This is a pre-copyedited version of Chapter 1 of *The Discourse of News Values* and is not identical to the post-production version to be published in the book (due for release in March 2017). To cite this pre-publication chapter please include all of the following information:

Bednarek, M. and Caple, H. (2016) 'Introduction' pre-publication version, available at: www.newsvaluesanalysis.com [Final version to be published as Chapter 1 in: Bednarek, M and Caple, H. (2017) *The Discourse of News Values: How News Organizations Create Newsworthiness*. New York: Oxford University Press.]

CHAPTER 1 INTRODUCTION

1. The discourse of news values

This book is about words, images and the construction of newsworthiness. By way of introduction, consider these three news items:

(1) Women feature in only 7 per cent of sports programming in Australia, representing a backwards step compared to a decade ago and highlighting a significant gender gap in a country where sport is king, a new report shows. (abc.net.au, published and accessed on 13 April 2015)

(2) Captain Adriano Binacchi, who manned the stranded, [sic] Carnival Spirit, is officially the world's most non-plussed sea captain. His ship took on 6-10m swells, but in taking questions from media his overall attitude seemed to be "no big deal". When asked if facing such violent sea conditions is rare he replied:
"Not really, it's not my first time."
Were there any injuries sustained on board?
"No injuries, just some minor sea sickness."
Damage to the ship?
"What damage? Maybe some glass window panes. Minor things."
(theguardian.com/au, published and accessed on 22 April 2015)[1]

(3) Figure 1.1

The Atlantic Photo: Jeff J Mitchell/Getty Images

Figure 1.1 A news photograph of migrants walking through Slovenia

In this book we are interested in how such verbal and visual texts provide an answer to the putative audience question *how is this news?* In other words, how do semiotic (meaning-making) devices justify the newsworthiness of reported events or issues? Let's look at example (1) first: This item mentions that the reported issue concerns the country in which the audience lives (*in Australia*), that it is negative (*a backwards step*) and of a large scale (*a significant gender gap*) and that it has only just come to light (*a new report shows*). In fact, if we read on, we realise that this item refers to a report published in 2010 (*Towards a Level Playing Field: Sport and Gender in Australian Media*) and therefore somewhat artificially constructs it as new or recent information.

Moving on to example (2), this is unusual in that it includes a news worker's interview questions in addition to the interviewee's answers. These questions appear designed to elicit statements that the event was unusual (*rare*) and had negative effects (*injuries*, *damage*), but such answers are not provided by the interviewee. Neither does he construct the event as of a large scale; on the contrary, he uses the adjective *minor* several times (*minor sea sickness*, *minor things*). This makes it difficult for the news worker to use his quotes to construct the event as newsworthy in terms of unusuality and major negative consequences. Rather, the news worker turns the captain (and the interview) into a newsworthy story – the captain is evaluated as *officially the world's most non-plussed sea captain* and an unexpected contrast is established between the size of the waves (*6-10m swells*) and his attitude (*no big deal*). Both of these examples show how news workers manipulate **linguistic** resources to construct events as newsworthy.

In example (3), a long line of people (the caption tells viewers that they are migrants) are depicted walking through farm land along a raised bank. The fact that the image frame crops out both the beginning and the end of this line of people suggests that their size or scale cannot be fully accounted for in this one image, or may even be beyond reckoning. Here **visual** resources have been manipulated to construct this happening as newsworthy, i.e. of extremely large scale or scope. In all three examples, semiotic resources are hence used to establish events as newsworthy, persuading the audience that an item is worthy of being published as news and worthy of their attention.

This book is about how news organisations – metaphorically speaking – 'sell' the news to us **as news** through verbal and visual resources, through what we might call the discourse of news values. News values are those values that have been recognised in the literature as defining newsworthiness. These include those constructed through discourse in examples 1, 2 and 3: Proximity (nearness to the audience), Negativity, Superlativeness (large scale/scope), Timeliness (e.g. recency, newness), and Unexpectedness (e.g. unusuality) as well as others. We will provide a comprehensive definition, a full overview and explanation of these news values in Chapters 2 and 3.

We need to point out here that the term (*news*) *values* is sometimes used by news organisations themselves, for example on their websites. Thus, the websites bbc.co.uk and ap.org (*Associated Press*) each have a section called 'our values' (*BBC*) or 'news values & principles' (*AP*). Sometimes similar values are included in sections labelled 'standards and ethics' (*The New York Times*) or in a code of practice (*Al Jazeera*).[2] The types of values or standards that these news organisations profess to the world include:

- Trust, independence, impartiality, honesty, focus on audience, quality and value for money, creativity, respect, diversity, team spirit (*BBC*);
- Truth, speed, accuracy, preciseness, honesty, integrity, fairness, independence, transparency, ethical behaviour, careful/unbiased/unaltered, transmitted in many ways (*AP*);
- Truth, fairness, impartiality, transparency, integrity, accuracy, independence (*NYT*);

- Truth, factuality, accuracy, clarity, honesty, courage, fairness, impartiality, balance, independence, credibility, diversity, respect of audience, transparency, diversity, support of colleagues (*AJ*)

Such journalistic values are also mentioned in introductions to newswriting (e.g. Bender et al 2009: 136-139), and some academics use the term *news values* to discuss them (e.g. Fuller 1996, Palmer 1998, Johnson and Kelly 2003). These values are clearly important for journalism, but it is also clear that they are very **different** to the 'newsworthiness values' that we have introduced in relation to examples 1-3 above. They are examples of moral-ethical (e.g. truth, impartiality, honesty, fairness) and commercial values (e.g. speed, access via multiple platforms). We have analysed elsewhere how news organisations create value for themselves through referencing these in marketing and publicity material (Bednarek and Caple 2015).[3] Such values can also be constructed through semiotic resources in news products – e.g. via speech/dressing styles, signature music, or set design (Bell and van Leeuwen 1994, van Leeuwen 1984, 1989, 2006b), but they are not the focus of this book. As mentioned above and further explained in Chapter 2, when we use the term *news values* we refer solely to 'newsworthiness' values. Our goal is to introduce readers to how we can systematically analyse how these news values are constructed discursively, that is, through verbal and visual resources. The shorthand that we use for our approach is *discursive news values analysis,* or *DNVA.*

2. Why study news values?

The key areas of enquiry that inform our research in this book are media linguistics, corpus linguistics, discourse analysis, multimodality, and social semiotics, with a focus on the professional context of journalism. We aim to provide new insights into journalistic texts as social and semiotic **practice**, which can inform how we teach and learn about such texts in first and additional language contexts – i.e. media literacy – as well as how we teach students to create such texts – i.e. journalism education. We are also interested in making a contribution to **research**, offering a new perspective on how to study news discourse.

There is a wealth of insightful linguistic research on news discourse, for example on ideology (e.g. van Dijk 1988a, b, Fowler 1991, Richardson 2007, Baker et al 2013a), audience design (e.g. Bell 1991, Jucker 1992), register and genre (e.g. White 1997, Biber et al 1999, Lukin 2010, Smith and Higgins 2013), newsroom practice (e.g. Cotter 2010, Perrin 2013), or the socio-historic development of news discourse (e.g. Conboy 2010, Facchinetti et al 2012) – to name but a few topics. New introductions to news discourse are also published (e.g. Bednarek and Caple 2012a, Busà 2014). All this illustrates the continuing importance and relevance of the semiotic practices of journalism today. However, the concept of news values has not figured prominently in most of these studies (see Chapter 2). While the body of research on news values is vast and diverse, this exists mostly within non-linguistic disciplines such as journalism and communications studies, which lack a systematic analysis of verbal and visual text.

But why should we study news values? As this book hopes to illustrate, DNVA aims to have both descriptive and explanatory potential, and means to answer a range of questions about news practice. This includes questions around the conventionalised resources or rhetoric of newsworthiness: DNVA can offer insights into what semiotic resources are repeatedly employed to establish particular news values (Bednarek and Caple 2014). In this way, DNVA can identify common practices, conventions and clichés of news reporting and offer insights into news as semiotic practice, either at a particular point in time or across news cycles (Potts et al 2015). Moving beyond this micro level of semiotic construction, it is also possible to use this type of analysis to explore if particular topics – such as indigenous news actors, asylum seekers/refugees, marriage equality, climate change – are associated with specific news values. Such repeated associations may then have ideological implications, and

DNVA can thus be used as a tool for Critical Discourse Analysis (see Bednarek and Caple 2014 for further discussion of the critical potential of DNVA). Again, it is possible to undertake such analysis diachronically and across cultures. The aim here is to see if specific news values are emphasised, rare or absent in reporting on particular topics or events, and in how far this is constrained by the event itself.

Further, DNVA can be used to analyse the packaging of news **as** news, for example in combination with attribution analysis (Bednarek 2016a). Such analysis makes it possible to see how news values are integrated and structured in the form of consumable news products and whether audience members engage with the voice and authority of the news organization or of sources (Bednarek and Caple 2012a: 214). Also in relation to packaging news, DNVA can be applied to examine the role that different (verbal/visual) components play – whether or not they reinforce, complement or contradict each other – and to identify un/successful practices for multimodal news stories. This fits with research interests in intersemiotic relations (Caple 2013a). All of the above types of analyses can be undertaken in relation to particular news outlets or outputs, including but not limited to differences between the so-called 'popular' and 'quality' press.[4] Such analyses can also bring in the notion of audience positioning, as each news outlet will have their own target audience.

Last, but not least, there are potential applications in journalism education: By analysing how news professionals construct newsworthy stories we can make explicit the tacit knowledge and experience that such professionals have and provide insights into contemporary journalistic norms and practices. Journalism students can then be made aware of these practices, for instance by deconstructing actual news stories for their construction of news values before constructing their own multimodal journalistic texts (Caple and Bednarek 2015). In so doing, students gain a fuller understanding of what news discourse is and how newsworthiness is created through different semiotic resources.

DNVA has been an ongoing research interest for both authors for a number of years. Bednarek and Caple (2012a, b) are our earliest publications on this – one is an introduction which we use with our students (2012a), while the other is an example analysis of one environmental online news story (2012b). We have explored the role of corpus linguistics in DNVA using small and large corpora (Bednarek and Caple 2014, Potts et al 2015, Bednarek in press). At the same time, Caple has been the lead researcher in publications where we focus on visual DNVA (Caple 2013a, Caple and Bednarek 2015). While most of this research focuses on print/online news, Bednarek (2016a) has started exploring broadcast news. This cumulative research experience has led us to the conclusion that the discursive approach to news values analysis deserves book-length treatment, where it can be more fully explored and accounted for.

3. Key terms

Before providing further information on the approaches that we will use in this book, it is necessary to briefly introduce some key terms: *news*, *discourse*, *multimodality* and *corpus linguistics*.

News

In everyday usage, the word *news* is frequently used to refer to new information. We might ask each other if there's any *news* or check our Facebook *news*feed. Here the source of the information (friends, family, strangers), its domain (public/private), and the type of information (gossip, opinion, announcement, cartoon) can vary. In this sense, the words *news* and *newsworthy* can be used to refer to new information presented in personal narratives or casual conversation (Sidnell 2010: 228). In other broad uses, the term *news* has been applied to all discourse around a particular hashtag including tweets by bloggers and activists

(Papacharissi and Oliveira 2012). In such and similar approaches, *news* becomes a broad concept that appears simply to refer to new content. Sometimes, the term *news* is used to refer to language as used in a newspaper and may include both editorials (opinion) and reportage – as is the case with Biber et al's (1999) news register, for instance.

In this book, we use *news* (and *newsworthy*) in a more specific way, as it relates to news reports disseminated by news organisations. As Fuller (1996: 6) states, most journalists would agree that 'News is a report of what a news organisation has learned about matters of some significance or interest to the specific community that news organisation serves'. Such a definition also brings into focus the notion of target audience (the specific community that a news organisation serves). As will become evident throughout the book, we argue that news values are dependent on target audiences and other contextual factors.

In relation to news, we also talk about reported events, broadcast news, and time and place of publication. When we use the term *event*, we use it as a cover term for events, issues and happenings, including elements or aspects of these. For example when we talk about how events are constructed as newsworthy, this includes the event's news actors or its location. Broadcast news may include audio and video published online or through podcasts, not just on radio or television. Thus, *publication* is used in a broad sense to cover the publication or transmission of stories online, on mobile devices, in print, on the radio or on television. Similarly, when we talk about *published* stories, we also mean broadcast stories.

In sum, this book is concerned with news reporting, including but not limited to hard news, soft news and research news.[5] We do not deal with other journalistic texts such as advice, opinion, reader emails, interviews, or quizzes. As fully explored elsewhere (Bednarek and Caple 2012a), news reporting exhibits unique semiotic characteristics, for example particular genre structures, uses of visuals, and lexical and syntactic features (e.g. nominalisation, evidentiality). In this book, we focus on exploring the semiotic resources of news discourse for their potential to construct news values, rather than providing a general introduction to these unique features.

Discourse and multimodality

Definitions of *discourse* are plentiful and have been discussed in different disciplines (e.g. Baker 2006: 3-5). One key distinction that is made in linguistics is between discourse as language in use and 'a more Foucauldian perspective, where discourses are seen as ways of looking at the world, of constructing objects and concepts in certain ways, of representing reality in other words, with attendant consequences for power relations' (Baker and McEnery 2015: 4-5). We align ourselves with the first perspective on discourse (language in use), but consider *discourse* as multi-modal. Strictly speaking, texts that are 'multi-modal' combine two or more modalities (e.g. visual, aural), whereas 'multi-semiotic' texts combine two or more semiotic (meaning-making) systems such as image or language (O'Halloran 2008). However, the term *multi-modal* has typically been employed to mean both. We will follow this convention in relation to both the adjective *multimodal* and the noun *multimodality*. Further, we use the term *semiotic mode* to refer to meaning-making systems (image, language), while the term *semiotic resource* is used to refer to linguistic devices and visual techniques. Thus, multimodality can be defined as 'the combination of different semiotic modes – for example, language and music – in a communicative artefact or event' (van Leeuwen 2005: 281).

Our multimodal approach distinguishes us from other researchers who only include language in the analysis of news discourse. But a multimodal perspective is clearly useful when considering today's news:

> By now, newspaper discourse cannot be viewed and studied exclusively or mostly as a monolithic verbal text; on the contrary, it is the multi-faceted polyhedron whereby

> image, image-caption, headline, column, lay-out, and positioning in the (web-)page simultaneously contribute to the meaning-making process of the piece in a compositional way. Thus, the 'news piece' has turned into a 'news package' that calls for a holistic interpretation in order to be fully grasped. (Facchinetti 2012: 183)

We are also interested in how such multimodal discourse is actually put to **use** and how it contributes to the construction of news. Hence, when we use the noun *discourse* and its derived adverb *discursively* we refer to semiotic resources in use – for instance, the use of specific linguistic or visual devices (see Chapters 4 and 5). In sum, our definition of discourse borrows from Halliday (1985) who states that text 'may be either spoken or written, or indeed any other medium of expression that we like to think of' (Halliday 1985: 10), and Halliday and Hasan (1976), who define text as 'a unit of language in use' (Halliday and Hasan 1976: 1).

Corpus linguistics

Corpus linguistics (CL) is an empirical approach to the analysis of linguistic data that makes use of computer technologies to analyse computerized collections of text (corpora), which are often carefully designed and of considerable size. A corpus linguistic investigation usually focuses on language use and typicality (repeated patterns), and may combine quantitative with qualitative analysis. In addition to developing a set of new techniques for the analysis of language, corpus linguistics has also developed new theoretical positions and concepts. It thus combines a methodological innovation with a particular approach to language (Lee 2007: 87). Introductions to corpus linguistics abound and include Hunston (2002), Baker (2006), McEnery et al (2006), and McEnery and Hardie (2012). In sum, researchers taking a corpus linguistic approach analyse an electronic dataset (corpus) with the help of computer software and using specific techniques, concepts and tools developed in corpus linguistics. We will introduce the main corpus linguistic techniques we use in this book in section 4.2.1.

4. Corpus-assisted multimodal discourse analysis (CAMDA)

4.1 A new topology for situating research

While the primary goal of this book is to introduce readers to DNVA, another goal is to promote research that brings together multimodality, discourse analysis and corpus linguistics – a combination of approaches that we have termed 'corpus-assisted multimodal discourse analysis (CAMDA)' (Bednarek and Caple 2014: 151).

The field of research that examines multimodality is vast (O'Halloran and Smith 2011), as are the approaches to multimodal discourse analysis. In a general sense, multimodal discourse analysis attempts to provide an 'integral and coherent picture of multimodal communication and all its resources, and all of the ways in which these are integrated' (van Leeuwen 2015: 108). The strand of multimodal discourse analysis that we are most aligned with is that of social semiotics (e.g. van Leeuwen 2005, Kress and van Leeuwen 2001, 2006), although we do not apply its metafunctional approach here (but see Caple 2013a).[6] In a more specific sense, multimodal analysis can be combined with particular approaches to the analysis of discourse, such as Critical Discourse Analysis (e.g. Machin and Mayr 2012, Machin 2013, Djonov and Zhao 2014). Other notable work that combines multimodality with discourse analysis includes contributions to Chouliaraki (2012), which examine the multimodality of new media discourse, including convergence journalism and social networking sites.

Discourse analysis and corpus linguistics have also developed a fruitful relationship over the last 25 years (Baker and McEnery 2015: 6-8). This includes corpus linguistic

research on discourse phenomena or discourse types as well as studies that combine in-depth discourse analysis with corpus linguistic techniques.[7] It includes both studies that are critical of analysed texts (combining CL and Critical Discourse Analysis, e.g. Mautner 2000, Baker et al 2008) and those that are not (e.g. Corpus Assisted Discourse Studies, see Partington et al 2013). However, only a few studies bring multimodality into the mix (e.g. Adolphs and Carter 2013, Bednarek 2015).

As yet, studies that combine all three – multimodality, discourse analysis and corpus linguistics – are rare. This is not surprising because such a combination of approaches is a highly complex undertaking. As will become clear, corpus-assisted multimodal discourse analysis involves a series of challenges that need to be negotiated before the analysis can proceed. News discourse, especially that which is rendered in the digital media of tablets and smart phones, is packaged in a complex verbal-visual display of images, graphics, typography, words, and navigational elements that guide the reader both within and away from the story page (e.g. through hyperlinks). Such multimodal richness leads to questions regarding what actually constitutes a multimodal analysis, and what should be the point of departure for the analysis. If readers (and researchers) engage with both the verbal and visual elements of a news story together, should the analyst treat the unit of analysis as a verbal-visual complex from the outset? Or is it possible for the analyst to separate out each semiotic mode (e.g. language, image) from its co-text and analyse each in isolation? How can corpus linguistics, which focuses on patterns **across** texts, be combined with multimodal discourse analysis, which focuses on patterns and relations between semiotic modes, often **within** texts? These are important methodological questions and need to be addressed in relation to both the context of analysis and the research paradigm being deployed.[8]

We see the value in a range of approaches to corpus-assisted multimodal analysis, depending on the type of research question the analyst poses and the type of data being examined. We have developed a topology (Figure 1.2) which maps the choices for both semiotic mode (horizontal axis) and unit of analysis (vertical axis). This topology shows four 'zones of analysis' where choices are made regarding the focus of analysis at any particular stage in the research process, allowing researchers to situate their research project in the most appropriate zone at each stage. Such an approach is useful whether the analysis is multimodal or not, corpus-assisted or not.

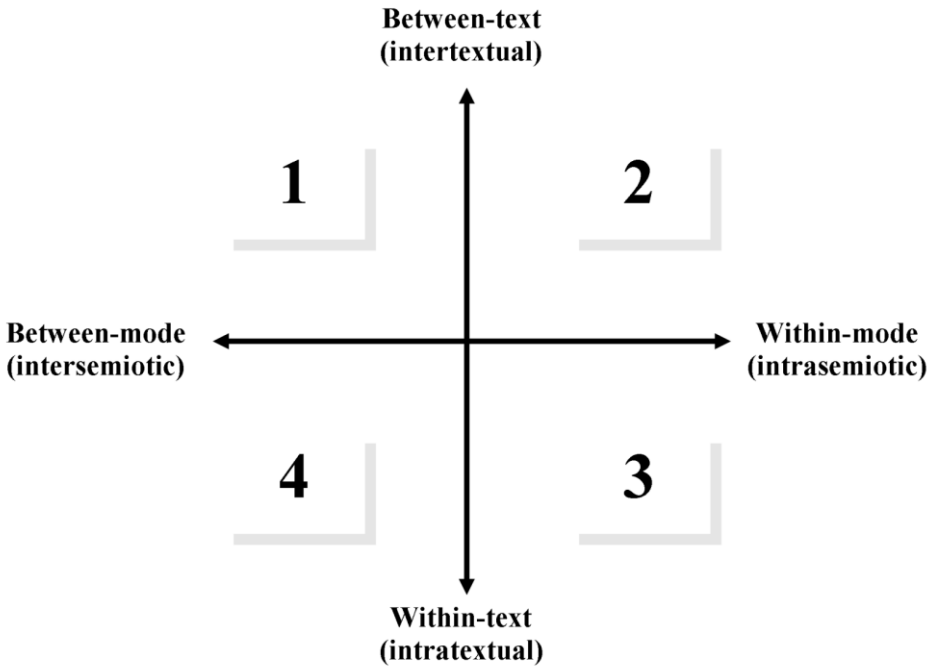


Figure 1.2 Zones of analysis in CAMDA

In relation to news values analysis, a researcher might ask, for example, how are news values discursively constructed in press photographs? Here the analyst is interested in understanding how a particular semiotic mode (image) construes news values. Such mono-modal analysis would be located in the right-hand side of the topology in Figure 1.2 (i.e. staying within-mode), and could examine the construction of news values in a photograph used within one text (and be situated in zone 3) or could examine the construction of news values in photographs used across a range of texts (and be situated in zone 2). One could then repeat this study with a different semiotic mode such as language and compare the results, bringing in a multimodal component through comparison of verbal and visual texts.

Researchers interested in how different semiotic modes **combine** to make meaning would locate their analyses in the left-hand side of the topology in Figure 1.2 (between-mode/inter-semiotic). In relation to news values analysis, the research question could be: how is newsworthiness constructed through the combination of semiotic modes? Such analyses could examine the contributions of both verbal and visual resources to the meaning of a single text (zone 4), or across a number of texts (zone 1).

Another way of viewing this topology is to consider the bottom half of the topology (zones 3 and 4) as concerning itself with logogenesis (Halliday and Matthiessen 1999: 17-18), the unfolding of meaning in text over time. Such analysis of logogenesis could either stay within-mode (e.g. looking at patterns of meaning as they unfold across a verbal text) or examining relations between-modes, e.g. how language and image co-contribute to the meaning of a particular text. Here issues such as discourse semantics or cohesion might be the focus of attention.

In contrast, the top half of the topology in Figure 1.2 (zones 1 and 2) is more interested in looking at patterns across a number of texts, where generalisations may be made about a particular language variety, looking for example at headline writing styles (within-mode, i.e. zone 2), or looking at how headlines and lead images interact with each other on digital news story pages (between-mode, i.e. zone 1).

Analyses located in different zones can also be combined: For example, one might analyse the unfolding of meaning (logogenesis) **across** a number of texts in order to make generalisations about the structure of a particular genre. This would combine zones 2 and 3 (if the analysis stays focused on one mode) or zones 1 and 4 (if the analysis considers more than one mode). As a summary, Figure 1.3 repeats the topology with example analyses.

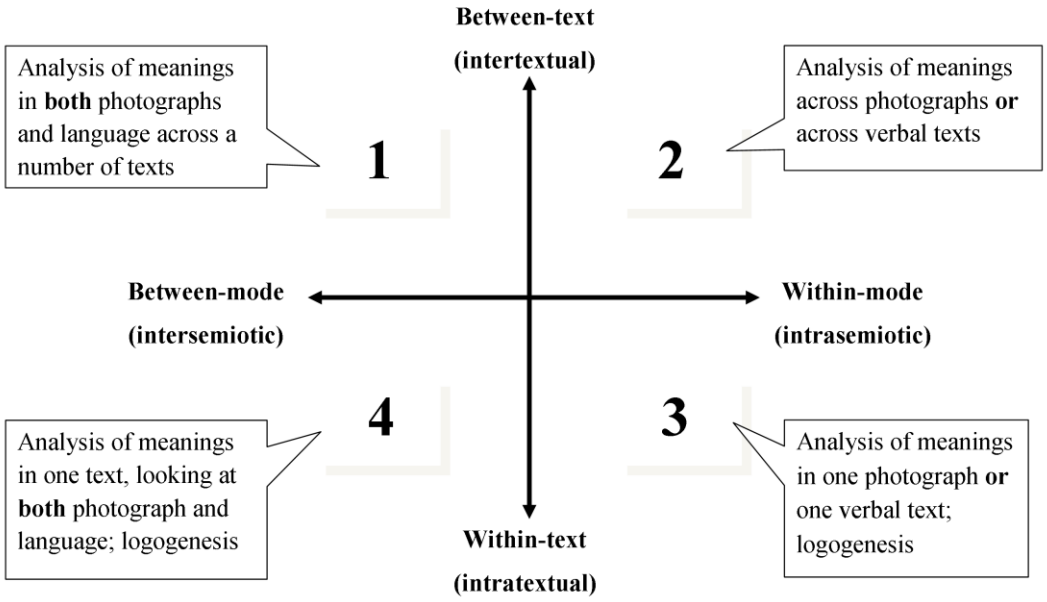


Figure 1.3 Zones of analysis with examples

In our previous studies on news values we have not yet used this topology to situate our research, but our data have ranged from one online news story (Bednarek and Caple 2012b) to analysis of a 9.65 million word corpus (Potts et al 2015). Some analyses focussed on images only (e.g. Caple 2013a), some only on language (e.g. Bednarek 2016a), and some combined analysis of both semiotic modes (e.g. Bednarek and Caple 2012a, b).

In this book, our empirical analyses are both within-mode and between-mode, and focus on between-text analysis: Chapter 6 presents a corpus-linguistic analysis of news about cyclists/cycling (zone 2, language); Chapter 7 analyses images disseminated by news organisations via social media (zone 2, image). Chapter 8 analyses language and photographs in a corpus of news stories shared via Facebook, first analysing each semiotic mode separately (zone 2) before bringing them together (zone 1). Since we do not focus much on the development of meaning **within** texts or logogenesis, we could call this type of analysis 'intertextual' CAMDA. We do not want to prescribe this as the only way of undertaking CAMDA, but rather encourage researchers to come up with different ways of doing so. In particular, we see the need to develop achievable and feasible approaches to the combination of between-text (intertextual) and within-text (intratextual) analysis, while also bringing together analysis of different semiotic modes. One of the outcomes of this book, we hope, is that other researchers will come up with creative ideas for such a combination of approaches.

4.2 Concepts, techniques and tools

In this section we introduce the key concepts, techniques, and tools that we apply in this book, starting with corpus linguistic analysis before moving on to visual analysis, and concluding with a brief mention of the tools (technologies) used in both.

4.2.1 Concepts and techniques for corpus linguistic analysis

A key component of CAMDA is corpus linguistic analysis (see section 3 above). In prior research on news values, corpus techniques such as lemma/word/n-gram frequency, key words/parts-of-speech/semantic tags, and collocation have been used in different ways (Bednarek and Caple 2012b, 2014, Potts et al 2015, Bednarek in press). Rather than repeating here what we say about these techniques there, we point interested researchers to these publications for further detail. In this section we briefly introduce the main corpus techniques we use in this book, without discussing debates around them (see e.g. McEnery and Hardie 2012, Hunston 2013).

Frequency, keywords and range

Most corpus linguistic software programmes, such as Wordsmith (Scott 2015), permit automatic frequency analysis, producing a list of items in a corpus together with the frequency with which each item occurs (frequency lists). One can distinguish between the frequency of *types* (different word forms) and *tokens* (all instances). For example, a corpus with 300,000 tokens may contain only 14,000 types, since many tokens will be repeated. Items in a frequency list can be lemmas (WALK), word forms (*walk, walks, walked, walking*) or longer structures (*I walked*). These longer structures are often called *n*-grams, referring to recurring combinations of *n*-words, for example bigrams (two words, e.g. *of the, you know*) or trigrams (three words, e.g. *at the end, you know that*). In any frequency list, grammatical words tend to be the most frequent and therefore fill the top of the list. It is possible to exclude such words by using what is called a *stop list* – a list of words that are ignored when compiling the frequency list. The stop list that we use in this book is a default English list

with 174 entries.[9] Frequency lists can be visualised in the form of *word clouds* where a larger size of a word represents a higher frequency (Figure 1.4).

Figure 1.4 Example of a partial word cloud (from Chapter 6)

Further, some corpus software allows users to sort items in a frequency list according to their distribution within or across files, which is also referred to as their *dispersion* (e.g. Gries 2008) or *range* (e.g. Nation and Waring 1997). In this book we use the term *range* to refer to the distribution of instances across individual corpus files, identifying in how many corpus texts an item occurs. This is important because some items with a relatively high frequency may only occur in a few texts in a corpus. Analysis of range – sometimes called consistency analysis – is useful for identifying the core features of a language variety (Bednarek 2012) and for analysing similarity more generally (Taylor 2013).

Frequencies can also be compared across two corpora, for instance through automatic keywords analysis. Here, the software compares the frequencies of items in one corpus (the node, target or study corpus) with their frequencies in a second corpus which provides a baseline (the reference corpus). The calculation takes into account the different sizes of the corpora and applies statistical tests – most often log likelihood (LL; G2). This test tells us if the difference between two corpora is statistically significant by providing a log likelihood value which corresponds to a particular p-value. A p-value of 0.05 (G2 = 3.84) means that we can be 95% confident that the results are not due to chance.[10] A keywords list then, is a list of items that are, statistically speaking, unusually frequent or unusually infrequent in the target corpus when compared to the reference corpus.

We also use a new software tool called ProtAnt (Anthony and Baker 2015a). This tool uses keywords to calculate which texts in a corpus are most and least prototypical of the corpus as a whole, when compared to a reference corpus.[11] To do so, ProtAnt first compiles a list of keywords for a corpus and then calculates how many of these keywords occur in each corpus file, ranking the files by the number of keywords they contain (Anthony and Baker 2015b: 278). Thus, the top ranked corpus texts will contain the most keywords (prototypical), while the lowest ranked corpus texts will contain the least keywords (atypical). The assumption behind this technique is that 'a text which contains a greater number of keywords from the corpus as a whole is also likely to be a more central or typical text in that corpus' (Anthony and Baker 2015b: 277). The primary motivation for this tool is to allow researchers to systematically identify texts for qualitative analysis – that is, as a down-sampling technique. It can also be used to identify what are the most 'typical' news values that are constructed in a corpus, which is the way we use it in Chapter 6.

Collocation and collocational networks

Another important corpus linguistic concept is that of collocation, which refers to the non-random association of words. It has been observed that some words 'go together', as it were – that is, they frequently occur in the vicinity of each other. Collocation analysis usually proceeds by taking a word (the *node*) and identifying which other words typically co-occur in a given co-textual span. These co-occurring words are called *collocates*. For example, *oh, sake, knows, thank, my* and *bless* are all collocates of *god* in British English. Typically, researchers examine a span of four or five words to the left and to the right of the node. Collocates can be grouped according to their meaning. Thus, some word forms co-occur with attitudinally negative collocates, and are said to have a *negative semantic prosody* (Louw 1993). In addition, one can identify collocational networks, i.e. networks of collocates. For instance, *spend* is a collocate of the node *time* and itself collocates with *money,* which in turn collocates with *pay* (Brezina et al 2015: 152-153). Such networks can be visualised using GraphColl (Baker and McEnery 2015; Brezina et al 2015), as seen in Figure 1.5. Each circle represents a word and the length of lines between words represents collocational strength (the shorter the stronger). Thus, we can see that *more* is a collocate of the node *cyclists* and itself collocates strongly with *than* and *people* (in the corpus described in Chapter 6).

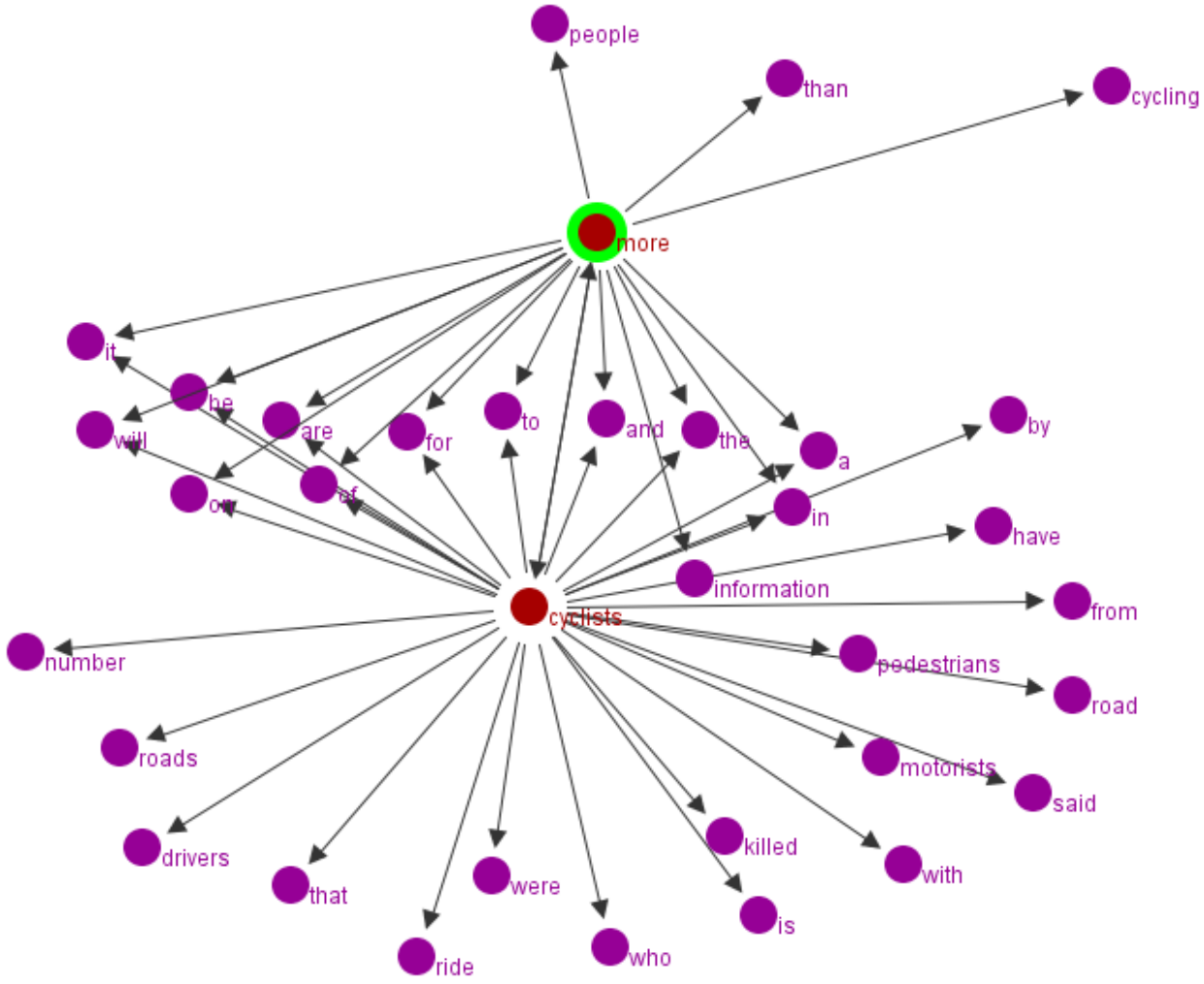


Figure 1.5 Example of a GraphColl network (from Chapter 6)

Collocates are automatically identified by most software tools using an in-built statistical collocation measure, with different statistics producing different results.[12] Most association measures identify collocates by comparing how often they are **expected** to co-occur with the node with how often they **actually** occur (Brezina et al 2015: 144). Unless otherwise stated, we generally use the MI3 statistic, a span of 5 words on each side of the node (5L:5R), with a minimum frequency threshold of 2, and do not calculate collocations across sentence breaks (when using Wordsmith). MI3 (Daille 1995) is the cubed variant of the mutual information statistic, which reduces its low frequency bias – it gives more weight to observed frequencies and ranks frequently occurring (typical) collocations much higher than those that are uncommon (Brezina et al 2015: 159-160). Other collocation measures that we will refer to are log likelihood and t-score.[13]

Semantic tags and word sketches

In addition to identifying word frequencies and word associations, corpus linguistic programs (taggers or parsers) can categorise words according to their likely meaning or grammatical function. For example, the UCREL Semantic Analysis System (USAS) tags words as belonging to particular semantic fields (Archer, Wilson and Rayson 2002). Each semantic tag stands for a semantic field such as 'Emotional Actions, States & Processes' or 'Time', with further subdivisions. For example, the items *recent, latest, new* might be tagged as belonging to the semantic field 'Time: Old, new and young; age'. With a tagged corpus, it becomes possible to create frequency lists of tags or word-tag combinations, for instance focusing on analysis of the most frequent semantic tags in a corpus.

Sketch Engine's (Kilgarriff et al 2014) word sketches bring together collocation analysis with grammatical analysis, by producing collocates for a node and grouping these collocates according to their grammatical relations (e.g. object of, subject of, modifier). In other words, Sketch Engine automatically identifies collocates as well as their likely grammatical relationship with the node (https://www.sketchengine.co.uk/word-sketch/).[14] In addition to simple word sketches for one lemma, Sketch Engine provides a functionality called word sketch differences, which allows the comparison of collocates for different lemmas or word forms by showing their shared and unshared collocates. In Chapter 6 we use this functionality for identifying common collocates of the singular (*cyclist*) and plural forms (*cyclists*) of the same lemma, focusing on similarity (Taylor 2013) rather than difference.

Concordances and search terms

The final technique to introduce here is concordancing – producing all occurrences for a particular search term (the node), together with its surrounding text (co-text). Concordancing is particularly useful for qualitative analysis, as the co-text can be expanded, and because concordances can automatically be sorted in different ways. For instance, Figure 1.6 shows 35 sample concordances of the word *Memphis* sorted alphabetically according to the right (again using the corpus described in Chapter 6).

Figure 1.6 Sorted concordances

Sorted concordances are particularly helpful for the identification of patterns, or recurring linguistic practices. Concordances can be produced for single word forms (e.g. *cyclist*) or combinations of word forms (e.g. *bike rider*; *cyclist death*) and * can be used as a wildcard to stand for one or more characters (e.g. a search for *cyclist** retrieves concordances for *cyclist, cyclists, cyclist's, cyclista, cyclistist*). A tool like Wordsmith also provides advanced search options such as the introduction of 'context words'. Using this function we can produce concordances for *cyclist* occurring in the co-text of *old* within five words to the left or right. Wordsmith can further calculate recurring 'clusters' for a given search term. These clusters are based on the concordance lines and are patterns of repeated phraseology within five words. Clusters can consist of two or more words (e.g. *cyclist deaths; cyclist was killed*; *death of a cyclist*).

In addition, some corpus tools offer information on the position of a search term in text files, showing if it occurs at the beginning, middle, end, or throughout the file. For instance, Wordsmith provides users with a dispersion *plot* (a visualization) and a dispersion value, which indicates the extent of uniformity of a search term's distribution. Generally, the dispersion value lies between 0 and 1, and the closer the value is to 1, the more uniform the dispersion.[15] Further, Sketch Engine allows users to view the distributional graph of

concordances, which shows the distribution of the search word across parts (slices) of the corpus.[16]

4.2.2 Concepts and techniques for visual analysis

Some of the terms that we draw on in this book for the analysis of images are borrowed from Kress and van Leeuwen (2006), although we use them somewhat differently and always with a focus on news values. Other concepts come from the work of Caple (2013a), especially regarding the relationship between compositional balance and aesthetic appeal. Additional terms are taken from technical handbooks on the workings of camera equipment. As with our previous research on the construction of news values in images (Caple and Bednarek 2015), we continue here to examine images in terms of their **content** (what is depicted in the image) and in terms of their **capture** (also glossed as 'camera technique'). The latter involves two strands of analysis: that of the composition of the image (how the information is arranged in the image frame) alongside analysis of technical affordances (e.g. shutter speed, aperture).

Content: Represented participants, attributes, activity sequence, setting

In examining image content we look primarily for who or what is represented: the *represented participant*. In Kress and van Leeuwen's (2006: 48) terms, represented participants are:

> the participants who constitute the subject matter of the communication; that is, the people, places and things (including abstract 'things') represented in and by the [...] image, the participants about whom or which we are [...] producing images.

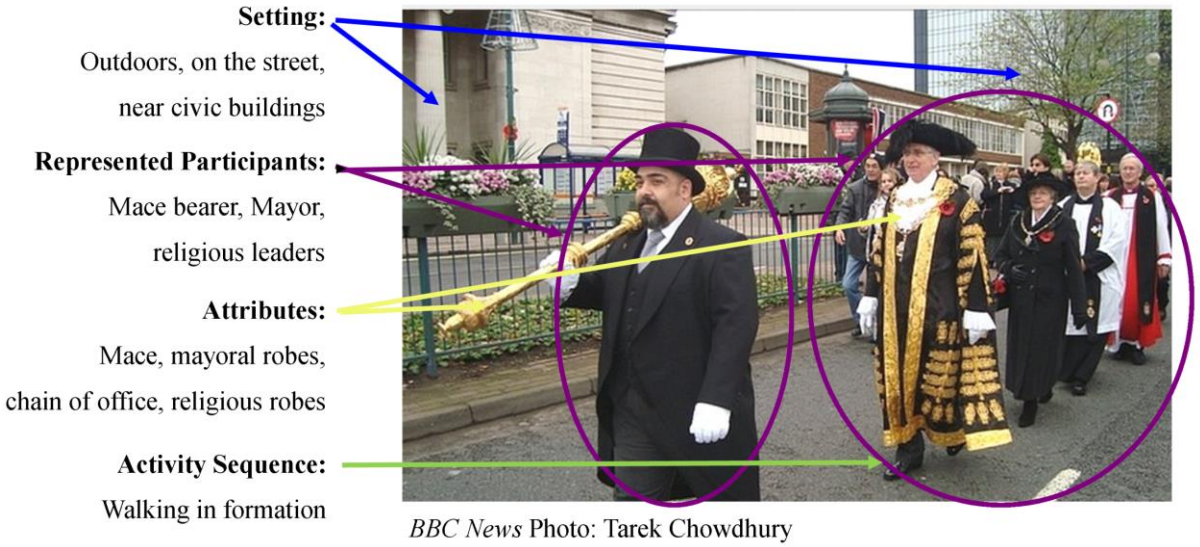
This allows us to identify who, where or what is the subject matter of the image, be it a widely-known famous politician, sports person, landmark or landscape, or an ordinary member of the public, or a victim of a negative happening.

We examine the different parts that constitute the represented participant, which in the case of people includes clothing or uniform, jewellery, medals, badges, equipment, and other regalia that they may be wearing, holding or using. We label these *attributes* ('Possessive Attributes' in Kress and van Leeuwen's 2006: 50 terms). An examination of attributes can help us to further distinguish what kind of person is being represented in an image, e.g. whether it is a regular police officer or a police commissioner.

We also take into consideration the activities the represented participants are engaged in. A person, for example, may be photographed being, thinking or feeling (e.g. posing for the camera as in a portrait shot with neutral, positive or negative facial expression and direct or indirect eye contact). People may also be photographed doing something, depicted as '"agents", the doers of that action' (van Leeuwen 2008: 142), e.g. firing a gun. They could be depicted as '"patients", the people to whom the action is done' (van Leeuwen 2008: 142), e.g. being fired at; and they may be photographed saying something or listening to something/someone, where eye-contact, gesture and body language can help to decode whether they are speaking or listening. We gloss analysis of such activities and the roles that represented participants play in them as *activity sequence*. Analysis of activity sequences can tell us more about what kinds of represented participants images depict and what they are doing (e.g. police arresting suspects).

We also examine the context or environment, glossed as *setting*, in which the represented participants are depicted (e.g. a court room, a government building, a lab, a person's living room). This may be non-existent, e.g. in an extreme close-up shot of a person's face, or maximally identifiable, e.g. in a very wide angle shot, or somewhere in between. The setting tells us where a news event takes place and may further help us to identify the kinds of people and activity sequences they are engaged in – for instance, a

person who is represented in a laboratory as filling a test tube with a syringe is most likely interpreted as a scientist engaged in some experiment. Figure 1.7 illustrates the visual resources we examine in relation to image content.



BBC News Photo: Tarek Chowdhury

Figure 1.7 Visual resources comprising image content

Capture: Composition and technical affordances

One strand of analysis in relation to image capture involves analysis of how the content (discussed above) is arranged in the image frame, which we gloss as *composition*.

Composition: Salience, shot length, cropping, angle, dynamic asymmetry, interrupted symmetry

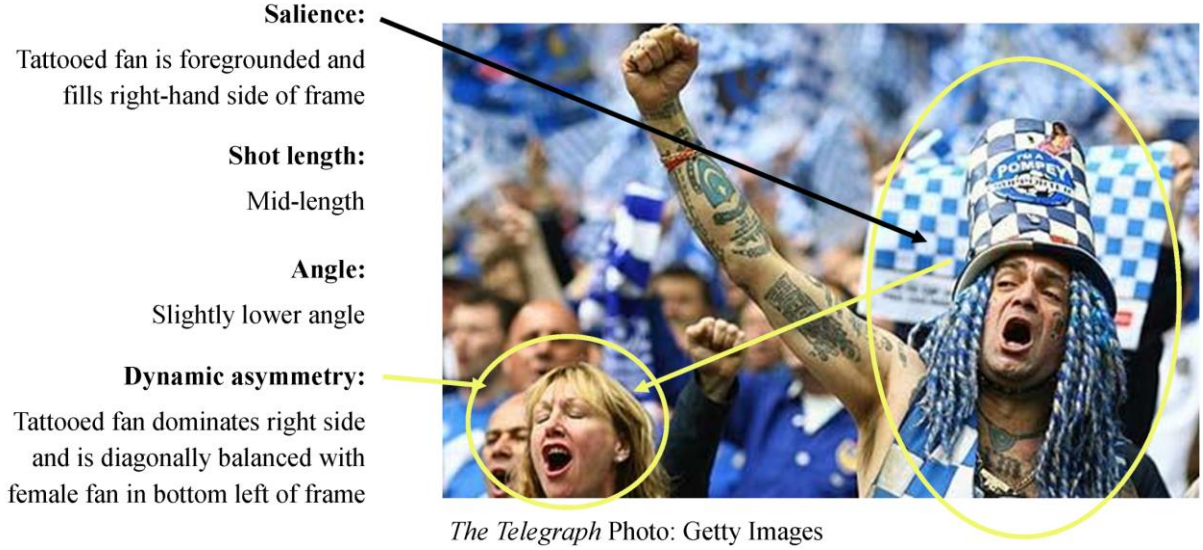
Represented participants may be placed towards the front of the image frame or moved backwards in the image frame. These are aspects of *salience* that can impact both on how the depicted elements relate to each other and on how closely or distantly viewers of the image relate to the image content. *Shot length* works in a similar way to salience in impacting on how closely or distantly viewers relate to image content. A long shot creates maximal disconnection between audience and represented participants ('*distanciation*' in van Leeuwen's 2008: 141 terms), but at the same time includes the setting in the image, thus informing audiences of where the depicted activity sequence is taking place. At the other end of the scale, an extreme close up eliminates the setting completely, but demands maximal engagement between audience and image content.

The concept of *cropping* an image relates closely to salience and shot length, as it also tells us something about what has been included or excluded from an image.[17] A photograph may show us, for example, a politician speaking from a lectern. However, the audience that her gaze and gestures are directed towards may be excluded from the image. By cropping out the audience, the image is focusing our attention on the politician and possibly aspects of her facial expression, gestures and body language. Cropping is also used in the image in example (3), given at the beginning of this chapter. We know that the heads at the front of the image are attached to torsos and legs and we are able to fill in this missing information. Cropping of this sort tells us that the size or scale of the event taking place in the image extends beyond what the image has captured.

Another concept that tells us about how information is arranged within the image frame is camera *angle* (horizontal and vertical). Represented participants may be photographed from eye-level, from a low angle (looking up towards the represented

participants), from a high angle (looking down on the represented participants), or from a frontal (face on) or oblique (from the side) angle.

Two further concepts concerning composition are *dynamic asymmetry* and *interrupted symmetry* (Caple 2013a). Dynamic asymmetry involves the use of the diagonal axis in composing an image and establishes unequal relations between represented participants. When the main represented participants are placed in the bottom left of the image frame, these may be counterbalanced by other participants (usually less salient) placed in the top right corner of the frame and vice versa. Equally, the remainder of the image frame may be left empty. In a symmetrically balanced image, all represented participants are shown in equal relation to each other (e.g. a line of soldiers on parade). Interrupted symmetry entails a slight 'defect' or flaw in the symmetry, e.g. if one of the soldiers was looking the wrong way, which interrupts rather than completely destroys the symmetrical balance of the composition. Figure 1.8 illustrates aspects of composition.



The Telegraph Photo: Getty Images

Figure 1.8 Visual resources comprising composition

Technical affordances: Movement, focus, noise

The second strand of analysis in relation to image **capture** involves analysis of the effects of camera settings on image content, e.g. whether all elements in the frame are in focus or not, whether all elements in the frame are well lit or not, whether elements are blurred or show movement, or whether they are static, frozen in time and space. We gloss this aspect of analysis as *technical affordances*.

While the researcher is not expected to know what shutter speed, aperture, ISO was selected in image capture, she can familiarise herself with the effects that such camera settings have on image capture.[18] A slow shutter speed, for example, can result in a blurring effect around moving objects. Thus, water can be made to look silky or smooth through the use of a slow shutter speed. The sense of movement in an athlete running or jumping can be enhanced by using a slow shutter speed combined with a panning action (moving the camera in sync with the movement of the subject) and a flash. A high shutter speed freezes action. It allows the viewer to see in great detail aspects of movement that she would not ordinarily be able to see with the naked eye, e.g. the shape of a water droplet or the contortions of the musculature of a diver performing an acrobatic dive. We gloss this as *movement*.

Depth of field, or how much of the image content is in focus, is an aspect of image capture that can be manipulated through changing the aperture in the camera settings. A drastically reduced depth of field will result in only a very narrow area of an image being in

focus or sharp, and will blur the rest of the image. Maximising the depth of field ensures that all elements in the image frame are in focus. We gloss this as *focus*.

Finally, a very high ISO (which is useful in very low lighting conditions and when a fast shutter speed is needed) will result in a very grainy effect in an image. A very low ISO will produce a clean, sharp, high quality image. We gloss this effect as *noise*. Aspects of technical affordances are illustrated in Figure 1.9.

The Guardian Photo: David Clapp/Getty Images

Figure 1.9 Visual resources comprising technical affordances

4.2.3 Tools/technologies

As has become apparent in Section 4.2.1, we use both classic and new corpus linguistic tools in our analysis, namely Wordsmith (Scott 2015), Sketch Engine (Kilgarriff et al 2014), ProtAnt (Anthony and Baker 2015a), and GraphColl (Brezina et al 2015). These tools allow us to undertake analysis of word/n-gram frequency, keywords and prototypicality, range, collocation and collocational networks, word sketches, and concordancing. In addition, we use UAM Corpus Tool (O'Donnell 2015), a software program that can be used for computer-assisted manual annotation, 'where a human annotates the text in terms of patterns that generally computers cannot recognize.' (O'Donnell 2007). This tool allows the researcher to upload texts, to create annotation schemes (e.g. valence: negative, positive or neutral/ambiguous), and annotate either the whole text or segments of the text accordingly, coding each sentence in turn if desired. It also allows complex queries and automatic processing of the annotated text data, for instance producing all text segments that were annotated in a particular way or providing comparative numbers and statistics (e.g. 55 of 99 texts are coded as negative, 17 as positive and 27 as neutral/ambiguous).

Further, we make use of a relational database (Microsoft Office Access) following an approach first applied in Caple (2009). We use this for the analysis of images and to bring the analysis of language and photographs together. While the initial design, construction and manual population of database fields is time consuming, it is a very efficient way of collating the analysis of a large data set (e.g. 1100 images in the case study in Chapter 7). The subsequent ability to query the inputted data is where the benefits of a database become clear. The query function works with the data that has been entered to not only show raw totals (e.g. how many instances of Negativity are present in the data?), but also to show any number of combinations of results. Complex questions can very easily be posed, such as: which news values are likely to be combined with Negativity and across how many items? Where our analysis spans both words and images (in Chapter 8) we can ask questions like: do items that

construct Negativity in the verbal text also construct Negativity in the visual text? Chapters 7 and 8 provide more details on the databases created for the case studies.

5. Summary and overview of chapters

In sum, this book brings together DNVA and CAMDA to explore the ways in which news values are constructed through the words and images that constitute journalistic discourse. The key features of this book are that it:

- is inter-disciplinary and multi-methodological, bringing together corpus linguistics and multimodal discourse analysis in analysing the news media;
- explores how the news is 'sold' (made newsworthy) to audiences through the semiotic modes of language and image;
- provides comprehensive analytical frameworks for the systematic analysis of multimodal news discourse, which can be used by other researchers in their own subsequent explorations;
- combines in-depth theoretical discussion with manifold analyses of authentic news discourse (language and images) from around the English-speaking world, including three chapters dedicated to new empirical case studies.

In this chapter we have provided a first introduction to DNVA and outlined the key terms, concepts and techniques that we use in this book. Readers unfamiliar with these approaches are advised to refer back to these sections when engaging with later chapters. Chapters 2 and 3 now provide an overview of the ways in which news values have been conceptualised in linguistics and other disciplines and outline our own approach (DNVA). Chapters 4 and 5 then describe the linguistic and visual resources that can be used to establish news values, providing comprehensive analytical frameworks. Chapters 6 to 8 present three new empirical studies.

While Chapter 6 illustrates how DNVA can be applied in the analysis of one particular topic (cyclists/cycling), both Chapters 7 and 8 focus on a range of different topics but have in common that they explore news in the context of social media. In an era where the new media have created a period of instability and digital disruption for journalism and its business models, where clickability, likeability and shareability have become crucial considerations in a new 'attention economy' (boyd 2012), and where technological advances have caused 'dramatic change in news making and news writing' (Facchinetti 2012: 152), it is crucial to analyse news reporting in this context. Chapter 7 examines the construction of news values in stories that news organisations post to their own Facebook feeds. Chapter 8 focuses on news values in news items that are widely shared by Facebook users. Chapter 9 then describes and illustrates two key areas for future research: diachronic and cross-cultural studies. Finally, Chapter 10 re-visits and reflects on each of the previous chapters.

---

[1] http://www.theguardian.com/australia-news/live/2015/apr/22/nsw-storms-wild-winds-and-flooding-as-sydney-weather-brings-chaos-rolling-report#block-5536dcd3e4b0e90b560ce55b, accessed 22 April 2015.

[2] These sections are available at http://www.news.com.au/more-information; http://www.bbc.co.uk/aboutthebbc/insidethebbc/whoweare/mission_and_values; http://www.ap.org/company/news-values; http://www.nytco.com/who-we-are/culture/standards-and-ethics/; http://www.aljazeera.com/aboutus/2006/11/2008525185733692771.html (all accessed 30 September 2015).

[3] Self-references can also be used to create value for a news organisation, for example crafting an image of itself as an investigative newspaper or watchdog (Marchi 2013).

[4] See Conboy (2006: 15-16) on tabloid news values, and Bednarek (2006a) on evaluation in UK tabloids and broadsheets.

[5] Hard news has been classified as time-bound (Bell 1991), as negative (Ungerer 1997) or destabilising (Feez et al 2008) in contrast to timeless (Bell 1991), positive (Ungerer 1997) or stabilising (Feez et al 2008) soft news. Soft news has also been connected to human-interest or entertainment (Bender et al 2009: 134), or hard news is contrasted with human interest rather than soft news (Piazza and Haarman 2011). In our experience, the suggested criteria cannot always be consistently applied. However, this book does not focus on news genre analysis or the linguistic differences between different kinds of news reporting.

[6] Such a metafunctional approach assumes that semiotic modes fulfil three major functions: an 'ideational' function, wherein they represent the world around and inside us; an 'interpersonal' function, wherein they enact social relations; and a 'textual' function, wherein message entities or 'texts' attempt to present a coherent whole (Kress and van Leeuwen 2006: 15).
[7] A bibliography is maintained by Costas Gabrielatos at http://www.gabrielatos.com/CLDA-Biblio.htm, accessed 30 September 2015.
[8] Another challenge concerns the management of researcher subjectivity or bias, since corpus and discourse analyses are subject to variation depending on a range of factors (Marchi and Taylor 2009, Baker and McEnery 2015). To manage subjectivity, researchers can adopt the principles of transparency and consistency (Baker 2009: 83), as we have aimed to do in this book.
[9] 'Default English stopwords list' downloaded from http://www.ranks.nl/stopwords/, accessed 12 November 2015.
[10] See further http://ucrel.lancs.ac.uk/llwizard.html, accessed 13 November 2015
[11] Since ProtAnt works with UTF-8 encoded text files, all files were first converted into UTF-8 using EncodeAnt (Anthony 2014).
[12] Debates about collocation revolve around criteria for identifying collocates, such as distance, frequency, exclusivity, directionality, dispersion and type-token distribution (Brezina et al 2015).
[13] While the MI3 statistic measures collocation strength, these two scores measure the confidence with which one can claim that there is a non-random association between two word forms (McEnery et al 2006: 57; Brezina et al 2015: 161). The t-score subtracts expected occurrences from observed occurrences and divides the results by the standard deviation (Hunston 2002: 70). The log likelihood formula compares the number of times two words occur together with how often they occur without each other, also taking into account when neither is the case (Dunning 2008). The top collocations produced with the t-score tend to consist of high-frequency pairs (Hunston 2002: 74; McEnery et al 2006: 57), often function words, while both MI3 and log likelihood provide a mixture of high-frequency function and lower-frequency content words (Baker 2006: 102).
[14] The statistic used by Sketch Engine to produce word sketches is logDice (Rychlý 2008). This measure is based on the Dice coefficient, and takes into account the frequencies of the node, of the collocate, and of the collocation within a particular grammatical pattern (Baker et al 2013a: 37). See further http://trac.sketchengine.co.uk/raw-attachment/wiki/SkE/DocsIndex/ske-stat.pdf, accessed 9 December 2015.
[15] http://lexically.net/downloads/version7/HTML/index.html?plotdispersionvalue.htm, accessed 8 January 2016.
[16] See https://www.sketchengine.co.uk/concordance/, accessed 9 December 2015.
[17] By *cropping*, we mean the way in which the image content has been framed and thus presented to audiences. A more appropriate term might be *framing*. However, since Kress and van Leeuwen (2006: 177) already use the term *framing* to mean something quite different, we have opted for the term *cropping* to avoid confusion.
[18] **Shutter speed** controls the length of time that light is allowed to make contact with the image sensors in the camera, by opening and closing the curtains (shutter) at the film plane. Lenses on a camera contain a diaphragm through which the light passes. The opening of the diaphragm (the **aperture**) can be made larger or smaller and determines the amount of focused light passing through the lens. **ISO** stands for *International Standards Organisation* and in photography it indicates levels of sensitivity to light. A low ISO (100), for example, means less sensitive to light and would be used under normal or bright lighting conditions to produce very clean/sharp high quality images. High ISO (e.g. 3200) means highly sensitive to light and would be used when lighting conditions are very poor, resulting in very grainy images that may appear less sharp. Fuller explanations of the technical aspects of camera settings can be found at any number of online photography sites, for example: http://www.all-things-photography.com/iso-settings/, accessed 22 March 2016.