HISTORICAL PERSPECTIVE: WHAT ARE CORPORA AND HOW HAVE THEY EVOLVED?

Michael McCarthy and Anne O'Keeffe

1. THE HISTORICAL ORIGINS

Corpus linguistics nowadays is perhaps most readily associated in the minds of linguists with searching through screen after screen of concordance lines and wordlists generated by computer software, in an attempt to make sense of phenomena in big texts or big collections of smaller texts. This method of exegesis based on detailed searches for words and phrases in multiple contexts across large amounts of text can be traced back to the 13th century, when biblical scholars and their teams of minions pored over page after page of the Christian Bible and manually indexed its words, line by line, page by page. Concordancing arose out of a practical need to specify for other biblical scholars, in alphabetical arrangement, the words contained in the bible, along with citations of where and in what passages they occurred.

The etymology of *concordantia* is the Latin *cum* meaning *with* and *cor* meaning *heart*, which ties in with the original ideological underpinning of this painstaking endeavour, namely to underscore the claim that the Bible was a harmonious divine message rather than a series of texts from a multitude of sources. Anthony of Padua, (1195-1231) is associated with the first known (anonymous) concordance of the bible, the

Concordantiae Morales, based on the Vulgate (the fifth century Latin version of the bible). A well-documented work around the same time was by Cardinal Hugo of St Caro (also referred to as St Cher), who in 1230, aided by a 500-strong team of Dominican monks at St James' convent in Paris, put together 'a word index' of the fifth century Vulgate (Bromiley 1997:757, see also Tribble this volume). Since then numerous other concordances of the bible have evolved, including Cruden's 1737 *A complete concordance to the holy scriptures* and Strong's 1890 *Exhaustive concordance of the Bible.* Nowadays, computer concordancing programmes replicate the work of 500 monks in micro-seconds.

The works of Shakespeare were also the subject of concordancing as a means of assisting scholars, for example Becket's 1787 *A concordance to Shakespeare*. As Tribble (this volume) illustrates by way of extract from Becket's concordance, the word and its linguistic context and location in the Shakespeare canon is given. For a literary scholar, this provides an immense resource. Though concordances from former times were laboriously compiled by hand, their spirit and intentions live on in the software programs we are now familiar with.

2. WHAT DROVE THE CREATION OF MODERN CORPORA?

While the process of concordancing and indexing has its origins in the painstaking work of biblical and literary scholars, the drive to create electronic corpora did not come from these quarters entirely. There was an influence from the work of Jesuit priest Roberto

Busa who created a electronic lemmatised index of the complete works of Saint Tomas Aquinas, *Index Thomisticus,* beginning in the 1950s and completing it in the late 1970s (see Tognini Bonelli, this volume). At least two other forces are more significant, namely the work of lexicographers and that of pre-Chomskyan structural linguists. In both cases, collecting attested data was essential to their work. Dr Samuel Johnson's first comprehensive dictionary of English, published in 1755, was the result of many years of working with a paper corpus, that is endless slips of paper logging samples of usage from the period 1560 to 1660. And perhaps the most famous example of the 'corpus on slips of paper' were the more than three million slips attesting word usage that the Oxford English Dictionary (OED) project had amassed by the 1880s, stored in what nowadays might serve as a garden shed. These millions of bits of paper were, quite literally, pigeon-holed in an attempt to organise them into a meaningful body of text from which the world-famous dictionary could be compiled.

As Leech (1992) points out, it was in the 1950s, in the era of American structuralists such as Harris, Fries and Hill, among others, when the notion of collecting real data came into its own. Where the work of the early biblical and literary scholars provides the background modus operandi of word searching and indexing, the structuralists were the forerunners of corpora not only in the sense of data gathering but in terms of the commitment to putting real language data at the core of what linguists study.

It is perhaps worth mentioning too that interest in first language acquisition based on transcribed data goes back a considerable way, with the earliest transcripts in the CHILDES Language Database dating back to the 1960s, even though the project was only formally established in 1984 (see its website). We should not forget, either, that literary scholars have for decades supplied useful concordances of the works of major authors. Already by 1979, Howard-Hill saw computer-generated concordances as a "general-purpose working tool for the study of literature" (Howard-Hill 1979:30). At least eight concordances of works by Conrad were published between 1979 and 1985, thanks to scholars such as Bender and Higdon, while other concordances for writers such as Gerard Manley Hopkins and T.S. Eliot were published around the same time.

The first computer-generated concordances had appeared in the late 1950s, using punched-card technology for storage (see Parrish 1962 for an early discussion of the issues). At that time, the processing of some 60,000 words took more than 24 hours. However, considerable improvements came about in the 1970s. Meanwhile, from as early as 1970, library and information scientists had developed a keen interest in Key Word In Context (KWIC) concordances as a way of replacing catalogue indexing cards and of automating subject analysis (Hines et al 1970), and many well-known bibliographies and citation source works benefitted from advances in computer technology. Such work was going on when the concerns of many of the contributors to the present volume were unarticulated and hardly conceived as jobs for the computer. It was the 1980s and 1990s which really saw the arrival of corpora as we know them now as tools for the linguist or applied linguist.

Before it found its way into the linguistic terminology, the term *corpus* had long been in use to refer to a collection or binding together of written works of a similar nature. The OED attests its use in this meaning in the 18th century, such that scholars might refer to a 'corpus of the Latin poets', or a 'corpus of the law'. The OED's first citation of the word *corpus* in the linguistic literature is dated at 1956, in an article by W. S. Allen in the *Transactions of the Philological Society*, where it is used in the more familiar meaning of "the body of written or spoken material upon which a linguistic analysis is based" (OED: 2nd Edition, 2009). McEnery, Xiao and Tono (2006) note that the more specific term *corpus linguistics* did not come into common usage until the early 1980s; Aarts and Meijs (1984) is seen as the defining publication as regards coinage of the term.

3. THE INFLUENCE OF TECHNOLOGY: FROM MAIN FRAME TO MODEM TO MULTI-MODALITY

By the time computers came to be usable by anyone other than a tiny group of specialists, the traditions of (a) trawling through texts to find all examples of a particular piece of language, (b) writing dictionaries based on attested usage, and (c) analysing language based on actual informant data were all well-established. It was the revolution in hardware and software in the 1980s and 1990s which really allowed corpus linguistics as we know it to emerge. For a start, the assumption that any large-scale computing required

a huge mainframe computer was to be challenged by the seemingly unstoppable increases in desktop computing power in the 1990s, enabling small teams and individuals to take on quite ambitious corpus projects. The parallel growth of the Internet and fast download speeds meant that data and results could be transferred easily from scholar to scholar, while the role of the clumsy text scanners of the early 1980s – some as big as household chest-freezers – could be replaced by instant access to vast quantities of text already in electronic form. In tandem, heavy and cumbersome reel-to-reel tape recorders were replaced by manageable analogue cassette recorders in the 1970s and later by miniature digital recorders and small but high-powered video and DVD recorders, with a consequent positive effect on the ability of scholars to create spoken corpora.

However, the first efforts of linguists to harness computational power to study language as evidenced in large volumes of text were hampered by the limitations of machines. Sinclair, for instance, in his earliest exploratory years of corpus analysis that were to culminate in the ground-breaking COBUILD project, used cumbersome punched-card systems for data-storage, a method which, in its most basic form, could be dated back to the 18th century! And many corpus linguists of the 'second generation' of computing will recall the unforgiving nature of early DOS-based proprietary software such as the *Oxford Concordance Program* (OUP 1987) popular in the late 1980s and early1990s, where the smallest error in writing the required string of commands could result in the hair-tearing frustration of a broken search. Such frustrations seemed to vanish forever with the advent of user-friendly GUI-based software suites such as Scott's *Wordsmith Tools* (1996-) and Barlow's *Monoconc* (1996-), which, along with other

programs mentioned by the authors in the present volume, have become the natural tools of today's applied linguists, powerful, easy to use and more than up to the tasks that researchers demand of them.

4. CORPUS DEVELOPMENTS: FROM MEGA-CORPUS TO MINI-CORPUS AND FROM MONO TO MULTI-MODAL

Technology has been the major enabling factor in the growth of corpus linguistics but has both shaped and been shaped by it. The ability to store masses of data on relatively small computer drives and servers meant that corpora could be as big as one wanted. In this regard, lexicographers led the way. Their aim has always been to collect the maximum amount of data possible, so as to capture even the rare events in a language. The early COBUILD corpora were measured in tens of millions of running words, other publishing projects soon competed and pushed the game up to hundreds of millions of words and, by the middle of the first decade of the 21st century, the Cambridge International Corpus (Cambridge University Press) had topped a billion running words of text. Very soon, researchers began to realise the potential of the entire world-wide-web as a corpus, with its trillions of words, a veritable treasure-trove of linguistic phenomena accessible at the click of a mouse (see Lee, this volume on the potential of the world-wide-web as a corpus).

However, precisely because of the ease with which data can be assembled and stored, the reverse of the coin of ever-bigger corpora has also manifested itself. Small, carefully targeted corpora (by which we commonly mean corpora of fewer than a million words of running text), have proved to be a powerful tool for the investigation of special uses of language, where the linguist can 'drill down' into the data in immense detail using a full armoury of software and shed light on particular uses of language. Several of the chapters in this volume report on relatively small corpus projects which have yielded invaluable information for their compilers (see chapters by Clancy, Evison, Farr, Koester, McIntyre and Walker, Thornbury, Vaughan, among others).

Technology also enabled the creation of multi-modal corpora, in which various communicative modes (e.g. speech, body-language, writing) could all be part of the corpus, all linked by simple technologies such as time-stamping and all accessible at one go. No longer did the spoken corpus linguist have to rely only on the transcript of a speech event; now there was the evidence of a video and audio stream tied to the transcript offering invaluable contextual and para-linguistic and extra-linguistic support to the analysis (see Adolphs and Knight, this volume).

Equally, linguists have had a role in shaping the technology in ways best-suited to their needs. Statistical operations such as Mutual Information scores were seen as ways of getting at the elusiveness of collocation, while benchmark statistical comparisons could be harnessed to tease out the significant 'fingerprints' of specialised uses of language (manifest in the Key Word function of Scott's *Wordsmith Tools*, for example, see Scott,

this volume). Such capabilities are not inherent in the computer's architecture and require the vision of linguists and applied linguists to see the potential for translating various types of counting operations that the computer can carry out into linguistically useful forms of informational output. More recently, Smith et al (2008) have drawn up *desiderata* from the linguist's point of view for the ongoing design of corpus tools which might better reflect linguists' needs for annotation and analysis.

5. THE MANY APPLICATIONS OF CORPUS LINGUISTICS

Corpus Linguistics (CL), for many, is an end it itself. That is, it provides a means for the empirical analysis of language and in so doing adds to its definition and description. This process has led to the refinement of our descriptions of lexis, leading to immensely enhanced coverage in dictionaries (as discussed above) and we have seen a proliferation of empirical studies about aspects of grammar (often in fine detail), as well as large-scale corpus-based reference grammars such as Biber et al (1999) and Carter and McCarthy (2006). Increasingly however CL is being used in the pursuit of broader research questions, that is, in areas such as language teaching and learning, discourse analysis, literary stylistics, forensic linguistics, pragmatics, speech technology, sociolinguistics and health communication, among others. As this volume testifies, CL has had much to offer other areas by providing a better *means* of doing things. In this sense, CL is a means to an end rather than an end in itself. That is, CL leads to insights beyond the realms of lexis or grammar by applying its techniques to other questions, some more easily answered by

computational analysis than others. In areas as diverse as second language acquisition and media studies, CL can be applied as a research tool.

In this volume, we have tried to bring together as diverse as possible a sample of the applications of CL so as to capture the state-of-the-art in terms of its how CL is being applied and might be applied in the future. Crucially for the development and vibrancy of CL, this process of application of CL to other areas has a *wash-back effect* for CL and in particular on how corpora and corpus software are designed, as we asserted above. As mentioned (see also Walter, this volume), the initial application of CL in our profession was in the area of lexicography, and software and corpora were co-designed so that lexicographers could make better dictionaries. Now the application of CL is diverse in the extreme, as are the needs of its users. While a lexicographer is interested in how best to profile a word semantically (see chapters by Walter and Moon, this volume), someone using CL in the study of second language acquisition may be interested in how aspects of language develop over time in one individual or a group of users (see Lu, this volume). These polar needs bring about divergent corpora and software design principles. The result is that there has never been a more fertile period in the discipline of CL. We now briefly survey some of the areas in which corpora have been adopted and audit the challenges and wash-backs that arise from these.

Language Teaching and Learning

Individuals such as Johns and Tribble have, for many years, championed the use of corpora in language learning in the form of Data-Driven Learning (DDL) (see chapters

by Tribble, Chambers and Sripicharn, this volume). Bringing corpora or corpus data into the classroom has brought many challenges over the years. By its nature, it turns the traditional order within the classroom upon its head. The corpus becomes the centre of knowledge, the students take on the role of questioner and the teacher is challenged to hand over control and facilitate learning. Chambers and O'Sullivan (2004) have shown the democratising effect of devolving the correction and remediation of student writing through the use of error tagging and follow-up student corpus investigation, for example.

As discussed in Chambers', Sripicharn's and Tribble's chapters in this volume, the teacher has to do a lot of preparation work in building up students' skills of investigation leading to hands on work with corpora or concordance print outs (see also Allan 2008). Reading a set of KWIC concordance lines, the key skill in DDL, is not something which can be assumed to be automatic. It demands the reader to abstract meaning through vertical reading of the node(s), and often through both left-to-right and right-to-left reading relative to a node on the concordance, and initially at the level of fragmented text (see chapters by Hunston and Tribble, this volume). It demands new micro-cognitive skills whereby the reader moves from phrase pattern to meaning by way of hypothesising and inference. This is a wash-back effect which has still to be properly addressed in DDL.

Another area of innovation within pedagogical applied linguistics which is directly related to CL is the development of learner corpora, that is, collections of spoken and written learner language. The work of Granger and her associates leads the way in this field (see Gilquin and Granger, this volume). This moves the focus of the corpus

from native speaker dominance. It brings the language of the learner into focus and allows, at a classroom level, a body of language which learners can both create and work with. Another step away from the monolithic native speaker corpus model has been the development of corpora of expert users such as the HKSCE (Cheng et al 2005) and the VOICE corpus (Seidlhofer 2004). These developments, along with the work of Granger et al, have challenged the notion of the corpus as a model of Standard English (or other language). The *English Profile* project (see its website), set up to provide empirical underpinning for the descriptions of the various levels of the Common European Framework of Reference (CEFR), also deals in learner data, such that the proficiency levels need not be defined solely in terms of the (usually unattainable) performances of native speakers. The ideological wash-backs of learner corpora have yet to be felt in their full force, but there is no doubt that CL has enabled researchers to ask new questions within new paradigms.

Other areas within pedagogical applied linguistics where we are seeing rapid development in the application and development of corpora include testing and teacher education. For both of these areas, the use of corpora can add to professionalisation in differing ways. The use of corpora in the area of testing, as detailed in Barker (this volume) can shed empirical light on issues of key standards and rating, manifested again in the research of the *English Profile* project. The project offers a core empirical framework upon which to base and score exams internationally, as well as potentially leading to new benchmarks for the design of teaching materials and curricula.

Professionalisation of the area of Language Teacher Education (LTE) through the use of corpora for reflective practice has been championed by Farr, and her chapter in this volume gives numerous insights into how CL can aid practice and professional development. A wash-back implication, in this area of application, is the need to make CL a core part of LTE programmes (see O'Keeffe and Farr 2003; McCarthy 2008).

Though it has been a slow process, more and more language teaching materials are now 'corpus-informed'. Increasingly, publishers are investing more in developing corpora, for example, major publishers such as Cambridge University Press, Oxford University Press, Pearson-Longman, Collins-COBUILD and Macmillan all closely guard multi-million word corpora and regularly launch new materials which are corpus-informed. The splenetic debates that raged in the pages of applied linguistics journals in the 1990s seem to have quelled to an acceptance that *corpus-informed* is not a bad or dangerous term (see Prodromou 1996, 1997a, 1997b; Owen 1996; Bernardi 2000; Widdowson 1991, 2000 Carter and McCarthy (1995), Aston (1995), Prodromou (1996), Owen (1996), Carter (1998), Cook (1998), Seidlhofer (1999), Sinclair (1991a, 1991b). The long running debates of the 1990s may have had a very positive spin off for CL in that more applied linguists and especially practising teachers became aware of corpora and wanted to learn more. More and more papers were presented at major conferences on the uses of corpora in language teaching. However, there still exists a gulf between the world of corpus linguistics and the everyday language teacher. As stressed by O'Keeffe et al (2007), more corpus linguists need to engage with applied linguists and language teachers, and vice-versa. Much of the purely descriptive research conducted by corpus

linguists into language use (that is, as an end it itself), would be of immense value to language teachers and materials designers if more widely disseminated. If CL is to have an optimum impact for language learners, this process of engagement between CL and pedagogical applied linguistics needs to be improved. In this volume, we include the work of many corpus linguists who are also language teachers and materials designers in an attempt to showcase the benefits of the synergy between CL and AL (see chapters by Chambers, Cheng, Conrad, Flowerdew, Hughes, Handford, Jones and Durrant, Thornbury, McCarten, Guilquin and Granger, Scripicharn, Vaughan, Walsh, among others).

Discourse analysis

Analysing discourse is another area where CL has been adopted as a means of looking at language patterns over much larger datasets. Existing models for above-sentence analysis such as Conversation Analysis (CA), Discourse Analysis (DA) and Critical Discourse Analysis (CDA) are all benefiting from the use of CL (see Thornbury, this volume, as well as chapters by Evison, O'Halloran and Walsh). CL can automate many (but certainly not all) of the processes of CA, DA and CDA through the use of wordlists, concordances and key word searches (see Evison, this volume). The process is not one-way however. CL on its own is not the basis for the analysis of discourse. It can provide the means for analysis but researchers invariably draw on theories and applications of either CA, DA or CDA. One example is the use of the CA notion of 'baseline', that is whereby the turn structure of an interaction, for example a telephone call opening, is compared to the 'canonical' or baseline interaction between 'unmarked'

interactants. For example, O'Keeffe (2006) compared the turn sequence of an opening of a call to a radio station with the canonical sequence of a call between people who are neither strangers nor intimately related (see Sacks et al, 1974). In the same way, CL uses 'reference corpora' against which results are compared (see Evison, this volume for an example of this).

Literary studies and translation studies

Comparison is also a key concern in the study of literature, poetry and drama. Burrows (2002) has noted that traditional and computational forms of stylistics have much in common in that they both involve the close analysis of texts and benefit from opportunities for comparison (see also Wynne 2005). The application of corpora to the study of literature, poetry and drama is surveyed in chapters by McIntyre and Walker and Amador Moreno. McIntyre and Walker show the application of *Wmatrix,* a software tool which greatly facilitates the comparison of texts. *Wmatrix*, in this case, is used to compare two volumes of poetry by William Blake as well as the texts of 12 blockbuster movie scripts. A function of the software which is illustrated very well in the chapter is its ability to assign semantic categories to key words in the corpora which are being compared. *Wmatrix* was developed to assign semantic tags by matching the text against a computer dictionary of semantic domains (see Rayson et al. 2004 for details of this procedure). This means that both key words and key semantic domains can be compared (see Rayson 2004, 2008). This offers immense scope for the automated study of stylistics (see Wynne 2005). Amador Moreno (this volume) gives an illustration of the usefulness of CL in analysis a whole novel. Because the novel is written in the first person, in Irish

English, she is able to draw on a one-million word corpus of the same variety (the Limerick Corpus of Irish English) as a reference for comparison.

Another area which has driven CL from outside has been that of translation. CL has much to offer this area in terms of aiding automatically the comparison of patterns across languages by comparing source and target texts. The constant need to better the tools of the trade has led to numerous innovations in corpus and software design. The challenge of how to align texts and their translations is discussed and illustrated in chapters by Kübler and Aston, and Kenning (this volume).

Forensic linguistics

Increasingly, linguists are being consulted within the legal sector to authenticate authorship. A number of case scenarios are provided in Cotterill (this volume). The corpus linguist is turned into an expert witness in the courtroom. This brings the challenge of communicating findings to a non-linguist audience. The adaptation of CL to this area is interesting to survey from the perspective of how CL is used or viewed. As Cotterill (this volume) notes " …forensic linguists tend to refer to [CL] as a tool or a resource since no method of analysis, corpus or otherwise, can guarantee the identification or elimination of authors". Clearly, CL, for forensic linguists, is a means to a very real end. In terms of wash-back effect, forensic linguists have added to the area of CL through their need to show succinctly and statistically how one or more texts contain features or patterns of typicality which prove beyond reasonable doubt that they were or were not written by the same author. This is referred to in terms such as *uniqueness* and

genuineness (cf. the seminal work of Coulthard 2004). The power of CL again here is its ability to automatically compare on a grand scale so as to corroborate evidence (or not) of uniqueness or genuineness in a text or texts. Cotterill (this volume) raises the important issue of whether forensic linguists can be called scientific (which ultimately washes back to the question as to whether CL can be called scientific). In the US court system, as Cotterill explains, scientific evidence, to be admissible, has to: 1. have a theory which has been tested; 2. have been subjected to peer review and publication; 3. have a known rate of error; and, 4. have a theory which is generally accepted in the scientific community (see Solan and Tiersma 2004 for a detailed discussion).

Pragmatics

Pragmatics is the study of language in use and so CL seems a logical ally to the field. However, much of the work in the area of pragmatics draws on elicited data from role-plays, interviews and Discourse Completion Tasks (DCTs), and early classic pragmatic studies relied on intuited data. The application of CL to this area has been slow and there are good reasons for this. Not least of all, there are relatively few corpora of spoken language (the main site for the study of pragmatics in use) and corpora are not designed with the study of pragmatics in mind. Pragmatic features such as speech acts, politeness, hedges, boosters, vague language and so on, are not automatically retrievable from a corpus. Rühlemann (this volume) discusses the many challenges for those interested in using a corpus to study pragmatics. Nonetheless, there are a number of insightful pragmatic studies which have used CL very successfully (see Rühlemann, this

volume). Schauer and Adolphs (2006) show how CL can work in tandem with existing methods, in their case, DCTs.

Many individual pragmatic features have been studied using CL. Pragmatic markers, including deictics, hedges, discourse markers, boosters, markers of shared knowledge (see Carter and McCarthy 2006) have been studied in both spoken and written contexts using corpora. Interestingly, a very fertile area has been the use of corpora to compare pragmatic features across different languages: Aijmer and Simon-Vandenbergen (2006) brings together chapters on pragmatic markers across a number of languages. Lewis (2006) examines adversative relational markers in French and English. Stenström (2006) explores Spanish pragmatic markers *o sea* and *pues* and their English equivalents while Downing (2006) looks at *surely* and its Spanish counterpart and Johansson looks at *well* and its equivalents in Norwegian and German.

Other areas which have amassed a considerable number of CL-based studies include hedging and politeness, vague language, irony, humour, hyperbole (McCarthy and Carter 2004), metaphor (Deignan 2005), deixis and modality, among others. Clearly, the strength corpus linguistics brings to the study of pragmatics is its power to automatically search for and retrieve particular items. Unfortunately this does not extend to all aspects of pragmatics. The wash-back effect from pragmatics has been the push for better capture and tagging of spoken language, in particular, the innovations in the area of muli-modal corpora have sprung from this demand.

Sociolinguistics, media discourse and political discourse

The interest in non-formal features of language provides a natural territory of expansion for CL into sociolinguistics and other areas of language in society such as media discourse and political discourse. Sociolinguistics is quintessentially concerned with language users, and here the question of metadata clearly raises itself in CL. It is not sufficient for a sociolinguist to work with a purely textual transcript; vital information about speakers such as age, gender, educational background, geographical origin etc. become integral features of the corpus-analytical process (see chapters by Andersen and Clancy, this volume). The wash-back on corpus design is most obviously in the kinds of metadata that must be gathered at the time of data collection, leading to elaborate questionnaire or interview demands on informants and a slew of new ethical considerations about data protection and privacy. These problems apart, there have been a number of successful corpus projects with a sociolinguistic motivation (e.g. the COLT corpus of London teenager language), as well as creative ways of using the existing demographic and morpho-syntactic information in corpora such as the BNC (see Andersen, this volume) and other tagged and heavily annotated resources. Detailed annotation and the ability to access and filter metadata are all-important in sociolinguistic versions of CL, and the wash-back effects on software design and use are already apparent.

The study of media discourse has as its natural (but not exclusive) ally critical discourse analysis (CDA). CDA attempts to expose the ideologies which inform and

underlie texts, and media texts are clearly a rich source for critical analysts. Benchmark analyses between media corpora and other, non-media corpora (where terms occurring with statistically significant frequency in particular media texts can be listed) can be used to focus on language choices which may be ideologically motivated. O'Halloran's chapter in this volume provides a discussion and examples, and looks further at the investigation of culturally significant key words using CL techniques. CDA has not been without its critics (see O'Halloran's chapter for a summary); the exploitation of CL and future refinements may make the case for CDA stronger by providing empirical evidence from sources such as corpora of media texts. In the same breath one might include the concerns of researchers into political discourse, where CL studies of language in contexts such as political speeches and parliamentary debates, as well as political news coverage, lead corpus linguists into areas such as key word analysis and comparisons across corpora. Ädel (this volume) provides extensive coverage of the field and its preoccupations.

All in all, CL can be argued to be a healthy vibrant discipline within the general umbrella of language study. Its origins were non-computational but its explosion and expansion in the fields of descriptive and applied linguistics are due mainly to the information revolution of the late 20th century, a revolution which continues, and from which CL will undoubtedly continue to benefit. In this handbook we have tried to capture the variety, the fluidity, the momentum and vision of CL as it exists at the time of publication, and to assess its contributions and applications within our several professions which all have language in common. The contributors to this volume are representatives of a large and

growing community of academics and professionals who have designed, used, adapted and applied corpora and associated software. Individually, their interests differ greatly; collectively, we hope that a single image, however grainy, will emerge to illustrate this fascinating field.

REFERENCES

Aijmer, K. and Simon-Vandenbergen, A.-M. (eds) (2006) *Pragmatic Markers in Contrast.* Amsterdam: Elsevier.

Allan R. (2008) "Can a Graded Reader Corpus Provide 'Authentic' Input?," *ELT Journal* 63: 23-32.

Bernardi, S. (2000) *Competence, Capacity, Corpus.* Bologna: CLUEB.

Burrows, J. (2002). "The Englishing of Juvenal: Computational Stylistics and Translated Texts," *Style* 36 (4): 677-679.

Carter, R. A. (1998) "Orders of Reality: CANCODE, Communication and Culture," *ELTJ* 52: 43-56.

Carter, R. A. and McCarthy, M. J. (1995) "Grammar and the Spoken Language," *Applied Linguistics* 16 (2): 141-58.

Carter, R. A. and McCarthy, M. J. (2006) *Cambridge Grammar of English.* Cambridge: Cambridge University Press.

Chambers, A. and O'Sullivan, Í. (2004) "Corpus Consultation and Advanced Learners' Writing Skills in French," ReCALL 16(1): 158–172.

Cheng, W., Greaves, C. and Warren, M. (2005) "The Creation of a Prosodically Transcribed Intercultural Corpus: The Hong Kong Corpus of Spoken English (prosodic)," *ICAME*, 29: 47-68.

Cook, G. (1998) "The Uses of Reality: A Reply to Ronald Carter," *ELT Journal* 52: 57-63.

Coulthard, M. (2004) "Author identification, Idiolect, and Linguistic Uniqueness," *Applied Linguistics* 25 (4): 431-447.

Deignan, A. (2005) *Metaphor and Corpus Linguistics*. Amsterdam: John Benjamins.

Downing, A. (2006) "The English Pragmatic Marker *surely* and its Functional Counterparts in Spanish," in K. Aijmer and A.-M. Simon-Vandenbergen

(eds), *Pragmatic Markers in Contrast.* Amsterdam: Elsevier, pp. 39-58.

Hines T. C., Harris J. L. and Levy, C. L. (1970) "An Experimental Concordance Program," *Computers and the Humanities* 4(3): 161-171.

Howard-Hill, T. H. (1979) *British Bibliography and Textual Criticism: A Bibliography.* Oxford: Clarendon Press.

Lewis, D. M. (2006) "Contrastive Analysis of Adversative Relational Markers using Comparable Corpora," in K. Aijmer and A.-M. Simon-Vandenbergen (eds) *Pragmatic Markers in Contrast*, Oxford: Elsevier, pp. 139-153.

McCarthy, M. J. (2008) "Accessing and Interpreting Corpus Information in the Teacher Education Context," *Language Teaching* 41(4): 563-574.

McCarthy, M. J. and Carter, R. A. (2004) "'There's millions of them': hyperbole in everyday conversation," *Journal of Pragmatics* 36 (2): 149-184.

O'Keeffe, A. (2006) *Investigating Media Discourse.* London: Routledge.

Owen, C. (1996) "Do concordances need to be consulted?," *ELT Journal* 50 (3): 219-224.

Parrish, S. M. (1962) "Problems in the Making of Computer Concordances," *Studies in Bibliography* 15: 1-14.

Prodromou, L. (1996) "Correspondence," *ELT Journal* 50 (4): 371-373.

Prodromou, L. (1997a) "Corpora: The Real Thing?," *English Teaching Professional* 5: 2-6.

Prodromou, L. (1997b) "From Corpus to Octopus," *IATEFL Newsletter* 137: 18-21.

Rayson, P., Archer, D., Piao, S. and McEnery, T. (2004) "The UCREL Semantic Analysis System," in *Proceedings of the Workshop on Beyond Named Entity Recognition Semantic Labelling for NLP Tasks, in association with the 4th International Conference on Language Resources and Evaluation* (LREC 2004), Lisbon: Portugal, pp. 7-12.

Rayson, P. (2008) *Wmatrix: A Web-based Corpus Processing Environment*, Computing Department, Lancaster University. (Available at: http://ucrel.lancs.ac.uk/wmatrix/)

Seildhofer, B. (1999) "Double Standards: Teacher Education in the Expanding Circle" *World Englishes* 18: 233-45.

Sacks H., Schegloff, E.A., Jefferson, G. (1974) "A Simplest Systematics for the Organisation of Turn-Taking for Conversation," *Language* 50(4): 696-735.

Schauer, G. A. and Adolphs, S. (2006) "Expressions of Gratitude in Corpus and DCT Data: Vocabulary, Formulaic Sequences, and Pedagogy," *System* 34(1): 119-134.

Seidlhofer, B. (2004) "Research Perspectives on Teaching English as a Lingua Franca," *Annual Review of Applied Linguistics* 24: 209-239.

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

Sinclair, J. (1991b) "Shared Knowledge," in J. Alatis, (ed) *Georgetown University Round Table on Languages and Linguistics*. Washington, D.C.: Georgetown University Press, pp. 489-500.

Smith, N., Hoffmann, S. and Rayson, P. (2008) "Corpus tools and methods, today and tomorrow: incorporating linguists' manual annotations," *Literary and Linguistic Computing* 23 (2): 163 - 180.

Solan, L. and Tiersma, P. (2004) "Author Identification in American Courts," *Applied Linguistics* 25(4): 448-465.

Stenström A. B. (2006) "The Spanish Discourse Markers *o sea* and *pues* and their English Correspondences," in K. Aijmer and A.-M. Simon-Vandenbergen (eds), *Pragmatic Markers in Contrast.* Amsterdam: Elsevier, pp. 155-172.

Widdowson, H. G. (1991) "The Description and Prescription of Language," in Alatis, J. (ed) *Georgetown University Round Table on Languages and Linguistics* Washington, D.C.: Georgetown University Press, pp. 11-24.

Widdowson, H. G. (2000) "On the Limitations of Applied Linguistics," *Applied Linguistics* 21(1): 2-25.

Wynne, M. (2005) "Stylistics: Corpus Approaches," in K. Brown, (ed) *Encyclopaedia of Language and Linguistics*. Oxford: Elsevier. Available at: http://eprints.ouls.ox.ac.uk/archive/00001003/01/Corpora_and_stylistics.pdf