MONIKA BEDNAREK & HELEN CAPLE

NEWS

NEWS

NEWS

THE DISCOURSE
OF NEWS VALUES

How News Organizations Create Newsworthiness

The Discourse of News Values

The Discourse of News Values

HOW NEWS ORGANIZATIONS CREATE NEWSWORTHINESS

MONIKA BEDNAREK AND HELEN CAPLE

OXFORD
UNIVERSITY PRESS

Contents

Part II   ANALYTICAL FRAMEWORKS

Part III   EMPIRICAL ANALYSIS

List of tables

List of figures

Acknowledgements

The Discourse of News Values

1

Introduction

1.1 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. (http://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." (http://theguardian.com/au, published and accessed on 22 April 2015)[1]

(3)

News photograph in figure 1.1 on page 2.

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

Figure 1.1 A news photograph of migrants walking through Slovenia (*The Atlantic* photo: Jeff J. Mitchell/Getty Images).

(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 realize 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 skilfully 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 farmland 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 organizations—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 recognized 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 organizations 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 organizations 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 organizations 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—for example, via speech/dressing styles, signature music, or set design (van Leeuwen 1984, 1989, 2006b; Bell and van Leeuwen 1994), but they are not the focus of this book. As mentioned earlier 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*.

1.2 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 conventionalized 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, or 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 (for further discussion of the critical potential of DNVA, see Bednarek and Caple 2014). Again, it is possible to undertake such analysis diachronically and across cultures. The aim here is to see if specific news values are emphasized, 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 2016). 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 joint 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 2016c). At the same time, Caple has been the lead researcher in publications where we focus on visual DNVA (Caple 2013a; Caple and Bednarek 2016). 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.

1.3 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*.

1.3.1 NEWS

In everyday usage, the word *news* is frequently used to refer to new information. We might ask each other if there is any *news* or check our Facebook *news*feed. Here the source of the information (friends, family, or strangers), its domain (public/private), and the type of information (gossip, opinion, announcement, or 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 organizations. 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 organization 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. nominalization, 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.

1.3.2 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 multimodal. Strictly speaking, texts that are 'multimodal' 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 *multimodal* 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).

1.3.3 CORPUS LINGUISTICS

Corpus linguistics 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 data set (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 1.4.2.1.

1.4 Corpus-assisted multimodal discourse analysis

1.4.1 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. Kress and van Leeuwen 2001, 2006; van Leeuwen 2005), 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 corpus linguistics 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 discourse 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). We use the term topology here in analogy to Martin and Matthiessen (1991) to refer to scalar rather than categorical distinctions which are typically represented in taxonomies. That is, these distinctions are best considered as clines, scales, or continuums. 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 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/intersemiotic). 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 generalizations 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 generalizations 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.

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 focused 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 organizations 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
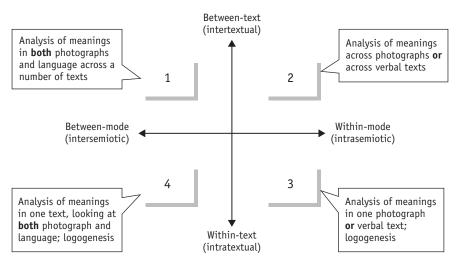

Figure 1.3　Zones of analysis with examples.

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.

1.4.2 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.

1.4.2.1 *Concepts and techniques for corpus linguistic analysis*

A key component of CAMDA is corpus linguistic analysis (see section 1.3). 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 2016c). 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 programs, 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 visualized in the form of *word clouds* where a larger size of a word represents a higher frequency (figure 1.4).

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

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

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.

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 visualized 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).
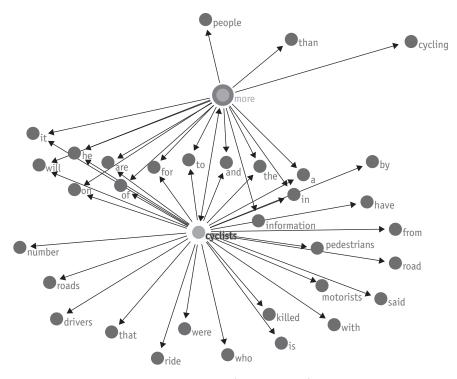

Figure 1.5 Example of a GraphColl network (from chapter 6).