stances. It is precisely the advantage of a corpus analysis that it "will allow us to

see which choices are privileged, giving evidence for mainstream, popular or entrenched ways of thinking." (Baker et al. 2013: 25).

The corpus research discussed above has also demonstrated the benefits of combining quantitative corpus techniques with qualitative analytical tools. A quantitative analysis based on statistically calculated data in form of keywords or collocations is a good point of entry that can provide "a general pattern map" (Baker et al. 2008: 295) pointing to salient discourses. Patterns identified in this way can be subsequently studied qualitatively, for example, by grouping them into semantic categories or by closely examining them in the text via concordance lines. This can help the researcher to discover much more subtle or nuanced devices and aspects of media discourse that a pure quantitative analysis cannot reveal. As Baker et al. (2008: 295) highlight, combining quantitative corpus techniques with, for example, qualitative CDA procedures can create "a virtuous research cycle", which invites the researcher to approach the studied phenomenon deductively and inductively at the same time. In this way, new findings might emerge leading to new research questions and new interpretations.

Despite the many benefits of a corpus analysis, there are still a number of methodological challenges that future corpus research concerned with language in the media would need to address. Firstly, media are inherently multimodal but corpus research rarely accounts for this multimodality. This is partially due to the fact that it is difficult to develop tools that could automatically tag features of multimodal representations. Given that images are integral part of media language, future corpus research would need to explore the ways in which to systemically account for the visual features accompanying media texts. Secondly, given the pervasiveness of online communication, especially Social Media in public domains, future corpus research would need to move out of the realm of the press and direct more attention to discourses disseminated via various online channels. A constructive step in this direction is work by O'Halloran (2012) on argumentation strategies in online commentaries of news stories, by Potts (2014) on gendered identities in gameplay videos on YouTube and McEnery's et al. (2015) research comparing representations of an ideologically motivated murder on Twitter and in the national press in the UK. Fourthly, Corpus Linguistics is not a perfect method; no one is. Although researchers often claim greater objectivity and generalisability of findings, subjective judgments are not excluded and made throughout the research process including the selection of topics to study, data sources, sampling, examples chosen to study in more depth and interpretations. As Marchi and Taylor (2009) show, two researchers working independently on the same corpus of media data and interested in the

same questions may produce findings that are convergent, dissonant and complementary. Thus, the researchers call for a stronger consideration of triangulation at all stages of the research process including theory, method and data. Triangulation does not guarantee complete objectivity (no procedure does), but it can offer "analytical depth and creative potential" (Marchi and Taylor 2009: 18) which, in turn, helps increase reliability and validity of research (cf. Baker and Levon 2015). Related to this is the issue of production and reception of media texts – areas that are hardly ever addressed in corpus-based analyses. An impact of identified discourses on wider audiences is often assumed, but rarely empirically validated. Corpus Linguistics is not alone in having a difficulty with the relationship between media producers, media representations and media recipients. Other approaches to media texts that as CL focus primarily on linguistic representations also suffer from what Breeze (2011: 508) calls a "naïve linguistic determinism". Discourses that are frequently repeated may not be as powerful as it might be assumed. Equally, discourses that are rare in a corpus may turn out to be very influential. Validating insights from corpus analyses with other methods, such as ethnographic or sociolinguistic approaches used to study media production and reception (e.g. Cotter 2010, also Cotter in this volume) could offer invaluable insights into how discourses really work in a given discourse community. Finally, the vast majority of corpus research has been concerned with the context of media communication in post-industrial Western (and English-speaking) societies. Although this research has been essential in uncovering overt and covert discursive patterns underlying xenophobic, racist and sexist views, it is also grounded in the particular Western ideology, which comes with its own taken for granted views about what is right and wrong. As Blommaert (2005: 36) rightly observes: "The world is far bigger than the Europe and the USA", but far too often, the First World views have been projected to the globe. Studying contexts outside Western societies, especially along the south-north dimension, could not only reveal different patterns of media communication. It could also help develop a critical stance towards the ideologies underlying the First World so that they are not assumed a universal validity.

FURTHER READING

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

This publication is a comprehensive account of the use of corpus tools and methods in discourse analysis with lucid explanations and illustrative examples easy to digest by those with little or no corpus experience.

Partington, A. et al. 2013. *Patterns and meanings in discourse. Theory and practice in corpus-assisted discourse studies (CADS).* Amsterdam: Benjamins.

This book is an excellent introduction to the methodology of Corpus-Assisted Discourse Studies with many case studies investigating language use in the media.

McEnery, T. and Hardie, A. 2012. *Corpus linguistics. Method, theory and practice*. Cambridge: Cambridge University Press.

This useful textbook introduces students to the key concepts and tools of Corpus Linguistics. Each section includes a range of practical tasks and questions to check students' understanding of the discussed matters.

RELATED TOPICS

lexical priming, evaluative prosody, corpus stylistics, media representations, Critical Discourse Analysis (CDA)

REFERENCES

Anthony, L. (2011) *AntConc*. Version 3.2.2. Tokyo, Japan: Waseda University. http://www.antlab.sci.waseda.ac.jp/ (Accessed: 12 May 2015).

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

Baker, P. and Levon, E. (2015) 'Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity', *Discourse & Communication*, 9(2), pp. 221-236.

Baker, P. and McEnery, T. (2005) 'A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts', *Journal of Language and Politics*, 4(2), pp. 97-226.

Baker, P., Gabrielatos, C. and McEnery, T. (2013) *Discourse analysis and media attitudes*. Cambridge: Cambridge University Press.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, 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.

Barnbrook, G., Mason, O. and Krishnamurthy, R. (2013) *Collocation. Applications and implications*. London: Palgrave Macmillan.

Bednarek, M. (2006) *Evaluation in media discourse: analysis of a newspaper corpus.* London: Continuum.

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

Biber, D. (2003) 'Compressed noun-phrase structures in newspaper discourse: the competing demands of popularization vs. economy', in Aitchison, J. and Lewis, D. M. (eds) *New media language*. London and New York: Routledge, pp. 169-181.

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

Blommaert, J. (2005) *Discourse*. Cambridge: Cambridge University Press.
Bongaerts, T., van Summeren, C., Planken, B. and Schils, E. (1997) 'Age and ultimate attainment in the pronunciation of a foreign language', *Studies in Second Language Acquisition*, 4, pp. 447-465.
Breeze, R. (2011) 'Critical Discourse Analysis and its Critics', *Pragmatics*, 21(4), pp. 493-525.
Caldas-Coulthard, C. and Moon, R. (2010) 'Curvy, hunky, kinky: using corpora as tools for critical analysis', *Discourse & Society*, 21(2), pp. 99-133.
Cotter, C. (2010) *News talk: Investigating the language of journalism*. Cambridge: Cambridge University Press.
De Mejía, A. (2002) *Power, prestige and bilingualism: international perspectives on elite bilingual education*. Clevedon: Multilingual Matters.
Duguid, A. (2010) 'Newspaper discourse informalisation: a diachronic comparison from keywords', *Corpora*, 5(2), pp. 109-138.
Ensslin, A. and Johnson, S. (2006) 'Language in the news: investigating representations of 'Englishness' using WordSmith Tools', *Corpora*, 1(2), pp. 153-185.
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, pp. 5-38.
Gabrielatos, C. and Marchi, A. (2012) 'Keyness: Appropriate metrics and practical issues.' *CADS International Conference 2012*. 13-14 September, University of Bologna, Italy, http://repository.edgehill.ac.uk/4196/1/Gabrielatos%26Marchi-Keyness-CADS2012.pdf (Accessed: 12 May 2015).
Hardt-Mautner, G.(1995) 'Only connect: critical discourse analysis and corpus linguistic', *UCREL Technical Paper* 6. Lancaster: University of Lancaster. http://ucrel.lancs.ac.uk/tech_papers.html (Accessed: 12 May 2015).
Jaworska, S. and Krishnamurthy, R. (2012) 'On the F-word: a corpus-based analysis of the media representation of feminism in British and German press discourse, 1990-2009', *Discourse & Society*, 23(4), pp. 1-31.
Johnson, S. and Ensslin, A. (2007) 'But her language skills shifted the family dynamics dramatically. Language, gender and the construction of publics in two British newspapers', *Gender and Language*, 1(2), pp. 229-54.
Kilgarriff, A., Rychlý, P., Smrz, P. and Tugwell, D. (2004) 'The Sketch Engine', *Proc EURALEX 200*. Lorient: France, pp. 105-116.
Koller, V. and Farrelly, M. (2010) 'Darstellungen der Finanzkrise 2007/08 in den britischen Printmedien', *Aptum* 6(2), pp.170-192.
Krishnamurthy, R. (1996) 'Ethnic, racial and tribal: The language of racism?', in Caldas Coulthard, C. and Coulthard M. (eds.) *Texts and practices: readings in critical discourse analysis*. London: Routledge, pp. 129-149.
Louw, B. (1993) 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies', in Baker, M., Francis, G. and Tognini-Bonelli, E. (eds) *Text and technology*. Philadelphia, Amsterdam: Benjamins, pp. 157-76.
Marchi, A. (2010) 'The moral in the story: a diachronic investigation of lexicalised morality in the UK press', *Corpora*, 5(2), pp. 161-189.
Marchi, A. and Taylor, C. (2009) 'If on a winter's night two researchers… A challenge to assumptions of soundness of interpretation', *Critical Approaches to Discourse Analysis across Disciplines*, 3(1), pp. 1-20.
Mautner, G. (2007) 'Mining large corpora for social information: The case of elderly', *Language in Society*, 36, pp. 51-72.

McEnery, T. and Hardie, A. (2012) *Corpus linguistics. Method, theory and practice*. Cambridge: Cambridge University Press.
McEnery, T., McGlashan, M. and Love, R. (2015) 'Press and social media reaction to ideologically inspired murder: The case of Lee Rigby', *Discourse & Communication*, 9(2), pp. 237-259.
Morley, J. (2003) 'The sting in the tail: persuasion in English editorial discourse', in Partington, A., Morley, J. and Haarman, L. (eds.) *Corpora and Discourse.* Frankfurt/Main: Peter Lang, pp. 239-255.
O'Halloran, K. (2012) 'Electronic deconstruction: revealing tensions in the cohesive structure of persuasion texts', *International Journal of Corpus Linguistics*, 17(1), pp. 91-124.
Partington, A. (2003) 'Rhetoric, bluster and on-line gaffes: the tough life of a spin-doctor', in Aitchison, J. and Lewis, D. M. (eds.) *New Media Language*. London and New York: Routledge, pp. 116-125.
Partington, A. (2004) 'Corpora and discourse, a most congruous beast', in Partington, A., Morley, J. and Haarman, L. (eds.) *Corpora and Discourse*. Frankfurt/M: Peter Lang, pp. 9-18.
Partington, A. (2014) 'Mind the gaps. The role of corpus linguistics in researching absences', *International Journal of Corpus Linguistics*, 19(1), pp. 118-146.
Partington, A., Duguid, A. and Taylor, C. (2013) *Patterns and meanings in discourse. Theory and practice in corpus-assisted discourse studies (CADS).* Amsterdam: Benjamins.
Potts, A. (2014) 'Love you guys (no homo): How gamers and fans play with sexuality, gender, and Minecraft on YouTube', *Critical Discourse Studies*, published online on 7 November 2014, pp. 1-24.
Potts, A., Bednarek, M. and Caple, H. (2015) 'How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina', *Discourse & Communication*, 9(2), pp. 149-172.
Scott, M. (2008) *WordSmith tools. Version 5*. Lexical Analysis Software. http://www.lexically.net/wordsmith/version5/index.html (Accessed: 12 May 2015).
Scott, M. (2010) 'Problems in investigating keyness, or clearing the undergrowth and marking out trails', in Bondi, M. and Scott, M. (eds.) *Keyness in texts*. Amsterdam: Benjamins, pp. 43-58.
Sinclair J. (1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Storjohann, P. and Schröter, M. (2011) 'Die Die Ordnung des öffentlichen Diskurses der Wirtschaftskrise - und die (Un-)ordnung des Ausgeblendeten', *Aptum*, 7(1), pp. 32-53.
Stubbs, M. (2001) *Words and phrases. Studies in lexical semantics*. London: Blackwell.
Taylor, C. (2010) 'Science in the news: a diachronic perspective', Corpora 5(2), pp. 221-250.
Taylor, C. (2014) 'Investigating the representation of migrants in the UK and Italian press: A cross-linguistic corpus-assisted discourse analysis', *International Journal of Corpus Linguistics*, 19(3), pp. 368-400.
Van Dijk T. A. (1995) 'Discourse Semantics as Ideology', *Discourse & Society*, 6(2), pp. 243-289.
Vessey, R. (2013) 'Challenges in cross-linguistic corpus-assisted discourse studies', *Corpora*, 8(1), pp. 1-26.

[i] The articles included in the corpus were downloaded from Nexis UK. The search terms were *multilingual** and *bilingual**. To ensure that bi- and multilingualism were topical and not mentioned in passing, only articles in which these terms occurred 3 times or more were included in the corpus.

[ii] http://www.theguardian.com/advertising/guardian-circulation-readership-statistics

[iii] The abbreviation SiBoL is a portmanteau of the University of Sienna and the University of Bologna. More information about the corpora can be obtained from: http://www3.lingue.unibo.it/blog/clb/