This volume is made up of 17 chapters which have developed out of papers and workshop sessions presented at the event entitled “Corpora: Seminar and Workshops”, held at the University of Padua, March 29-31, 2007. It maintains the straightforward, practical approach which characterized that event, meant as an introduction to the use of corpora even for novices. At the same time it goes into a wide range of different applications for corpora in language teaching and language research in higher education. One of these involves the creation and use of learner corpora. Another application involves corpus-assisted research into political discourse in the media. Language for special purposes is also focussed on as a research topic, an academic discipline, and language to be translated. Multimodal corpora are also considered. Proposals are made for corpus-based research into the language of films, and into translation (and mediation) universals. A corpus-based study of text complexity in reading tests is also presented. Large-scale corpora commercially available are also discussed. An online module for translator training is presented, as is an Internet-accessible corpus of Old English poetry.

The three editors are tenured academics teaching English in the degree courses in languages at the University of Padua, Italy. Carol Taylor Torsello is full professor while Katherine Ackerley and Erik Castello are researchers. They share an interest in corpus linguistics, systemic functional linguistics, discourse analysis, language learning with new technologies, and language testing.
Contents

CAROL TAYLOR TORSELLO
Dedication to John Sinclair.................................................................11

GUY ASTON
Remembering John Sinclair..............................................................13


CAROL TAYLOR TORSELLO, KATHERINE ACKERLEY AND ERIK CASTELLO
Corpora for University Language Teachers: An Introduction.........23

Part I
Getting started

ELENA TOGNINI BONELLI
Corpora and LSP: Issues and Implications...........................................31

MARIA TERESA PRAT ZAGREBELSKY
Learner Corpora at the Crossroads of Computer Corpus Linguistics, Foreign Language Pedagogy and Second Language Acquisition Research.................................................................49

MARIA TERESA MUSACCHIO AND GIUSEPPE PALUMBO
Shades of Grey: A Corpus-driven Analysis of LSP Phraseology for Translation Purposes .................................................................69

FRANCESCA COCCETTA
Multimodal Corpora with MCA .......................................................81
Part II
Ideas and Suggestions for Corpus Work

ALAN PARTINGTON
The Armchair and the Machine: Corpus-assisted Discourse Research .......................................................... 95

CAROLINE CLARK
A CADS Analysis of Television Reports from Iraq: Were Embeds ‘in Bed’ with the Coalition? ........................................... 119

SARA GESUATO
Linguistic Research with Large-scale Corpora .................................................. 129

MARGHERITA ULRYCH AND AMANDA MURPHY
Descriptive Translation Studies and the Use of Corpora: Investigating Mediation Universals ........................................... 141

CHRISTOPHER TAYLOR
Predictability in Film Language: Corpus-assisted Research ............ 167

ERIK CASTELLO
A Corpus-based Study of Text Complexity .................................................. 183

LARISSA D’ANGELO
Creating a Corpus for the Analysis of Identity Traits in English Specialised Discourse ........................................... 199

Part III
Corpora in University Foreign Language Teaching

GUY ASTON
Using the XML Edition of the British National Corpus ............ 213
GIUSEPPE BRUNETTI
Tagging the Lexicon of Old English Poetry .....................................237

ANTHONY BALDRY
Turning to Multimodal Corpus Research for Answers
to a Language-course Management Crisis.................................247

KATHERINE ACKERLEY
Using Comparable Expert-writer and Learner Corpora
for Developing Report-writing Skills .............................................259

SILVIA BERNARDINI
“What Students Want”…? Practical Suggestions
for Corpus-aided Translator Education .........................................275

FIONA DALZIEL AND FRANCESCA HELM
Exploring Modality in a Learner Corpus of Online Writing ............287

Notes on Contributors.......................................................................303
Remembering John Sinclair

To describe John Sinclair’s contribution to linguistics in a couple of pages would be ridiculous, so I shall not pretend to attempt it. I will instead opt for a personal view, remembering some of the ways he contributed, in my experience, to improving the description and teaching of English.

John Sinclair was an empiricist. In a famous anti-Chomskyan comment, he said “usage cannot be invented, it can only be recorded” (Sinclair 1987a: xv). Sometime in the 1960s he came across a computer, which allowed him to start systematically exploring actual usage (as hilariously recorded in David Lodge’s (1984) novel Small World). To understand what that means, go back to the sixties, when he wrote:

Any stretch of language has meaning only as a sample of an enormously large body of text; it represents the results of a complicated selection process, and each selection has meaning by virtue of all the other selections which might have been made, but have been rejected. (Sinclair 1965: 76-77)

Putting together an enormously large body of text in order to understand that selection process became the major objective of his research.

Towards an analysis of discourse (Sinclair/Coulthard 1975) drew on what we would now call a corpus of transcribed classroom interaction to describe the hierarchical organisation of discourse within which choices were made by teachers and pupils. In contrast with this substantially top-down perspective, whose innovative features were at the higher levels of the transaction and exchange, but which was effectively relegated to oblivion by the conversational analysis approaches then in vogue, his own research increasingly assumed the bottom-up approach which was to remain its distinguishing characteristic, with a primary focus on lexis, moving
progressively upwards towards wider grammatical and discourse structuring.

In the words of Antoinette Renouf (personal communication), in the 1970s John was effectively “biding his time until computing caught up”. Then the COBUILD project – an agreement between the Glasgow publishers Collins and the University of Birmingham – was finally born. The first COBUILD corpus was, at a mere seven and a half million words, the largest in the world, whose processing apparently required the whole of Birmingham University’s computing resources at weekends. And from the COBUILD corpus there emerged the first COBUILD dictionary (Sinclair 1987a), whose philosophy was explained in two books every language teacher should read: Looking Up (Sinclair 1987b) and Corpus, Concordance, Collocation (Sinclair 1991). The dictionary was revolutionary in:

- its choice of headwords and word-senses, partly if not exclusively based on their frequency in the corpus, and the detailed attention consequently paid to common words and their many uses (the latest edition of COBUILD lists no less than 32 uses of the verb come, excluding phrasal verb forms);
- its definitions, which took a discursive form that was used to capture pragmatic features of what a word could be used to do (“if you describe someone as apathetic, you are criticising them because they do not seem to be interested in or enthusiastic about doing anything”);
- its examples, all taken from the corpus, and selected to display typical collocational and colligational behaviour.

Under its ‘real English’ slogan, the COBUILD dictionary was perhaps not as commercially successful as it deserved. But it certainly put the fear of God into other publishers of learner dictionaries, several of whom pooled their resources to back a rival corpus initiative, the British National Corpus. Meanwhile, the COBUILD project went on developing. The COBUILD grammar appeared (1990), followed by various other works on ‘pattern grammar’, which grouped words sharing similar colligations. The corpus itself grew into the Bank of English, reaching a total of 400 million words by the turn of the century. Size was key to John’s reasoning. If you adopted the tenet
that every word has its own grammar (Halliday 1966), this meant that you needed hundreds or thousands of occurrences of each word form in order to work out what that grammar was. And if you were to keep up-to-date with changes in the language, you needed to update the corpus continually with new texts.

The intention to develop COBUILD as a monitor corpus – one whose content would change over time – posed problems from other perspectives. Continually moving the goalposts as the corpus changed would make research results irreplicable over time – by the time you went to look again, the corpus might have changed. Replicability was also problematic for the reason that the corpus was not publicly available. It belonged to HarperCollins and Birmingham, and you couldn’t obtain a static copy of it. If you were lucky enough to work at Birmingham or paid for the privilege, you could use Jem Clear’s query interface to do searches in the current version of the corpus. But you needed a current account with the Bank, and the questions you could ask were limited to those the interface allowed. Another problem was that in the search for quantity, less attention could be paid to quality. Like its ‘web as corpus’ successors today, the Bank found itself drawing predominantly on texts which were readily available in electronic form – particularly newspapers and radio transcripts. It thus became legitimate to question whether its range of texts and text-types was sufficiently balanced. As a standard point of reference for linguistic research, the more conservative approach adopted by the BNC started to win out.

John’s retirement from Birmingham in 1995 was followed a couple of years later by Harper Collins’ decision to cease funding the COBUILD project. But in any case, John had moved on beyond conventional dictionaries and grammars at that point. He was interested in larger lexical units than dictionaries generally deal with, the collocational, colligational, semantic and pragmatic regularities in phraseologies, and the grammars which fitted them together. In the paper he presented at the AIA conference in 1995 (Sinclair 1996) he provided an overview of his view of the lexical unit which was highly influential, and one of his last projects, Phrasebox (2008), was designed to provide ways of extracting these units from corpora, and to make them available to teachers and learners. This area of his work has had an extraordinary impact on ELT. Teachers and learners have
begun to change their idea of the lexical unit, and to see it as generally involving a coselection of several words following the ‘idiom principle’ (Sinclair 1991). The content of ELT courses has increasingly become more sensitive to the idea that, in terms of the companies it keeps, every word has its own grammar.

What has changed less in ELT is the way we teach that content. In 2002, at the Teaching and Language Corpora conference in Bertinoro, there was a memorable encounter between John Sinclair and Henry Widdowson (Aston/Bernardini/Stewart 2004). Sinclair’s position was that corpus analysis (plus a profound linguistic training) could show you what you should teach. Widdowson’s position was that this told you nothing about how to teach it, in the absence of a theory of learning. In retrospect, I feel both were right. Sinclair’s work has changed our view of how language works – I think we can reasonably term it a Kuhnian paradigm shift in the study of usage – from a predominantly grammatical focus to a lexical one. But that poses rather than solves problems as to how we can teach learners to use the language. To maintain a Widdowsonian metaphor (Widdowson 2000), from an applied linguistic perspective (rather than one of linguistics applied), findings from corpus analysis cannot tell you how to teach. But corpus findings have provided data which you cannot avoid taking into account – you can’t teach English today and not face up to John’s work. His final book, co-authored with Anna Maaranen (Sinclair/Maaranen 2006), initiated a formal description of a grammar which would be able to combine the lexical units on which he had been working for the previous fifteen years.

As this is a personal comment, let me add a final note concerning the British National Corpus. Obviously, the BNC wasn’t John’s favourite corpus. I suspect he felt it was unfair competition, a rival which had obtained a million pounds of government funding after he and his colleagues had spent years raising private finance to construct COBUILD – and moreover, without COBUILD, the BNC would probably never have happened. He felt that a corpus which was continually updated would be intrinsically superior to the static snapshot which the BNC provided. He also didn’t like the BNC’s part-of-speech tagging, which he saw as imposing an interpretation rather than ‘trusting the text’ – though on this point he subsequently relaxed his position, accepting proposals to part-of-speech tag the Bank of
English, and indeed to parse it syntactically at a shallow level. On one point, however, he didn’t modify his position, and that was trusting the text as far as tokenisation was concerned. The BNC had opted to treat certain word sequences as single grammatical units – things like up to, in spite of, as soon as, matter of fact, at least, as opposed to, in addition to, one another, at all, in touch, at random, in favour of, for example. I think the reason he was so opposed to these multiword tokens was because his own idea of the lexical item was so much more sophisticated, going far beyond the simple fixed collocations that the BNC had decided to treat as single units. And on this he was surely right. John changed our view of the lexical item – and the latest edition of the BNC (2007) has eliminated these multiword tokens, returning to trust the spacing in the orthographic text to define words. By ironic coincidence his death occurred the same day the new XML edition of the BNC was announced, so maybe he never knew that he had finally taught the ‘opposition’ something.

But John was generous with his wisdom. All of us in Italy were immensely lucky when Elena and he chose to come and live here and set up the Tuscan Word Centre. Not a place that I will ever forget – or the journey to get there. It was set up as an international centre running residential courses for young researchers, to teach whom John was able to attract scholars from all over the world. His generosity with time, thought, and encouragement for the next generation was a model for us all. As well as his unfailingly benign co-ordination of seminars at TWC, he also lectured widely in Italian universities, presented papers in national conferences, and collaborated on a number of national research projects, not least the construction of the CORIS/CODIS corpus of written Italian at the Language Centre in Bologna. After all the support he gave us over a decade, he will be sorely missed by English linguistics in Italy.

But then he will be sorely missed all over the world. Scholars of far greater authority than I have declared that John Sinclair was the most authoritative and influential British linguist of his generation. I am honoured to have been allowed these few words to affirm that from my own limited perspective. We all share an immense debt of gratitude to him.
References


