August 1995

'Only Connect.'
Critical Discourse Analysis and Corpus Linguistics

Gerlinde Hardt-Mautner
Institut für Englische Sprache
Wirtschaftsuniversität Wien
Augasse 9
A – 1090 Wien
Austria
ghm@isis.wu-wien.ac.at

. . . I continue to believe that one should not characterize linguists, or researchers of any kind, in terms of a single favorite tie to reality. (. . .) I would like to see the day when we will all be more versatile in our methodologies, skilled at integrating all the techniques we will be able to discover for understanding this most basic, most fascinating, but also most elusive manifestation of the human mind.
(Chafe 1992: 96)

I. Introduction

A brief look at the genesis of this paper will help to explain its structure and orientation. Originally, the project it developed from — an analysis of the EC/EU discourse of the British press — was to draw solely on the theoretical foundations and descriptive resources of the framework known as critical discourse analysis, or CDA for short (cf. Fairclough 1989, 1992, 1993, 1995a, 1995b; van Dijk 1991, 1993; Wodak 1990, Wodak et al., 1990). However, the mainly qualitative methodology used in CDA proved ill-suited to handling the sizeable corpus that formed the basis of the study.[1] It was this mismatch between the chosen framework and the nature of the data that led to the development of an alternative analytical procedure, combining the use of concordance programmes with CDA's traditional qualitative analysis.

Both my research project and the present paper have been designed primarily in accordance with the agenda of CDA, not that of corpus linguistics. With an audience of critical discourse analysts (rather than corpus linguists) in mind, this paper is not concerned with the computer's rôle in lexicography or grammatical description but with its potential in helping to unravel how particular discourses, rooted in particular socio-cultural contexts, construct reality, social identities and social relationships (cf. Fairclough 1992: 64). The choice of priorities for this paper meant that the technicalities of computer

[1] For details of the corpus see Appendix A.

processing would remain in the background and not themselves become the object of investigation. Readers with expertise in corpus linguistics should not be disappointed by the lack of novelty or the step-by-step account, as the main idea is to describe what can be done by using existing programmes that are widely available, user-friendly and will run on a PC. The intended audience are linguists who work within a CDA framework and in whose general research routine the computer may so far have played a part only as a word processor.

Both the programmes featuring in this paper (*Longman Mini Concordancer* and *Wordcruncher*) satisfy the criteria of availability, PC-compatibility and user-friendliness, which was why they seemed an obvious choice as tools. These are still the programmes most likely to be available to researchers who have limited or no access to the more sophisticated (and generally unmarketed) software being developed in specialized research centres.

Wordcruncher and *Longman Mini Concordancer* (LMC) each have their respective merits and drawbacks, which is why it seemed sensible to use them concurrently. LMC is more user-friendly, and in many ways more versatile, especially in the area of KWIC ('keyword-in-context') concordances;[2] but it is quite limited as far as file sizes are concerned. *Wordcruncher*, on the other hand, has no difficulty coping with larger files, but you pay the price of having to come to terms with an unwieldier programme.[3]

The application of corpus linguistics has so far been mainly in two areas: lexicography, on the one hand, and more general linguistic research, on the other, whether with the 'pure' aim of description or with language teaching in mind. It is not (yet) common practice to harness the computer in the service of some form of 'critical' inquiry. There are a few notable exceptions, though, including Caldas-Coulthard (1993), Fox (1993), Louw (1993), as well as, most pertinently, Stubbs (1992) and Stubbs and Gerbig (1993).

Because of the nature of my current research interest I am here only concerned with the analysis of written text. I am therefore not addressing any of the complex issues connected with representing speech in computer-readable format (cf. Leech/Myers/Thomas eds., 1995). Even so, the approach outlined below would, *mutatis mutandis*, also be applicable to critical discourse analyses of spoken material.

Finally, and emphatically, I want to make the point that the approach discussed in this paper is intended to supplement, not replace, the methods normally used in CDA. Qualitative and quantitative techniques need to be combined, not played off against each other.

[2] An example of a KWIC concordance is given in the appendix. For a glossary of basic corpus-linguistic terminology see Sinclair (1991: 169-176).

[3] My comments on *Wordcruncher* refer exclusively to the DOS version. The UK launch of *Wordcruncher* for Windows came too late to be taken account of here, though the new version will be used in the final stages of the project.

II. Critical Discourse Analysis and Larger Corpora: Squaring the Methodological Circle

Critical discourse analysis is not an obvious candidate for computer applications. Its methodological tradition — including an essentially holistic approach to text as well as a concern for the discourse/society interface — does not augur well for the integration of computer-aided analysis. Fowler and Kress, in their seminal paper in *Language and Control*, made the point still valid today that 'there is no analytic routine through which a text can be run, with a critical description issuing automatically at the end' (Fowler and Kress 1979: 197). More recently, Fowler also stressed that '[c]ritical interpretation requires historical knowledge and sensitivity, which can be possessed by human beings but not by machines' (Fowler 1991: 68).

The main reason why there isn't an 'automatic' discovery procedure is, of course, that 'there is no constant relationship between linguistic structure and its semiotic significance' (Fowler 1991: 90). It is impossible — or at least misguided — to ascribe a particular, invariable ideological effect to any one form. You cannot say, for example, 'Passives always do X in a text, and I've found lots of passives in my text, so my text is doing X.' Such simplistic reasoning would be an example of what Simpson (1993) refers to as 'interpretative positivism':

> 'Where the problem of interpretative positivism arises is where a *direct* connection is made between the world-view expounded by a text and its linguistic structure. Amongst other things, this step will commit an analyst to the untenable hypothesis that a particular linguistic feature, irrespective of its context of use, will always generate a particular meaning.'
>
> (Simpson 1993: 113; italics in the original)

The quantitative 'dissection' of text appears to be at odds with CDA's commitment to analysing coherent discourse at all linguistic levels. 'To isolate specific forms', Fowler and Kress argue (1979: 198), 'to focus on one structure, to select one process, in fact to lift components of a discourse out of their context and consider them in isolation would be the very antithesis of our approach'.

However, by opting for qualitative analysis, what is gained in terms of depth is usually lost in terms of breadth: the more detailed and holistic the method, the less data one can reasonably hope to cope with. Hence, this approach is 'especially relevant to detailed analysis of a small number of discourse samples' (Fairclough 1992: 230). Because critical discourse analysis is best suited to deal with small corpora the question of representativeness obviously looms large. There may be a temptation *to proclaim* features as typical rather than build up the notion of 'typicality' on the basis of frequency. The hidden danger is that the reason why the texts concerned were singled out for analysis in the first place was precisely that they were not typical, but in fact quite unusual instances which aroused the analyst's attention.

Given an appropriate institutional structure and adequate funding, the problem can be partly overcome by involving teams of researchers who cooperate on a single project. In-depth qualitative work on different text types by different team members can be collated to validate the overall interpretation of the data. The authority, plausibility and reliability of the analysis can be further enhanced if the team members come from varied disciplinary backgrounds and bring diverse conceptual worlds and analytical tools to bear on the discourse (cf. Wodak et al. 1990: 55). However, the rich and varied potential of team work is not available to the researcher working individually, so alternative ways of broadening the empirical base must be found.

In *News Analysis* (1988) the solution proposed by van Dijk (though again involving some — scrupulously acknowledged — assistance from teams) is to combine quantitative and qualitative analysis, with the quantitative component being limited to 'surface' indicators like coverage frequency and size as well as basic content analytic categories like the presence/absence of certain topics and value judgements, or the frequency of quotations. The insights gained by such a 'superficial content analysis', van Dijk argues, are 'useful but incomplete', while 'more sophisticated discourse analysis methods, such as the description of thematic, schematic, local semantic, stylistic or rhetorical structures' elude quantification and 'must still be limited to a few sample items'. He concludes, prophetically, that '[o]nly the work of large teams or, in future, of computers would enable the qualitative analysis to be quantified' (van Dijk 1988: 66).

One of the problems with 'superficial' quantification (i.e. taking account of formal linguistic categories rather than semantic ones) is that the coding and counting procedures distance the analyst from the source text. Once a linguistic phenomenon has become a tick on a coding sheet, to be processed by statistics software, the co-text, so vital for interpretation, is lost, and very often irretrievably so.

Developing the research design for my own project, entitled *The EC/EU Debate in the British Daily Press*, I was facing precisely the kind of methodological dilemma just outlined. As I favoured a bottom-up approach rooted firmly in textual evidence, I wanted to work from a larger, potentially more representative empirical base. The data originally consisted of the newspaper coverage on the EC/EU in four daily newspapers (the *Guardian*, the *Daily Telegraph*, the *Daily Mirror* and the *Sun*) and from selected periods between 1971 and 1994. Even with the focus narrowed down to newspaper editorials, the corpus still amounted to approx. 168,000 words. While this is small fry by the standards of corpus linguistics, where corpora (like the British National Corpus at Lancaster and the COBUILD Corpus in Birmingham) are now in the 100 and 200 million range, it is a formidable corpus to take on from a discourse analytic perspective, and one definitely too large to be tackled by conventional methods only.

There was never any doubt that simple concordancing programmes, especially when put to work on untagged corpora, would be unable to perform most of the sophisticated analytical procedures mentioned by van Dijk in the above quote. Simple concordancers like the ones I shall discuss below, are quite

literally 'word' crunchers in that they will only capture phenomena tied to individual lexical units. It is all too obvious that such computer programmes cannot by themselves produce a meaningful analysis. The essential 'historical knowledge and sensitivity' referred to by Fowler (1991: 68) is not, or at least not yet, within the computer's reach. Yet, as I hope to demonstrate, even the crudest techniques of corpus linguistics can make useful contributions to the study of discourse from a critical perspective.

III. Preparing the Data

A. The Procedure

The decision to enlist the services of the computer is inevitably followed by the sobering realization that a host of rather boring and time-consuming preliminaries have to be got over before the analysis proper can begin. Transforming raw text into a computer file or files readable by a concordancing programme involves the following steps:

- Transferral into electronic form by scanning or typing, depending on the quality of the hard-copy originals. This stage is arguably the most labour-intensive. Inputting is obviously not necessary if the data already exists in machine-readable form (as is the case with newspapers on CD-Rom, for example). The prior availability of an electronically stored version will naturally make certain data more attractive to work on, as it enables the researcher to all but by-pass the inputting phase. Although this is certainly worth bearing in mind when drawing up a research design, the temptation ought to be resisted to limit corpora *a priori* to texts published electronically. This may be sound time management but it is obviously questionable methodologically. Even so, it is also true that electronic communication is continuously spreading to new walks of life so that an increasing range of genres and registers are now being produced in electronic form. Informal conversation on the Internet is a case in point.

- The scanned and keyed-in texts have to be checked meticulously — probably by running them through a spell-checker — to eliminate typos and any errors the OCR[4] software may have made in text recognition. Even the least meaning-distorting orthographic error, easily decodable by a human reader, will result in the computer happily recording a new word form. If the error occurs at or near the beginning of the word, the correct and incorrect spellings will not even be placed close to each other in the alphabetical wordlist. The smaller the corpus, the more serious this kind of mistake is, because if the overall frequency of individual items is quite low anyway, losing (or rather 'mislaying') even a single occurrence through wrong spelling affects the accuracy of the analysis.

- Finally, the inputted and corrected texts have to be fitted with whatever codes the software requires to be able to locate individual occurrences.

[4]OCR = Optical Character Recognition.

Longman Mini Concordancer can process only one kind of tag, an eight-character code, between pointed brackets, which identifies the source text. If the corpus consists of newspaper articles, for example, each article would be headed by a tag giving the newspaper and the date.[5] These tags can then be displayed at the beginning of each line in the concordance so that occurrences can be located and if necessary traced back easily to their hard-copy originals. *Wordcruncher* allows tagging on three levels so that items can be located even more precisely. The levels are originally called 'book', 'chapter' and 'paragraph' but can be redefined to suit other kinds of data.

These three steps suffice to start off the simplest form of computer analysis. For more complicated procedures, more elaborate tagging (syntactic and/or semantic) would be required (cf. Leech and Fligelstone 1992: 124-127; McEnery and Wilson [forthcoming]). Whether any such additional investment of time and resources is justified will depend on the kind of information the analyst hopes to extract from the corpus. For a project in critical discourse analysis in which computer processing is not the only analytical tool, the best policy is probably to start with the modest (and theory-independent) tag set needed for source identification, and to re-edit the corpus if and when more ambitious forms of tagging are thought to be essential.[6]

The prospect of having to go through this elaborate preparation phase may be a powerful deterrent. However, unless instant gratification is the prime objective a thorough cost-benefit analysis will invariably come down on the side of involving the computer. Theoretically perhaps, some of the procedures I shall describe below can be done with the help of the proverbial shoebox and 6 by 4 inch index cards. In practice, however, the card-flipping approach is doomed to fail once the corpus has outgrown the kind of size that can be tackled manually with ease. Empirical work will always require saintly patience and dogged determination, but there is no particular virtue in wasting these qualities unnecessarily — nor in dodging a promising line of inquiry because it can only be pursued with the help of computers.

B. Semiotic Impoverishment

The account so far may have given the impression that making data machine-readable is entirely a mechanical chore. As a matter of fact, the transferral from paper to disk needs to be problematized with regard to the implications it has for the nature of the data. That the inputting process involves stripping the text of most of its non-verbal properties (such as layout, typography, pictures, graphic elements, etc.) may not matter when the intended goal is of a lexicographic character, but to anyone working within a CDA framework it is far from a trifling matter. The impact of discourses depends crucially on their 'multi-modality', and to confine the analysis to the verbal component is

[5] Cf. the sample concordance in Appendix B.

[6] Cf. Sinclair's advice (1991: 28-29) 'to refrain from imposing analytical categories from the outside until we have had a chance to look very closely at the physical evidence'. Stubbs and Gerbig (1993) also advocate (and follow) a 'clean text' policy.

to exclude many other elements vital to the meaning-making process.[7]

Furthermore, creating a machine-readable corpus involves decisions about what is to be considered a higher-level textual unit. It may seem intuitively obvious that a newspaper article, for example, should be treated as a unit and consequently be fitted with a highest-level tag (the 'book' level in *Word-cruncher* terminology). But what is the basis of that decision, that is, what do we recognise as boundary markers between an article and others around it? Should it be headlines, typographic variation, or lines and boxes? All of these may play a part in delimitation, but none needs to be conclusive. Even if, as indeed in the majority of cases, individual articles are clearly delineated, there are still varying degrees of thematic coherence, implicit or explicit, between different articles on a page and within a section of the paper. As long as our data comes in the form of a complete newspaper page, or, preferably still, a complete newspaper, we can take account of these phenomena if and when they appear revelant because we have not prematurely imposed a structure on the data. In computer processing, on the other hand, we cannot capture any pattern of organisation higher than the one that is recognised by the coding system. If 'article' is the highest level, then this is all we can hope to make statements about; opaque intertextual links between, say, an editorial and an adjacent commentary, will not become apparent. To pick these up, it is still necessary to refer to the original — an easy enough task thanks to the source identification made possible by the concordancers. Incidentally, the need to view texts in their authentic co-textual environment is a strong argument for not relying exclusively on data in their electronic form, such as newspapers on CD-ROM.

Although at the moment concordancers are restricted to processing strings of characters it is not hard to envisage programmes able to cope with a multi-media environment. 'Hypertext' systems are a step in this direction (Burnard 1992: 17-20). Ultimately, concordance programmes should work on the scanned image in its entirety, allowing the user to search for elements in all codes and thus putting an end to the semiotic reductionism currently besetting computational analysis. Any search process involving the non-verbal will have to go beyond the automatic matching of identical character strings and move on to the kind of fuzzy logic that is needed to identify similarities in non-linear configurations, such as pictures. To accomplish the latter, the text processing software needs to be endowed with human-like discernment, flexibility, and above all, a learning capacity that it does not at present have. However, given the rate of technological development, utopias in this field have a habit of turning into yesterday's news at a phenomenal pace — if, that is, there is a sufficiently strong commercial motive to turn technological feasibility into marketable software.

[7] Cf. Kress (1993: 188): '...the most pressing issue is the recognition of the increasing role of the visual and semiotic in all forms of communication. It is no longer possible to avoid this issue in critical analyses, on the assumption, explicitly or implicitly held, that all (relevant) meaning in a text is, as it were, fully glossed in the verbal component of the text.'

IV. How Can a Concordancer Contribute to CDA?

A. Where to Start

Once the corpus is up and running, the analysis must be focused, and the initial hypotheses operationalised and homed in on. Unlike the lexicographer, the discourse analyst working with a dedicated and thematically homogeneous corpus will rarely be interested in the complete range of forms that occur in it but will concentrate on those that are frequent and salient enough to permit making meaningful statements about the particular discourse being investigated.

The researcher's knowledge about the genre and the topic concerned clearly plays an important role in finding a suitable starting point. In the case of my project dealing with newspaper discourse on the European Union, it was, for example, reasonable to expect that key terms would include *Europe, Brussels, federalism* and *sovereignty*, to name but a few. Previous experience with news language would further suggest that news actors were worth looking at, just as it would be safe to assume that personal pronouns, especially *we* and *you*, would be central to newspapers' constructing their own and their readers' identity and the rapport between the two. Stubbs and Gerbig (1993), in their computer-aided analysis of geography textbooks, concentrate on the representation of change, causation and agency; accordingly, they concentrate on linguistic features that studies of factual writing have shown to be involved in the linguistic encoding of these notions, namely passives, ergative verbs and subject nominal groups.

In other words, employing a concordancer as a relatively new and perhaps unusual tool does not mean that the analyst starts off with a *tabula rasa* — a point also made by Stubbs and Gerbig (1993: 78). The background research and hypothesis-building that the critical discourse analyst would normally engage in remain indispensable guides.

At the same time, the concordancer does provide new ways of kick-starting the analysis because it enables researchers to pursue even the most tentative leads. Wordlists and the accompanying data on frequency provide just such leads. For example, studying the wordlists (in descending order of frequency) for my four sub-corpora of newspaper editorials (from the *Telegraph*, the *Guardian*, the *Sun* and the *Mirror* respectively),[8] I noticed that in the *Sun* and the *Mirror* corpora the names of the papers were among the 20 most frequent lexical (as opposed to grammatical) items, and that this was not the case with the *Guardian* and the *Telegraph* corpora. Nor did the occurrences of *this paper('s)* and *this newspaper('s)*, also shown in Table 1, make up the difference. These absolute frequencies are all the more staggering because the sizes of the four sub-corpora are in fact inversely related: the corpus of leading articles from the *Telegraph* is three times as big as that from the *Sun*, and nearly four times as big as that from the *Mirror*. A calculation of relative frequency — say per 1000 words — would thus have yielded the same overall result. What we have here is a very simple initial clue that the editorials in

[8] See Appendix A.

	this (news)paper('s)	title of the paper
Daily Telegraph	6	3
Guardian	1	7
Sun	1	35
Mirror	2	38

Table 1: References to the newspaper (by the expression *this [news]paper* or by the respective title of the paper) in a corpus of leading articles

	the people absolute frequency	*the people* frequency per 1000 words
Daily Telegraph	2	0.03
Guardian	11	0.14
Sun	22	1.2
Mirror	6	0.4

Table 2: Frequency of *the people* in a corpus of newspaper editorials. The counting was done on the basis of a KWIC concordance, and the figures only include *the people* followed by a group boundary, thus excluding occurrences involving post-modification (as in *the people of Europe*, for example).

these four papers take a different approach to self-reference. The two tabloids refer to themselves by their names, the broadsheets do not. It would now be up to some further qualitative and, once more, quantitative probing to ascertain what happens in coherent discourse; what, if any, substitute techniques the broadsheets employ to refer to themselves (*we* being a good hunch, for example) and what effect these choices have on the papers' discoursal self-positioning vis-à-vis their readers.

Among the lexical items that appear high up on the frequency list in the tabloids but not the broadsheets we also find *people*. This coincides with anecdotal observation that the *Sun*'s editorials in particular claim to be speaking for 'the people'. *That is what the people want* — used as the thundering closing sentence of a leader on 20 January 1995 — is an example that certainly feels very typical. With the help of computing, it becomes possible to substantiate this 'feeling', and, more importantly, the notion of 'typicality'. Table 2 gives the raw frequencies of *the people* in column 1 and the average number of occurrences per 1000 words in column 2.

The quantitative evidence thus confirms that there really is a case for regarding the *Sun*'s use of *the people* as distinctive compared to the other papers investigated. The next step would be to examine the individual examples thrown up by the KWIC concordance to see, for example, what transitivity patterns *the people* is commonly bound up with. Finally, the complete texts in which *the people* occurs will need to be accessed to gauge the full ideological significance of this expression. It is at this stage that the critical discourse analyst's traditional toolkit will once more come into its own. (For examples of how the quantitative and qualitative approaches may be combined, see Section IV.B.)

The notion of 'wordlist', quite surprisingly perhaps for the computing novice, includes not only 'words' (i.e. word forms) in the traditional sense but also

	Question marks	Question marks per 1000 words
Daily Telegraph	44	0.8
Guardian	195	2.5
Sun	115	6.3
Mirror	52	3.5

Table 3: Absolute and relative frequencies of question marks in four corpora of newspaper editorials

punctuation marks. It is easy to see why these may be worth exploring. In written discourse, the number of question marks, for example, gives at least a rough measure of the number of questions contained in the corpus. This, in turn, is an important stylistic feature which, among other things, creates an impression of interactivity. In newspaper editorials questions are clearly an important rhetorical device for engaging the reader in the argument. What the punctuation wordlist reveals is that some papers employ this technique rather more often than others (Table 3).

Naturally, this quantitative information is only the first step — not least because of the crude (and, of course, linguistically inaccurate) method of locating questions through question marks rather than on the basis of illocutionary function. Still, once the trail has been laid, we can use concordances to refine the analysis, and we can look at complete newspaper articles to give a detailed, qualitative account of how questions function in argumentative prose. It must be remembered that the full co-text of any item we are looking at, whether punctuation mark, lexical or grammatical item, is always accessible at a single keystroke, so that the integrity of the discourse, and the interpretative potential that comes with it, are not in jeopardy.

Information on frequencies, though dealt with here under the heading of 'Where to Start', remains useful throughout the analysis. In fact, the more we know about the texts in the corpus as well as the discursive and social practices of which they are a part (cf. Fairclough 1992: 73), the more specific and hence the more efficient our computer-aided searches and calculations become. While the full wordlist, as the previous examples have shown, may itself provide interesting clues at an early stage, we are also likely to return to the question of frequency later on when we know more precisely which items are particularly relevant.

When examining frequency data, it is important to realize that a wordlist based on raw text, without part-of-speech or semantic tagging, is rather a crude affair — hence the need in many cases to double-check frequencies manually with the help of KWIC concordances. Working on an untagged corpus, all the programme can do is, after all, count the occurrences of all strings with a space on either side, thus lumping together homographs, different word classes (e.g. *make* [v.] vs. *make* [n.]) and different senses (e.g. *plan*t [tree, flower, etc.] vs. *plant* [factory]) as well as separating the elements of composite (but discontinuous) items such as phrasal verbs (e.g. *set in*, *bring about*), complex prepositions (e.g. *according to*, *in relation to*, etc.) and noun-noun constructions (e.g. *welfare state*, *Trade and Industry Secretary*, etc.).[9]

[9]However, the programmes do allow the user to search for such elements. *Wordcruncher*

Finally, when we talk about frequency it is necessary to stress, at the risk of labouring the obvious, that this is a relative notion. We must be clear about what frame of reference we are using. When the data consists of generically homogeneous sub-corpora — as was the case with my own project on European coverage in the press — frequency can be interpreted corpus-internally in relation to the different sub-corpora. Ideally, an external standard of comparison should also be sought, so that individual texts, as Stubbs and Gerbig point out, can be 'located in diatypic space' not only in relation to other texts but also in relation to other 'text types and text corpora' (Stubbs and Gerbig 1993: 64).

B. Practical Applications

In this section I shall give practical examples of how computer-generated concordances can be used in critical analysis. The first is concerned with the representation of news actors, the second explores an aspect of pronoun usage, and the third illustrates how even the relatively narrow environment of a single concordance line can point to larger-scale discursive processes. Taken together, these examples ought to reassure critical discourse analysts that the use of concordancing does not compromise their agenda. Far from it: a certain amount of purposefully applied word-crunching can enhance the investigation by offering new vistas.

1. The Representation of News Actors

News actors are an old favourite with linguists working on the media. Who is shown to be involved in an event and how, and what labels are used to refer to them, are tell-tale signs of how the event as a whole is interpreted by the media outlet concerned. The reason why people — more specifically, people belonging to élite groups — should feature so prominently in news stories in the first place is that 'personalization' is one of the news values responsible for the selection and structuring of news (cf. Galtung und Ruge 1973). For the tabloid press in particular, personalization serves to achieve 'a metonymic simplification of complex historical and institutional processes' (Fowler 1991: 15).

In the press coverage on the European Union, one of the central news actors until recently was Jacques Delors, long-time President of the EU Commission. The analysis which follows is based on a KWIC-concordance for the occurrences of *Delors* in leading articles which appeared in the four papers investigated during selected periods between September 1988 and November 1992.[10] The concordance was printed out[11] as well as called up on-line, so that the citations could be expanded whenever necessary. In the absence of

(DOS) has a 'combined search' facility (with three options, namely 'within n characters', 'within same paragraph', and 'within same chapter' ['chapter' being the level between 'Book' and 'Paragraph']. *Longman Mini Concordancer* asks automatically for 'word' *or phrase* to concordance'.

[10] For a profile of the corpus see Appendix A.

[11] See Appendix B.

anaphoric annotation, references to Delors in the form of expressions other than the proper name could not be captured in this initial round of analysis. However, as far as pronominal references were concerned, these were likely to be found in close vicinity of the proper name and would therefore show up in the expanded context. A few instances of Delors being referred to by different news actor labels (such as *the Commission President*) might have slipped the net, it is true, but on the whole it seemed unlikely that an editorial would talk about Jacques Delors without at least mentioning his name once.

If we compare, first of all, the *Sun* and the *Mirror*, the most obvious difference is that the latter hardly mentions Delors. There are only two occurrences in the *Mirror*'s editorials, and none at all in '91 and '92, when Jacques Delors was very much in the limelight because of the negotiations concerning the Treaty on European Union (commonly referred to as *the Maastricht Treaty*). One of these two occurrences is in fact positive (*EC PRESIDENT Jacques Delors is right to be suspicious of John Major*). For the *Sun*, on the other hand, Delors is both a major news actor and a chief target of editorial acrimony (cf. Hardt-Mautner 1995). The KWIC concordance enables us to see at a glance what negative news actor labels (nominal or adjectival) are used: *Eurodud* (l.3),[12] *bureaucrat* (l.7), *back stabber* and *shifty* (l.11). By implication, Delors is also referred to as a 'penpusher' (l.12 [*penpushers like Jacques Delors*]), a 'Eurocrat' (l.15 [*empire-building by Eurocrats like Jacques Delors*]) and as a 'Socialist' (l.8 [*the darkest recesses of the Socialist mind of Jacques Delors*]). (To appreciate the negativity of the last label one needs to remember that in the *Sun*'s value system *socialist* is unequivocally pejorative.) Looking at the right half of the concordance — to the right of the 'node', that is (Sinclair 1991: 175) — we find the activities that Delors is described as engaging in. Again there is a clear preponderance of negatively loaded expressions: Delors *pipes up* (l.9), *threatens* (l.9) and *prattles on* (l.10), and a 'deal' with the US is said to have been *sabotaged by Delors* (l.4). Further activities, in nominalised form, are *complaint* (l.2) and *sops* (l.13). In other cases we have to look beyond the main verb (and indeed beyond the length of the original concordance line) to detect the negative evaluation: *bought two giant balloons* (l.3) is given a negative slant by the reference to taxpayers' money in the following sentence, and *subsidies* in l.4, like *Socialist* in l.8, has negative connotations in the *Sun*'s essentially Thatcherite universe of values. This is compounded by the French being the beneficiaries, and by the impression of indiscriminateness conveyed by *handed out*. Finally, in l.12 (*Jacques Delors setting our taxes, making our laws* ...), it is not the verbal processes themselves that have negative connotations but the juxtaposition of an 'out-group' news actor (*Delors*) and an 'in-group' pronoun (*our*).

The data for the *Telegraph* and the *Guardian*, predictably, present a more complex picture. Certainly the news actor labels do not leap off the page as they do in the *Sun*. That in itself is worth noting as it shows that the two broadsheets do not peddle their value judgements as blatantly as the tabloids.

The number of occurrences of *Delors* in the *Telegraph* and the *Guardian* is imbalanced, though not as strikingly as between the *Sun* and the *Mirror*. Although the *Guardian* corpus is larger than the one from the *Telegraph* it

[12]Line references are to the concordance in Appendix B.

contains only three quarters (21 in all) of the *Telegraph*'s occurrences of *Delors* (28). The *Guardian* is on the whole defensive about Delors, referring to him as *a passionate and committed man* (l.1), and arguing elsewhere (l.11) that *Delors is absolutely right*. There are only two negative statements about him; one mitigated adverbially (l.2: *an allegedly overweening Delors*) and the second one blunted by a shift of *Delor*s from head of the noun group to pre-modifying position, with *blueprint* as the head (l.5: *The Delors blueprint is seriously defective in too many ways*). It is not Delors who is 'defective' but only his blueprint — a subtle but important difference. Also, criticism of Delors is shown to be ridiculously exaggerated:

- l.10: *the sceptics wanting Jacques Delors for breakfast*
- l.13: *Jacques Delors is not the cartoon menace of English fulminations but sometimes his tongue serves him ill.* (Note the mitigation in and tentativeness of the second clause.)
- l.21: [*The argument (...) will (...) become subsumed in*][13] *ritual blasts against Mr Delors*

The *Daily Telegraph*, on the other hand, is more critical of Delors than the *Guardian*. There are, first of all, several negative predications about him:

- l.8: *Jacques Delors has admitted as much* (with admit[14] carrying the presupposition that some negative act has been committed.)
- l.10 *As both a socialist and a committed federalist M Delors has never disguised his vision of a European superstate* (*Superstate*, *socialist* and *federalist* are stock ingredients of Conservative anti-European discourse, and they all have overwhelmingly bad connotations.[15])
- l.11 *The element of cynicism arises when we compare what M Delors has written, in his role of long-sighted international* [*technocrat, about the need to bring France out of its shell, with his efforts on behalf of the recalcitrant and often riotous French farmers*.] (This very obliquely accuses Delors of being a political opportunist.)
- l.20 *M Delors, not a self-effacing man*
- l.26: *the Delors vision of federalism, and of monetary union, is plainly* [*unrealistic and unacceptable in Britain and probably Denmark and other states also*.] (Note how the paper bolsters up its own position by citing support from 'outside'.)

[13] Square brackets are used to indicate quotations from outside the range of the standard KWIC format. Exploring the concordance on-line, one can view the larger context at any time by hitting the ENTER key.

[14] Sans serif face is used to indicate lemmata.

[15] When Conservative Eurosceptics criticise European integration as a 'socialist' idea they ignore the fact that Germany, which favours the idea of European federalism, has had Conservative governments since 1982. Stephen Hill, in an essay introducing a collection of articles by anti-European Conservatives, talks about Mitterand and Kohl having assembled 'the necessary machinery to turn Europe into a Superstate' and then refers to '[t]heir socialist objectives' (Hill 1993: 4). In the concoction of an ideology out of disparate elements, factual accuracy is clearly insignificant.

The labels *international technocrat* (albeit a long-sighted one) and *international civil servant* also have negative connotations if we see them in the wider context of the polemics against 'unelected Brussels bureaucrats'. It is another stock ingredient of anti-European discourse to emphasise that EU officials are not elected politicians and should therefore not have the same authority as MPs and Cabinet ministers.

These are just the negative statements made in the *Telegraph*'s own voice. In addition, there are other negative statements which are either attributed to or 'seconded' by other voices (named or implied):

- l.12 ...*M. Jacques Delors. Here is a man who is widely believed to be abusing his position...*
- l.15 *For the francophobe, M Delors is a handy personification of all that is most* [*reprehensible about France and the French*.]
- l.16 [*It seems clear from the testimony of the*] *Agricultural Commissioner, Mr Ray MacSharry, that M Delors is indeed guilty of exploiting his position as an international civil servant to manipulate the Gatt talks more in the interests of the French electoral timetable than of international trade.*
- l.22: *Britain is not alone in finding the Delors programme for EMU far too ambitious.* (Cf. the comment on l.26 above.)

As all these (except l.22) come from the same editorial, it makes sense to follow this lead by abandoning the KWIC format for the moment and look at three coherent paragraphs, thus taking the whole range of referring expressions into account, not just the proper name.

> (1) THESE are high times for francophobes. (2) Not only do the French seem intent on thrusting a widely unloved treaty down our throats; (3) now they are threatening to drag the Community into a trade war which would undermine world commerce during a recession. (4) Such behaviour confirms the darkest suspicions of those who fear that France's ideal Community would be a beggar-my-neighbour, inward-looking bloc.
>
> (5) Conveniently for the anti-Gallic tendency, all these evils seem traceable to the single figure of M Jacques Delors. (6) Here is a man who is widely believed to be abusing his position as President of the European Commission in defence of French agricultural interests, just as he is suspected of scheming to bring about a federal Europe, which he might one day co-rule, as President of France. (7) For the francophobe, M Delors is a handy personification of all that is most reprehensible about France and the French, combining protectionism and parochialism with neo-Napoleonic ambition.
>
> (8) The problem today is that the caricature shows worrying signs of coming to life. (9) France's behaviour over the Gatt negotiations is ruthlessly selfish. (10) It seems clear from the testimony of the Agricultural Commissioner, Mr Ray MacSharry, that M Delors is indeed guilty of exploiting his position as an international civil servant to manipulate the Gatt talks more in the interests of the French electoral timetable than of international trade.
>
> (*Daily Telegraph*, leading article, 921107[16])

[16]In this and all future references to dates, the format used is YYMMDD.

The negative evaluation of Delors is attributed to *francophobes*; to *the anti-Gallic tendency*, to *those who fear that France's ideal Community would be a beggar-my-neighbour, inward-looking bloc*, and to *the Agricultural Commissioner*. In addition, there are the agentless passives in Sentence (6),*is widely believed* and *is suspected; traceable*, though adjectival, also falls into this category, because semantically and derivationally it is related to the passive ('x is traceable' = 'x can be traced'). These passives also help to distance the author from the propositions expressed. *Widely*, on the other hand, works in the opposite direction, reinforcing the idea that this opinion is generally accepted. (Note another *widely* serving the same function in Sentence [2], in *a widely unloved treaty*.) Further hedging is achieved by *seems*, used once in each paragraph (Sentences 2, 5 and 10).

However, a closer look at the modality in this passage reveals that the rhetorical distancing is counter-balanced by 'high-affinity epistemic modality' (Fairclough 1993: 148). *Seem intent* in Sentence (2), expressing a degree of doubt, is followed by *are threatening* in Sentence (3). In (3) and (4), the two clauses describing the implications of French policy both use *would*, the strongest of the non-factive modals (stronger than *could* or *might*): *a trade war which would undermine ...; France's ideal Community would be* .... Sentence (7), though beginning with *[f]or the francophobe*, also continues with two verbs in categorical mode: *M Delors is a handy personification of all that is most reprehensible* .... The same is true of the third paragraph. While Sentence (8) talks about a *caricature*, Sentence (9) states, without any mitigation, *France's behaviour over the Gatt negotiations is ruthlessly selfish*. In Sentence (10), *seems* is combined with an attribution (*from the testimony of* ...), only to be followed by the categorical *is guilty*, given further reinforcement by *indeed*. This passage thus illustrates what Simpson calls the 'non-harmonic' combination of modal patterns, 'where modal operators exhibiting conflicting degrees of commitment are combined' (Simpson 1993: 153).

As far as the analytic procedure is concerned, this example demonstrates that the KWIC concordance format is partly revealing in its own right and partly serves to direct our attention to extended passages to be investigated as 'sites' of discoursal processes. Conversely, a close-up on a longer stretch of text may reveal lexical items or phrases that do a lot of the rhetorical work (e.g. *seem, is believed* and *is suspected*) and which can then be the subject of another round of searches through the computerised data. That search, in turn, may reveal further 'sites' on which certain argumentative patterns are being developed; these can again be analysed qualitatively. There is thus a constant movement between close-up and wide-angle views of the data, the results of each being fed back into the other.

In addition to the 'right sort' concordance just explored, we can also compile a left-sort one. This is more suitable for studying the honorifics used for news actors, for example. In the case of Jacques Delors, the four papers differ quite markedly in their practices. The *Sun* is the only one that uses the unadorned, and consequently very impolite, *Delors*. The two instances in the *Mirror* are both of *Jacques Delors*. The *Guardian* switches between *Jacques Delors* and *Mr Delors* (*Maitre Delors* occurs once, but facetiously, in an extended restaurant metaphor). The *Telegraph* on the other hand never uses *Mr Delors* or *Jacques*

Delors, but always uses the French form, either *M[onsieur] Delors* or *M Jacques Delors*.

Finally, it is interesting to note that the two broadsheets, but not the *Sun*, use *Delors* in pre-modifying position, in *the Delors Plan* (G,[17] l.16), *the Delors proposals* (G, l.17), *the 30 per cent 'Delors Two' increase* (G, l.20), *[t]he Delors compromise* (DT, l.6), *the Delors programme* (DT, l.22), *the Delors project* (DT, l.23), *anti-Delors rhetoric* (DT, l.24), and *the Delors vision of federalism* (DT, l.26). It could be argued, somewhat speculatively, that this is because such noun phrases are rich in presupposition (note the definite article in all but one of the examples) and hence make heavy demands on the kind of political background knowledge that only the 'quality' papers expect their readers to have.

2. Studying Pronoun Usage

Personal pronouns play a crucial role in the construction of social identities and social relations. Accordingly, critical linguists and discourse analysts have always paid a great deal of attention to them. 'Personal pronouns always deserve notice', Fowler and Kress point out (1979: 201), and indeed pronouns are invariably included in checklists of which linguistic features to target for analysis. The role of personal pronouns in discourse is thus well documented, and the literature contains plenty of illuminating insights to inspire further analyses (cf. Fairclough 1989: 179-182; Wilson 1990: 50-76; Diller 1994: 100-104; Johnson 1994). Inevitably investigations conducted without the help of computing facilities have had to confine themselves to looking at individual texts. Using a concordancer, on the other hand, puts the researcher in a position to survey a much larger amount of text and to compare patterns of pronoun usage in different corpora.

What follows is part of an investigation into the rôle of personal pronouns in newspaper editorials. For presentation here, the use of *you* has been selected — though the picture obviously remains incomplete unless *we* and *one* are given equal attention. Also, an in-depth qualitative account of pronoun usage, omitted here for the sake of brevity, remains highly relevant, because without it, the results gained through computer-aided analysis cannot be interpreted properly.

For the survey of *you* in editorials, the concordances were scanned through twice; first to set up the different categories of *you* relevant to the genre under investigation; and a second time to allocate each instance to one of these categories. Altogether, five types of *you* were identified:

1. Indefinite *you* (as in *the answer depends on where you start from* [*Guardian*, 920919]);

2. *you* addressing the reader (as in *Why the Sun believes you should vote yes* [*Sun*, 750604]);

[17] G stands for *The Guardian*, DT for *The Daily Telegraph*.

	(1) indefinite *you*	(2) *you* addressing the reader	(3) *you* addressing someone other than the reader	(4) indefinite *you* in a quotation	(5) *you* addressing a person, in a quotation
Telegraph (Total: 6)	3	—	—	1	2
Guardian (Total: 76)	58	1	2	2	13
Sun (Total: 63)	9	25	27[a]	2	—
Mirror (Total: 25)	—	12	11	—	2

[a] 16 of these come from one editorial which is 'addressed' to John Major. Although this partly distorts the sum total, it is significant in its own right because it shows that direct address of a person other than the reader need not be a one-off but may be used consistently over a longer stretch of text. On this evidence there is a good case for assuming a hybridization of genres ('editorial' and 'personal letter'). On hybridization see Fairclough (1995: 142; 211).

Table 4: *you* in newspaper editorials

3. *you* addressing someone other than the reader (as in *Are you paying attention, Tony Benn?* [*Daily Mirror*, 750602]);

4. indefinite *you* in a quote (as in *his declaration that 'you cannot bully Britain'* [*Daily Telegraph*, 921013]);

5. direct-address *you* in a quote (as in: *In answer to the question: 'Are you personally in favour of joining the Common Market?'* [*Daily Mirror*, 711026]).

Note that the distinction between (2) and (3), though hardly relevant to the grammarian, is very important if the analysis proceeds at the level of discourse. Clearly, whether *you* in an editorial is used to directly address the reader or someone other than the reader — a public figure, usually — has important implications for the way in which papers position themselves vis-à-vis their readers and vis-à-vis the élites they report on. The case for separating out *you* in quotations rests on the obvious significance that quoting has for the blending of different 'voices' in the discourse.

The figures for the four papers investigated are shown in Table 4.

Both types (2) and (3) emerge unequivocally as features of tabloid editorial style. They are particular favourites with the *Sun* (not just in terms of absolute frequency, but also if the differing corpus sizes are taken into account). The second striking divergence between the figures concerns indefinite *you* in the broadsheets. Although at 78,379 words the *Guardian* corpus is only about a third bigger than the *Telegraph* corpus of 56,169 words, the *Guardian* uses indefinite *you* nearly 20 times as often as the *Telegraph*. How are we to interpret this imbalance? Textbook grammars describe *you* in its generic use as 'informal' (Greenbaum and Quirk 1990: 115; Leech and Svartvik 1994: 58). In contrast to *one*, Fairclough argues (1989: 180), *you* is used 'to register

solidarity and commonality of experience in working-class speech'. Also, it is associated with 'the "formulation of morals and truisms". (. . .) occasionally we will employ "you" (. . .) to reflect upon a kind of conventional wisdom as opposed to actual experience' (Wilson 1990: 57). It would clearly be over-simplistic (and indeed counterintuitive) to attribute the high frequency of indefinite *you* in the *Guardian* exclusively to any one of these interpretations. The figures do not 'prove' that the *Guardian* is fond of peddling 'morals and truisms'; nor do they show the paper to be 'working-class' — with 52 per cent of its readers in the AB social class and a further 27 in C1 (Harrop and Scammell 1992: 181) that label is hardly appropriate, in spite of readers' predominantly left-wing political affiliations.[18] The effect created by a high frequency of indefinite you is probably a combination of all these factors. Intuitively, *Guardian* editorials do indeed read as more informal than those in the *Telegraph*; they appeal, if not exactly to a 'working-class' consciousness, so at least to a commonality of experience among those without substantial unearned incomes, and their relaxed, often conversational style often involves breaking down cause-and-effect relationships or otherwise complex states of affairs into 'truisms' personalized through the use of *you*:

- *You cannot have one foreign policy with 10 Foreign Ministers; nor can you have one tax policy and one economic policy with 10 Treasuries.* (G, 711014)

- *Whether you agree with her stand depends on whether you regard the Community budget as comparable with a national one* (G, 870702)

- *You cannot impose a strong exchange rate. The right to that has, as in Germany and Japan, to be earned.* (G, 920917)

- *as the coal debacle underlines, you can only ask for sacrifices if you know where you are going;* (G, 921016)

The low frequency of indefinite *you* in the *Sun*, and the total absence in the *Mirror* also needs to be treated with a fair amount of caution — not least because both 'informality' and 'working-class' would be perfectly compatible with tabloid style. However, the significantly smaller corpus sizes — 18,245 and 14,959 words respectively — mean that the corpora are not as representative as those of the broadsheets and it is certainly wise not to jump to conclusions. (It might be argued, admittedly on a rather speculative note, that the high incidence of 'direct address' *you* in these papers somehow blocks its use in the indefinite sense.)

Whatever our final interpretation, if it was not for the concordancer we could not even start weighing up different possibilities for want of a sound empirical base. The computer-aided investigation informs and enhances the traditional qualitative analysis, and vice-versa.

[18] According to MORI figures quoted in McNair (1994: 128), 59% of *Guardian* readers support Labour and 22% the Liberal Democrats.

3. Contestation of Meaning and Semantic Profiling

My final examples give further evidence that the use of concordancing programmes does not necessarily mean restricting oneself to below-sentence-level linguistic description but can in fact open a window into larger-scale discoursal processes. What we do need, of course, is a word-based 'peg' to hang the analysis on — a discrete, searchable item around which the higher-level phenomenon that we are after is expected to cluster.

In issue-centred studies, as many CDA projects indeed are, a close reading of a small but reasonably representative sample of texts will usually reveal key terms that are central to the issue concerned. In my own study on the British press and European integration, these key terms include, not surprisingly, *Europe* and *European*, as well as (given public perceptions of the EU and its institutions) *bureaucrat* and *bureaucracy*; in connection with the Maastricht Treaty negotiations, *federal/ism* and *sovereignty* play a key role, just as *France/French* and *German/y* are pivotal in the representation of Britain's relationship with these countries within the EU.

To explore the significance of *European* as a key term, KWIC concordances, once again, provided a good starting point. Scanning through the lines individually soon made it clear that *European* (adjective and noun) had a wider range of meanings than conventional dictionary definitions suggested. Also, in its immediate environment there were the following tell-tale signs suggesting that its meanings were not simply diverse but in fact 'contested' (cf. Fairclough 1992: 186):

- The occurrence of pre-modification, indicating that there are various types of 'Europeanness' — Table 5.
- The occurrence of grading, indicating that 'Europeanness' is regarded as a matter of degree rather than a categorical, either-or concept — Table 6.
- Inverted commas as a meta-linguistic distancing device, indicating a non-literal usage — Table 7.

If we probe further and look beyond editorials we find further evidence that the concepts of 'Europe' and 'Europeanness' are indeed a matter of contention. *European*, as the examples from the corpus show, can mean 'pro-European', and it is this meaning for which the anti-EU camp needs to compete if it wants to shed its image of petty nationalism. The pro- and anti-EU factions thus vie for 'possession' of the term.[19] In a House of Commons Debate in June 1991, Margaret Thatcher tackled this semantic struggle head-on:

> ...we should not let those who support a federal Europe pretend that they are somehow **more European** than the rest of us. They

[19]The appropriation of key terms by rival factions is a recurring feature of political struggle. The volume entitled *Begriffe besetzen*, edited by Liedtke, Wengeler and Böke (1991), contains a number of papers exploring this phenomenon.

Telegraph, 880922	It is good, pragmatic politics for the French to allow us to incur the odium of being half-hearted	**Europeans**	while Paris pursues her national interests under a cloak of Euro-pietism.
Telegraph, 941214	Even those of us who remain instinctive	**Europeans**	find it difficult to avoid intemperance
Guardian, 790314	Not even the most starry-eyed	**European**	of the religious sort can argue that the CAP is an Ark of any particular Covenant.
Sun, 790605	The Liberals, too, are totally committed	**Europeans**	
Sun, 921019	The Prime Minister can no longer hide behind his Mr Nice Guy image, playing the good	**European.**	

Table 5: The occurrence of pre-modification

Guardian, 790606	Mr Heath, that most	**European**	of British politicians,
Telegraph, 920921	But with France, allegedly the most	**European**	country in Europe, splitting down the middle on Maastricht

Table 6: The occurrence of grading

Telegraph, 711025	such a passionately	**“European”**	Government as that of Mr Heath
Guardian, 840619	even in such conscientious	**“European”**	countries as Holland and West Germany

Table 7: Inverted commas as a meta-linguistic distancing device

> are not; they are just more federal. There is nothing specifically European about a federal structure — indeed, the opposite: it is the nation state which is European. (...) The **true Europeans** are those who base themselves on Europe's history and traditions rather than on constitutional blueprints.
>
> (*Hansard*, 6th series, vol. 193, Session 1990-91; my emphasis)

Eurosceptics aim to deflect the accusation of being 'Little Englanders' by interpreting the label *European* in accordance with their own, non-integrationist agenda. 'What is it to be European?' asks the MP Bill Cash, a leading Conservative 'Euro-rebel', at the beginning of an essay entitled, ominously, *A Brave New Europe* (Cash 1993). In his answer (part of which is quoted below) Cash deploys a whole cluster of positive-value catchphrases: *historically sovereign* (ll.2f.), *Christian culture* (l.3), *glorified in its diversity and talented competition* (ll.4f.).[20] It is only in the second paragraph (ll.6–16 below) that the European Union is mentioned, and it is embedded in a corresponding cluster of negative-value terms: *dark origins* (l.7),*a forbidding and atavistic concept of the Volk* (ll.7f.), *racism and fascism* (l.9), *a Europe grey, dull and uniform* (l.13):

> What is it to be European? We Europeans live geographically in historically sovereign countries on a Continent in which, to a greater or lesser extent, we share or have so far shared a Christian culture, which glorified in its diversity and talented competition from the Renaissance to the nineteenth century.
>
> But there is another Europe. A Europe which has fallen prey to a European internationalism which owes its dark origins to a forbidding and atavistic concept of the *Volk* which, at its worst, has spawned racism and fascism. Twelve of these countries for nearly forty years have drawn together with some success but within an increasingly exclusive legal structure known as the European Community, bound together now by the 'Treaty of European Union'. This other Europe is a Europe grey, dull and uniform, driven now into a Hall of Mirrors by obsolete yet powerful political ambitions grafting on old solutions to new problems, particularly since the re-unification of Germany and the collapse of the old USSR.
>
> (Cash 1993: 57)

That Cash claims in-group membership as a 'European' (in *We Europeans*, l.1) is a semantically marked choice, given that, as evidence from the corpus suggests, *Europe* and *European* frequently refer to mainland Europe excluding Britain.[21]

[20] Cf. also a rather similar passage in Cash (1992: 15): 'Our history and our involvement in Europe in war and peace for centuries — our role in the last two World Wars — prove our European credentials, precisely because of our commitment to freedom and democracy now under threat from the Maastricht treaty. This commitment makes us, as it has made us in the past, Great Europeans —not Little Englanders' (my emphasis).

[21] We find the same markedness in Margaret Thatcher's Bruges speech (Thatcher 1988), in which *we Europeans* occurs only once, whereas *we British, we in Britain* and *in Britain, we* occur no less than six times. (On other aspects of pronoun usage in Thatcher's Bruges speech, see Diller 1994).

Europe and *European* have become labels that are defined and hence appropriated by political lobbies. So when Cash writes about the rebels' quest, *[f]ar from being a campaign against Europe, it is a campaign for Europe* (Cash 1993: 57), he is laying claim to a particular interpretation of 'Europe'. Definition, as always, emerges not only as a semantic but also as a profoundly political act.

While the contestation of meaning is clearly a discoursal process involving longer stretches text, it does tend to crystallise around the item in question so that it can be captured by a concordancer. The computer programme shows up the lower-level linguistic reflexes, indicating to the human analyst where to look for the higher-level process. The larger the corpus, the more obviously useful the computer's help becomes, and the more versatile the program, the more discovery procedures are at the analyst's disposal. The software currently in use at Cobuild in Birmingham, for example, builds 'word pictures' of collocational patterns (and calculates their statistical significance). If a given node was shown to co-occur frequently with a metalinguistic expression such as *word* or define, it would seem safe to conclude that the meaning of the node was somehow at issue. The collocations of *federal* — a key term in the European debate — are a case in point. In the Cobuild corpus from the *Today* newspaper (comprising roughly 10 million words), *word* turns out to be the lexical item most likely to be found in the environment of *federal*.

Another area in which corpus linguistics and CDA can be expected to have mutual interests is that of 'semantic prosody', defined by Louw (1993: 157) as '[a] consistent aura of meaning with which a form is imbued by its collocates'. Louw's concern in that particular paper is to explain irony as a departure from an expected collocation. It is only through looking at a sufficiently large corpus that principled statements can be made about what is 'expected'. Examining a multitude of examples from a 37 million word corpus held at Birmingham, Louw shows not only that some forms have an overwhelmingly 'good' or 'bad' prosody (*bent on* and *symptomatic of* are two cases of a 'bad' prosody) but also that 'the prosodies based on very frequent forms can bifurcate into "good" and "bad", using a grammatical principle like transitivity in order to do so' (Louw 1993: 171). His example, *build up*, has a good prosody when it is used transitively with a human subject (as in *build up better understanding*), but a negative one when it is used intransitively (it is toxins and armaments, for example, that are said to build up).

It is immediately obvious that 'semantic prosody' is a very exciting concept for the critical discourse analyst. If we can establish, on the basis of hard corpus evidence, what semantic company words regularly keep and what readers' or listeners' expectations are therefore likely to be, we are in a much better position to assess the particular semantic choices that are made in the text or texts under scrutiny. Pursuing these ideas is also a major theoretical challenge within critical discourse analysis. Drawing on corpus evidence fundamentally redefines the nature of 'interpretation', turning it from an introspective undertaking into an empirical one.

V. A Final Caveat

Although the techniques of corpus linguistics offer exciting possibilities for critical discourse analysis, three warnings are in order.

Firstly, the limitations of an untagged corpus and simple concordancing software must be clearly understood and taken account of at every stage. Phenomena that are beyond the reach of crude concordancing methods must not fall by the wayside. There is a danger, as Stubbs and Gerbig (1993: 78) also remind us, of 'counting only what is easy to count'. This is true, in particular, of many syntactic phenomena and of discoursal patterning. However, as I have demonstrated, we may sometimes find that these higher-level phenomena do have lexical reflexes and in such cases the concordancer can at least steer us towards those sites in the discourse where we can expect to find an instance of a particular argumentative pattern to occur.

Secondly, the level of generality which we claim for our results must be in strict proportion to the size and composition of the corpus investigated. A corpus of more than 160,000 words like the one I have been working on may seem large to the discourse analyst, but is clearly peanuts for the lexicographer or the grammarian who is used to dealing with corpora of millions of words. It is one of the home truths of corpus linguistics that massive corpora are indeed necessary to make any reliable statements about a language in general. The smaller the corpus, the more modest the claims must be. However, to a certain extent, a shortfall in words can be partly offset by internal homogeneity (in terms of genre, social, regional and historical variation) — provided that generalization beyond the genre and/or variety represented in the corpus is approached with great caution.

Thirdly, constant vigilance is in order so as not to fall into the trap of childish fascination with computers, numbers, and statistics. The very neatness of a condordance printout can be seductive. A *l'art pour l'art* attitude is easily indulged in, which can create an illusion of achievement when very little of substance has been revealed. At the same time, it is important to allow some room for 'playing around' with the data, because this can, and often does, generate useful ideas. In front of the keyboard and screen the right balance needs to be struck between unleashing and curbing the impulses of the *homo* or *femina ludens*. To prevent the computer's tail from wagging the analyst's dog, it pays to bear in mind that 'in the last analysis, the best machine for grinding general laws out of large collections of facts remains the same as Darwin's and Jespersen's — the human mind' (Svartvik 1992: 12).

VI. Summary

To recapitulate, a concordance program can contribute to qualitative analysis in the following ways:

Firstly, it allows the researcher to describe syntactic and semantic properties of key lexical items *exhaustively* rather than selectively. With the computer's

help, both in retrieving and displaying the data, the analyst can look at a large number of occurrences rather than generalise in an undisciplined fashion on the basis of a few purposely selected examples.

Secondly, it can function as a heuristic tool, raising questions to be followed up, and drawing analysts' attention to phenomena that they can then investigate with the help of their qualitative apparatus.

Thirdly, the concordancer produces 'results' in its own right. The frequency of a particular form, or the occurrence of certain collocates, may in itself be relevant from a critical perspective. However, quantitative evidence of any kind rarely speaks for itself, which is why care must be taken to put into perspective the bare facts the concordancer has provided (hence the inverted commas around 'results'). Even when the computer has entered the fray, triangulation remains a valuable methodological principle (cf. Miles and Huberman 1994: 266-267).

Fourthly — and this is the most 'mechanical', but none the less important, application — the concordancer is an extremely useful search tool, allowing the analyst to retain a much firmer grip on the corpus than would otherwise be possible. This is likely to pay off handsomely both in the research as well as in the writing-up stage.

To sum up, concordancing effectively heralds a breaking down of the quantitative/qualitative distinction, providing as it does the basis for quantitative analysis without 'deverbalising' the data, that is, without transferring it, through human intervention, to the numerical mode. The difference is, precisely, that between number crunching and word crunching. The latter leaves the co-text intact, while the former obliterates it.

To integrate the traditional qualitative analysis with the computer-aided component, I propose the following procedure:

a. On the one hand, the qualitative analysis of individual texts reveals 'loaded' items whose collocational behaviour (including its aura of meaning, or 'semantic prosody'; see Section IV.B.3) can then be investigated using the larger corpus held in the computer.

b. On the other hand, 'roaming' in the computerized corpus draws the analyst's attention to certain items or collocational patterns which can then also be studied qualitatively in their larger textual environments.

c. In addition, the findings resulting from both (a) and (b) can be compared with evidence from larger corpora such as newspapers on CD-ROM, the COBUILD corpus (Birmingham) or the BNC (Lancaster).

The idea is to move constantly between these different views of the data, rather than working in a 'quantitative' and a 'qualitative' compartment respectively. Let us briefly recall van Dijk's views on methodology in *News Analysis*. At the end of the passage quoted from earlier, he says: 'Generalizations from qualitative analysis must be based upon more intuitive knowledge of the data or upon convergence in small sets of analyses' (van Dijk 1988: 66). From

a mid-1990s perspective, with the potential of concordancing in mind, we are now in a position to argue that the analyst's intuition, though still an important tool, at last has a powerful and versatile ally on its side.

The availability of and access to computing infrastructure is improving rapidly and dramatically and it is getting progressively easier to tackle fairly ambitious projects. Even where human and financial resources are limited, however, a simple, off-the-peg concordance program can open up quite amazing vistas. A 'home-grown' corpus will invariably be much smaller than one that has been built as part of a major academic or commercial venture. Yet having a machine-readable corpus at all creates new ways of finding answers to what Biber and Finegan call 'Mount Everest questions — questions arising because the corpora are available but otherwise practically impossible to imagine' (Biber and Finegan 1991: 205).

Acknowledgements

For comments on the draft version of this paper I am indebted to Carlos M. Gouveia, Norman Fairclough, Geoffrey Leech, and Sari Pietikäinin.

References

Bell, A. (1991). *The Language of News Media*. Oxford/Cambridge, MA.: Basil Blackwell.

Biber, D. and Finegan, E. (1991). On the exploitation of computerized corpora in variation studies. In: K. Aijmer and B. Altenberg (eds), *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London and New York: Longman, pp. 204-220.

Burnard, L. (1992). Tools and techniques for computer- assisted text processing. In: C.S. Butler (ed.), *Computers and Written Texts*. Oxford and Cambridge, MA: Blackwell, pp. 1-28.

Caldas-Coulthard, C.R. (1993). From Discourse Analysis to Critical Discourse Analysis: The Differential Re-Presentation of Women and Men Speaking in Written News. In: J.M. Sinclair, M. Hoey and G. Fox (eds), *Techniques of Description. Spoken and Written Discourse. A Festschrift for Malcolm Coulthard*. London and New York: Routledge, pp. 196-208.

Cash, W. (1991). *Against a Federal Europe. The Battle for Britain*. London: Duckworth.

Cash, W. (1992). *Europe. The Crunch*. London: Duckworth.

Cash, W. (1993). A brave new Europe. In: S. Hill (ed.), *Visions of Europe. Summing up the Political Choices*. London: Duckworth, pp. 57-72.

Chafe, W. (1992). The importance of corpus linguistics to understanding the nature of language. In: J. Svartvik (ed.), *Directions in Corpus Linguistics. Pro-*

ceedings of the Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlin/New York: Mouton de Gruyter, pp. 79-97.

Diller, H.-J. (1994). Thatcher in Bruges. A study in Euro-rhetoric. *Journal for the Study of British Cultures* 1[2], 93-109.

Fairclough, N. (1989). *Language and Power.* London and New York: Longman.

Fairclough, N. (1992). *Discourse and Social Change.* Cambridge: Polity Press.

Fairclough, N. (1993). Critical discourse analysis and the marketization of public discourse: the universities. *Discourse and Society* 4[2], 133-168.

Fairclough, N. (1995a). *Critical Discourse Analysis. The Critical Study of Language*. London and New York: Longman.

Fairclough, N. (1995b). *Media Discourse*. London: Edward Arnold.

Fowler, R. (1991). *Language in the News. Discourse and Ideology in the Press*. London and New York: Routledge.

Fowler, R. and Kress, G. (1979). Critical linguistics. In: R. Fowler, B. Hodge, G. Kress and T. Trew (eds), *Language and Control.* London/Boston/Henley: Routledge and Kegan Paul, pp. 185-213.

Fox, G. (1993). A Comparison of 'Policespeak' and 'Normalspeak': A Preliminary Study. In: J.M. Sinclair, M. Hoey and G. Fox (eds), *Techniques of Description. Spoken and Written Discourse. A Festschrift for Malcolm Coulthard*. London and New York: Routledge, pp. 183-195.

Galtung, J. and Ruge, M. (1973). Structuring and selecting news. In: S. Cohen and J. Young (eds), *The Manufacture of News. Deviance, Social Problems and the Mass Media*. London, pp. 62-72.

Greenbaum, S. and Quirk, R. (1990). *A Student's Grammar of the English Language*. Harlow: Longman.

Hardt-Mautner, G. (1995). *Up Yours Delors*. Zur EG-Berichterstattung von The Sun aus diskursanalytischer Perspektive. In: G. Held (ed.), *Verbale Interaktion*. Hamburg: Verlag Dr. Kovacs, pp. 158-182.

Harrop, M. and Scammell, M. (1992). A tabloid war. In: D. Butler and D. Kavanagh (eds), *The British General Election of 1992*. London: Macmillan, pp. 180-210.

Hill, S. (1993). Introduction to the 13th nation. In: S. Hill (ed.), *Visions of Europe. Summing up the Political Choices*. London: Duckworth, pp. 1-10.

Hodge, R. and Kress, G. (1993). *Language as Ideology*[2]. London and New York: Routledge.

Johnson, D.M. (1994). Who is we? Constructing communities in US-Mexico border discourse. *Discourse and Society* 5[2], 207-231.

Kress, G. (1993). Against arbitrariness: the social production of the sign as a

foundational issue in critical discourse analysis. *Discourse and Society* 4[2], 169-191.

Leech, G. and Fligelstone, S. (1992). Computers and Corpus Analysis. In: C.S. Butler (ed.), *Computers and Written Texts*. Oxford/Cambridge, MA: Blackwell, pp. 115-140.

Leech, G. and Svartvik, J. (1994). *A Communicative Grammar of English*[2]. Harlow: Longman.

Leech, G., Myers, G., and Thomas, J. (eds). (1995). *Spoken English on Computer. Transcription, Mark-up and Application*. Harlow: Longman.

Liedtke, F., Wengeler, M., and Böke, K. (eds). (1991). *Begriffe besetzen. Strategien des Sprachgebrauchs in der Politik*. Opladen: Westdeutscher Verlag.

Louw, B. (1993). Irony in the text or insincerity in the writer? In: M. Baker, G. Francis and E. Tognini-Bonelli (eds), *Text and Technology. In Honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins, pp. 157-176.

McEnery, A. and Wilson, A. (forthcoming). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McNair, B. (1994). *News and Journalism in the UK*. London and New York: Routledge.

Miles, M.B. and Huberman, M.A. (1994). *Qualitative Data Analysis*. Thousand Oaks/London/New Delhi: Sage Publications.

Simpson, P. (1993). *Language, Ideology and Point of View*. London and New York: Routledge.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, M. (1992). Institutional Linguistics: Language and Institutions, Linguistics and Sociology. In: M. Pütz (ed.), *Thirty Years of Linguistic Evolution. Studies in Honour of René Dirven on the Occasion of his 60th Birthday*. Amsterdam: Benjamin.

Stubbs, M. and Gerbig, A. (1993). Human and Inhuman Geography: On the Computer-Assisted Analysis of Long Texts. In: M. Hoey (ed.), *Data, Description, Discourse. Papers on the English Language in honour of John McH Sinclair on his sixtieth birthday*. London: Harper Collins, pp. 64-85.

Svartvik, J. (1992). Corpus linguistics comes of age. In: J. Svartvik (ed.), *Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin/New York: Mouton de Gruyter, pp. 7-13.

Thatcher, M. (1988). *Text of the Prime Minister's Speech at Bruges on 20th September 1988*. London: Conservative Political Centre.

van Dijk, T.A. (1988). *News Analysis. Case Studies of International and National News in the Press*. Hillsdale, NJ: Lawrence Erlbaum Associates.

van Dijk, T.A. (1991). *Racism and the Press. Critical Studies in Racism and Migration*. London and New York: Routledge.

van Dijk, T.A. (1993). Principles of Critical Discourse Analysis. *Discourse and Society* 4[2], 249-283.

Wilson, J. (1990). *Politically Speaking. The Pragmatic Analysis of Political Language*. Oxford/Cambridge, MA: Blackwell.

Wodak, R. (1990). Discourse analysis: problems, findings, perspectives. *Text* 10[1/2], 125-132.

Wodak, R. et al. (1990). *Wir sind alle unschuldige Täter. Diskurshistorische Studien zum Nachkriegsantisemitismus*. (= Suhrkamp Taschenbuch Wissenschaft 881). Frankfurt am Main: Suhrkamp.

Appendix A. Corpus Profile

The overall corpus consisted of the newspaper coverage of events related to the EC/EU debate, from four daily papers and between 1971 and 1994. For the computer-aided part of the analysis only the leading articles were used, whereas for the qualitative analysis the entire coverage was taken into account.

Composition of the Corpus of Editorials

Newspaper titles	Number of leading articles	Number of words
The Daily Telegraph	107	56,169
The Guardian	121	78,379
The Sun	94	18,245
Daily Mirror	48	14,959
TOTAL	**370**	**167,752**

The imbalance in the sizes of the four sub-corpora was an inevitable consequence of the differences in the amount of editorial coverage devoted to the political topic under investigation.

Periods and Events Included in the Corpus

Periods used as the basis for data collection	Dates / Events
4 - 30 Oct. 71	4 - 8 Oct.: Labour Party Conference; 13 - 16 Oct: Conservative Party Conference; 21 - 28 Oct.: House of Commons debate on entry to EC.
20 - 24 Jan. 72	22 Jan.: Prime Minister Heath signs accession treaty
1 - 6 Jan. 73	First week of membership
26 May - 11 June 75	5 June: referendum on EC
12 - 15 March 79	13 March: official launch of European Monetary System
5 - 12 June 79	7 June: European Elections
27 Nov. - 4 Dec. 79	29 - 30 Nov.: EC summit in Dublin
11 - 19 June 84	14 June: European Elections
23 - 28 June 84	25 - 26 June: EC summit in Fontainebleau
2 - 5 Dec. 85	2 - 4 Dec.: EC summit in Luxembourg
18 Feb. 1986	17 Feb.: Single European Act is signed
1 - 2 July 87	1 July: Single European Act comes into force.
21 - 24 Sept. 88	20 Sept.: Margaret Thatcher's Bruges speech
12 - 29 June 89	15 June: European Elections; 26 - 27 June: EC summit in Madrid
26 Oct. - 10 Nov. 90	27 - 28 Oct.: EC summit in Rome; 1 Nov.: Geoffrey Howe resigns; The Sun starts its campaign against Jacques Delors.
12 - 18 Dec. 90	14 - 15 Dec.: EC summit in Rome
9 - 14 March 91	11 March: summit meeting of John Major and Helmut Kohl

27 - 29 June 91	28 - 29 June: EC summit in Luxembourg
2 - 14 Dec. 91	9 - 11 Dec.: EC summit at Maastricht
7 - 30 Sept. 92	Start of six-month UK presidency of European Council; 8 Sept.: speech by John Major; 16 Sept.: withdrawal of sterling from EMS; 20 Sept.: French referendum on Maastricht Treaty; 28 Sept.: meeting of EC finance ministers in Brussels.
1 - 31 Oct. 92	1 Oct.: conflict with Bundesbank; 5 - 9 Oct.: Conservative Party Conference in Brighton; 16 - 17 Oct.: EC summit in Birmingham.
2 - 7 Nov. 92	4 Nov.: Vote on Maastricht in House of Commons
13 - 24 April 1993	Scandal over European Bank for Reconstruction and Development; 16 April: Maastricht Bill in House of Commons; 20 April: delay of Euro Tunnel opening; 22 April: Hurd rejects referendum
19 July - 1 Aug. 1993	20 July: High Court ruling on judicial review of Maastricht; Maastricht Bill passes Third Reading in House of Lords and receives Royal Assent; 22 July: vote on social chapter in House of Commons; 23 July: Major faces vote of confidence over Maastricht; 26 July: leak of 'Bastards!' tape; 29 July: ERM crisis
22 - 26 March 1994	Dispute over blocking vote
Sept. - Dec. 1994	7 Sept: Major at Leiden (speech on Europe); 12 Oct: Conservative Party Conference (including speeches by Lamont and Portillo); 17 Nov.: Queen's Speech; 28 Nov.: Commons Debate and vote on European Finance Bill; 10 Dec.: EU summit in Essen; 12 Dec.: Jacques Delors announces he will not run for the French presidency; 13 Dec.: renewed debate on Euro-referendum; from 14 Dec.: dispute with Spain over fishing; 21 Dec.: House of Lords ruling concerning part-time workers.

Appendix B. Concordances for ‘Delors’

3. Discourse and Discourse Analysis in Corpus Linguistics In her criticism of sociolinguistics, Hasan (2004) emphasizes the importance of data driven research within the field that investigates the interrelations between the linguistic and the social. Only when the sociolinguistics allows â€œdata to speak to itâ€, it becomes obvious that language has to be viewed as meaning potential. Below I introduce the view of discourse in corpus linguistics which, being a strongly data driven approach, can not only be complementary for conducting discourse analysis in Applied Linguistics and CDA but also can The terms Critical Linguistics (CL) and Critical Discourse Analysis (CDA) have been frequently used interchangeably. Recently, however, the term CDA seems to have been preferred and is being used to denote the theory formerly identified as CL. Thus, I will continue to use CDA exclusively in this paper (see Anthonissen 2001 for an extensive discussion of these terms). The roots of CDA lie in classical Rhetoric, Text linguistics and Sociolinguistics, as well as in Applied Linguistics and Pragmatics (see also Wodak & Meyer 2009; Fairclough 2003; Wodak 2004, 2007; Renkema 2004; Blommaert 2005) Discourse analysis is used to study language in social context. It focuses on the purposes and effects of written and spoken communication.Â Unlike linguistic approaches that focus only on the rules of language use, discourse analysis emphasizes the contextual meaning of language. It focuses on the social aspects of communication and the ways people use language to achieve specific effects (e.g. to build trust, to create doubt, to evoke emotions, or to manage conflict).