ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Corpus linguistics in the study of news media

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Corpus linguistics in the study of news media / Anna Marchi. - STAMPA. - (2022), pp. 40.576-40.588. [10.4324/9780367076399-40]

Availability:
This version is available at: https://hdl.handle.net/11585/835833 since: 2022-03-10

Published:

DOI: http://doi.org/10.4324/9780367076399-40

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

19 April 2024

This is the final peer-reviewed accepted manuscript of:

[Anna Marchi, Corpus linguistics in the study of news media. In A. O'Keeffe and M. McCarthy *The Routledge Handbook of Corpus Linguistics, 2nd edition*, London and New York, Routledge/ Taylor & Francis, 2022, pp. 576 - 588**]**

The final published version is available online at: **[***https://dx.doi.org/10.4324/9780367076399-40***]**

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

40. Corpus linguistics in the study of news media

1. Why study news discourse using corpora?

The definition of "news discourse" can be rather broad and accommodate various forms of journalistic output. Bednarek and Caple describe it as: 'the kind of discourse we encounter when we turn on the television, when we open the newspapers, when we go online or when we switch on the radio to get our dose of daily happenings' (Bednarek & Caple 2012:1). The qualities which make news discourse an ideal territory for corpus linguistics are inbuilt in the definition: its relevance and its abundance.

The consumption of news discourse is a regular routine fulfilling a common need, as journalism represents 'one of the ways society tells itself about itself' (Dickinson 2008: 1), therefore journalism is a powerful agency of symbolic control, also because of its ubiquity. Every day news outlets produce an enormous amount of text. The *New York Times* published online approximately 230 pieces of content daily in 2016 (Meyer, 2016), a print edition of (pre-tabloid, i.e. pre January 15th 2018) *Guardian* contained, on average, 90,000 words on a weekday and more than twice as much on Sundays, the word-count of a 55 minutes *BBC World news* program is around 11,000 words. In the past two decades the emergence of news aggregator databases, such as *LexisNexis* or *Factiva*, has meant that newspaper content – and, to a lesser extent, broadcast transcribed content – is easy to access in a ready-to-search format, making it possible to compile comprehensive news corpora relatively quickly and with a reasonable level of accuracy.

Given the influence and the availability of news discourse, it not surprising that the majority of studies adopting various combinations of corpus and discourse approaches (see Marchi & Taylor 2019: 5)

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

are based on news corpora. Nartey and Mwinlaaru (2019) reviewed 121 studies combining corpus linguistics and Critical Discourse Analysis between 1995 and 2016 and found that the distribution of 'domains of engagement' was dominated by media 53% (prominently newspapers), followed by politics (38%) and social media (14%).

The work presented in this chapter will mainly focus on print, or newspaper journalism, as it represents the target of the vast majority of corpus studies of news discourse to date. The term newspaper journalism might seem rather outdated in an age of fluid formats and proliferation of platforms: some of the journalism classified as such might have never even been published on paper. Selling copies represents a negligible source of revenue for most news outlets. The *Independent* discontinued its print edition in 2016, since 2011 the *Guardian* has adopted a 'digital first' (Jarvis 2011) strategy; the readers of the print version represent nowadays a tiny portion of the readership of a newspaper, and in some cases (for instance with the *Daily Mail* compared to the *Mail Online*) a very different readership altogether. What can be problematic for a corpus linguist is that digital versions of news items often have different headlines and other variations, as well as including much more material, such as podcasts or live blogs. Under the simplified label 'newspaper journalism' I include the static, text-only output of mainstream traditional press; while I am convinced that traditional journalism is still alive and very relevant, I am also well aware that the definition is reductive and this will be discussed broadly both in the next section and in the final one.

Section 2 will be dedicated to methodological aspects, including corpus design, particularly with reference to corpus use. In sections 3 and 4 I will review key studies using newspaper corpora to investigate news form (its structure and function) and content (representations and ideologies).

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Finally, section 5 will reflect on the limitations of doing corpus work on news media texts as well as look into perspectives and potential developments.

2. News texts and news corpora

Newspapers are used as corpora, or in corpora, in different ways. They can be part of general corpora compiled to study language at large. For example newspapers and magazines constitute 30% of the British National Corpus (BNC 1994) and the BROWN family corpora include different types of press texts (44 reportage samples for each corpus: 1931, 1961, 1991 and 2006). There are then specialised corpora, either entirely composed of newspapers or where newspapers represent a main partition that can be compared with other types of text. One of the largest (non topic-specific) newspaper-only corpora is the SiBol corpus (or the Siena-Bologna family of newspaper corpora). Accessible on *SketchEngine*,[i] SiBol is a 650 million word collection of the whole output of British broadsheets (1.5 million articles) for the years 1993, 2005, 2010 and 2013 (the 2013 corpus was extended to include British tabloids and English language newspapers from different countries).

Studies of media corpora can be of two main kinds: studies of news structure and discourse function (reviewed in section 3) and studies of media representation/construction of specific topics (reviewed in section 4). Sometimes a research project will contain a mixture of the two types.

The corpus we use and what we want to do with it are the fundamental elements of any research: 'the corpus data we select to explore a research question must be well matched to that research question' (McEnery & Hardie 2012: 2). In other words, the question has to suit the corpus, *or*

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

the corpus has to suit the question. This may seem self-evident, but the issue is far from trivial and is perhaps worth some explicit reflection. The architecture of a corpus has enormous repercussions on what we can do with it. Clearly, if we are interrogating a pre-existing corpus we have to be aware of the limitations that come with adapting a data-set to purposes which are most likely different from those the corpus was originally built for. Let us take as an illustration a large, balanced, general corpus created for a broad range of linguistic research; one such example is the Corpus of Contemporary American English (COCA), which is freely accessible online and, at the time of writing, contains more than 1 billion words coming from a variety of sources, including newspapers over the past 20 years. The newspapers partition contains about 123 million words (approximately 4 million per year), sampled from all sections of ten US newspapers of very different nature (in terms of readership or geographical distribution); this corpus is an extraordinary resource for synchronic and diachronic genre analysis of news discourse, however, because of its composition, it is not particularly suitable for topic specific studies, say, for instance, an investigation of changing representations of the Muslim world in the US press. We *can* use general corpora to study specific social, cultural and political issues, but the generalisations we can make are constrained by external criteria (e.g. sampling balanced on the basis of source and news section) which may not match the scope of our research. This is why in order to study a specific topic, such as the one hypothesized, researchers prefer to create specialised corpora, compiled following criteria which attempt to strike the best balance between recall (i.e. including all relevant texts) and precision (i.e. including only texts that are relevant) – see Gabrielatos (2007: 6).

Obviously the definition of relevance depends not only on the specific topic at hand, but also on how we intend to approach that topic and this is reflected in the two main kinds of custom-

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

built newspaper corpora: search- term(s) based and whole output. Search-term(s) based corpora rely on the completeness and unambiguity of the query to achieve an accurate coverage of the topic. A well-formed query optimises precision and recall and, by reducing the overall volume of the data collection (compared to complete output), it also makes it easier to download and store data for a continuous stretch of time (as opposed to sampling points in time, e.g. years). On the other hand, with this type of design the re-usability of the corpus is rather limited, moreover we lose relevant information, most notably about the weight and distribution of the topic under scrutiny with respect to the overall content of the source. The size of newspapers has changed considerably over time and different publications differ substantially in the number and size of pieces they publish, therefore it may be important to be able to observe coverage of topic/event relative to the volume of the whole newspaper coverage. I discuss this in detail in Marchi (2018), in particular reference to diachronic research, and reach the conclusion that the ideal solution in terms of corpus compilation would be 'collecting the complete output of a source of data for a continuous stretch of time' (Marchi 2018: 178).

A minimization *a priori* impact of corpus-design still does not efface the fact that any corpus is *made* (no matter how "intelligent" the design) and the choices on which it is based not only will, but should determine its use. Even if we are compiling custom corpora using the entire output of a source, compilation requirements depend on the purpose of the research and rest on reasoned choices which may be arranged on three levels: source (i.e. where do data come from?), substance (i.e. what data are included?), and structure (i.e. how are data structured?).

Source

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

An initial distinction on data collection procedures and platforms depends on whether we are compiling a corpus of contemporary news or of historical newspapers.

There are established historical newspaper corpora, such as the *Zurich English Newspapers corpus* (or ZEN, a 1.2-million-word collection of newspapers, covering 120 years of British press history, between 1671 and 1791). Some corpora allow for long term diachronic research, for example the *Rostock Newspaper corpus*, comprising ten samples drawn from six British newspapers (10,000 words per newspaper) from 1700 to 2000 at 30 or 40-year intervals. However, for historical research as well, it has become relatively easy to compile custom corpora using newspaper archives (e.g. the *Times Digital Archive*) or commercial databases storing digitalised newspapers (such as *ProQuest*). The British Newspapers Archive stores almost 36 million pages dating from the 1700s; Library of Congress has an online repository of newspapers printed in the U.S. between 1789 and 1963. The main challenge for some historical corpus research can be making the format suitable for corpus analysis, i.e. converting scanned images to full-text, an issue which is widely discussed in Digital Humanities.

For contemporary datasets, that is corpora of newspapers published in the digital age, obviously there is not a problem of conversion from analogic format. There might however be an issue of variability of content depending on the point of access. Before *Nexis* some newspapers sold editions on CD-ROM, which were used to extract data for large newspaper corpora from the late 1990s and early 2000s (for example SiBol 1993 and SiBol 2005). News aggregators rapidly made disks obsolete, but reliance on aggregators still determined which sources would be used for research, depending on availability and level of accessibility. Not all newspapers are available on news-aggregators and the newspaper content directed to a database is never complete: some may be omitted (Leetaru 2011) and

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

the same newspaper may be available on different databases but the content may not coincide. Nevertheless the content stored from databases tends to be relatively stable over time. The alternative is harvesting online editions of the newspapers using web-crawlers (such as BootCat), or web-scraping scripts (using programming languages such as Python), in this case the content may be more complete, but it is subject to greater variability: articles may be (and frequently are) removed, edited or updated over time.

Substance

Web-based corpora grant greater richness but at the cost of greater noise, such as ads; some of the noise is easy to detect and remove, for example duplicates or near-duplicates, other "hybrid" material less so, for example paid for (journalistic) content. As Baroni and Ueyama point out though 'it is not correct to refer to the problems above as "problems of Web corpora"; rather they are problems of large corpora built in short time and with little resources, and they emerge clearly with Web corpora since the Web makes it possible to build "quick and dirty" large corpora' (Baroni & Ueyama 2006). Decisions need to be made concerning cleaning the corpus and how much cleaning to do. What to do, for example, with "boilerplate" information, that is repeated formulaic strings such as:

> It is the policy of the Guardian to correct significant errors as soon as possible. Please quote the date and page number. Readers may contact the office of the readers' editor
>
> [repeated 317 times in the Guardian 2005]

Independently from the source used to collect the data, decisions must be made in terms of what to include in the collection and what to exclude from it. For example registry articles (birth, marriages

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

and deaths), legal and court announcements, church services, or weather forecasts, all present very repetitive language which may skew the analysis.

Many issues may be solved by annotating the corpus. Mark-up allows us to include everything in the corpus and at the same time selectively isolate elements from the analysis. Exploiting the HTML code of web-pages or the editorial and descriptive metadata in the database storage makes some level of mark-up relatively easy to encode (for example encoding sections), but annotating a corpus remains a time-consuming task (see Chapter X [Reppen], this volume).

Structure

The substance of a corpus and its structure are strictly related, as they define the corpus architecture. Perhaps the single most important decision we make when designing a corpus is defining the organising principle of its structure: what will be the unit of analysis. The way data are stored determine the way they will be analysed. The level of aggregation of documents (e.g. 1 file = 1 daily edition or 1 file = 1 article) determines the granularity with which we can analyse the data (unless texts are marked-up in a way that gives us access to multiple levels). This is particularly relevant in the case of online newspapers, as texts published online are increasingly a blend of locally produced texts and material from other sources, but it has always been a concern since, as Hundt points out, 'more variation within traditional text categories (such as "newspapers") exists than between different text categories' (Hundt 2008: 171) and – she continues – 'the question is whether one year's worth of The Guardian or The Times can be considered a single-register corpus or not' (*ibid.* 179).

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

As Egbert and Schnur (2018) elegantly explain: 'just as neither leaves nor forests can exist without trees, textual features and corpora cannot exist without texts' (2018: 161). According to the authors the fact that the text is the sampling unit in corpus design and construction ensures that the corpus 'represents the way naturally occurring discourse is produced' (*ibid.*: 165). Individual texts then may or may not be the observational unit. This depends on the research question at hand – and we may need to consider different units for different purposes. However, failing to carefully consider what should constitute the building bricks of our corpus will dramatically limit what we can do with it. Some corpora are often exploitable for a long time after their origin. and we cannot envisage future research needs. Because of the 'serendipitous nature' (Partington 2009: 282) of corpus-assisted research, we cannot even predict the ramifications of our ongoing analysis. Therefore, the corpus structure has to be deliberate and flexible (see Cirillo et al. 2009).

The discussion of corpus design is limited by space constraints, but I hope I have managed to convey its importance: 'Data collection is often regarded as the easiest step of the entire analytical process, yet it is actually one of the most complex, with the quality of the collection and the preparation process impacting on every other stage of the project' (Leetaru 2011: 7). It may be true that newspapers are the favoured source for corpus-assisted research because they are easy to collect, but apparent convenience quickly fades when we reflect on the implications of corpus compilation. I will come back to a reflection on methodological and practical issues at the end of the chapter. The next two sections will focus on applications of corpus research in the analysis of news discourse, with a brief review of crucial work in the area.

3. News form

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Newspaper corpora are used in a variety of linguistic areas of interest, such as stylistics (Semino & Short 2004), pragmatics, historical linguistics (Studer 2008), sociolinguistics and, of course, discourse analysis (see section 4). Studies of formal features of news discourse, such as broad-spectrum lexico-grammatical studies and studies of textual position, provide an understanding of the discourse type (or types), which is fundamental for any study of newspaper corpora.

Biber (1988) developed a typology of texts on the basis of a model of variation across registers, where each dimension of variation comprises a set of lexico-grammatical features that occur frequently in texts and reflect shared communicative functions (e.g. exposition or persuasion), establishing essential knowledge about the structure and function of news. Dimensional studies, such as Biber's, describe the characteristics of a specific register (e.g. newspaper reportage) by comparing it with other registers (e.g. academic prose, editorials, conversation) or comparing different variables with respect to frequent co-occurring lexical and syntactic features (e.g. modal verbs, stance adverbs). Exploring variation, Biber et al. (1998) observed, for example, the high frequency of temporal adverbs in news and that temporal references such as *yesterday*, *last night* and *tomorrow* are more frequent in British news than in American news, while the latter shows higher frequency of mentions of days of the week. This difference might not exist anymore, as there probably was (but it would be worth checking and it would make a neat little study) an overall move to days of the week or specific dates when content was made available online, where readers can access it at any time and potentially forever.

Following Biber, Morley (2004) analysed persuasion in British newspaper editorials, by comparing them with news reports. Morley used comparison across text types, setting editorials against news

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

reports, as well as comparison within texts, contrasting the last paragraph of editorials with the rest of the text. He found an increased frequency in the use of modal verbs, adverbs of epistemic stance and the construction *it is* followed by an evaluative adjective followed by *that*/*to* in the conclusive paragraph, thus demonstrating how text-final position ('the sting in the tail') accomplishes a specific persuasive goal, and showing the correlation between the structure of editorial discourse and its function of telling the relevant authority what it must do.

Textual position in hard-news reports was the focus of a large scale project developed at the University of Liverpool (see Hoey and O'Donnell 2015) aiming, among other things, at testing Hoey's theory of lexical priming (2005), which states that words can be textually primed and that an aspect of priming is the position within the text where a word tends to appear. The researchers investigated frequency and usage of lexical-items in text-initial or paragraph-initial position, in a corpus of the Guardian Home-news section between 1998 and 2004. The corpus was subdivided into sub-corpora on the basis of structural sub-divisions (e.g. a sub-corpus of text-initial sentences, one of non-initial sentences, etc.) and the sub-corpora were compared in order to access positional frequency of lexical items and to analyse words that are primed to appear in text-initial position. The analysis proved strong associations of words and phrases with textual positions and showed recurrent patterns of usage depending on position within a text, in accordance with linguistic as well as extra-linguistic claims about the relationship between form and meaning. For example the function of the *nucleus* (White 1997) / headline-lead as the textual 'anchor point' (White 1997: 116) in newspaper discourse: i.e. the gravitational centre towards which subsequent text links back (see O'Donnell et al. 2012).

As mentioned earlier, broad studies such as these offer fundamental corpus-derived knowledge about the specificity of the texts under examination, for example, the fact that texts in a newspaper are not

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

homogeneous and neither are different stretches of text within an article. This kind of evidence is relevant to all research done on newspapers as it provides empirical anchoring for further investigations. For instance, starting from the set of register-defining categories (as identified by Biber et al. 1998), Duguid (2010) found a diachronic increment (between 1993 and 2005) in the frequency of lexico-grammatical items that are typical of spoken language, such as intensifiers and markers of vagueness, indicating an informalisation of newspaper discourse and a tendency to adopt common evaluative terms suggesting conversational style. This tendency has been progressing in the course of the eighteenth and nineteenth century and, according to Biber, 'over the past few decades, these changes towards more oral styles in newspaper language have accelerated' (Biber 2003: 170). This change over time is well documented in media studies and is known as "tabloidisation" (see for example Connell 1988 or McLachlan & Golding 2000), with reference to the classic distinction in the British press between popular newspapers or tabloids and quality newspapers or broadsheets. The comparison of broadsheet and tabloid corpora has been a fertile ground in stylistic research, see for example Takami (2004) on the morphological and semantic analysis 'broadsheet adjectives' and 'tabloid adjectives', or (Semino) 2006 on metaphorical uses of speech activity expressions. And this is just one example of convergent findings between corpus work (in particular in the area of corpus stylistics) and non-linguistic approaches and shows how a linguistic analysis of news discourse can yield valuable corroboration as well as insights for a broader understanding of how journalism operates.

4. News content

Corpus-Assisted Discourse Studies (CADS, as coined by Partington in 2004) and the constellation of variously labelled Corpus & Discourse approaches (see Marchi & Taylor 2018) interested in

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

investigating language from a social as well as from a structural perspective find a natural habitat in journalistic discourse corpora.

The paper that iconically started CADS is a study of newspaper editorials. Hart-Mautner (1995) investigates the discourse on the EU in the British press in a corpus of editorials from four British papers, dealing with the European Union/Community from 1971 to 1994. The corpus she uses is very small (168,000 words) but each article included in the corpus is not only relevant to the research question, but also has a high relevance in terms of journalistic status (since editorials encapsulate the official voice of a newspaper). This
ensures homogeneity (comparability across articles) and generalizability (representativeness of the findings with respect to press attitudes at large). One of the aspects she investigates in her work is the representation of participants, by comparing across newspapers positioned on the opposite sides of the political spectrum. Presumably aiming at optimal comparability (both for size and typology), she contrasted the views of two tabloids (shorter articles, personalised style and popular audience): the *Sun* (Conservative-leaning) and the *Daily Mirror* (Labour-leaning), and of two broadsheets (longer articles and more information-rich content covering a broader range of topics): the *Telegraph* and the *Guardian*. On top of being one of the earliest attempts to integrate corpus linguistics and CDA, which served as springboard for subsequent research and for reflection on the methodology, Mautner's paper is also an example of sound and interesting work with small-scale but carefully designed corpora.

Europe-related issues have been a rather popular topic in CADS (see for example Bayley & Williams 2012 on media representations of European identity, citizenship and institutions in a 23 million words corpus of newspapers and broadcast news from four European countries) and recently found renewed interest with Brexit (see for example Islentyeva 2021).Lutzky & Kehoe 2019 investigate the discourse

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

of Brexit in a corpus containing every article published on the *Guardian* website since 2000. Bednarek (2006a) examined patterns of evaluation surrounding the European constitution in British broadsheets and tabloids. From this study she constructed a parameter-based theory of evaluation (Bednarek 2006b) and described the composite manifestations of evaluative style in the British press and the pervasiveness of evaluation in news discourse.

Evaluation and evaluative prosody are a major focus and/or a major outcome of much CADS research done on media texts. Discourse studies (corpus-assisted or otherwise) are often concerned with issues of ideology (e.g. Taylor 2008 on metaphors of anti-Americanism in the British, Italian and American press), stereotyping (e.g. Jaworska, & Krishnamurthy 2012 on the representation of feminism in the British and German press), and the shaping of perceptions and attitudes through language (e.g. Bevitori 2014 on the social construction of 'knowing and believing' in editorial opinion discourse about climate change). An influential set of studies on stereotyping and polarization has been carried out at Lancaster University: Baker and McEnery (2005) looked at the discourses surrounding refugees and asylum seekers in two kinds of texts, newspaper articles and United Nation documents. They created a news corpus containing all articles mentioning the terms *refugee(s)* and/or *asylum seeker(s)* in 2003 and a corpus of the UNHCR documents on refugees for the same year. This piece of work is particularly interesting from a methodological point of view because, rather than comparing the corpora directly (e.g. using keywords analysis), the researchers analysed collocational patterns in the two sets of texts, they identified categories, such as 'quantification', 'movement', 'tragedy', 'aid' and 'crime', and they compared findings for the two corpora. This kind of analysis of collocates makes similarity visible as well as difference and allows for a comprehensive view on the representation of identities, that are often complex and intertwined, rather than being realised in clear cut discourses. This research was followed up and expanded in the project *Discourses of Refugees and Asylum*

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Seekers in the UK Press 1996–2006, known as RASIM (Baker et al. 2008). The researchers compiled a 140 million words corpus, spanning over a nine years period, which included a large selection of British broadsheets and tabloids; all articles dealing with refugees, asylum seekers, immigrants and migrants (Gabrielatos 2007) were included in the corpus. RASIM (Gabrielatos & Baker 2008) is a key in modern-diachronic CADS for the richness of its approach as it employs a variety of ways into the data (combining methods from corpus linguistics and CDA), a range of different perspectives onto the corpus (e.g. looking at diachronic change or stability of representations, at political stance differences and comparing styles between quality and popular newspapers) and a wide spectrum of tools. The researchers look at the distribution of RASIM stories identifying spikes of prominence, they do an extensive collocation analysis of the target terms and group collocates in semantic categories (giving access to discourse prosodies, stereotypes, analysis of metaphors, and so on), they look at "consistent-collocates" (that is, collocates that remain stable over time) showing how some discourses are progressively confirmed and reinforced. A similar corpus and a similar approach is adopted in Baker, Garbrielatos & McEnery (2012) (see Further reading section) to study the representation of Muslims and Islam in the British press between 1999 and 2005.

Another key set of studies in diachronic corpus-assisted analysis of newspapers (Partington et al. 2013) is the work done on the SiBol corpora (presented earlier in this chapter). Rather than tracking the evolution of discourses over a continuous period of time, with SiBol researchers can compare corpora from different points in time in order to identify change or stability. Since the SiBol corpora contain the whole output of the newspapers for the years under investigation, a wide range of different topics can be researched. Additionally, since the particular newspapers were the left-leaning *Guardian*, the right-leaning *Telegraph* and the centrist *Times*, socio-political, issues can be viewed and contrasted from different perspectives. For example Partington (2012) examined changing

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

patterns in the discourses relating to antisemitism from 1993 to 2009. Partington (2010) identified which social concerns were labelled 'moral panics' by the left and by the right in 1993 and then in 2005, to see which ones remained and which appeared and disappeared (see also Further reading section).

Mutation over time is a fundamental aspect of research on newspaper discourse. Another core aspect, which however has remained somewhat underexplored until recently, is the imprint of journalistic practices on news discourse. As Cameron says:

> If we are interested in the way language is used in media texts, it is useful to bring our analysis to bear not only on the texts themselves, but also on the rules journalists, broadcasters and their editors follow in the production of those texts; and where possible, on the institutional processes whereby those rules are promulgated, amended and enforced.
>
> (Cameron 1996: 332)

One way to do this is to investigate self-reflexive language in the media. Moschonas and Spitsmuller (2010), for example, analyse how *media language* is discussed in a corpus of Greek and German newspapers, focussing on the period from the mid-1990s to the early 2000s. Despite the small scale of the analysis (based on only 80 texts for each language, selected individually on the basis of relevance to the topic) and the consequent lack of a quantitative grasp, this study represents a rare example of analysis of metalanguage in journalistic texts. Journalistic self-reflexivity is at the core of my own work (see Marchi 2019 in Further reading section). The intersection between the study of journalistic products and journalistic practices is the new frontier of CADS research. Because of its intrinsic interdisciplinary vocation and 'serendipitous' nature (Partington, 2009: 286), CADS should

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

foster an holistic approach that takes into account the complex interaction between production, message and reception. A recent corpus-based attempt to assess reception is Keohe and Gee (2019) analysis of the comment section on the *Guardian* website; their take is however solely textual. Wright et al. (2020) look at the interaction between readers and producers combining a quantitative content analysis of comments with in-depth interviews with practitioners; but their study entirely disregards language. Most corpus work to date remains grounded on the message part and the discursive level. The most influential and innovative example of discursive approach to the analysis of production to date is Bednarek and Caple's (2017) discursive news values analysis (see Further reading section). The authors analyse how the newsworthiness of events is mediated through language and image and how newsworthiness values are constructed discursively through the repetition of patterns, both verbal and visual. This brings us to the limitations and perspectives of corpus-assisted research on newspaper corpora.

5. Limitations and perspectives

I reserved a large portion of this short chapter to corpus compilation, because we tend to think that our research begins once we have data, whereas it starts well before then. I believe it is important to emphasise this to make a more general point, that is, that we need to pay more attention to what is often overlooked, if not in the research process itself, then in its account: the 'blind spots' and the 'dusty corners' (see Marchi & Taylor 2018).

There are things we may miss when studying journalistic discourse using corpora (or, rather, using just corpora) and greater awareness about these limitations will help in overcoming them. When utilising large-size corpora, for example, we may fail to consider the diversity of text types a newspaper contains, as well as other textual aspects which would be available to a finer-grained

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

investigation, such as intertextuality or attribution and averral. This kind of drawback is relatively easily solved; in fact corpus-assisted approaches encourage a constant shunting between the 'big picture' and close reading. However a 'corpus can only reveal its own contents' (Baker 2006: 181) and text corpora solely contain text (and sometimes annotation for para-textual features). Because of the rapid evolution beyond the printed paper, the growing importance of multimodal texts and the overall impact of digital technologies on the news media industry, a mono-semiotic and mono-source analysis is reductive (i.e. limited) and forcedly reductionist (i.e. unduly oversimplified). We need a more comprehensive approach, one that takes into account multiple signs, but also one that goes beyond the message and takes into account aspects of production and reception. We need a holistic approach which is 'practice oriented' (Richardson 2008: 153) and contextualized. The linguistic analysis should be informed by extra-linguistic elements, incorporate an investigation of the broader context and, ideally, embed an interdisciplinary approach. Luckily an 'intrinsic interdisciplinary vocation' (Marchi et al. 2017: 174) and its 'omnivorous interests' (*ibid.*) are CADS's strong suits and therefore we are well equipped to work in this direction.

The chapter you just read comes with its own baggage of flaws, perhaps the most obvious is that the work presented, as well as the considerations made about journalism, are self-evidently Anglo-centric. This is in part due to space constraints, but it is also a warning on the need to be aware of, and reflect on, our partial perspectives and our individual primings.

For further reading I chose three monographs, a collection of edited articles and a recent journal article, which, in my opinion, combine methodological soundness and innovation. All choices also share the characteristic of being at the same time inward looking (i.e. they encourage awareness of and reflection on the impact of research practices) and outward looking (i.e. they encourage attention towards the extra-linguistic context and journalistic practices).

Further reading

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Baker, P., Gabrielatos, C. & McEnery, T. (2012) *Discourse Analysis and Media Attitudes: The British Press on Islam*, Cambridge: Cambridge University Press. (This book is a fundamental reference for diachronic corpus analysis and an achievement in terms of social relevance of the research and of its impact beyond linguistics and beyond academia. It tackles questions about the British media and society, looking at the representations of Islam and Muslims in the press and revealing how prejudice is discursively mediated).

Bednarek, M. & Caple, H. (2017) *The Discourse of News Values. How Organizations Create Newsworthiness*, Oxford: Oxford University Press. (This volume overcomes many of the limitations associated to corpus approaches to discourse as well as to text-only approaches in general. The authors propose a multisemiotic approach to the study of media discourse and show how the newsworthiness of events is constructed discursively through words and images).

Hansen, K. R. (2016) News from the future: A corpus linguistic analysis of future-oriented, unreal and counterfactual news discourse, *Discourse & Communication*, 10(2): 115-136. (This study is 'a grammatically founded approach to future-oriented journalism' [p. 116], i.e. of journalism reporting on planned events and discussing consequences, expectations, and agendas. By analysing patterns of use of modal verbs, verb tenses and speech acts, this longitudinal study of four Danish newspapers describes the shift towards a less event-centred and more analytical kind of journalism).

Partington, A. (ed.) (2010) *Corpora Special Issue. Modern Diachronic Corpus-Assisted Studies*. Edinburgh, Edinburgh University Press. (Another fundamental reference for diachronic analysis of newspaper corpora. The six articles in this edited collection cover different types of corpus analysis

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

as well as a broad range of topics, using the same dataset: the *SiBol* corpora).

Marchi, A. (2019) *Self-Reflexive Journalism. A Corpus Study of Journalistic Culture and Community in the Guardian,* London: Routledge. (This is a book-length study of how journalists define the boundaries of their community and set the standards of good vs. bad journalism in constructing and negotiating representations of their profession. The study has a strong methodological focus, employing and encouraging eclecticism and flexibility in the use of tools and constant scrutiny of the impact of research practices).

References

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

Baker, P & McEnery, T. (2005) A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts, *Journal of Language and Politics* 4(2): 197-226.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T. & Wodak, R. (2008) A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press, *Discourse & Society* 19(3): 273-305.

Baker, P., Gabrielatos, C. & McEnery, T. (2012) *Discourse Analysis and Media Attitudes: The British Press on Islam*, Cambridge: Cambridge University Press.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Baroni, M. & Ueyama, M. (2006) Building general- and special-purpose corpora by Web crawling, *Proceedings of the 13th NIJL International Symposium*, 31-40.

Bayley, P. & Williams, G. (eds.) (2012) *European Identity: What the Media Say*, Oxford: Oxford University Press.

Bednarek, M. (2006a) Evaluating Europe - Parameters of evaluation in the British press. In C. Leung & J. Jenkins (eds.). *Reconfiguring Europe - the Contribution of Applied Linguistics (British Studies in Applied Linguistics 20)*, London: BAAL/Equinox, pp. 137-156.

Bednarek, M. (2006b) *Evaluation in Media Discourse*, London: Continuum.

Bednarek, M. & Caple, H. (2012) *News Discourse*, London: Continuum.

Bednarek, M. & Caple, H. (2017) *The Discourse of News Values*. *How Organizations Create Newsworthiness*, Oxford: Oxford University Press.

Bevitori, C. (2014) 'Values, Assumptions and Beliefs in British Newspaper Editorial Coverage of Climate Change,' in C. Hart & P: Cap (eds.) *Contemporary Critical Discourse Studies*, London: Bloomsbury Academic, pp. 603-625.

Biber, D. (1988) *Variation across Speech and Writing*, Cambridge: Cambridge University Press.

Biber, D. (2003) 'Compressed noun phrase structures in newspaper discourse: The competing demands of popularization vs. economy,' in J. Atchison & D. Lewis (eds.) *New Media*

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Language, London and New York: Routledge, pp. 169-181.

Biber, D., Conrad, S. & Reppen, R. (1998) *Corpus Linguistics. Investigating Language Structure and Use*, Cambridge: Cambridge University Press.

Cameron, D. (1996) Style policy and style politics: a neglected aspect of language of the news, *Media Culture Society* 18: 315-333.

Cirillo, L. Marchi, A. & Venuti, M. (2011) 'The Making of the Cordis Corpus. Compilation and Markup,' in J. Morely & P. Bayley (eds.), pp. 13-33.

Connell, I. (1998) Mistaken identities: tabloid and broadsheet news discourse, *The Public* 5(3): 11-31.

Dickinson, R. (2008) Studying the sociology of journalists: the journalistic field and the news world, *Sociology Compass* 2: 1-17.

Duguid, A. (2010) Newspapers discourse informalisation: a diachronic comparison from keywords, *Corpora* 5(2): 109-138.

Egbert, J. & Schnur, E. (2018) 'The role of the text in corpus and discourse analysis: missing the trees for the forest,' in C. Taylor & A. Marchi (eds.), pp.159-173.

Gabrielatos, C. (2007) Selecting query terms to build a specialised corpus from a restricted-access database, *ICAME Journal* 31: 5-43.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Gabrielatos, C. & Baker, P. (2008) Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005, *Journal of English Linguistics* 36(1): 5-38.

Hoey, M. (2005) *Lexical Priming: a New Theory of Words and Language*, London and New York: Routledge.

Hoey, M. & O'Donnel, M. B. S. (2015) 'Examining associations between lexis and textual position in hard news stories, or according to a study by…,' in N. Groom, M. Charles & J. Suganthi (eds.) *Corpora, Grammar and Discourse. In Honour of Susan Hunston*, Amsterdam: John Benjamins, pp. 117-144.

Hundt, M. (2008) 'Text corpora,' in A. Lüdeling & M. Kytö (eds.) *Corpus Linguistics: An International Handbook. Vol. I*, Berlin: de Gruyter, pp. 168-186.

Islentyeva, A. (2021) *Corpus-Based Analysis of Ideological Bias: Migration in the British Press*, London and New York: Routledge.

Jarvis, J. (2011) Digital first: what it means for journalism, *The Guardian*, 26th June 2011.

Jaworska, S. & Krishnamurty, 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 and Society* 23(4): 401-431.

Kehoe, A. & Gee, M. (2019) "Thanks for the donds". A corpus linguistic analysis of topic-based

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

communities in the comment section of The Guardian, in U. Lutzky & M. Nevala (eds.) *Reference and Identity in Public Discourse*, Amsterdam: John Benjamins.

Leetaru, K. (2011) *Data mining methods for content analysis: An introduction to the computational analysis of content*, London and New York: Routledge.

Lutzky, U. & Kehoe, A. (2019) 'Friends don't let friends go Brexiting without a mandate,' in V. Koller, S. Kopf & M. Miglbauer (eds.) *Discourses of Brexit*, London and New York: Routledge.

Marchi, A. (2018) 'Dividing up the data. Epistemological, methodological and practical impact of diachronic segmentation,' in C. Taylor & A. Marchi (eds.), pp. 174-196.

Marchi, A. (2019) *Self-Reflexive Journalism. A Corpus Study of Journalistic Culture and Community in the Guardian,* London: Routledge.

Marchi, A., Lorenzo-Dus, N. and Marsh S. (2017) Churchill's inter-subjective special relationship: a corpus-assisted discourse approach. In S. Marsh and A. P. Dobson (eds.) *Churchill's special relationship: commemorating the 70th anniversary of the Fulton Iron Curtain speech*. Oxford: Oxford University Press.

Marchi, A. & Taylor, C. (2018) 'Introduction,' in C. Taylor & A. Marchi (eds.), pp. 1-15.

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

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

McLachlan, S. & Golding, P. (2000) 'Tabloidization in the British press: a quantitative investigation into changes in British newspapers, 1952–1997,' in C. Sparks & J. Tulloch (eds.) *Tabloid Tales: Global Debates over Media Standards*, Maryland: Rowman & Littlefield Publishers, pp. 75-90.

Meyer, R. (2016) How Many Stories Do Newspapers Publish Per Day?, *The Atlantic*, 26th May 2016.

Morley, J. (2004) 'The sting in the tail: persuasion in English editorial discourse,' in A. Partington, J. Morley & L. Haarman (eds.) *Corpora and Discourse*, Bern: Peter Lang, pp. 239-255.

Morley, J & Bayley, P. (2009) *Corpus-Assisted Discourse Studies on the Iraq Conflict*, London and New York: Routledge.

Nartey, M. & Mwinlaaru, I. N. (2019) Towards a decade of synergising corpus linguistics and critical discourse analysis: a meta-analysis, *Corpora* 14 (2), 203-235.

O'Donnel, M.B., Scott, M., Mahlberg, M. & Hoey, M. (2012) Exploring text-initial words, clusters and concgrams in a newspaper corpus, *Corpus Linguistics and Lingustic Theory* 8(1): 73-101.

Partington, A. (2004) 'Corpora and discourse, a most congruous beast,' in A. Partington, J. Morley & L. Haarman (eds.) *Corpora and Discourse*, Bern: Peter Lang, pp. 11-20.

Partington, A. (2009) 'Evaluating Evaluation and Some Concluding Thoughts on CADS.' in J.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Morely & P. Bayley (eds.), pp. 261-303.

Partington, A. (ed.) (2010) *Corpora Special Issue. Modern Diachronic Corpus-Assisted Studies*, Edinburgh: Edinburgh University Press.

Partington, A. (2012) The changing discourses on antisemitism in the UK press from 1993 to 2009: A modern-diachronic corpus-assisted discourse study, *Journal of Language and Politics* 11(1), 51-76.

Partington, A., Duguid, A. & Taylor, C. (2013) *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*, Amsterdam: John Benjamins.

Richardson, J. E. (2008) Language and journalism. An expanding research agenda, *Journalism Studies* 9(2): 152-160.

Semino, E. (2006) 'A corpus-based study of metaphors for speech activity in British English,' in A. Stefanowitsch & S. Gries (eds.) *Corpus-Based Approaches to Metaphor and Metonymy*, Berlin: Mouton de Gruyter, pp. 36-62.

Semino, E. & Short, M. (2004) *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing,* London and New York: Routledge.

Studer, P. (2008) *Historical Corpus Stylistics: Media, Technology and Change*, London and New York: Continuum.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Takami, S. (2004) 'A corpus-driven identification of distinctive words: "Tabloid adjectives" and "Broadsheet adjectives" in the Bank of English,' in N. Inoue & T. Tabata (eds.) *English Corpora under Japanese Eyes*, Amsterdam: Rodopi, pp. 93-113.

Taylor, C. (2008) Metaphors of anti-Americanism in a corpus of UK, US and Italian newspapers, *ESP Across Cultures* 5: 137-152.

Taylor C. & Marchi, A. (ed.) (2018) *Corpus Approaches to Discourse: A Critical Review*, London and New York: Routledge

White, P. (1997) 'Death, disruption and the moral order: the narrative impulse in mass-media "hard news" reporting,' in F. Christie & J. R. Martin (eds.) *Genre and institutions. Social processes in the workplace and school*, London: Continuum, pp. 101–133.

Wright, S. Jackson, D. & Graham, T. (2020) When Journalists Go"Below the Line": Comment Spaces at The Guardian (2006–2017), *Journalism Studies*, 21:1,107-126.

[i] https://www.sketchengine.eu/sibol-corpus/

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.