Corpus Linguistics: what it is and what it can do

Alan Partington

1. Introduction

Corpus Linguistics, whether it be classified as a discipline, a methodology, a theoretical approach, a conceptual frame or a new paradigm (there is considerable disagreement, confusion even, amongst practitioners, see Taylor 2008, Gries 2009), entails in essence the compilation of very large archives of running texts for subsequent analysis of many various types. When, in the 1960s, Nelson Francis built what was the first "general-purpose" (as he termed it) language corpus to be widely employed, namely the *Brown* Corpus (after the University where it was compiled), he was aware of some of the uses to which it was being put: he lists:

> […] a Swedish scholar has used it to make counts of letter frequencies in printed English […]; a philosopher in Hong Kong is studying the collocations of the word *good*; a scholar in Jerusalem is studying word-families; and my own students have used it in many studies, including the English modal auxiliaries […] and the progressive aspect of English verbs (1982: 8).

When the Lancaster-Oslo/Bergen better known as the *LOB* Corpus was compiled shortly after, mirroring the *Brown* Corpus in structure but containing texts of British English, the first corpus-based cross-language and cross-cultural studies became possible (Hofland, K. and Johansson, S. 1982; Mair 1998). Leech and Hofland conclude their keyword (see section 3 below) cross-cultural analysis of two corpora in painting (rather over-confidently given the small data-sets empolyed) as follows:

> "a picture of US culture in 1961 – masculine to the point of machismo, militaristic, dynamic and actuated by high ideals, driven by technology, activity and enterprise – contrasting with one of British culture as more given to temporizing and talking, to benefiting from wealth rather than creating it, and to family and emotional life, less actuated by matters of substance than by status" (1992: 44-45).

The *International Corpus of English* (*ICE*) project began in 1990 with the primary aim of collecting material for comparative studies, including cross-cultural studies, of Englishes worldwide. Twenty-four research teams around the world are preparing electronic corpora of their own national or regional variety of English, including India, Canada, Hong Kong, Ireland and East Africa each with a format similar to that of *Brown* and *LOB* (Greenbaum ed. 1996, for updates see[1])

From these beginnings, at the same time both modest and heroic - the *Brown* corpus may only count one million words but they were the days of computer input by punch card and paper tape - the following fifty years have seen the development of countless language corpora of different types for a vast variety of purposes, by no means all of which within the domain of linguistics proper. They are used *inter alia* in literary studies, philosophy, theology and political science whilst many professionals such as lawyers or journalists work with text archives which might be considered rudimentary forms of language corpora.[1] Workers in the field of Artificial Intelligence took an early interest in the role corpora might play in teaching machines to comprehend and produce natural language. A vital use, particularly relevant to this journal, is to improve translation techniques, both human and machine. But it is in linguistics proper that they have achieved their particular flowering and had their greatest influence, furthering our knowledge of how language is structured and how humans use it to communicate meaning, to express evaluations and to influence the behaviour and beliefs of their interlocutors.

Corpora can be either heterogeneric or monogeneric, that is, they may contain texts of many different types, subject to the constraints on what the compilers can practically and legally obtain, or they may contain texts of a single type. The former, heterogeneric corpora, are thus intended to

[1] See http://ice-corpora.net/ice/

be in some way representative of the language in question as a whole. The latter, on the other hand, are compiled as a means of studying a particular discourse type, for example, the language of law, of economics, of Parliamentary debates, and so on. As we shall see, discovering the particular characteristics of one discourse type can only be reliably ascertained and evaluated by contrasting that type with others.

Heterogeneric corpora tend to be very large, nowadays typically at the very least 100 million words in size. Their compilation can be complex and expensive and tends to be carried out by special organisations attached to Universities or large institutions, such as publishing houses. Monogeneric corpora, on the other hand, can be relatively easy to compile and are often created by individual researchers with a special interest in a particular discourse type.

2. Corpus editing or annotation

Corpora are sometimes edited, a process often referred to as 'annotation', either by the compilers or by third-party users. There are two principal forms of annotation known, respectively, as *part-of-speech* (or *POS*) *tagging* and *mark-up*. In the first of these each lexical element in the corpus or segment thereof is assigned a *tag* or label indicating its grammatical status (noun, determiner, qualifier, and so on) in the context in which it appears. This is usually performed semi-automatically; the software makes a preliminary assignment but human post-editing is normally essential. Tagging is generally carried out for linguistic purposes, either as a precursor to parsing the text or to the check the accuracy (and therefore grammatical understanding) of the tagging system.

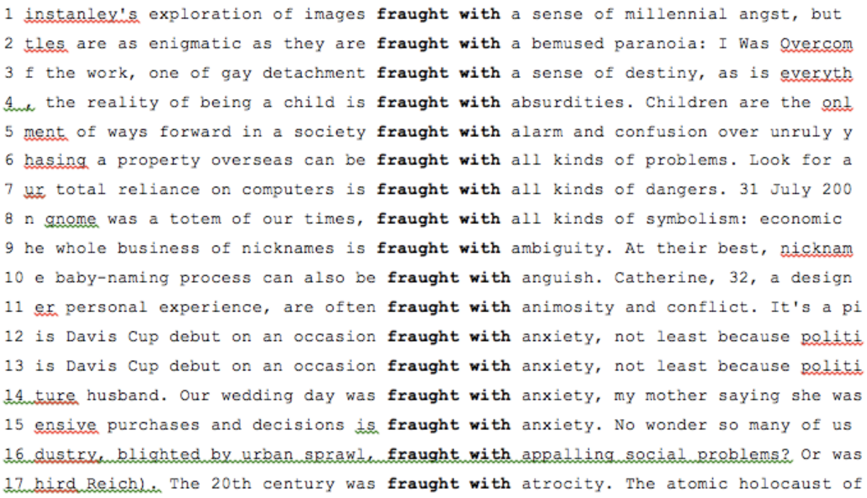
Editors may choose to mark-up an almost infinite variety of items. They may wish to indicate structural units of texts, such as introductions and closing sequences, or passages of transaction and interaction, or even shifts in the topic of discussion. In spoken texts they may wish to add information about the sex, age, occupation, and so on, of speakers. Or they may wish to indicate the occurrence of foreign words, slang, personal names, place names, dates, or almost anything an analyst might conceivably be interested in. Standardised editing protocols have been devised which enable marked-up texts to be machine-read in any platform environment. The most commonly used is the Text Encoding

Initiative (T.E.I.). Such editing/annotation is clearly highly painstaking and can require considerable investments of time and financial resources.

3. Instruments for analysing corpora

A corpus by itself is simply an inert archive. However, it can be 'interrogated' using dedicated software. The most important interrogation tools include, first of all, the *concordancer*, then calculators of *frequency*, *keywords*, *clusters* and *dispersion*.

The concordancer extracts as many examples as the analyst wishes of the word or expression under analysis - usually known as the searchword - and arranges them in a concordance, that is, a list of unconnected lines of text that have been summoned by the concordance program from a computerised corpus, with the searchword located at the centre of each line. The rest of each line contains the immediate co-text to the left and right of the searchword. It is generally possible to specify the number of characters of co-text from around, say, 40 to, realistically, around 500 on each side. For example:



```
1 instanley's exploration of images fraught with a sense of millennial angst, but
2 tles are as enigmatic as they are fraught with a bemused paranoia: I Was Overcom
3 f the work, one of gay detachment fraught with a sense of destiny, as is everyth
4 , the reality of being a child is fraught with absurdities. Children are the onl
5 ment of ways forward in a society fraught with alarm and confusion over unruly y
6 hasing a property overseas can be fraught with all kinds of problems. Look for a
7 ur total reliance on computers is fraught with all kinds of dangers. 31 July 200
8 n gnome was a totem of our times, fraught with all kinds of symbolism: economic
9 he whole business of nicknames is fraught with ambiguity. At their best, nicknam
10 e baby-naming process can also be fraught with anguish. Catherine, 32, a design
11 er personal experience, are often fraught with animosity and conflict. It's a pi
12 is Davis Cup debut on an occasion fraught with anxiety, not least because politi
13 is Davis Cup debut on an occasion fraught with anxiety, not least because politi
14 ture husband. Our wedding day was fraught with anxiety, my mother saying she was
15 ensive purchases and decisions is fraught with anxiety. No wonder so many of us
16 dustry, blighted by urban sprawl, fraught with appalling social problems? Or was
17 hird Reich). The 20th century was fraught with atrocity. The atomic holocaust of
```

Figure 1: A concordance of the expressions *fraught with* from a corpus of UK newspapers.

Such a list enables the analyst to look for eventual patterns in the surrounding co-text, which proffer clues to the use of the searchword. In the example given in Table 1, it can be seen how the expression *fraught with* very generally premodifies something bad, especially of three categories, namely, danger, problems and anxiety (but counterexamples are possible as in line 3). Concordances allow the observer to discover patterns of *collocation*, that is, how any particular word or expression co-occurs with other words or sequences of words with a particular frequency. These patterns are often not available to introspection alone.

The frequency calculator - often called the word-list tool - supplies a list of the words in the corpus in order of frequency. The frequency lists of two or more corpora can also be compared using the *Keyword* facility to show up *relative* frequency, or *key*-ness of vocabulary in a corpus. In practice, this tool produces lists (one alphabetical and one ordered by significance) of all words which are significantly *more* frequent in the first corpus than the second and also of those which are significantly *less* frequent. The frequency word list thus gives an indication of absolute frequency of lexis in a corpus, whilst the keyword list indicates relative frequency. They can both provide considerable information about both the particular grammatical structures found in the kind of discourse contained in the corpus and the sort of topics dealt with therein (for an extended example, see section 7 below).

Clusters are multi-word units, that is, sequences or strings of words which "are found repeatedly together in each other's company" in sequence (Scott 2007). The software user can specify the length of the string s/he is interested in, generally from two to, realistically, ten words. They are a kind of extremely tight "extended collocation". Clusters are an intriguing phenomenon in themselves. Partington and Morley (2004) suggest they "constitute 'missing links' on the chain or cline from the linguistic morass to the abstraction we call grammar" and their study will "tell us a great deal about how speakers go about the construction of discourse". In discourse terms, they reveal typical ways of saying things and therefore typical author/speaker messages. The software generally allows the user to cluster items in three ways: from the Concordance programme by clicking directly on the cluster menu option, by preparing cluster lists from *WordList* (by activating and specifying cluster length in the *settings* menu option) and finally key-cluster lists can be compiled by the keywords software by comparing cluster lists. The key-cluster lists become efficient when very large corpora are being examined.

Finally, the dispersion tool plots where an item occurs within a text. It can display in graphic fashion where an item or set of items typically occur in a large number of texts. One may wish, for instance, to discover whether editorial exhortative modals like *should* or *ought to* generally appear at the beginning middle or end of newspaper editorials or at which point during press briefings particular issues tend to be discussed - which may well reflect the relative degree of importance the participants endow them with. Figure 1 is a dispersion plot of the item 'laughter' from a transcription of a series of press briefings from the Clinton era. It is noticeable how bouts of laughter tend to cluster together and also to occur towards the end of a briefing (Partington 2003):

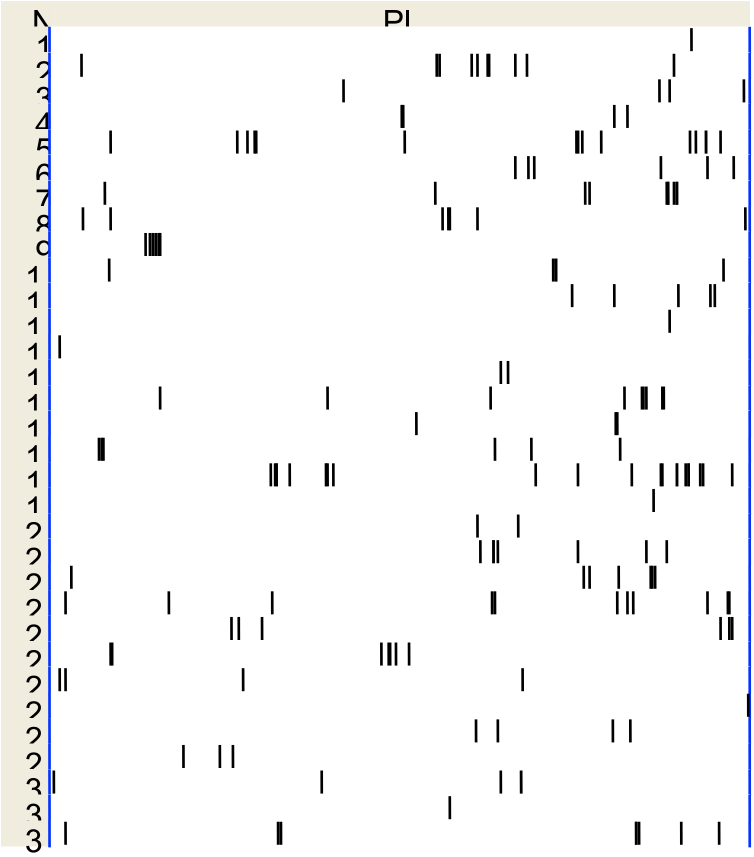


Figure 1. A plot over time (a dispersion lot) of the incidence of bouts of laughter in 32 White House press briefings held during the Clinton administration (*WordSmith Tools*, version 4.0).

4. The study of language

Heterogeneric corpora, by enabling researchers to take into account vast quantities of language data and therefore obtain an overview of the authentic behaviour of language users not otherwise readily available to the 'naked ear', have helped provide a mass of new information about the grammar and lexis of languages, and have led to the compilation of a new generation of dictionaries, of grammatical descriptions, as well as language-teaching materials. Given its global importance, the lion's share of corpus research has been conducted into English, but the field is in expansion as regards other languages, especially German, Portuguese, French, Polish, Japanese and the Scandinavian languages.

Examples of lexicological and grammatical research which corpora have enabled include the following.

Grammar

Before the advent of corpora, grammarians had a good idea of the grammatical structures possible in a language, but it was impossible to judge their relative frequency. Using corpora make it possible to see which structures are fairly common and which are extremely rare in the language as a whole (very useful for language pedagogy) and also how different types of discourse 'prefer' different modes of grammatical expression (for example, transitivity).

Several competing grammatical descriptions, particularly of English, have been tested using computers. These include Transformational Grammar (TG), Valency Grammar (VG) and Systemic-Functional Grammar (SFG). In this process, the system is first 'taught' the rules of the grammar and then exposed to actual sentences to ascertain whether it responds appropriately (the working definition of 'understanding'). SFG has proved to be highly effective in correlating grammatical description and meaning in natural language texts (Tucker 2006, 2007).

Corpus-assisted analysis has shown how grammatical distinctions are much more intricate and complex than previously thought. To highlight this, Francis examines concordances of the word *possible*. This item "has a wide range of environments which make it unique among adjectives" (Francis 1993: 147). It appears in the pattern "*the* + a superlative adjective + *possible* + head noun", as in *the highest possible level*, *the worst possible outcome* as well as in the single unit *as soon as possible*. It also appears, combined with *as*, after a wide range of adjectives, adverbs and quantifiers: *as early as possible*, *as often as possible* etc. Other *possible* patterns include *where/wherever possible*, *when/whenever possible*, *if possible*. This range of environments makes *possible* grammatically unique, but it is by no means unusual in being so:

> If we take any one of a huge range of the more frequent words in English, and examine its citations *en masse*, it will emerge that it too has a unique grammatical profile, which certainly cannot be encapsulated by calling the word in question an adjective or a noun or a preposition (Francis 1993: 147).

In correlation, corpus study has also shown how grammatically 'creative' human language users are. A study of the use of *if* constructions in newspapers, for example, has shown how they were used in a far greater variety of forms than listed in any current grammar of English (Partington 1998: 79-88).

Synonymy

Corpora can shed light on the precise relations and subtle distinctions of use among members of a set of similar items, at first glance synonymous, such as, for instance, *completely*, *entirely*, *utterly*, *absolutely*, *perfectly* etc. This is important information, especially for non-native speakers.

False or true friends

For translation purposes, by interrogating parallel corpora of two languages, it is possible to test the reliability as translation equivalents of cognate items, such as, - taking English and Italian - *just* and *giusto*, *correct* and *corretto* etc.

Evaluative language

Lexical items do not have just denotational meaning but also connotational or evaluative meaning. Corpus research is revealing that many more items than was previously suspected express a speaker's favourable or unfavourable attitude to the object of discourse, often unbeknown to the user. This can only be seen in the combinatorial behaviour of items, the kinds, the sets of other words/phrases it collocates with. It has been suggested that the study of these so-called hidden *semantic prosodies*, also known as *evaluative prosodies* (Morley and Partington 2009), can reveal instances of both irony and insincerity in the user, particularly in suasive discourses such as advertising and politics (Louw 1993).

Historical studies

So-called diachronic linguistics compares language from different periods in time to gather information on language change (Kytö and Rissanen 1990; Mair 1998, Mair et al 2003). See the section below on Modern-Diachronic Corpus-assisted Discourse Studies.

5. The study of discourse

Corpus-assisted research in the field of discourse analysis generally entails the comparison of two or more corpora of different discourse types and very often also the comparison of the contents of a monogeneric corpus with that of a heterogeneric one. In fact, discourse study is necessarily comparative or contrastive in two separate but related ways. Firstly, within an individual discourse type, only by comparing the choices being made by speakers or writers at any point in a discourse with those which are normal, that is, usual within the genre, can we discover how *meaningful* those choices are. Testing observations and findings against corpus data can provide 'background information' against which particular events can be judged.

Secondly, if we are also interested in the characteristics and content of the discourse type itself, it is vital to be able to compare and contrast its particular features and patterns with those of other discourse types. In this way we discover *how* it is special, and can go on to consider *why*. All genre or discourse-type analysis is thus properly comparative. In the wider field of discourse studies, this requirement has unfortunately not always been observed in practice. Corpora provide the means and methodology to enable rigorous and principled comparative/contrastive study to be performed.

The types of research possible using monogeneric corpora include the following:

Style and authorship studies

These generally attempt to identify distinctive characteristics of a particular author's writings. A recent development in this area is forensic linguistics which analyses written documents or transcripts in the attempt to provide evidence in legal cases of disputed authorship (Coulthard 1993, 1996).

Such studies are by their nature comparative, the particular characteristics of one author are only evident and available for evaluation when their work is compared to that of others. The choice of comparative texts is clearly an important one; comparing texts from entirely different fields of discourse could well result in a surfeit of information, too much noise. With this in mind, Morley (2007) in order to evaluate the features of Wordsworth's poetry, itself downloaded into a corpus, constructs a comparative corpus of nearly three million words containing the poetry, novels, essays, private letters and magazines from the period of roughly 1780 to 1820, downloaded from the Gutenberg website. Fischer-Starcke (2010) contrasts a corpus containing Jane Austen's novels with another of the novels of her contemporary writers and a third comprising Gothic novels. Partington (2008) contrasts the prose style of P.G.Wodehouse's work from the 1910s and 1920s with two other corpora compiled from Gutenberg, one of ordinary fiction written during the same period of circa 1.5 million words, and another of comic writing, mostly from the same time but, given the relative paucity of material, also from a slightly earlier period, containing circa one million words. Such studies generally entail at an early stage the compilation and analysis of keyword and key-cluster lists (section 3) of items which are significantly *more* frequent in one corpus compared to the others and also of those which are significantly *less* frequent, as described above.

Political science

The use of corpora in studies into politics fall into two camps. The first type is similar in its aims to discourse and conversation analysis and uses corpus techniques to investigate a particular political / institutional discourse type, exactly as with any exemplar of discourse, to uncover and analyse non-obvious patterns of language or aspects of linguistic interaction. Partington (2003) is an attempt to devise corpus-assisted discourse studies (CADS, section 6) methodologies to investigate the communicative strategies used by speakers in a particular form of institutional conflict talk between politicians and journalists, and treats issues of general linguistic interest such as facework, participation roles, attribution and metaphor. Clearly intercultural studies, which focusses heavily on the cultural elements in communication strategies and conflict could profit from this approach.

Corpus-assisted studies of (im)politeness in press briefings and judicial inquiries and have shown how participants in institutional settings operate with not just one set of positive face needs, but two, namely, competence face and affective face. The former is bolstered by appearing to be competent, authoritative and in control whilst the latter is enhanced by persuading our peers that we are, first of all, non-threatening, but also congenial and good to be around. The problem is that, since affective face is closely related to belonging to an in-group, the two forms of face are generally incompatible at any one time. Different participants can be seen to give different weight to the two types and adopt different strategies in maintaining them (Partington, 2006, pp.97-98, pp169-170; Taylor, 2009).

The second type of research is more overtly engaged with the political, social and cultural aspects of the set of texts under study and attempts to uncover any non-obvious ideological meanings and messages they may contain. Teubert (2001) has studied the language of Euroscepticism in the UK employing a corpus of texts deriving from various self-proclaimed Eurosceptic websites. Johnson et al (2003) is a diachronic study of the varied and changing ways in which PC terms (*politically correct*, *political correctness* etc) were employed in three corpora of different UK quality newspapers from 1994-1999, particularly in reference to how Labour party policies were perceived. Several authors have studied patterns of language in an individual politician's public addresses or those of a political party (for example, Fairclough 2000) with the aim of shedding more detailed light on ideological positions and how they are communicated.

Other corpus-assisted studies into politics include, among many others:

- How four UK newspapers, that is, two tabloids, the *Mirror* and the *Sun* and two so-called qualities, the *Guardian* and the *Telegraph* (the first in each pairing being left-leaning, the second right-leaning) evaluated EC/EU news actors (Hardt-Mautner 1995).
- The rhetoric of Berlusconi's electoral speeches (Garzone and Santulli 2004);
- How prediction is effected in economic texts, that is, how economic forecasts are presented and hedged (Walsh 2004);

- The language of representative assemblies, or parliaments and the question of special discourse communities working within specific political institutions (Bayley ed. 2004);
- Baker et al (2008) analysed a 140-million-word corpus of British news articles about refugees, asylum seekers, immigrants and migrants (collectively RASIM). It tested how collocation and concordance analyses were able to identify common categories of representation of RASIM, as well as directing analysts to representative texts in order to carry out more detailed qualitative analysis.
- The *CorDis* project (Morley and Bayley [eds.] 2009) investigated the intertextuality, that is, the interconnectedness of political discourse types. *CorDis* is a composite corpus or, alternatively, a collection of subcorpora of around 6 million words of transcribed spoken (c. 4.5 million) and written (c. 1.5 million) texts from UK and US sources of varying types but all relating to the post-2003 conflict in Iraq. It was devised to reflect the temporal progression from sources of news creation such as the UK House of Commons and the US House of Representatives to news negotiation in press briefings to news reporting and commenting in the media.

6. Corpus-Assisted Discourse Studies

Most of the corpus studies into politics can be seen as emanating from the previously mentioned CADS, in which aspects of the methodology and instruments commonly used in corpus linguistics are applied in the study of features of discourse. In other words, CADS combines the *quantitative* types of analysis used in corpus linguistics (i.e. large quantities of texts and statistical analysis) with the *qualitative* methods more typical of discourse studies, which examine particular stretches of discourse in detail, stretches whose particularly interesting nature may well have been identified by the initial quantitative overview. In this school of thought, research is ”a dynamic process which links together problems, theories and methods” (Bryman and Burgess 1994:4) and the researcher is free to shunt back and forth among hypotheses, data-collection, analysis, evaluation and even speculation, as long as these phases are kept separate and the movements among them are closely charted.

The aim of the CADS approach is the uncovering, in the discourse type under study, of what we might call *non-obvious meaning*, that is, meaning which might not be readily available to naked-eye perusal. Much of what carries meaning in texts is not open to direct observation: "you cannot understand the world just by looking at it" (Stubbs [after Gellner 1959] 1996: 92). We use language "semi-automatically", in the sense that speakers and writers make semi-conscious choices within the various complex overlapping systems of which language is composed, such as those of transitivity, modality and lexical sets, such as among "synonyms" (*freedom*, *liberty*, *emancipation*, *deliverance*), modification, and so on. Authors themselves are, famously, generally unaware of all the meanings their texts convey (an extreme expression of this notion being the "intentional fallacy", Wimsatt and Beardsley, 1946). By combining the *quantitative* and *qualitative* approaches it may be possible to better understand the processes at play in the discourse type. It may be possible, in other words, to access such non-obvious meanings.

Given that the aim of CADS research is to acquaint oneself as much as possible with the discourse type(s) in hand, CADS researchers typically engage with their corpus in a wider variety of ways than is traditional in other forms of mainstream corpus linguistics. As well as via wordlists and concordancing, intuitions for further research can also arise from reading or watching or listening to parts of the data-set - a process which can help provide a feel for how things are done linguistically in the discourse-type being studied. CADS work also frequently combines what can be learned from corpus analysis with other sources of information on the topic in hand, be this linguistic or socio-cultural. For instance, Partington (2003) viewed a number of the Web press briefings. These audio-visual transcripts formed a corpus which could then give information on the kinetics (gestures, expressions and so on) involved. Duguid (2010) makes systematic comparisons of dictionary definitions of the terms she investigates (intensifiers and emotionally laden items) with what the corpora disclose about them, and finds there is much more to say about the evalautive weight, whilst Taylor (2010) looks at websites and popular science books to compare what they have to say about attitudes to science with what her corpus data reveals (the SiBol corpus, see section 8).

7. Cross-cultural studies

Another field in which corpus work has made a contribution is that of cross-cultural (or intercultural) studies, in several different ways. What follows is a small sample.

Various cultural differences in discourse practices, often of interest in translation studies, have been analysed. Williams (2010), for instance, compares a corpora of research articles in English and Spanish, and includes statistical methods to highlight cultural differences in academic discourse; in particular in the ways writers in the two languages use first person verbs in the 'Methods' section of research pieces. In particular, he argues that the almost exclusive choice of *mostrar* to translate the English verb *show*, which is very frequent in research articles, is both linguistically and culturally inappropriate, given that *mostrar* is used to mean 'show' in the sense of 'put on view', 'display' and not in the sense of 'indicate' or 'prove', which it has in scientific metatext.

Cheng, Greaves and Warren (2008) have conducted cross-cultural studies on differences in conversational discourse practices between native and non-native speakers of English using the Hong Kong Corpus of Conversational English (*HKCCE*), compiled at the Hong Kong Polytechnic University, which contains around 50 hours of transcribed natural conversations involving 340 participants, with eleven different occupations, of around half a million words; 48% of the conversations being produced by native speakers and 52% by non-native speakers. The transcriptions were marked up (section 2) with prosodic information based on the discourse intonation system devised by Brazil (1997). Much of their work focused on differences and similarities between the two groups in intonation patterning, for example, the intonation of declarative-mood questions, of yes-no and *wh*-questions, of disagreement, of extended phrases and of vague language, and in differences in speakers tone choices (rise and rise-fall tones) to exert dominance and control.

Chi-Chiang (2005) compiled two corpora, one of 80,000 Chinese corpora and one of 33,000 English words, each containing news reports of a similar sort of events, namely fire incidents. From a statistical overview analysis, particularly of concordances of typical phraseologies adopted in these reports, she uncovers different strategies of emotional involvement on the part of the reporter:

> An English reporter is usually emotionless and invisible from the report. A Chinese reporter […] is frequently evaluative and often betrays his feelings about the incident, [for instance] using words and phrases like (*suoxing*, "fortunately") or (*xinghao*, "luckily"), (*buxing zhong zhi daxing*, "a fortune among misfortunes"), and so on (Chi-Chiang 2005: 220).

There are 98 instances of *fortunately*, and 31 instances of ("luckily") in the Chinese news corpus, both of which have almost always followed by a "no casualty" or "minor casualty" clause in this context. In the English fire corpus, in contrast, only two instances of *fortunate* and five instances of *lucky* were found, four of them attributed by the reporter to another voice in indirect speech no instances of *fortunately* or *luckily* – so called "attitudinal adverbs" (2005: 220).

The media-linguists working group, composed of corpus linguists and discourse analysts, from the European Union-funded *IntUne* research project (2005-9) on the theme of Citizenship, set themselves the task of monitoring how the print and television media represented problems of citizenship in the four countries with which the group was concerned - France, Italy, Poland, U.K. - and to analyse the ways in which evidence of this attention was presented to the public. The entire project was therefore explicitly intercultural in a socio-political sense. To this end, two sets of corpora were compiled from each country, one of news programmes from major TV channels, one of articles from national and local newspapers of differing political leanings. Topics studied included how the press in different countries report and evaluate the politics and customs of other member states, a comparison of how European institutions are presented and evaluated in the English and French presses (Dugalès and Tucker, forthcoming)

In a comparison of how any sense of a common historical European identity is presented in the Italian and English presses, Marchi and Partington (forthcoming) found that the UK press tends to stress the history of intra-European conflicts. Any sense of cooperative common identity can only really be found in discussion of the culture and the arts, such as "European music", "European morals" (whatever they may be) and even the "European unconscious", or when "Europe" is seen in contraposition to another cultural or geographic entity such as America or Africa or the World. The Italian press tended instead to concentrate on European history since WW2, and the processes of political, economic and cultural integration. It reports these in an entirely uncritical, almost supine, fashion (in stark contrast to the UK press which delights in reports of European dis-union), often citing pro-integration politicians verbatim.

In a comparison of how various different groups of immigrants are presented in the Italian and English presses, Morley and Taylor (forthcoming) found first of all that representations were by no means always or even generally negative. If newspapers criticised immigration it was attributed to some outside source, a vox-pop or a reader's letter, but positive aspects such as work and contributions to the economy were also emphasised. In the Italian press, Chinese immigrants appeared to be the object of greatest criticism, largely however due to a certain antipathy to China itself as an economic competitor. The Italian press was also aware of how current criticism of immigrants was often exactly the same as that heard not so long ago about Italian emigrants abroad.

Finally, here, a number of the *CorDis* project studies (section 5) were also overtly cross-cultural. A number of differences in the way in which the Iraq conflict was presented by the US and the UK media were noted, many of which were due to differing cultural practices in news production. US TV news anchors tend to recount the news in a more narrowly reporting voice style, whilst UK news presenters include more comment analysis, are more explicitly evaluative and include more varied voices and opinions. Headlines in opinion articles in UK newspapers are both more dialogic and idiomatic than their US counterparts, using a variety of syntactic and grammatical constructions, including the frequent use of personal pronouns, to involve the reader. The UK press is also more litigious, with newspapers much more frequently overtly attacking other papers or TV channels for their political stances (Morley and Bayley 2009).

In the past, the field of cross-cultural studies, or at least areas of the field, periodically come under criticism for a lack of systematicity (for example, McSweeney 2002). Now, though, the kind of comparative statistical analyses which corpus techniques makes available constitute an extremely valuable way of providing quantitative evidence for claims regarding differences in cultural practice reflected in language. They also make possible new avenues of research. In fact, both the *CorDis* and *IntUne* projects would have been inconceivable without the inclusion of corpus analysis.

8. Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS)

Another very particular form of cross-cultural studies is the recent offshoot of CADS, named modern diachronic corpus-assisted discourse studies (MD-CADS; see Partington [ed.] 2010 for the first collection of papers in the field). Much CADS work employs comparison of some kind and this sub-field is entirely predicated upon a very particular kind of comparison, namely, that between discourse practices and between attitudes to cultural and social issues in different moments of recent time.

Researchers at the Universities of Siena and Bologna in Italy and Portsmouth in the UK compiled the *SiBol* Corpus, consisting of three (sub)corpora from different but contemporary periods in time, designed and compiled so as to be as alike as possible to eliminate potential maverick variables. The first, *SiBol 93* contains all the articles published by the three main UK quality newspapers, namely *The Times*, the *Telegraph* and the *Guardian* in the year 1993. The second and third, namely *SiBol 05* and *Port2010*, contain articles which appeared in the same three newspapers in the years 2005 and 2010. They contain around 100, 145 and 135 million words respectively. Articles in the two corpora were marked up to permit the retrieval of metalinguistic information about the different political orientation of the three newspapers, date and specific source.

One important finding regards changes in the cultural practices involved in UK newspaper production. Duguid (2010) found many indications of an increase in the personalisation or familiarisation of newspaper register over the thirteen years between the two corpora. Indeed, UK 'quality' newspapers appear to be adopting some of the language practices once thought typical of their downmarket counterparts, the tabloid papers. This finding is consonant with other studies: Fairclough (1995) has written on what he terms the *conversationalisation* of media discourse; others talk of *political cross-discourse* (Alvarez-Cáccamo and Prego-Vásquez, 2003). As regards newspapers in particular, McNair (2003) describes what he calls the *tabloidisation* of UK so-called quality newspapers.

As lexical evidence of this process of conversationalisation, Duguid observed the very clear presence of intensifying and emotional words in the 2005 and 2010 lists, e.g. *fantastic*, *amazing*, *hugely*, *loving*, *fabulous*, *iconic*, *compelling*, *gorgeous* and many more. This must be contrasted with the almost complete absence of such items from the 1993 data; the only evaluations in the keywords of *SiBol 93* which could be regarded as remotely hyperbolic are *distinguished* and *necessary*.

As grammatical evidence of this process of conversationalisation, the most striking feature of the 2005 and 2010 keyword lists compared with the 1993 data is the degree of salience of first and second person personal pronouns. The items *you I*, *your*, *my*, *we*, *me* and *us*, as well as *yourself* and *myself* are all high in the lists.

They also include a large number of verb contractions, including *it's*, *I'm*, *that's*, *he's*, *there's*, *you're*, *I've*, *we're* and *I'd*, as well as a large variety of negative contractions, including *don't*, *didn't*, *doesn't*, *can't*, *wasn't*, *isn't*, *won't*, *couldn't*, *wouldn't*, *aren't* and *hasn't*.

The most salient verb in the recent keyword lists is *get,* followed by *can, think, want*, *got*, *know* and *like*. We find the frequent use of the progressive form: *going*, *getting*, *looking*, *doing*, *playing* and *drinking*, and question words *where*, *when*, *why*, *how* and *what*. All these items are commonly found in conversational forms of the language (Leech and Smith 2006)

Turning to the 1993 keyword list, we come across a good number of formal terms of address or personal appellation, all of which disappear from the 2005 and 2010 lists. These include *Mr*, *Mrs*, *Lord*, *Dr*, *Sir*, *Lady*, *Rev*, *Herr*, *Signor* and even *President*. The UK press seems to have curtailed its use of courtesy forms.

Another significant lexical-grammatical change is in the relative frequency of the type of linkers present in the lists. In 1993 we find relatively literary items such as *therefore*, *moreover*, *nevertheless*, *indeed* and *whilst*. The 2005 keywords instead include the more everyday *and*, *but*, *because*, *also* and *while*.

A number of other specific cultural and social issues have been the object of MD-CADS analyses. Marchi (2010) looks at changes in the use of the item *moral* and its related forms (*morality*, *immoral*, and so on) to see what the UK qualities construed as moral issues in 1993 and in 2005, and how they are evaluated. Although the main focus is diachronic she also looks at differences across individual papers. Overall she finds that morality is increasing viewed as a personal rather than a social question. Taylor (2011) has analysed the use of the lexemes boy(s) and girl(s) in the UK newspapers from 1993 to 2010 and was able to highlight the ways in which girl is consistently associated with sexual contexts and, in a more detailed analysis described the ways in which this is both the result of female children being described in adult terms and also the result of adult females being infantilised. Finally, Partington (forthcoming) examines the discourses relating to antisemitism in the three leading UK national "quality" newspapers from 1993 to 2009. Considerable changes were noted between the discourses in the earlier corpus compared to the later ones. In the first, the majority of discourses were either historical or they were discussions of potential or reported antisemitism outside the UK. In the later corpora, however, there is much more discussion about a perceived resurgence of violent antisemitism in the UK and Western Europe. Traditional right-wing strains have been joined by leftist versions, characterised by the conflation of Jews with Israel and conspiracy theories of Jews controlling America, and by a strain of Islamist antisemitism which has reportedly entered Western European culture.

9. Conclusion

Given the number and variety of studies which have been performed using corpora in recent times, any brief overview such as this has to be a highly personal selection and an almost random bucket from the ocean.

I have attempted to emphasise the particular potential of employing corpora in comparative and contrastive studies, including cross-cultural studies, given how '[a] key way that we make sense of things is by casting them in relationship to something else' (Baker, 2010: 125). These have included comparisons across languages, across different forms of the same language such as, say, US and UK varieties (which can also presuppose cross-cultural study) and different discourse types. We have even seen how, in MD-CADS the 'same' form of language and the 'same' culture can be contrasted with their incarnations in other different periods of time.

Corpus linguistics and comparative studies, including the kind of comparison and contrasts inherent in cross-cultural studies, are, in fact, natural partners. An analyst who wishes to compare one set of data as expressed in texts with another such set would do well to consider compiling corpora containing tokens of the texts in question. New and different kinds of information may well arise from the quantitative statistical analyses which this permits, which may even lead to new and different types of research questions which may be asked of the data. Conversely, due to the relative ease in recent times of compiling corpora of at least some types of discourse, along with the availability of software to perform preliminary statistical comparisons, very many corpus linguists, as witnessed by the research outlined here, have become *comparative* and *contrastive* corpus linguists.

[1] For a variety of examples see the Oxford Text Archive site:

[2] Text Encoding Initiative (http://www.tei-c.org/index.xml)

[3] http://www.intune.it/

[4] The UK newspapers chosen were the *Guardian*, the *Daily Telegraph* (national), and the *Scotsman*, and the *Western Mail* (local). The Italian newspapers were *La Repubblica*, *Il Corriere della Sera* (national), *La Gazzetta del Sud* and *Il Giornale di Brescia* (local).

References

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

Baker, P. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh University Press

Bayley, P. (ed) 2004. *Cross-Cultural Perspectives on Parliamentary Discourse*. Amsterdam and Philadelphia: John Benjamins.

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

Brazil, D. 1997. *The Communicative Role of Intonation in English*. Cambridge: Cambridge

University Press.

Bryman, A. and Burgess, R. (eds) 1994. *Analyzing Qualitative Data*. London: Routledge.

Chi-Chiang, S. 2005. 'Fixedness in genre-specific language and intercultural differences: Comparing English and Chinese fire news corpora'. *International Journal of Corpus Linguistics* 10:2, 199–225.

Cheng, W., Greaves C. and Warren, M. 2008. *A Corpus-driven Study of Discourse Intonation*. Amsterdam: John Benjamins.

Coulthard, D. 1993. 'Beginning the study of forensic texts: Corpus, concordance and collocation'. In Hoey, M. ed. *Data*, *Description*, *Discourse*. London: Harper Collins, 86-97.

Coulthard, M. 1996. 'The official version: Audience manipulation in police records of interviews with suspects'. In Caldas-Coulthard, C. and Coulthard, M. eds. *Texts and Practices*. London: Routledge, pp. 166-178.

Dugalès, N. and Tucker, G. (Forthcoming). 'Representations of Representation: European Institutions in the French and British Press'. In Bayley, P. and G. Williams (eds.). *European Identity: What the Media Say*. Oxford University Press.

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

Fairclough, N. 2000. *New Labour, New Language?* New York: Routledge.

Fischer-Starcke, B. 2010. *Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries*. London: Continuum.

Francis, G. 1993. 'A corpus-driven approach to grammar – principles, methods and examples'. In Baker, M., Francis, G. and Tognini-Bonelli, E. (eds) 1993. *Text and Technology: In Honour of John Sinclair*. Amsterdam and Philadelphia: John Benjamins, pp. 137-56.

Francis, W. Nelson 1982. 'Problems of assembling and computerizing large corpora'. In S. Johansson (ed.) *Computer Corpora in English Language Research*. Bergen: The Norwegian Computing Centre for the Humanities

Garzone, G. and Santulli, F. 2004. 'What can Corpus Linguistics do for Discourse Analysis?' In A. Partington, J. Morley and L. Haarman (eds), pp. 351-368.

Greenbaum, S. (ed.) (1996) *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Gries*,* S. 2009. 'What is corpus linguistics?' *Language and Linguistics Compass* 3. 1-17.

Hofland, K. and Johansson, S. 1982. *Word Frequencies in British and American English*. Bergen: The Norwegian Computing Centre for the Humanities.

Johnson, S., Culpeper, J. & Suhr, S. 2003. 'From "politically correct councillors" to "Blairite nonsense": discourses of "political correctness" in three British newspapers'. *Discourse and Society*, 14:1, 29–47.

Kytö, M. and M. Rissanen. 1990. The Helsinki Corpus of English texts: Diachronic and dialectal. *Medieval English Studies Newsletter* 23, 11-14.

Leech, G. and R. Fallon. 1992. 'Computer corpora: what do they tell us about culture?' *ICAME Journal*, 16, 1-22.

Leech, G. and N. Smith. 2006. 'Recent grammatical change in written English 1961-1992: some preliminary findings of a comparison of American with British English'. In A. Renouf and A. Kehoe, A. (eds) *The Changing Face of Corpus Linguistics.* Amsterdam: Rodopi, pp. 186-204.

Louw, B. 1993. 'Irony in the text or insincerity in the writer – The diagnostic potential of semantic prosodies'. In M. Baker, G. Francis and E. Tognini-Bonelli (eds), *Text and Technology.* Amsterdam and Philadelphia: John Benjamins, pp. 157-176.

McSweeney, B. 2002. 'Hofstede's model of national cultural differences and their consequences: A triumph of faith - A failure of analysis'. *Human Relations*, 55:1, 89-118.

Mair, C. 1998. 'Corpora and the study of the major varieties of English: Issues and results'. In H. Lindquist et al (eds) *The Major Varieties of English.* Växjö: Växjö University Press, pp. 139-157.

Mair, C. and Hundt, M. and Leech, Geoffrey and Smith, N. 2003. 'Short term diachronic shifts in part-of-speech frequencies: a comparison of the tagged *LOB and F-LOB corpora.' International Journal of Corpus Linguistics*, 7:2, pp. 245-264.

Marchi, A. and Partington. A. (Forthcoming). 'Does 'Europe' have a common *historical* identity?' In P. Bayley and G.Williams (eds) *European Identity: What the Media Say*, Oxford: Oxford University Press.

Morley, J. and Bayley P. (eds) 2009. *Corpus-assisted discourse studies on the Iraq conflict: Wording the war*, London: Routledge.

Morley, J. and Partington, A. 2009. 'A few *Frequently Asked Questions* about semantic – or *evaluative* – prosody'. *International Journal of Corpus Linguistics* 14:2, 139-158.

Morley, J. and Taylor, C. (Forthcoming). 'Us and Them: How immigrants are constructed in British and Italian newspapers'. In P. Bayley, and G.Williams (eds) *European Identity: What the Media Say*, Oxford: Oxford University Press.

Partington, A. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam and Philadelphia: John Benjamins.

Partington, A. 2003. *The Linguistics of Political Argument : The Spin-doctor and the Wolf-pack at the White House*. London: Routledge.

Partington, A. 2006. *The Linguistics of Laughter: A Corpus-Assisted Study of Laughter-talk*. London: Routledge.

Partington, A. 2008. 'From Wodehouse to the White House: A Corpus-Assisted Study of Play, Fantasy and Dramatic Incongruity in Comic Writing and Laughter-Talk'. *Lodz Papers in Pragmatics* 4:2, 189-213.

Partington, A. and Morley, J. 2004. 'At the heart of ideology: Word and cluster / bundle frequency in political debate'. In Lewandowska-Tomaszyk, B. ed. *Practical Applications in Language and Computers (PALC 2003).* Bern: Lang, 1179-192.

Partington, A., Morley, J. and Haarman, L. (eds) 2004. *Corpora and Discourse.* Bern: Lang.

Partington, A. (Forthcoming). 'The changing discourses on antisemitism in the UK press from 1993 to 2009: A Modern-Diachronic Corpus-Assisted Discourse Study'. *Journal of Language and Politics.*

Scott, M. 2007. *WordSmith Tools Users' Manual Version 5.0*. Oxford: Oxford University Press.

Taylor, C. 2008. 'What is corpus linguistics? What the data says'. *ICAME Journal*, 32, 143-164.

Taylor, C. 2009. 'Interacting with conflicting goals: facework and impoliteness in hostile cross-examination'. In J. Morley & P. Bayley (eds) *Corpus-assisted discourse studies on the Iraq conflict: Wording the war*, London: Routledge, pp.208-233.

Taylor, C. 2010. 'Science in the news: a diachronic perspective.' *Corpora* 5:2, 221-250.

Taylor, C. 2011. 'Searching for similarity: the representation of boys/s and girl/s in the UK press in 1993, 2005 and 2010'. Talk given at *Corpus Linguistics 2011*, University of Birmingham, July 22nd 2011.

Teubert, W. 2001. 'A province of a federal superstate, ruled by an unelected bureaucracy. Keywords of the Euro-sceptic discourse in Britain'. In A. Musolff, C. Good, P. Points, & R. Wittlinger (eds), *Attitudes towards Europe. Language in the unification process*. Aldershot, UK: Ashgate, pp.45-88.

Tucker, G. 2006. 'Systemic in*corpora*tion: on the relationship between corpus and Systemic Functional Grammar'. In G. Thompson and S.

Hunston (eds) *System and Corpus: Exploring Connections*. London: Equinox, 81-102

Tucker, G. 2007. 'Exposure, expectations and probabilities: implications for language learning'. In A. McCabe, M. O'Donnell, and R. Whittaker (eds) *Advances in Language and Education*. London: Continuum, 239-253.

Walsh, P. 2004. 'Throwing light of prediction: Insights from a corpus of financial news articles'. In A. Partington, J. Morley and L. Haarman (eds), pp. 335-348.

Williams, I. 2010. 'Cultural differences in academic discourse: evidence from first-person verb use in the methods sections of research articles'. *International Journal of Corpus Linguistics* 15:2, 214-239.

Wimsatt, W. and Beardsley, M. 1946. 'The Intentional Fallacy'. *Sewanee Review*, 54, 469-470.