doi:10.1093/ijl/ecm018

SUMMER 2007

EURALEX NEWSLETTER

Editor: Paul Bogaards
Email: p.bogaards@let.leidenuniv.nl

The EURALEX Newsletter

This quarterly Newsletter is intended to include not only official announcements but also news about EURALEX members, their publications, projects, and (it is hoped) their opinions, and news about other lexicographical organizations. Please try to support this by sending newsletter contributions to Paul Bogaards at the email address above. The deadlines for spring (March), summer (June), autumn (September), and winter (December) issues are respectively 15 January, 15 April, 15 July, and 15 October annually.

The EURALEX Web Site

The URL of the EURALEX web site is www.euralex.org

John Sinclair (1933–2007)

John Sinclair, who died of cancer on 13 March 2007 aged 73, was a pioneer in discourse analysis and corpus linguistics, and the founder and chief editor of the Cobuild series of dictionaries, grammars, and aids for foreign learners. He was a corpus linguist par excellence – in fact, he virtually invented the discipline and led the field in developing the corpus-based analysis of collocations. His influence on lexicography, on the study and teaching of language, and on linguistics in general was profound. It is far too early to attempt a balanced assessment of this influence. It may well be that, in years to come, the empiricism of scholars such as Sinclair and Halliday will come to be recognized as the late 20th-century mainstream in linguistics in the English-speaking world – a mainstream that flowed from Saussure through European structuralism – rather than the syntactocentric American school that arrogated to itself the name 'mainstream' in the 1970s. Certainly, linguists, lexicographers, and language teachers as diverse as Mona Baker, Geoff Barnbrook, Malcolm Coulthard, Alice Deignan, Gwyneth Fox, Michael Hoey, Susan Hunston, Patrick Hanks, Tim Johns, Ramesh Krishnamurthy, Bill Louw, Anna Mauranen, Rosamund Moon, Alan Partington, Antoinette

International Journal of Lexicography, Vol. 20 No. 2.

Renouf, Ute Römer, Michael Stubbs, Wolfgang Teubert, and Geoffrey Williams (to name but a few, and in alphabetical order as befits an obituary in a journal of lexicography) are all proud to acknowledge their intellectual debt to Sinclair – as, of course, is John's widow, Elena Tognini Bonelli, an important corpus linguist in her own right.

John McHardy Sinclair was born in Edinburgh in 1933, the son of moderately prosperous middle-class parents. His elder sister, Beryl T. ('Sue') Atkins, was to achieve fame as a bilingual lexicographer and frame semanticist. John Sinclair attended George Heriot's School in Edinburgh and went on to read English language and literature at the University of Edinburgh. After graduating with first-class honours in 1955 he was called up for national service as a clerk in the RAF, where his natural independence of mind did not consort well with military discipline, although he was well served by his natural taciturnity, keeping his more subversive thoughts to himself. In 1958 he returned to the University of Edinburgh as a research student. In 1965, at the tender age of 31, after only a few years as a lecturer in Edinburgh, he was elected to the Chair of Modern English Language at the University of Birmingham. He had still not completed his doctorate. With quiet but focused determination, he set about turning Birmingham into a major centre for the empirical study of the English language. As such it was to become famous, attracting scholars from all over the world. Sinclair was not a prolific writer. His main modus operandi was influencing and inspiring others. He was a gifted teacher, a master of the 'silent method', in which students are encouraged to debate a selected issue and work things out for themselves, with minimal, judicious, and unobtrusive prompting from the teacher, rather than being exposed to the unremitting monologues that used to be characteristic of the lecture hall. He had an extraordinary gift for the *mot juste* and for laconically outlining solutions to problems in a way that could seem unremarkable at first but that somehow stuck in the mind and came back, hauntingly, to guide the recipient in later years. He knew how to use language as well as how to study it, and he cared about people.

In his early years he wrote a grammar textbook, *A Grammar of Modern Spoken English* (1972). During the mid 1980s, when work on the first edition of the Cobuild project was in full swing, he led a series of seminars for lexicographers and other research staff that were instructive, memorable, and enjoyable. He made studying fun. In 1991, with members of the Cobuild team, he published the *Collins Cobuild English Grammar*. This was the forerunner of a more extensive and more radical work, *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English* (2000) by Susan Hunston and Gill Francis, in the series *Studies in corpus linguistics*, edited by Elena Tognini-Bonelli. One of his last published works was the *Linear Unit Grammar* (2006), written with Anna Mauranen, which shows how

the 'idiom principle' (see below) can be integrated into a lexically based theory of grammar.

Despite all this, Sinclair's main claims to fame are not based on his contributions to grammar, but rather on his work in discourse analysis and lexicography. Like his mentor, fellow Firthian, and friend Michael Halliday, he recognized the importance of syntax but did not suffer from syntactocentrism. He identified and analysed complex levels of linguistic structure at levels both above and below the sentence: the discourse and the lexicon. His early years in Birmingham were devoted in part, with Malcolm Coulthard and others, to the study of discourse based on recordings of classroom discussions. A report to the Social Sciences Research Council, *The English Used by Teachers and Pupils* (1972), defined exchanges in the classroom in terms of the tactics that speakers (both children and teachers) systematically use to achieve their discourse goals. This work was subsequently elaborated in *Towards an Analysis of Discourse* (1975) and *Teacher Talk* (1982). Its importance is that it demonstrated very clearly that human conversation – or at any rate classroom discourse – has a fundamentally regular structure. It is not a free-for-all, and structure is not confined to the sentence.

At the same time as studying discourse structures, Sinclair was working out new approaches to lexis. In an extraordinarily prescient paper, 'Beginning the Study of Lexis', which appeared in 1966 in a memorial volume for J. R. Firth, Sinclair mapped out the agenda for the work on lexis that was to preoccupy him for the rest of his life, in particular addressing 'problems that are not likely to yield to anything less imposing than a very large computer'. His understanding of the complex, *probabilistic* nature of lexical collocation was already profound, and his desire to find out what actually goes on intense. His warning against 'inflating expectations into preconceptions' has come back to haunt those of his contemporaries who disregarded it and whose theoretical speculations are now history. His identification of aspects of collocational analysis (which, with typical modesty, he describes as 'extremely crude') are the bedrock of corpus-driven lexical analysis today. These aspects may be briefly summarized as: (a) the mutual predictive strength of the collocating nodes; (b) their distance apart; (c) the nature of what lies between them; and (d) their grammatical roles.

In January 1970 Sinclair delivered to the Office for Scientific and Technical Information in London (OSTI) a report entitled 'English Lexical Studies' (now universally known as *The OSTI Report*). The importance of this was not at first recognized. It was not until 2004, largely thanks to the indefatigable efforts of Ramesh Krishnamurthy, that it was published for a wider audience. Sinclair had collected an electronic corpus of 135,000 transcribed words of spontaneous conversation and compared it with the Brown Corpus of 1 million words of written English, which had become available a few years earlier.

On this modest foundation of evidence, compiled with laborious effort using the latest technology of the time (which would nowadays be regarded as hopelessly primitive), he was able to illuminate some core issues, chief among them:

- The nature of collocation and lexical patterning;
- The nature of the lexical item (including 'multiword items' – there is an illuminating discussion of the term *red herring*);
- The relationship between grammar and lexis;
- The Zipfian distribution of word frequencies;
- Some differences between spoken and written language.

When *Collins English Dictionary* (1979) was almost complete, a nervous publisher hired John Sinclair as consultant General Editor. This was done for marketing reasons – so late in the project that it was not possible for him to have any significant effect on the content. He made a couple of friendly visits in 1978 to Aylesbury, where the project was being completed, and to the publisher's offices in Glasgow. He kept his own counsel, made a few encouraging remarks, and duly sprinkled holy water on the publication. But he had come face to face with the traditional preoccupations of lexicographers writing for an audience of native speakers: for example, coverage of unfamiliar terminology (especially rare and scientific terminology), definitions worded for substitutability, and fear of overrestriction (resulting in massive overuse in definitions of hedges such as 'etc.') All of these and other features were to be radically different in the new learners' dictionary that he was designing: Cobuild.

One minor but amusing and illuminating consequence of this 1978 interaction involved me personally. Incredible as it may sound, in those days I was unfamiliar with both John's extremely low-key style and the ways of the academic world, and this unfamiliarity resulted in a horrible misunderstanding. John asked me whether I would like to 'come to Birmingham one day and talk to us about dictionaries'. I imagined that he had in mind some sort of informal chat about future projects. It was not until I was ushered into a room full of eager-eyed strangers that it dawned on me that 'talk to us about dictionaries' meant 'lead a seminar' and 'us' meant 'English department staff and research students'. Somehow I survived. Four years later, after learning something about EFL teaching (Sinclair organized a short-term job for me in Sweden) and more about the academic world (at the University of Essex), I joined Cobuild as project manager.

It is granted to very few people to be involved in something totally original. The lexicographers who worked on the first edition of Cobuild were among those privileged few. For the first time ever, lexicographers had evidence that enabled them to begin to see how words actually fitted together.

Opening a fresh concordance for all uses of a given word in a 7.3 million word corpus was like opening a window on a landscape of fresh snow on a sunny winter's morning. Very often, a simple right-sort of concordances revealed patterns of usage that were at the same time unexpected and yet obvious (once seen). Other patterns needed more effort to reveal. Significant collocations were everywhere to be observed in the data and nowhere to be found in dictionaries. Two modest examples are the relationship between ***storm*** and ***protest*** or ***torrent*** and ***abuse***. (Why do we say 'a storm of protest' and 'a torrent of abuse' rather than 'a storm of abuse' and 'a torrent of protest'? The latter phraseologies are perfectly possible and grammatically well-formed, but they do not occur in any corpus with statistical significance, whereas the first two are clichés or (almost) lexical items in their own right.)

Another anecdote may be mentioned here, as it illuminates both John's low-key approach and his profound effects on the listener. One of the first words in the Birmingham 7.3 million word corpus that I examined was ***lap***. I discussed it with John.

'Hmm,' I said. 'Not much here about going once round a track.'

John's response was: 'I'm more interested in all those punctuation marks.'

I looked again. Now I noticed that more often than not the word ***lap*** is followed by a comma or a full stop. From this tiny fragment of evidence, after a few years of brooding, I was able to conclude what John no doubt had already noticed:

- ***lap*** usually occurs in a prepositional phrase in clause-final position (*in his lap, on her lap*).
- Its meaning is intimately interwoven with this syntagmatic fact. (It is questionable, for example, whether the lap should be classed a body part, for you don't have a lap when you stand up.)
- Lexical items are typically associated with preferential syntagmatic patterns such as the clause-final occurrence of ***lap***, and lexicographers need to be sensitive to them.
- A sense or use of a word that our intuitions tell us is frequent probably isn't.

Lexicography is a slow-moving field, quick to copy minor competitive features, but slow to accept radical innovation. The first edition of *Collins Cobuild English Learner's Dictionary*, published in 1987 (which was to undergo various changes of title in subsequent editions), was nothing if not radical. It was the first corpus-driven dictionary of any language; it was the only learner's dictionary to give examples drawn from actual usage for every sense of every word; and it focused on helping learners to use English words idiomatically (it was an 'encoding dictionary' rather than a 'decoding dictionary'). It adopted a controversial new approach to definition writing. On publication, it was, in equal measure, hailed by students ('At last,

a dictionary that we can understand') and reviled by those of their teachers who were old-school metalexicographers. Instead of assuming that the meanings of *all* words can be captured in substitutable definitions, Cobuild gave due prominence, where appropriate, to explaining and exemplifying pragmatic, conversation-organizing, text-organizing, and discourse-organizing uses of words.

The sincerest form of flattery, they say, is imitation. Many (though of course not all) of the innovations introduced by Sinclair and his colleagues in Cobuild were subsequently quietly imitated by other leading learners' dictionaries. Sinclair's unacknowledged influence was not restricted to learners' dictionaries. Among dictionaries of English for native speakers, for example, the *(New) Oxford Dictionary of English* ((N)ODE, 1998, 2003) is the only one that takes corpus evidence and syntagmatics seriously, and these aspects of that dictionary can be traced back indirectly to the influence of Sinclair – though NODE was not nearly radical enough for his taste.

Perhaps Sinclair's greatest single contribution to the study of language is his recognition of the importance of collocation as an organizing principle of language. His book *Corpus, Concordance, Collocation* (1993) discusses such issues as corpus creation, word frequencies, the nature of language in use, sense and structure, phraseology, the relationship between lexis and grammar, and above all, collocation. He identifies very clearly the tension between what he called the **open-choice principle**:

> 'a way of seeing language as the result of a very large number of complex choices. At each point where a unit is complete (a word or a phrase or a clause), a large range of choices opens up and the only restraint is grammaticalness'

and the **idiom principle**:

> 'Many choices within language have little or nothing to do with the world outside. . . . a language user has available to him or her a large number of semi pre-constructed phrases that constitute single choices.'

In 1995 John and Elena set up the Tuscan Word Centre in a peaceful agricultural setting in the hills above Florence. Researchers came from all over the world, as they had once come to Birmingham, to study corpus linguistics and to work out, under the guidance of John and Elena, empirical approaches to the study, description, and teaching of language. One product of this was another book, *Reading Concordances* (2003), a volume outlining 18 tasks in corpus linguistics, using concordances to enlarge understanding of language. They include: how to make meaning distinctions, how to identify underlying regularity, and how to deal with literal and metaphorical phraseology.

Trust the Text: Language, Corpus, and Discourse (2004) is a collection of papers which challenge a number of commonly held assumptions about language. In Part I Sinclair explains and discusses the notion of 'prospection': the predictability of what is likely to follow on the basis of what has just been said. He challenges the notion of the lemma as a stable unit, pointing out that collocates are not evenly distributed across all forms of a lemma, and some forms of a lemma may even have a different meaning from the others. Part II expounds Sinclair's distinction between the 'interactive' and the 'autonomous' planes of discourse: language users interact in structured ways, but they also organize their contributions autonomously, i.e. by drawing upon their stored experiences. Part III is probably the most interesting for lexicographers, discussing the nature of the lexical item (a limited set, which can express an unlimited set of meanings), the 'empty lexicon', and lexical grammar.

Sinclair's deft touch and his predilection for simplicity masked a mind and personality of ferocious complexity. He could make complex issues seem simple, and he disliked pretentious terms like 'predilection'. He had a profound moral sense (which he did not parade openly but which one occasionally bumped into) and he was an extremely subtle judge of human nature. His quietness masked a steely determination. He would say what he had to say with great mildness of manner, but he meant what he said and very rarely changed his mind – especially not under duress. He did not go out of his way to pick a fight – but woe betide anyone (academic or bureaucrat) who chose to pick a fight with him. At departmental meetings he spoke only when necessary, and then often at the very end of a meeting, when people were pushing back their chairs and getting up to go. Such interventions, being impeccably logical, radical, and tough-minded, could be disconcerting. He published comparatively few papers in academic journals. This was partly due to the fact that so much of his time was devoted to the teaching, encouragement, and guidance of others, but partly also to his rooted objections to the peer-group refereeing system, which was, in his view, all too often used as an opportunity for reviewers to attempt to impose their dissenting views – or worse still, their ignorance – anonymously on the writer.

The foregoing few sentences may make him sound rather humourless, but the reverse was the case. He could be serious when dealing with serious issues, but in everyday life he was delightful company: convivial, persuasive, funny, and occasionally wickedly impish. He was much loved as a teacher and a colleague, with a gift for a neat turn of phrase. He described the semantic lightness of frequent words as 'the blue jeans principle': the more you use them and wash them, the more the colour washes out. On the use of made-up examples as a source of evidence, he commented: 'One does not study botany by making artificial flowers.'

Patrick Hanks

Euralex 2008

The 13th EURALEX International Congress will be held 15-19 July 2008 in Barcelona, Spain. The Congess will be organized by the InfoLex Research Group at Pompeu Fabra University.

The EURALEX Congresses bring together professional lexicographers, publishers, researchers, software developers, and others interested in dictionaries of all types. The programme will include plenary lectures, parallel sessions on the topics listed below, software demonstrations, pre-congress tutorials and specialized workshops, a special session for students and work-in-progress, a book and software exhibition, and social events for participants and their guests. The congress in Barcelona will run from Tuesday afternoon through Saturday midday. The sessions will be held on the main campus of Pompeu Fabra University, which is centrally located and easily reached by public transportation.

Please see the congress website for further information: http://www.iula.upf.edu/agenda/euralex_08/index.htm.

Deadline for submitting proposals: 31 October 2007.

Contact: euralex2008@upf.edu.

LEXICOM-EUROPE 2007

A Workshop in Lexicography and Lexical Computing, Masaryk University, Brno, Czech Republic, June 4th-8th 2007, http://nlp.fi.muni.cz/lexicom2007.

Led by Sue Atkins, Adam Kilgarriff and Michael Rundell of the Lexicography MasterClass, Lexicom is an intensive one-week workshop, with seminars on theoretical issues alternating with practical sessions at the computer. There will be some parallel 'lexicographic' and 'computational' sessions. Topics to be covered include:

- Corpus creation
- Corpus analysis

  (1) Software and corpus querying
  (2) Discovering word senses, recording contextual information

- Frame Semantics and its application to lexicography
- Writing entries for dictionaries and lexicons
- Using web data

Applications are invited from people with interests and experience in any of these areas. Over the last seven years Lexicom workshops (in Europe and in Asia) have attracted well over 200 participants from 32 countries, including lexicographers, computational linguists, professors, research students,

translators, terminologists, and editors, managers and technical support staff from dictionary publishers and information-management companies.

The venue, Brno, the beautiful and ancient capital of Moravia, is the Czech Republic's second city. To register for Lexicom, go to: http://nlp.fi.muni.cz/lexicom2007. Early registration is advised. The workshop has been over-subscribed in previous years. Further details, including draft programme and reports of past events can be found at: http://www.lexmasterclass.com

Sue Atkins, Michael Rundell & Adam Kilgarriff

The Lexicography MasterClass

Forthcoming events

2007

June

13–16, Dictionary Society of North America, DSNA XVI, University of Chicago. Information: Erin McKean, 4907 N. Washtenaw Ave, Chicago IL 60625.

September

12–14, Ivanova State University, Russia: VII International school on lexicography. Information: Prof.. Dr. Olga Karpova, Ivanovo State University, English Philology Department, Ermak St., 39, Ivanovo, 153025, Russia, or Conference Coordinator Katerina A. Shaposhnikova. Tel.: +7 (0932) 37 54 02, fax: +7 (0932) 37 54 02, e-mail: lexico2005@ivanovo.ac.ru or omk@ivanovo.ac.ru.

2008

April

4, Troisième Journée québécoise des dictionnaires (Québec, Canada), sur le thème ''Les dictionnaires de langue française: de la Nouvelle-France au Québec contemporain''. Pour tous renseignements: monique.cormier@umontreal.ca.

July

15–19, Barcelona, Spain: 13th International EURALEX Conference. The conference will be hosted by Pompeu Fabra University. Deadline for receipt of abstracts: October 31, 2007. Please see the EURALEX website for details. Email: euralex2008@upf.edu.