12

Using corpora in discourse analysis

Alan Partington and Anna Marchi

1 Introduction: an outline of corpus-assisted discourse studies (CADS)

In the linguistics literature, "discourse" is often defined in two, not mutually exclusive, ways, namely, structurally, for instance, "language above the sentence or above the clause" (Stubbs 1983: 1) and functionally, for example, "language that is doing some job in some context" (Halliday 1985: 10). We shall privilege the functional viewpoint here, though analyzing the structures of discourses is important for shedding light on the jobs being done. It has to be stressed that discourse is not a special form of language, but a perspective upon it, language described not only as a set of interacting units and systems, but also precisely that implied by Halliday, as an instrument put to work. The work which it does is the attempt by one participant or set of participants to influence the ideas, opinions, and behavior of other participants. Such work can be studied in a single text or in a number of tokens of similar texts to try to infer generalities of behaviors and responses (which may well then in turn serve as background to studying particular language events for particular, special meanings).

Most forms of traditional non-corpus-assisted discourse analysis have practiced the close-reading (that is, "qualitative analysis") of single texts or a small number of texts in the attempt to highlight both textual structures and also how meanings are conveyed. Some types, such as much work in critical discourse analysis (CDA), use few concepts from linguistics proper, tending to rely on the analyst's knowledge and experience (and prejudices) of similar texts, in a manner reminiscent of literary analysis (though with a politically driven purpose). Other traditional discourse analysis is more linguistically grounded. Thompson (1996a: 108–112), for instance, demonstrates the power of functional grammar, in particular transitivity analysis, in displaying how meanings, including what we might call *non-obvious* meanings, are communicated. Analyzing a nursing recruitment advertisement, he demonstrates how the grammar itself in subtle ways positions the reader as a certain type of personality (caring and compassionate) and that, as a nurse, your transitivity role, as actor/agent would be to "help patients recover their [own] normal transitivity roles (their normal function as human beings)" (1996a: 112), but without neglecting that a nurse must also have a sensible role as Mental Senser with regard to the real-world Phenomenon of money: "So what sort of money can you expect as a nurse?" (1996a: 111, 112).

In what follows we will attempt to outline ways in which corpus-assisted discourse studies (CADS) can help build upon traditional qualitative linguistic analysis, what "added value" it can bring. We contend that it can contribute in two ways. First, by combining close reading with statistical "overview" analysis, very generally of a large number of tokens of the discourse type under scrutiny, which can enable the analyst to build up a detailed picture of how work is typically performed in that type of discourse. Second, by integrating into the analysis a number of insights into how discourses function which have developed within the field of corpus linguistics.

The three most commonly employed statistical overview techniques are the following. First, frequency listing of words and clusters (that is, strings of words which "are found repeatedly together in each other's company" (Scott 2004), also known as *n*-grams and lexical bundles), which tells us which items and clusters are common in a particular set of texts of a certain discourse. Second, keyword and key-cluster listing, which tells us which items and clusters are more or less frequent in one set of texts, representative of one discourse type, relative to another set, perhaps of another discourse type or of "general," that is, heterogeneric, English (or any other language). Finally, the concordance which searches through large quantities of texts and can collect together and display recurring patternings of words surrounding the search (or "node") item stipulated by the analyst.

The main linguistic insights arising from or developed using corpus-linguistic research include the lexical grammar notion of co-selection, the psychological but also textual theory of lexical priming, and, finally, evaluative cohesion.

The principle of co-selection or co-occurrence states that a far greater proportion of the language of most discourse types is made up, not of the accretion of individual items chosen from the mental lexicon, but of prefabricated or semi-prefabricated collections of items; "chunks" if we prefer. These include simple collocations (defined as two items which regularly co-occur in texts), such as *roaring fire*, proper names like *the Houses of Parliament*, set phrases like *as a matter of fact*, *by all means*, idioms like *never a dull moment*, semi-idiomatic templates, for instance, LIVE *to a [ripe/grand] old age* (Stubbs 2000: 2–3), templates containing fixed parts but also elements of considerable variability, for example, (Locality A) BE *a* number + time-word + vehicle + (*journey, trip, voyage, flight*, etc.) *from* (Locality B), which can be realized as *a twenty-minute bus ride from, a two-hour train journey from, a five-day bike-hike from*, and so on (Hoey 2005: 16–17), and also other abstract items which have what we might call lexical-grammatical satellites orbiting around them, such as *brook* + negative + modal, which can be realized in a wide variety of ways *will brook no* . . ., *determined not to brook*, and so on (Sinclair 2004a: 36–37). The case study in Section 3 below examines how participants can use such chunking in discourse production to their own ideological advantage.

Lexical priming (Hoey 2005) is a self-reproducing *mental* phenomenon whereby the normal language user learns, by repeated acquaintance with a lexical item and by processes of analogy with other similar items, the typical behavior of that item in interaction. In particular, we learn which other lexical items it co-occurs with regularly (*collocation*), which semantic sets it occurs with (*semantic association*; other authors would favor the term *semantic preference*; see Sinclair 2004a: 32–33, 142), which grammatical categories it co-occurs with or avoids and which grammatical positions it favors or disfavors (*colligation*), which positions in an utterance or sentence or paragraph or entire text it tends to prefer or to avoid occurring in (*textual colligation*), and whether it tends to participate in cohesion or not.

The user then reproduces this behavior in their own linguistic performance. By metaphorical extension, the lexical item itself is said to be primed to behave in these particular ways, and so lexical priming is also regarded as a *textual* phenomenon. Thus, for example, the item *winter* is said to be primed to collocate with *in, that, during the*, etc. As regards colligational behavior, Hoey's most complex examination is of the colligational behaviour of the item *consequence*. He looks at 1,809 occurrences in his corpus data (a 100-million-word corpus of *Guardian* newspaper texts) and discovers first of all that it displays a clear aversion to appearing as part of the object of a sentence (4% of occurrences) but no such aversion to appearing as part of a verb complement (24%). Given that, exactly as with many types of research, the relevance of such individual findings can only be evaluated by comparison with the behavior of other items, he also looks at four other abstract nouns, *question, preference, aversion*, and *use*, none of which exhibits the same absence from object position (occurring there, respectively, on 27%, 38%, 38%, and 34% of occurrences).

The principle of evaluative cohesion, very closely associated with coselection, states that, in normal circumstances, speakers and writers will attempt to maintain consistency or "harmony" of evaluation – of the evaluative polarity, good or bad – at local moments in discourse production. Evaluation is here intended as "the indication of whether the speaker thinks that something (a person, thing, action, event, situation, idea, etc.) is good or bad" (Thompson 1996a: 65). It is not difficult to demonstrate how items with the same evaluation tend to cluster together. The item *fraught with* is normally chosen precisely because it "fits" with other items of negative evaluation and a concordance offers the following illustrations of clustering of items of similar (negative) evaluation (examples from the SiBol newspaper corpus, see Section 2):

(1) The seven-year journey from that dazzling sales pitch in the Far East to the reality of 2012 will be *complicated* and *arduous*, and after Thursday we must *fear* it will be **fraught with** the *rawest of hazards* for ordinary citizens. (SiBol 05)

(2) But *appearances can be deceptive* – these funds can be **fraught with** *danger*. The managers buy *riskier* bonds to add to the mix to boost the income. (SiBol 05)

Such evaluative harmony is normally taken for granted and only becomes apparent when, for dramatic or ironic effect, a speaker/writer chooses to upset it by combining items of opposing evaluative polarity within the same text, for example, *an outbreak of honesty, the onslaught of goodwill and attention* (SiBol), when both *outbreak of* and *onslaught of* very generally co-occur with negative items. The concordancer is an excellent way of locating examples of such prosodic clash. It has also proved invaluable, through its ability to collect large numbers on instances of use in context, as a means of uncovering the evaluative polarity of many items which was not previously apparent to the naked eye, such as *set in, dealings, utterly, potentially, sit through, orchestrate, true feelings*, and *par for the course* (generally negative) and *flexible, persevere, provide, career, my place, make a difference*, and *brimming with* (generally positive).

The main function of evaluative cohesion and the consistency of evaluation at local points in the discourse is to help maintain comprehensibility for the listener, since it meets rather than upsets primed expectations. A discourse needs to make sense not only ideationally but also at the evaluative level.

Both lexical priming and evaluation will be highly relevant to the case study in Section 3.

2 A survey of previous CADS research

2.1 Corpus-assisted studies of sociopolitical discourses: (im)migration, race, gender

In this survey we concentrate on sociopolitical CADS, that is, a branch of linguistics in which corpora are employed to help study how social and political phenomena are represented and constructed in the cultural products of a society. A pioneer in the formulation of this type of research is Stubbs (1996 and 2001b), who was one of the first scholars to propose and promote the synergic use of corpus linguistics and discourse analysis. The core idea is that a mixed-approach "has the empirical data and the hermeneutic methods to try out some new approaches to long-standing problems, and should therefore try to move from the descriptive to explanatory adequacy" (Stubbs 2006: 34).

Comparative analysis of lexical patterns is a powerful tool to investigate how social, cultural, and political representations, such as gender or race, are constructed and reinforced by the accumulation of linguistic patterns. Scholars have used collocation analysis to study the discourse of sexual and gender difference. Stubbs (1996) analyzes striking differences in Baden-Powell's messages to guides and scouts and shows, for example, how the former are "full of references to men," as well as family (*husband, children*), whilst the latter make "no mention of women" or family (1996: 84). Pearce (2008) examines the differences between the lemmas *man* and *woman* in the BNC, classifying their collocates in semantic domains, including physical appearance, attitudes and interests, psychological traits, social relations, and occupations, and finding in all domains a reinforcement of gender stereotypes. Baker (2006 and 2008) compares the terms *spinster* and *bachelor*, also in the BNC and shows, for example, the cultural stigmatization of spinsters that emerges from the collocates, for instance, we find *eligible* bachelors, but *frustrated* spinsters.[1] Baker (2010) looks at gendered terms (such as *girl* and *boy* and male and female pronouns and titles) in the four British diachronic corpora of the Brown family, demonstrating the usefulness of diachronic corpora for getting social snapshots of an age. Taylor (2013) examines differences and similarities in the representation of *boy/s* and *girl/s* in the British press over time, from 1993 to 2010. The representation of females in infantilized/sexualized ways was seen to be a continuous feature of newspaper discourse over the time period but she also found that *boy/s* were increasingly characterized in these ways over time. Macalister (2011) examines gender constructions in children's books over a ninety-year period. Whilst he finds confirmation of the gender stereotypes which emerged in previous research he also registers an increased visibility and emphasis on individuality regarding girls. Baker (2005) compares the discourses surrounding *gay(s)* and *homosexual(s)* and other related terms in various corpora, including parliamentary debates and articles from two politically opposed British tabloids, the *Daily Mail* and the *Mirror*, and transcripts of the sit-com *Will & Grace*. Examining the collocates of *gay(s)* and *homosexual(s)*, he shows how the word *gay* presents an element of self-definition that is not there in the term (of medical derivation) *homosexual*. He also argues, through a review of collocates, that homosexuality in general is presented as a behavior (and a definitely negative one) rather than an identity in the tabloid press.

Other collocational research has focused on the representation of minorities. In his seminal paper Krishnamurthy (1996) analyzes how the words *ethnic, racial*, and *tribal* are used, juxtaposing three different kinds of materials: individual newspaper articles, dictionaries, and a large corpus of English. What is particularly interesting in Krishnamurthy's paper in terms of CADS methodology is the combination of perspectives and windows onto the data. He begins with the close reading of newspaper articles that inspire the analysis, then looks at dictionary definitions, and finally he examines the collocational profiles of the target words in a general corpus. From the comparative profile of collocates, we learn, for example, that *ethnic* and *tribal* tend to be associated with specific groups, with the second focusing on a particular group's members, while *racial* is used in a more abstract way. The main activity related to *racial* is *discrimination*, where the *race* is the "Done-to"; with *ethnic* it is *violence*; and with *tribal* the violence is made specific as with *killing* – in both of the latter cases the named group is more likely to be "Doer."

[1] This does not imply that *bachelor* is characterized in a univocally positive way. Baker highlights also how the term (and consequently the identity) is, for instance, associated with a man's inability to take care of himself and with loneliness.

A further influential example of corpus-assisted discourse analysis, also related to ethnicity, is the Lancaster University project on the representation of refugees, asylum seekers, and immigrants (or RASIM). Baker and McEnery (2005) looked at the discourses surrounding refugees and asylum seekers in two kinds of texts, UK newspaper articles and United Nations documents. Rather than comparing the corpora directly (for instance, using keywords analysis) the researchers analyzed collocational patterns in the two sets of texts, they identified categories of how RASIM were portrayed, such as "quantification," "movement," "tragedy," "aid," and "crime," and they compared findings for the two corpora. This kind of analysis of collocates (that is, of commonly co-occurring lexis) makes similarity visible as well as difference and allows for a comprehensive view of the representation of identities, that are often complex and intertwined. This research was followed up and expanded in the project *Discourses of Refugees and Asylum Seekers in the UK Press 1996–2006*,[2] where two groups of scholars, one consisting of corpus linguists and the other of critical discourse analysts (see Section 1 above) attempted to combine theoretical backgrounds and methods traditionally associated with their respective fields (Baker *et al.* 2008). The study fails to fulfill its promise of methodological integration between CDA and corpus linguistics given that, during the course of the research, the two teams operated largely independently from one another. While not a convincing example of "synergy," the RASIM work is, nevertheless, particularly interesting from a methodological point of view because it describes a variety of ways of entering the data and adopts a range of different perspectives, such as looking at diachronic change or stability of representations, at political stance differences and comparing styles between quality and popular newspapers (Gabrielatos and Baker 2008). The researchers used a number of tools: they traced the temporal distribution of RASIM stories, identifying "spikes" and "troughs" of press interest; they performed an extensive

[2] For further details see www.ling.lancs.ac.uk/activities/285/.

collocation analysis of the target terms and grouped collocates into semantic categories and they looked at "consistent-collocates" (that is, collocates that remain stable over time and are therefore unlikely to be the product of a particular event), showing how some discourses are progressively confirmed and reinforced. Among the discourses reinforced by repeated co-occurrence there is, for example, a moral panic about quantity prompted by the widespread use of negative metaphors of *flooding, streaming, pouring*, or the discourse of illegality created by collocates such as *caught, detained, smuggled*.

While in the case of RASIM the researchers track the evolution of discourses over a continuous period of time, another series of diachronic studies compares corpora from different points in time in order to identify change or stability. An instance of this is the SiBol set of CADS which employs comparable newspaper corpora from 1993, 2005, and 2010 (see Partington 2010a; Partington *et al.* 2013). Since the SiBol corpora contain the whole output of the newspapers for their years, a wide range of different topics can be researched, sociolinguistic as well as purely linguistic (lexical and grammatical), and since the particular newspapers were the left-leaning *Guardian*, the right-leaning *Telegraph*, and the centrist *Times*, sociopolitical issues can be viewed and contrasted from different perspectives. Partington (2010b) identifies which social concerns were labeled "moral panics" by the left and by the right in 1993 and then in 2005, to see which ones remained (for example, juvenile crime), which disappeared (for example, lone parents, trade unions, abortion), and appeared (for example, immigration, binge-drinking, obesity). Duguid (2010a) examines the word prefix *anti* in order to track the changes in the items it premodifies and the changes in social and political concerns they reflect, noting, for instance, the rise in mentions of *anti-capitalism, anti-money*, and *anti-globalisation*, and also of *anti-gun, anti-bullying*, and *anti-slavery*. Taylor (2010) uses the same corpora to look at the ways *science* is represented in the press over time, observing how the "other" to science projected by the UK press changes from "culture" and "the arts" in 1993 to "religion" in 2005. Marchi (2010) looks at what the press portrays as pertaining to the moral domain and how this changes over time. She finds a general decrease in the use of the label *moral*, a growing reference to the notion of *moral relativism*, and a tendency to see *morality* as belonging to the personal rather than social sphere. She also highlights how the inductive data-driven "funnelling" process commonly used in CADS, which consists of looking at the data, finding patterns, restricting the analysis to that phenomenon/portion is prolific in generating new questions (2010: 164). Partington (2012) examines discussions of anti-Semitism. In the earlier material it was seen largely as a historical phenomenon or restricted to Eastern Europe but concerns about a resurgence of the phenomenon in certain religious and political circles in Western Europe are widespread in the recent data. Duguid (2010b) analyzes the evolution of UK broadsheet journalistic style over

time, noting an increase both in informal style (for instance, use of vague lexis) and in overt expressions of evaluation (for example, a much greater use of hyperbole and positive evaluation, very probably reflecting an ever-growing inclusion of PR material in the magazine sections), all evidence that the so-called "quality" press is increasingly adopting linguistic practices typical of their tabloid rivals.

2.2 Comparison across discourse-types

Discourse analysis is, of course, inherently comparative; it is only possible to both uncover and evaluate the particular features of a discourse type by comparing it with others. We are not deontologically justified in making statements about the relevance of a phenomenon observed to occur in one discourse type unless, where it is possible, we compare how the phenomenon behaves elsewhere. Several corpus techniques, for example, keyword and key-cluster tools, have the specific aim of facilitating comparison. We can compare between corpora or within a corpus (for example, for different speaker roles such as questioner and responder) and we can compare a specialized corpus to a general (or "heterogeneric") one. We have discussed diachronic comparisons, but the parameters and entities to be compared can be various. Bondi (2008), for example, analyzes the role played by stance markers in academic journals for two different disciplines, namely history and economics. Bednarek (2006) compares the evaluation of the European constitution in the British broadsheets and tabloids, finding pervasive and consistent negative evaluation of the EU in the latter, while encountering a less monolithic and not univocally Eurosceptic attitude in the former. We can also compare across languages and/or different geographical and political entities. Bayley *et al.* (2012) also examines attitudes towards the EU, focusing on the semantic construction of citizenship and identity in the British, French, and Italian media, where the researchers confirm the stronger sense of "Europeness" and sometimes perhaps acritical positive evaluation of European citizenship expressed in the press in France and Italy compared to the UK. If the corpus is compiled and marked-up to identify speaker turns, we can compare different speakers, as in Taylor (2009) on friendly and hostile examination in the Hutton Inquiry, or Bachman (2011) on the debate in the British Parliament over same-sex marriage where speeches in favour and speeches against are compared. In addition to comparison between corpora and subcorpora, we can also compare different words within a corpus, for example, near-synonyms (e.g. *climate change, global warming*, and *greenhouse effect* in Grundman and Krishnamurthy 2010), or related concepts, as in the previously mentioned research on gender. Undoubtedly the most influential work in corpus-assisted comparative discourse analysis is Biber *et al.* (1999), which describes in systematic and exhaustive detail the lexical-grammatical features of four very different macro-discourse types, namely conversation, fiction, newspaper language, and academic prose, and uses the analyses to infer a general grammar of English. It is the most data-grounded and data-rich description of the language every produced.

Other CADS work looks at the interaction between discourse types. Duguid (2009) investigates that between political discourse and news discourse by analyzing the presentation of "voices" in a multi-genre corpus[3] about the Iraq war in 2003. Employing Thompson's (1996b) framework, she studies how speech events are embedded in one another by means of attribution and, following the traces of speech representation, she demonstrates empirically how messages move from the political arena via the media to the public.

From Stubbs on, several CADS researchers have found it useful to combine linguistics with other disciplines and Bednarek's recent work on TV dialogue (Bednarek 2010) and on language in different news media (see Bednarek and Caple 2012) integrates notions from the fields of semiotics, media studies, and sociolinguistics in the study of multimodal data (vision and speech). Cotter (2010) is an interdisciplinary study of journalistic language from the perspective of the newsworkers, combining linguistics and newsroom ethnography, and it employs a wide range of data, ranging from news stories themselves to interviews with news practitioners and episodes of communication among them. By integrating input from different sources as well as combining analytical tools, Cotter offers a comprehensive account of the context of newsmaking and bridges the gap between traditional analysis of the message (discourse as product) and production research (discourse as process).

2.3 Reflections on the methodologies of corpus-assisted discourse studies

According to Johnson's (2012) content analysis of the *International Journal of Corpus Linguistics*, the interest in using corpora to study discourse has grown in recent years. There is also increasing reflection on methodological aspects, with studies addressing issues of accuracy, accountability, and potential research shortcomings, a discussion on good practices that seems particularly important for a territory of research in expansion. Taylor (2013) points out that there is an embedded tendency in CADS research to focus on difference, while overlooking similarity. Critics of corpus linguistics claim that corpus research is not well equipped to identify what is absent; Taylor (2012) illustrates ways to overcome this potential failing. Some experimental work on replicability (Marchi and Taylor 2009 and Baker 2011) has been carried out in order to investigate ways of ensuring that corpus-assisted procedures, given the sometimes very large numbers of text tokens it handles, are as transparent and as accessible to other researchers as possible. Stubbs (2001b: 124) and Partington (2009: 293–294) discuss the question of what the latter terms *para-replicability*, that is, the replication of an analysis with either a fresh set of texts of the same discourse type or of a related discourse type, "in order to see whether [findings] were an artefact of one single data set" (Stubbs 2001b: 124) or whether they can be considered more generally valid, clearly an important scientific procedure in the analysis of features of any discourse type.

3 A case study: *forced primings* in White House briefings

By means of this case study on political discourse we wish to demonstrate a number of ways in which "added value" can be brought to discourse analysis by the integration of corpus techniques. In particular, we wish to show first how the concordancer's ability to collect examples of a similar linguistic phenomenon, as contained in repeated word strings or clusters, can lead to insights into the intentions of discourse participants, second how corpus techniques can enable the tracking of discourse features over time, and third how, contrary to charges from some quarters, corpora can shed light on what is absent from a dataset under examination and what this might signify. At the same time we will illustrate the typical CADS methodology of moving back and forth between statistical overview analysis (keywords and concordancing in this case) and close textual reading.

The overall topic is a study of the discourse type of White House press briefings during the opening period of the Arab Uprisings. It examines both the phenomenon of *forced primings* (Duguid 2007) – that is, the strategic flooding by linguistic means of messages favorable to speakers or their clients into an ongoing discourse – and also the related phenomenon of competition amongst speakers to have their messages, their reading of events, accepted by either interlocutors or an audience of beneficiaries (the party for whom the language event is taking place; Halliday 1994: 144; Partington 2003: 57–58). "Priming" is of course a term borrowed from Hoey (2005), who argues that individuals are primed to ingest a semi-conscious knowledge of the properties and meanings of lexical items by repeated exposure in interaction with others (Section 1 above). We contend that a roughly similar priming process can occur when individuals are repeatedly exposed to clusters and word patterns expressing the same underlying meaning, and that the process can be effected deliberately, a form of semi-subliminal persuasion.

[3] The *CorDis* corpus: an XML marked-up collection of subcorpora, including sources of news creation (British House of Commons and US House of Representatives debates), negotiation between news creators and news mediators (White House press briefings), news messages (British and American TV news programs, news reports, and comment articles from British and US quality and popular press), and the Hutton judicial inquiry (Morley and Bayley 2009).

3.1 The corpus used in the analysis: White House press briefings

White House press briefings are press conferences held on a regular basis, in normal times, daily. They are a particular type of *institutional talk* (Drew and Heritage 1992), which is defined as talk between professionals and lay people, but the definition can be stretched, as here, to include talk between two groups of professionals with an audience of lay persons (the TV and internet audience). Briefings are a particularly fascinating genre of institutional talk in that they combine features of informal talk, given that the participants meet so often and know each other well, and confrontational or "strategic" talk. The two parties involved – the spokesperson or Podium (officially known as the White House Press Secretary) and the press – have very different interests and aims, which are in conflict on several levels. The Podium wishes to project his political ideas and particular view of the world, the press to test that view, often suggesting more critical alternatives. The press hopes to uncover ever more information, including any evidence of weakness, malpractice, internal dissension, and so on, whereas the Podium ideally wants to give as little away as possible outside the official line of his employers (Partington 2003). Moreover, the stakes are very high. Not only are the Podium's words often treated by the press as White House policy, but they risk interpretation by non-American bodies as official US policy. Since they are broadcast both on television and on the internet, "any misstep can be beamed instantaneously around the world" (*CNN-allpolitics*).

The corpus of briefings employed here is called WH-Obama, and contains all the briefings of the Obama administration in the year from December 2010 to the end of November 2011 (*c.* 1,300,000 words, compiled by Franconi 2011). For comparison purposes, we also use WH-Bush, containing briefings from the George W. Bush administration (*c.* 3,400,000 words, compiled by Riccio 2009).

3.2 Asserting the administration's message, imposing primings in briefings

From the point of view of the White House, the whole raison d'être of briefings, the reason they were instituted in the first place, is to affirm the administration's favored view of events to the press and through them to the public. To this end, the Podium's discourse is replete with repeated phrases, often with minor variation. This was first noted in earlier research whilst both watching briefings, broadcast by C-Span public service TV, and by reading a good number of transcripts (Partington 2003). To study this phenomenon, we prepared lists of relatively long clusters – 4, 5, 6, and 7 items in length – using the WordSmith Wordlist Tool. Individual items which reoccurred in these clusters could then be concordanced in the hope of throwing light on the nature of messages being launched. For instance, the concordance of *realize* in the WH-Bush corpus yields 25 occurrences of HELP *the Iraqi people realize a better future* / *a better and brighter future* / *a free and peaceful future*. A concordance of *job* in the same corpus reveals how the Podium uses it to praise some party, typically some member of the government or service personnel abroad. Of the 250 occurrences, 23 are of the form, *do* an* [intensifier] *job* (for instance, the President / Secretary Rice / our troops are *doing an outstanding/superb/terrific job*). Others are of the form *we greatly appreciate the job they are doing* or we will *make sure our troops have all the tools/resources they need to get the job done* / *do their job*. This use of *job* in expressing messages of praise – of positive evaluation for some person or people – is entirely absent from the press's language; indeed commendation of any kind is much rarer in journalists' turns.

In the first six months of the WH-Obama corpus, on the other hand, in times of severe economic crisis, *job* collocates 101 times in the Podium's speech with *grow* and *growth* in constructions like [our aim is to] *drive* / *increase job creation and economic growth* and, in fact, *grow the economy and increase job creation* is the most common long cluster in the corpus.

It is true that, as Biber *et al.* remark, spoken discourse is particularly characterized by an abundance of (semi-)prefabricated phrases (see Section 1):

> Time pressure makes it more difficult for speakers [compared to writers] to exploit the full innovative power of grammar and the lexicon: instead they rely heavily on well-worn, prefabricated word sequences, readily accessible from memory. *(Biber* et al. *1999: 1049)*

But the kind of repeated sequences we see here are very often longer and syntactically more complex than Biber *et al.*'s examples of prefabs (or "lexical bundles," as they term them) from conversation, such as, *Can I have a . . . ?*, *Do you know what . . . ?*

The cluster lists of WH-Obama contained several clusters containing WORK and concordancing this item showed that, in the same six months, the Podium uses *we* + WORK a total of 198 times, often accompanied by a positively evaluating intensifier: *we are working avidly, we have worked assiduously* / *diligently* / *aggressively* / *very hard* / *every day*. In a keyword list comparing WH-Obama with the one-million-word spoken section of the BNC Sampler (a collection of diverse discourse types) the following items all appeared among the top 200 keywords: *continue* (as in *continue our efforts, continue to work on . . .*), *forward* (*move the economy forward, as we go forward to create an America that . . .*), *action, progress, effort/efforts, measures, steps, commitment, decision/decisions*. The attempt is made to portray and evaluate the White House and their political affiliates generally as active to the point of workaholism. However, the impression (and positive evaluation) is not shared by at least one journalist in the room:

(1) Q: Why does the Congress and the President and Washington generally act like a college kid and wait until the last minute to get everything done? (20/12/2010)

It is sometimes possible, moreover, with the benefit of corpus techniques, to observe how White House messages evolve, how the exact nature of the primings flooding into the discourse changes over time, which provides strong evidence of deliberate attempted linguistic engineering. This temporal tracking is possible since each briefing is contained in a separate file which is named by date.

For instance, in a study of how the Arab revolts were debated in the briefings, the first step was to concordance the names of some of the countries involved, namely, *Libya/Libyan(s)*, *Syria/Syrian(s)*, and *Egypt/Egyptian(s)*, along with the names of the countries' leaders, *Qaddafi*, *Assad*, and *Mubarak*. In January 2011, *Libya* or *Libyan* is not mentioned in the briefings room. In February, both the Podium and press are comfortable in discussing the Libyan *government*, which is mentioned 32 times, but in March only 9 times, and after that never at all (except a couple of times in the context of freezing Libyan government assets). In the same February, there are 6 mentions of the Libyan *regime* and 6 of the *Qaddafi regime*. By March, however, *regime* is used a total of 58 times, 37 co-occurring with *Qaddafi* and 21 with *Libyan*. In the final six months of the year, we find only *Qaddafi* with *regime* and never *Libyan*. The evaluatively neutral *Libyan government* has rapidly been replaced in briefings discourse with the negative *Qaddafi regime* in a priming shift to create diplomatic distance between the White House and the Libyan administration. Perhaps the most interesting aspect is that *Libyan government* disappears from the journalists' speech almost as quickly as from the Podium's. They clearly acquiesce to the White House's message and evaluations on this issue or, if we prefer, the administration's priming flooding has been successful.

There is a similar process of diplomatic distancing regarding Syria in WH-Obama, but the process is slower and not complete. In the first six months of 2011, we find 48 occurrences of *Syrian government* and only 3 of *Syrian regime*. In the second six months, there are 34 references to *Syrian regime* but it is still called *government* 18 times, all by the Podium. The Podium's language towards the Syrian leader is also gentler than that used about Qaddafi. Throughout the year he continues to be called "*President* Assad," whilst Qaddafi moves from "*Colonel*" or "*Muammar* Qaddafi" to predominantly just "Qaddafi." When the White House in August 2011 finally deems the Syrian leader to have *lost his legitimacy*, he is asked to *step aside* (previously simply to *change course* or *cease the violence*) and there is no talk of *remove/removal* from power, as for Qaddafi.

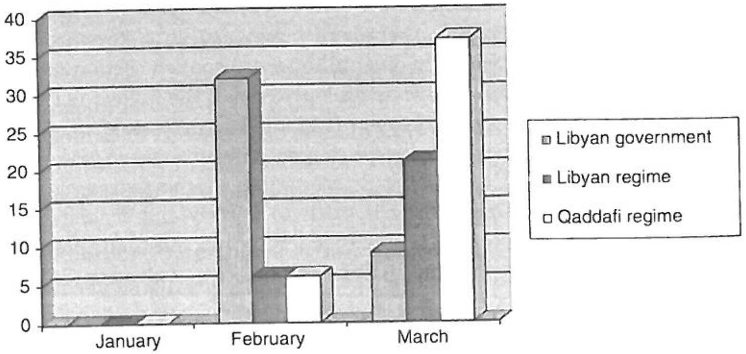

Figure 12.1 How the briefings participants refer to the Libyan administration in the first three months of 2011

There is also an evolution in the administration's evaluative messages regarding Egypt, but once again a different one. The administration in general and the Podium in particular are clearly wrongfooted and embarrassed by events. The item *Mubarak* was concordanced month by month (this is possible since, we might recall, each briefing is contained in a separate file which is named by date). Although occasionally simply "Mubarak" for the press, the Podium refers throughout the year to "*President* Mubarak." We then passed from the concordance to close reading of the co-text around occurrences of *Mubarak*, a very common process in corpus-assisted discourse studies. He is initially praised as "a close and important partner with our country" (27/01/2011) and the President even praises the Egyptian army just before news breaks of its failure to protect protestors (02/02/2011). The Podium cannot bring himself to condemn the President:

(2) Q: And as you stand today, you still back President Mubarak?
MR.: GIBBS Again Egypt is a strong ally. (26/01/2011)

As the violence grows, much of it reportedly committed by supporters of the President, the Podium is asked the straight question:

(3) Q: Do you think Mubarak is a dictator?

Realizing this phrasing might afford the Podium some room for footing evasion (perhaps "it's not important what I think"), the journalist switches the target or *recipientship* (Partington 2003: 51) of the question to the President:

(4) Q: More importantly does the President think Mubarak is a dictator?

But they are still not obliged with a straight answer:

(5) The administration believes that President Mubarak has a chance to show the world exactly who he is by beginning this transition which is so desperately needed in his country and for his people now. (02/02/2011)

Instead the Podium's repeated message is that of urging (maximum) *restraint/nonviolence*, sometimes *on all sides*. One journalist at least becomes frustrated at this priming flooding:

(6) Q: "Deeply concerned," "urging restraint" – to this point, from my knowledge, no US official has come out and condemned the violence. Is it time to condemn the violence?

This seems indeed to shame the Podium into slightly stronger language:

(7) MR GIBBS: Let's be clear Mike. Urging restraint and then seeing violence is obviously very counter to what we believe should be had. And we would strongly condemn the use of violence on either side during this situation, absolutely. (28/01/2011)

But note *on either side*. And *condemn the violence* is not the same as directly condemning the perpetrators of the violence. At this point, we concordanced month-by-month the items *violence* and *side(s)* and the co-text of the resulting occurrences were read. No mention is forthcoming, regarding Libya, about the violence committed by opponents of the regime, nor reference to condemning violence *on both/all/either side(s)*, as was the case with Egypt. In fact a concordance of *restraint* and another of *violen** in an 8-word span of *on either/both/all sides* yielded altogether 18 results, all of them contained in the Podium's turns. The countries where, according to the White House Podium, *both* sides in a conflict – that is, the government/regime and its opponents – need to refrain from violence and exercise restraint are principally Bahrain (7 occurrences) and Yemen (5).[4] The message is, in contrast, used twice about Egypt, just once about Syria, and is completely absent in discourses about Libya, where there is only *one* side seen as perpetrating violence, as the following typical statement implies:

(8) MR CARNEY: Well, let me just say that the President strongly condemns [...] the bloodshed perpetrated by the Libyan government in Libya. (23/02/2011)

This is an indication (admittedly limited given the small sample) that the administration is unwilling to take sides against the comparatively "friendly" governments of Bahrain and Yemen, but exhibits no such qualms about the rulers in Libya and Syria, who had long been seen as less than congenial by many in the West.

By the beginning of February, alongside *restraint* in Egypt, the White House also begins to flood the discourse with calls for (*orderly*) *transition* (67 occurrences in the month) which at times must *begin now*, and for *progress*, *change* and *free and fair elections* (this last phrase is used by the Podium 39 times, from 28 January to the end of February):

(9) MR GIBBS [...] I think this underscores precisely what the President was speaking about last night, and that is the time for a **transition** has come and that time is **now**. The Egyptian people need to see **change**. We know that that **meaningful transition** must include opposition voices and parties being involved in this process as we move toward **free and fair elections**. But that process must **begin now**. (02/02/2011)

[4] The remaining three refer not to Arab Spring protests but to the Israelis and Palestinians.

Analysis of keywords and key-clusters lists, using WH-Bush as reference corpus, reveals other repeated Middle East foreign policy messages. The Podium affirms that in Libya, Syria, and Egypt *people* have *legitimate aspirations* (19), *legitimate grievances* (16) (but "Bahrain is a very different case," 21/03/2011). Because a *humanitarian crisis* is unfolding in Libya, the US provides *humanitarian assistance/relief*, which includes, rather bizarrely, *non-lethal* humanitarian relief:

(10) MR DONILON: [...] And again, it would have to be worked out with the opposition group that control various aspects in the east, but they could also be air provisions of non-lethal humanitarian relief. (10/03/2011)

This would make air provisions of *lethal* humanitarian relief an intriguing euphemism indeed, a very novel evaluative clash (see end of Section 1). To stress that the US is not acting alone, in the keyword lists we find *partners*, *allies*, *together with*, *international*, *coalition*, and *multilateral*, but to avoid any accusation of lack of initiative or of weakness, also *US leadership*. Finally, *each/every country* co-occurs 40 times with *different*, for example:

(11) MR. CARNEY: Well Dan, as we've said, **each country** that has been affected by this unrest is **different. Each country** in the region is **different. Each country** has **different** traditions, political systems and relationships with the United States and other countries around the world. (24/02/2011)

This illustrates the Podium's stock response to questions on why the administration's evaluative reaction to the various regional uprisings was so different, from bombing in Libya, to diplomatic pressure in Egypt, to kid gloves over Syria and Bahrain (see also Franconi 2011).

Throughout this section, the methodology followed is the recursive "shunting" between concordancing and close reading then back to concordancing salient items noted during the close reading, very common practice in CADS work.

3.3 Institutions and forced priming

Our knowledge, use, and expectations of language are, of course, determined by our exposure to language in context but, as we see in briefings, not all exposure is the result of random personal experience. The above episodes constitute what we have called *forced priming*. Frequently repeated phraseologies – *getting the job done*, *each country is different*, and so on – result in evaluative messages being deliberately flooded into the discourse for a particular purpose. Institutions and enterprises spend considerable investment in encouraging priming through planned repetition, a process Fairclough has called the "technologisation" of discourse (1996: 71–83), and for this reason it can be illuminating to employ concordancing and key-item comparison to examine frequency data in institutional discourse.

In an age of mass communication and near instant reproduction of multimedia material, there is increased care and attention paid by institutions to how their desired messages are conveyed. Those who have the information gatekeeping role, including the Podiums in the White House, as well as government special advisers (Duguid 2009; Taylor 2009), are, of course, professional discourse technicians.

3.4 Tracking appearance and disappearance of items: governments and regimes in White House briefings

There is a general methodological point which emerges from these investigations. It has been claimed by non-corpus-assisted discourse analysts that "the corpus-based analysis tends to focus on what *has* been explicitly written, rather than what *could have been* written but was not" (Wodak 2007 quoted in Baker *et al.* 2008: 296). However, it is only by having this set of briefings texts compiled into a corpus and accessible to keyword and concordancing software that we were able first of all to assert with some confidence the complete absence of discussions of these countries prior to a certain date or that, say, Assad is never referred to as a "dictator" in the entire SiBol 2010 newspaper datasets (see Section 2 on SiBol).

Indeed, corpus linguistics techniques actually allow us to cross-compare the discourses around the different countries and their leaders and, furthermore, to *identify* absences, to *quantify* the relative absence or presence of certain messages, and to *track over time* how certain messages can move into or out of the ongoing discourse between the White House and the press.

For example, we were able to identify the absence of any mention of Libya or Qaddafi in the briefings before February 2011; it just was not on the press's map.

We were able to track changes in the denomination of the various Arab country administrations and in the honorifics or dishonorifics applied to the leaders, for instance that *Libyan government* rapidly disappears – becomes absent – whilst *Qadaffi regime*, previously absent, becomes the normal appellation.

In terms of quantifying *relative* absence versus presence we can contrast the fact that the *Colonel* of *Colonel Qaddafi* quickly disappears – becomes absent – whilst the honorific *President* continues over the period to be applied to *Mubarak*. The concordance of *all/both sides* allowed us to identify the countries where both government and opposition were urged to *show restraint* and those where only the government was being blamed for the violence.

One general observation is also highly pertinent here. In the first part of this study outlining forced priming keyword and key-cluster analyses were conducted contrasting WH-Obama with both the BNC and WH-Bush. Of course the entire raison d'être of keywording, a vital tool in the corpus linguistic kit, is to ascertain and quantify the relative presence in and absence from a target corpus of lexical items – that is what "keyness" means – usually as a first step in investigating what that relative presence/absence may infer.

It is hard to see how, without the corpus techniques or some extremely time-consuming substitute for them, any firm, objective statements on these matters could be made. Before debating "what is implied, inferred, insinuated or latently hinted at" (Baker *et al.* 2008: 296), one needs to ascertain what actually *was* and *was not* said or written, and the corpus-assisted discourse analyst would appear to be in a good position to do so.

4 Conclusion

The most obvious advantage of integrating corpus resources into discourse analysis is the potential it offers for analyzing large numbers of tokens of any particular discourse type, which enables the analyst to study typical discourse structures, typical ways of saying things, and typical messages, alongside the local structures, meanings, and messages available to traditional close reading. It also provides a way of locating potentially interesting linguistic features – for instance, sites of unusual evaluation – in a large body of texts, which the analyst can then home in upon. Additionally, it facilitates comparison among discourse types, highlighting the relative frequency and the possible different roles of the linguistic features they display, for instance, differences in collocational patterning or "profile" of the "same" lexical item or set of items.

There remains one final methodological-theoretical consideration. As in all research, there is much in corpus linguistics that is subjective, including the choice of research question and of the procedures and software to employ, not to mention the interpretation of the output data. However, there is one phase at least, namely the statistical analyses performed by the machine, where the analyst cannot either consciously or unconsciously predetermine the output and, when it arrives, s/he must deal with whatever it contains, including, and especially, things previously unexpected. These latter may include "known unknowns"; for instance, Marchi (2010) knew that some issues would be discussed in moral terms in the UK press in 1993 and 2005, but she did not, until the data analysis, know which. Or they may include "unknown unknowns"; for example, Duguid's (2010b) discovery of large numbers of explicitly evaluative keywords in the 2005 SiBol newspaper data, as compared to that of 1993 (see above), was entirely unanticipated, and an explanation needed to be sought and categorizations of the lexis developed. In this sense then corpus-assisted research, including that into discourse, is partly data-driven; intuitions themselves deriving serendipitously from data observation can often direct the course of the research, especially by encountering the unexpected, including counterexamples which challenge the initial research hypothesis. It is very generally when the possible insights of the analyst – their chances of finding things – are not entirely predetermined, and therefore constrained, by his or her initial theoretical framework and also perhaps by the paucity of data available, that advances in knowledge can be made.

13

Pragmatics

Brian Clancy and Anne O'Keeffe

1 Introduction

Corpus pragmatics is a methodological framework that allows for the interpretation of spoken or written meaning, with an emphasis on providing empirical evidence for this interpretation (see O'Keeffe *et al.* 2011). It is a relatively recent development within the field of corpus linguistics and interest in this "subfield" has blossomed as spoken corpora have become more readily available. Meaning is an elusive concept to say the least but what is clear is that participants in interaction, especially those engaged in spoken discourse, negotiate meaning through a series of sometimes almost barely perceptible "clues" that are supplied by the participants themselves, their shared knowledge (both personal and cultural) and the situation in which the interaction takes place. Given that classical pragmatics has its roots in the philosophy of language, traditionally, the study of pragmatics has employed an interpretative methodology in order to account for this negotiation of meaning. Therefore, many of the illustrative examples are invented rather than "attested" or "in use." Corpus linguistics has emerged as a sympathetic methodological companion for the study of pragmatics providing researchers with representative samples of real-life language in use, and an attendant empirical tradition.

Corpus pragmatics is distinct from other fields in corpus linguistics. However, in common with other fields, corpus pragmatics investigates the co-textual patterns of a linguistic item or items, which encompasses lexico-grammatical features such as collocation or semantic prosody. However, where corpus pragmatics' "added value" lies is in its insistence that these patterns be considered in light of the *context* – the situational, interpersonal, and cultural knowledge that interactional participants share. Through an iterative process, corpus pragmatics therefore moves beyond important but surface observations of lexico-grammatical patterns to allow a more nuanced interpretation of these patterns taking into