

THE SEARCH FOR UNITS OF MEANING:

A Tribute to John McHardy Sinclair (14 June 1933 - 13 March 2007)

Michael Stubbs, Universität Trier, Germany.

Abstract

This Laudatio was held at the University of Erlangen, Germany, on 25 October 2007, on the occasion of the posthumous award of an honorary doctorate to John McHardy Sinclair.

The Laudatio briefly summarizes some facts about his career in Edinburgh and Birmingham, and then discusses the major contributions which he made to three related areas: language in education, discourse analysis and corpus-assisted lexicography. A major theme in his work throughout his whole career is signalled in the title of one of his articles: "the search for units of meaning". In the 1960s, in his early work on corpus analysis, he studied the relation between objectively observable collocations and the psychological sensation of meaning. In the 1970s, in his work on classroom discourse, he studied the prototypical units of teacher-pupil dialogue in school classrooms. And from the 1980s onwards, in his influential work on corpus lexicography, for which he is now best known, he studied the kinds of patterning in long texts which are observable only with computational help. The Laudatio provides brief examples of the kind of innovative findings about long texts which his work has made possible: his development of a "new view of language and the technology associated with it". Finally, it situates his work within a long tradition of British empiricism.

John Sinclair is one of the most important figures in modern linguistics. It will take a long time before the implications of his ideas, for both applied linguistics and theoretical linguistics, are fully worked out, because many of his ideas were so original and innovative. But some points are clear. Most of us would be pleased if we could make one recognized contribution to one area of linguistics. John made substantial contributions to three areas: language in education, discourse analysis, and corpus linguistics and lexicography. In the case of discourse analysis and corpus-assisted lexicography, he created the very areas which he then developed.

The three areas are closely related. In the 1960s, in Edinburgh, he began his work on spoken language, based on the belief that spoken English would provide evidence of "the common, frequently occurring patterns of language" (Sinclair et al 1970 / 2004:

19). One central topic in John's work on language in education was classroom discourse: authentic audio-recorded spoken language. The work on classroom language emphasized the need for the analysis of long texts, as opposed to the short invented sentences which were in vogue at the time, post-1965. In turn, this led to the construction of large text collections, machine-readable corpora of hundreds of millions of running words, and to studying patterning which is visible only across very large text collections. This allowed the construction of the series of COBUILD dictionaries and grammars for learners of English. The COBUILD dictionaries, which were produced from the late 1980s by using such corpus data, were designed as pedagogic tools for advanced learners of English.

There were always close relations between his interest in spoken language, authentic language and language in the classroom, and between his theoretical and applied interests.

EDUCATION AND CAREER

John was a Scot, and proud of it. He was born in 1933 in Edinburgh, attended school there, and then studied at the University of Edinburgh, where he obtained a first class degree in English Language and Literature (MA 1955). He was then briefly a research student at the University, before being appointed to a Lectureship in the Department of English Language and General Linguistics, where he worked with Michael Halliday. His work in Edinburgh centred on the computer-assisted analysis of spoken English and on the linguistic stylistics of literary texts.

In 1965, at the age of 31, he was appointed to the foundation Chair of Modern English Language at the University of Birmingham, where he then stayed for the whole of his formal university career. His inaugural lecture was entitled "Indescribable English" and argues that "*we use language rather than just say things*" and that "*utterances do things rather than just mean things*" (Sinclair 1966, emphasis in original). His work in the 1970s focussed on educational linguistics and discourse analysis, and in the 1980s he took up again the corpus work and developed his enormously influential approach to lexicography.

He took partial retirement in 1995, then formal retirement in 2000, but remained extremely active in both teaching and research. With his second wife Elena Tognini-Bonelli, he founded the Tuscan Word Centre, and attracted large numbers of practising linguists to teach large numbers of students from around the world. [NOTE 1.]

Throughout his career he introduced very large numbers of young researchers and PhD students - who were spread across many countries - to work in educational topics, discourse analysis and corpus analysis. John himself travelled extensively in many countries, but nevertheless spent most of his career based in Birmingham. One reason, he once told me, was that he had built up computing facilities there, and, until the 1990s, it was simply not possible to transfer such work to other geographical locations.

He died at his home in Florence in March 2007.

"THE SEARCH FOR UNITS OF MEANING"

It is obviously rather simplistic to pick out just one theme in all his work, but "the search for units of meaning" (Sinclair 1996) might not be far off. This is the title of one of his articles from 1996. In his first work in corpus linguistics in the 1960s, he had asked: "(a) How can collocation be objectively described?" and "(b) What is the relationship between the physical evidence of collocation and the psychological sensation of meaning?" (Sinclair et al 1970 / 2004: 3). In his work on discourse in the classroom, he was looking for characteristic units of teacher-pupil dialogue. In his later corpus-based work he was developing a sophisticated model of extended lexical units: a theory of phraseology. And the basic method was to search for patterning in long authentic texts.

Along with this went his impatience with the very small number of short invented, artificial examples on which much linguistics from the 1960s to the 1990s was based. The title of a lecture from 1990, which became the title of his 2004 book, also expresses an essential theme in his work: "Trust the Text" (Sinclair 2004b). He argued consistently against the neglect and devaluation of textual study, which affected high theory in both linguistic and literary study from the 1960s onwards (see Hoover 2007).

LANGUAGE IN EDUCATION

One of his major contributions in the 1960s and 1970s was to language in education and educational linguistics.

In the 1970s, he was very active in developing teacher-training in the Birmingham area. He regularly made the point that knowledge about language is "sadly watered down and trivialized" in much educational discussion (Sinclair 1971: 15), and he succeeded in making English language a compulsory component of teacher-training in BEd degrees in Colleges of Education in the West Midlands.

Also in the early 1970s, along with Jim Wight, he directed a project entitled *Concept 7 to 9*. This project produced innovative teaching materials. They consisted of a large box full of communicative tasks and games. They were originally designed for children of Afro-Caribbean origin, who spoke a variety of English sometimes a long way from standard British English, but it turned out that they were of much more general value in developing the communicative competence of all children. The tasks focussed on "the aims of communication in real or realistic situations" and "the language needs of urban classrooms" (Sinclair 1973: 5).

In the late 1970s, he directed a project which developed ESP materials (English for Specific Purposes) for the University of Malaya. The materials were published as four course books, entitled *Skills for Learning* (Sinclair ed 1980). In the early 1990s, he became Chair of the editorial board for the journal *Language Awareness*, which started in 1992. One of his last projects is *PhraseBox*. This is a project to develop a corpus linguistics programme for schools, which John worked on from around 2000. It was commissioned by Scottish CILT (Center for Information on Language Teaching and Research) and funded by the Scottish Executive, Learning and Teaching Scotland and Cànan (the Gaelic College on Skye). The software gives children in Scottish primary schools resources to develop their vocabulary and grammar by providing them with

real-time access to a 100-million-word corpus. The project is described in one of John's more obscure publications, in *West Word*, a community newspaper for the western highlands in Scotland (Sinclair 2006a). The materials are due to go on-line for schools in early 2008.

In a word, John did not just write articles about language, but helped to develop training materials for teachers and classroom materials for students and pupils.

DISCOURSE ANALYSIS

His contribution to discourse analysis continued his early interests in both spoken language and in language in education. The work started formally in 1970 in a funded research project on classroom discourse. The work was published in 1975 in one book written with the project's co-director, Malcolm Coulthard, *Towards an Analysis of Discourse* (Sinclair & Coulthard 1975), and then a second book in 1982 with David Brazil *Teacher Talk* (Sinclair & Brazil 1982).

It was through this work that I got to know John personally. I'd been doing my PhD in Edinburgh on classroom discourse, and joined John and Malcolm on a second project on discourse analysis, which studied doctor-patient consultations, telephone conversations, and trade union / management negotiations. My job was to make and analyse audio-recordings in the local car industry. The work from this project was never published as a book, but individual journal articles and book chapters appeared.

It is difficult to remember how little was available on discourse analysis at the time of these projects, and therefore how innovative John's approach was. J. R. Firth had long ago pointed out that "conversation is much more of a roughly prescribed ritual than most people think" (Firth 1935: 66). But in the early 1970s, the published version of John Austin's *How to Do Things with Words* was still quite recent (published only ten years before in 1962). John Searle's *Speech Acts* had been published only two or three years before (1969). Paul Grice had given his lectures on *Logic and Conversation* some five years before (in 1967), but they were formally published only in 1975 and seem not to have been known to the project. This Oxford (so-called) "ordinary language" work was, however, not based on ordinary language at all, but on invented data: anathema to John's approach.

Michael Halliday's work was known of course, and provided a general functional background, but even *Language as Social Semiotic* was not published until several years after the project (in 1978). Anthropological and sociological work also provided general background: Dell Hymes on the ethnography of speaking (available since the early 1960s), Erving Goffman on "behaviour in public places" (from the 1960s); William Labov on narratives and ritual insults (from the early 1970s). Harvey Sacks' lectures were circulating in mimeo form; I arrived in Birmingham with a small collection, and had heard him lecture in Edinburgh around 1972; but little had been formally published. Otherwise, in the early 1970s, work on classroom discourse (by educationalists such as Douglas Barnes) provided insightful observation, but little systematic linguistic description.

There was a general feeling that discourse should somehow be studied, but there were few if any attempts to develop formal models of discourse structure. The two Birmingham projects were the first of their kind, but just a few years later, "discourse analysis" had become a clearly designated area, with its own courses and textbooks. One of the first student introductions was by Malcolm Coulthard (1977). John was following the principle proposed by J. R. Firth in the 1930s, that conversation is "the key to a better understanding of what language really is and how it works" (Firth 1935: 71), but John's work on discourse was some ten years ahead of the avalanche of work which it helped to start.

The aspect of the Birmingham discourse model which everyone immediately grasped was the stereotypical teacher-pupil exchange. In classic structuralist manner, John proposed that classroom discourse is hierarchic: a classroom lesson consists of transactions which consist of exchanges which consist of moves which consist of acts. It was probably the prototypical exchange structure which everyone immediately recognized: an IRF sequence of initiation - response - feedback (Sinclair & Coulthard 1975: 64):

- I Teacher: What is the name we give to those letters? Paul?
- R Pupil: Vowels.
- F Teacher: They're vowels, aren't they.

Nowadays, the IRF model is widely taken for granted, though I suspect that many people who use it no longer know where it comes from.

In addition, John never abandoned an interest in literature, and his work on text and discourse analysis always included literary texts. In the 1970s, along with the novelist David Lodge, who was his colleague in the English Department at Birmingham, he developed a course on stylistics. The title of an early article was "The integration of language and literature in the English curriculum" (Sinclair 1971). For the course, they selected extracts of literary texts in which a specific linguistic feature was foregrounded: such as repetition, verbless sentences, complex noun phrases, and the like. From one end, they taught grammar through literature, and from the other end, they showed that grammatical analysis was necessary to literary interpretation.

The analysis of literary texts was part of John's demand that linguistics must be able to handle all kinds of authentic texts. He argued further that, if linguists cannot handle the most prestigious texts in the culture, then there is a major gap in linguistic theory. Conversely, of course, the analysis of literary texts must have a systematic basis, and not be the mere swapping of personal opinions. In an analysis of a poem by Robert Graves, he argued that "the role of linguistics is to expose the public meaning of texts in a language" (Sinclair 1968). He similarly argued that "if literary comment is to be more than exclusively personal testimony, it must be interpretable with respect to objective analysis." (Sinclair 1971: 17)

In all of this work there is a consistent emphasis on long texts, authentic texts, including literary texts, and on observable textual evidence of meaning.

CORPUS LINGUISTICS AND LEXICOGRAPHY

(1) The "OSTI" Report

Post-1990, John was mainly known for his work in corpus linguistics. This work started in Edinburgh, in the 1960s, and was informally published as the "OSTI Report" (UK Government Office for Scientific and Technical Information, Sinclair et al 1970/2004). This is a report on quantitative research on computer-readable corpus data, carried out between 1963 and 1969, but not formally published until 2004.

The project was in touch with the work at Brown University: Francis and Kučera's *Computational Analysis of Present Day American English*, based on their one-million-word corpus of written American English, had appeared in 1967. But again, it is difficult to project oneself back to a period in which there were no PCs, and in which the university mainframe machine could only handle with difficulty John's corpus of 135,000 running words of spoken language.

Yet the report worked out many of the main ideas of modern corpus linguistics in astonishing detail. This work in the 1960s formulated explicitly several principles which are still central in corpus linguistics today. It put forward a statistical theory of collocation in which collocations were interpreted as evidence of meaning. It asked: What kinds of lexical patterning can be found in text? How can collocation be objectively described? What size of span is relevant? How can collocational evidence be used to study meaning? Some central principles which are explicitly formulated include: The unit of lexis is unlikely to be the word in all cases. Units of meaning can be defined via statistically defined units of lexis. Homonyms can be automatically distinguished by their collocations. Collocations differ in different text-types. Many words are frequent because they are used in frequent phrases. One form of a lemma is regularly much more frequent than the others (which throws doubt on the lemma as a linguistic unit).

It proposed that there is a relation "between statistically defined units of lexis and postulated units of meaning" (Sinclair et al 1970/2004: 6). As John puts it in the 2004 preface to the OSTI Report, we have a "very strong hypothesis [that] for every distinct unit of meaning there is a full phrasal expression [...] which we call the canonical form". And he formulates one of his main ambitious aims: a list of all the lexical items in the language with their possible variants would be "the ultimate dictionary" (Sinclair et al 1970/2004: xxiv).

In a word, the OSTI Report makes substantial progress with a question which had never had a satisfactory answer: How can the units of meaning of a language be objectively and formally identified? It is important to emphasize that this tradition of corpus work was concerned, from the beginning, with a theory of meaning.

The work then had to be shelved, because the machines were simply not powerful enough in the 1970s to handle large quantities of data. It was started again in 1980s as the COBUILD project in corpus-assisted lexicography.

(2) The COBUILD project

In the 1980s, John became the Founding Editor in Chief of the COBUILD series of language reference materials. He built up the Birmingham corpus, which came to be called the Bank of English, and along with a powerful team of colleagues - many of whom have made important contributions to corpus linguistics in their own right - the first COBUILD dictionary was published in 1987: the first dictionary based entirely on corpus data. The team for this and later dictionaries and grammars included Mona Baker, Joanna Channell, Jem Clear, Gwynneth Fox, Gill Francis, Patrick Hanks, Susan Hunston, Ramesh Krishnamurthy, Rosamund Moon, Antoinette Renouf, and others. The first dictionary, *Collins COBUILD English Language Dictionary* (Sinclair ed. 1987) was followed by a whole series of other dictionaries and grammars, plus associated teaching materials, including *Collins COBUILD English Grammar* (Sinclair ed 1990), and *Collins COBUILD Grammar Patterns* (Francis et al, 1996, 1998).

A recent account of the COBUILD project is provided by Moon (2007), one of the senior lexicographers in COBUILD, who worked with the project from the beginning. She analyses why the "new methodology and approach" of the project had such "a cataclytic effect on lexicography". When the project started there simply was no "viable lexicographic theory", whereas lexicography is now part of mainstream linguistics. Yet, it was in some ways just too innovative to be a total commercial success. There was a clash between commercial priorities and academic rigour, and the purist approach to examples turned out to be confusing and not entirely right for learners. Only the most advanced learners and language professionals could handle the authentic examples. By 1995 other major British dictionary publishers (CUP, Longman and OUP) had copied the ideas. This was imitation as the sincerest form of flattery, but they subtly changed the attitude to modifying attested corpus examples and made the dictionaries more user-friendly. It remains true however that it is the COBUILD project which developed the lexicographic theory. Many of the principles of corpus compilation and analysis are set out in *Looking Up* (Sinclair ed 1987). The title is of course a play on words: you look things up in dictionaries, and dictionary making is looking up, that is, improving with new data and methods.

Many of John's main ideas are formulated in what is now a modern classic: *Corpus Concordance Collocation* (Sinclair 1991). The corpus, the concordance and the collocations are chronologically and logically related. First, you need a corpus: a machine-readable text collection, ideally as large as possible. Second, you need concordance software in order to identify patterns. Third, these patterns involve collocations: the regular co-selection of words and other grammatical features.

We have had paper concordances since the Middle Ages. But modern concordance software can search large corpora very fast, re-order the findings, and help to identify variable extended units of meaning. It is difficult to illustrate the power of this idea very briefly, because it depends on the analysis of very large amounts of data. But a simple example is possible. In both the OSTI Report and in a 1999 article John points out that *way* is "a very unusual word", and that "the very frequent words need to be ... described in their own terms", since they "play an important role in phraseology" (Sinclair 1999: 157, 159). It doesn't make much sense to ask what the individual word *way* means, since it all depends on the phraseology:

- all the way to school, half way through, the other way round, by the way, a possible way of checking, ... etc

Few of the very common words in the language "have a clear meaning independent of the context". Nevertheless, "their frequency makes them dominate all text" (Sinclair 1999: 158, 163).

Here is a fragment of output from some modern concordance software: all the examples of the three words *way - long - go* co-occurring in a 6 million word corpus [NOTE 2]. The concordance lines were generated by software developed by Martin Warren and Chris Greaves (Cheng et al 2006), in a project that John was involved in.

Six-million-word corpus: all examples of *long - way - go*

01 added that there was still a long way to go in overcoming Stalinist structure
 02 ges in 1902 there was still a long way to go: A. M. Fairbairn warned Sir Alfre
 03 e," he said. There is still a long way to go, however, to reach the 1991 high
 04 handicapped. There is still a long way to go before the majority of teachers i
 05 on peas. But we still have a long way to go. Next we imagine our blizzard rag
 06 hem to Church-we still have a long way to go to reach our African Church stand
 07 other we still have an awful long way to go. TRADES UNIONS AND THE EUROPEAN
 08 real good time Gemma's got a long way to go before she gets to eighty You're
 09 s demonstrates that we have a long way to go before we have true democracy in
 10 ice, but I'm afraid we have a long way to go before we catch up to the Japanes
 11 dications are that there is a long way to go before the Algerian problem is fi
 12 f small atomic reactors has a long way to go before it becomes a commercial pr
 13 seventy. I mean he's, he's a long way to go. Co could you cut me a slice of t
 14 aeli occupation, Israel has a long way to go to convince anyone that it is ser
 15 ichael Caine." But he's got a long way to go. "David Who? Never heard of him,"
 16 teen. My turn to draw. A long long way to go though. Difficult. Well look wher

 17 y the earth's gases will go a long way toward bolstering or destroying cosmic
 18 nd and motion, fountains go a long way toward selling themselves in showrooms
 19 ntion. Vaccinations also go a long way toward eliminating the spread of more v
 20 and ready to act, would go a long way toward making the 'new world order' mor
 21 s father owned, it might go a long way toward explaining why she was reluctant
 22 these very simple cases go a long way towards explaining puzzling features of
 23 duction in wastage would go a long way to easing the manpower problems. In gen
 24 and guineas. That should go a long way to easing the strain on an amateur team
 25 ll, if voters are loyal, go a long way to ensuring the election of one or even
 26 s in partnership, we can go a long way. You do believe that?" The expression i
 27 ius scarf, and a boy can go a long way with those things. You got a job yet? W
 28 Conolly said. "We could go a long way on this. I didn't know Major Fitzroy wa

 29 Of course we go back an awful long way don't we? Yes. Yeah. Are you going to t
 30 I go to London. And we go any damn way I please, as long as I don't interfere

Data of this kind make visible the kinds of patterning which occur across long texts, and provides observable evidence of the meaning of extended lexical units. Several things are visible in the concordance lines. They show that *way* "appears frequently in fixed sequences" (Sinclair 1970/2004: 110), and that the unit of meaning is rarely an individual word. Rather words are co-selected to form longer non-compositional units. Also, the three words *way - long - go* tend to occur in still longer sequences, which are not literal, but metaphorical.

There are two main units, which have pragmatic meanings: [a] is used in an abstract extended sense to simultaneously encourage hearers about progress in the past and to warn them of efforts still required in the future; [b] is also used in exclusively abstract senses.

- [a] there BE still a long way to go before ...
- [b] (modal) go a long way to(wards) VERB-ing ...

Now the word *way* is just one point of origin for a collocation, and it is shown here with just two collocates. Imagine doing this with say 20,000 different words and all their frequent collocates in a corpus of 500 million words, and you have some small impression of the ambitious range of John's aim of creating an inventory of the units of meaning in English.

In a series of papers from the 1990s (Sinclair 1996, 1998, 2005), he put forward a detailed model of semantic units of a kind which had not previously been described. In these articles, John argued consistently that "the normal carrier of meaning is the phrase" (Sinclair 2005), and that the lack of a convincing theory of phraseology is due to two things: the faulty assumption that the word is the primary unit of meaning, and the misleading separation of lexis and grammar. The model is extremely productive, and many further examples have been discovered by other researchers. It's all to do with observable empirical evidence of meaning, and what texts and corpora can tell us about meaning.

The overall finding of work is that the phraseological tendency in language use is much greater than previously suspected (except perhaps by a few scholars such as Dwight Bolinger, Igor Mel'čuk and Andrew Pawley), and its extent can be quantified.

PUBLICATIONS

John's work was for a long time not as well known as it deserved to be. This was partly his own fault. He often published in obscure places, not always as obscure as community newspapers from the Scottish Highlands, but nevertheless frequently in little known journals and book collections, and it was only post-1990 or so that he began to collect his work into books with leading publishers (OUP, Routledge, Benjamins). He once told me that he had never published an article in a mainstream refereed journal. I questioned this and cited some counter-examples, which he argued were not genuine counter-examples, since he had not submitted the articles: they had been commissioned. He was always very sceptical of journals and their refereeing and gate-keeping processes, which he thought were driven by fashion rather than by standards of empirical research.

He was also particularly proud of the fact that, when he was appointed to his chair in Birmingham, he had no PhD and no formal publications. His first publication was in 1965, the year when he took up his chair: it was an article on stylistics entitled "When is a poem like a sunset?", which was published in literary journal (Sinclair 1965). It is a short experimental study of the oral poetic tradition which he carried out with students. He got them to read and memorize a ballad ("La Belle Dame Sans Merci" by Keats) and

then studied what changes they introduced into their versions when they tried to remember the poem some time later.

His last book *Linear Unit Grammar*, co-authored with Anna Mauranen, is typical Sinclair (Sinclair & Mauranen 2006). It is based on one of his most fundamental principles: if a grammar cannot handle authentic raw texts of any type whatsoever, then it is of limited value. The book points out that traditional grammars work only on input sentences which have been very considerably cleaned up (or simply invented). Sinclair and Mauranen demonstrate that analysis of raw textual data is possible. On the one hand, the proposals are so simple as to seem absolutely obvious: once someone else has thought of them. On the other hand, they are so innovative, that it will take some time before they can be properly evaluated. I will not attempt this here, and just note that the book develops the view that significant units of language in use are multiword chunks. But here, the approach is via a detailed discussion of individual text fragments as opposed to repeated patterns across large text collections. Either way, it is a significant break with mainstream linguistic approaches.

IN SUMMARY ...

1. First, John's work belongs to a long tradition of British empiricism and British and European text and corpus analysis, derived from his own teachers and colleagues (especially J. R. Firth and Michael Halliday), but represented in a broader European tradition (for example by Otto Jespersen) and in a much more restricted American tradition (for example by Charles Fries). This work, based on the careful description of texts, is very different from the largely American tradition of invented introspective data which provided a short interruption to this empirical tradition. As John pointed out in a characteristically ironic aside, "one does not study all of botany by making artificial flowers" (Sinclair 1991: 6).
2. Second, the description of meaning has always been at the centre of the British Firth-Halliday-Sinclair tradition of linguistics. It is in John's work that one finds the most sustained attempt to develop an empirical semantics. As he said in a plenary at the AAAL (American Association for Applied Linguistics), "corpus research, properly focussed, can sharpen perceptions of meaning" (Sinclair 2006b).
3. Third, he is one of the very few linguists whose work has changed the way we perceive language. In the words of one of his best known observations: "The language looks rather different when you look at a lot of it at once" (Sinclair 1991: 100).
4. Fourth, John is one of the very few linguists who have made substantial discoveries. As E. O. Wilson (1998: 61) has argued: "The true and final test of a scientific career is how well the following declarative sentence can be completed: *He (or she) discovered that ...*" John's work is full of new findings about English, things that people had previously simply not noticed, despite thousands of years of textual study. But then they are only observable with the help of the computer techniques which he helped to invent, and which the rest of us can now use to make further discoveries. These include both individual phraseological units, but also methods of analysis - how to extract patterns from raw data - and principles: in particular the extent of phraseology in language use.

John's vision of linguistics was always long-term: "a new view of language and the technology associated with it" (Sinclair 1991: 1). He developed some of his main ideas in the 1960s, and then waited till the technology - and everyone else's ideas - had caught up with him. As he remarked with some satisfaction:

"Thirty years ago [in the 1960s] when this research started it was considered impossible to process texts of several million words in length. Twenty years ago [in the 1970s] it was considered marginally possible but lunatic. Ten years ago [in the 1980s] it was considered quite possible but still lunatic. Today [in the 1990s] it is very popular." (Sinclair 1991: 1)

John Sinclair's work has shown how to use empirical evidence to tackle the deepest question in the philosophy of language: the nature of units of meaning.

Like many other people, I owe a very large part of my own academic development to John's friendship and inspiring ideas. I knew him for over thirty years: from 1973 when he appointed me to my first academic job (on the second project in discourse analysis) at the University of Birmingham. In October 2007 in Erlangen, John was due to receive his honorary doctorate personally, and then take part in a round table discussion, where he would have responded, courteously but firmly, to our papers; and shown us when we had strayed from his own rigorous standards of empirical research. I was so much looking forward to seeing John again in Erlangen, and to continuing unfinished discussions with him. I will miss him greatly, as will friends and colleagues in many places in the world. But I am very grateful that I had the chance to know him.

NOTES

1. Some biographical details are from the English Department website at Birmingham University and from obituaries in *The Guardian* (3 May 2007), *The Scotsman* (10 May 2007), the *International Journal of Corpus Linguistics* (12, 2: 2007), and *Functions of Language* (14, 2; 2007). I am grateful to Susan Hunston and Michaela Mahlberg for comments on a previous version of this paper.
2. The corpus consisted of Brown, LOB, Frown and FLOB plus BNC-baby: five million words of written data and one million words of spoken data.

REFERENCES

- Cheng, W., Greaves, C. & Warren, M. (2006) From n-gram to skip-gram to congram. *International Journal of Corpus Linguistics*, 11, 4: 411-33.
- Coulthard, R. M. (1977) *An Introduction to Discourse Analysis*. London: Longman.
- Firth, J. R. (1935). The technique of semantics. *Transactions of the Philological Society*. 36-72.
- Francis, G., Hunston, S. & Manning, E. (1996, 1998) *Collins COBUILD Grammar Patterns*. 2 vols. London: HarperCollins.

- Hoover, D. (2007) The end of the irrelevant text. *DHQ: Digital Humanities Quarterly*, 1, 2. <http://www.digitalhumanities.org/dhq/vol/001/2/index.html>. [Accessed 3 Nov 2007].
- Moon, R. (2007) Sinclair, lexicography and the COBUILD project. *International Journal of Corpus Linguistics*, 12, 2: 159-81.
- Sinclair, J. M. (1965) When is a poem like a sunset? *A Review of English Literature*, 6, 2: 76-91.
- Sinclair, J. M. (1966) Indescribable English. Inaugural lecture, University of Birmingham. [Abstract in Sinclair & Coulthard 1975: 151.]
- Sinclair, J. M. (1968) A technique of stylistic description. *Language and Style*, 1: 215-42.
- Sinclair, J. M. (1971) The integration of language and literature in the English curriculum. *Educational Review*, 23, 3. [Page references to reprint in R. Carter & D. Burton eds (1982) *Literary Text and Language Study*. London: Arnold.]
- Sinclair, J. M. (1973) English for Effect. *Commonwealth Education Liaison Newsletter*, 3, 11: 5-7.
- Sinclair, J. M. ed (1980) *Skills for Learning*. Nelson: University of Malaya Press.
- Sinclair, J. M. ed (1987) *Looking Up*. London: Collins.
- Sinclair, J. M. ed (1987) *Collins COBUILD English Language Dictionary*. London: HarperCollins.
- Sinclair, J. M. ed (1990) *Collins COBUILD English Grammar*. London: HarperCollins.
- Sinclair, J. M. (1991). *Corpus Concordance Collocation*. Oxford: Oxford University Press.
- Sinclair, J. M. (1996) The search for units of meaning. *Textus*, 9, 1: 75-106.
- Sinclair, J. M. (1998) The lexical item. In E. Weigand ed. *Contrastive Lexical Semantics*. Amsterdam: Benjamins. 1-24.
- Sinclair, J. M. (1999) A way with common words. In H. Hasselgård & S. Oksefjell eds. *Out of Corpora*. Amsterdam: Rodopi. 157-79.
- Sinclair, J. M. (2004a) New evidence, new priorities, new attitudes. In J. Sinclair ed. *How to Use Corpora in Language Teaching*. Amsterdam: Benjamins. 271-99.
- Sinclair, J. M. (2004b) *Trust the Text*. London: Routledge.
- Sinclair, J. M. (2005) The phrase, the whole phrase and nothing but the phrase. Plenary, *Phraseology 2005*, Louvain-la-Neuve, October 2005.
- Sinclair, J. M. (2006a) A language landscape. *West Word*. January 2006 Issue. <http://road-to-the-isles.org.uk/westword/jan2006.html>. (Accessed 3 Nov 2007).
- Sinclair, J. M. (2006b) Small words make big meanings. Plenary, AAAL (American Association for Applied Linguistics).
- Sinclair, J. M. & Brazil, D. (1982) *Teacher Talk*. Oxford: Oxford University Press.
- Sinclair, J. M. & Coulthard, R. M. (1975) *Towards an Analysis of Discourse*. London: Oxford University Press.
- Sinclair, J. M., Jones, S. & Daley, R. (1970 / 2004) *English Collocation Studies. The OSTI Report*. Edited R. Krishnamurthy. London: Continuum. [Original mimeoed report 1970.]
- Sinclair, J. M. & Mauranen, A. (2006) *Linear Unit Grammar*. Amsterdam: Benjamins.
- Wilson, E. O. (1998) *Consilience: The Unity of Knowledge*. London: Abacus: