International Journal of Lexicography Advance Access published July 5, 2008
doi:10.1093/ijl/ecn028

CORPUS-DRIVEN LEXICOGRAPHY

Ramesh Krishnamurthy: *Aston University, Birmingham, UK*
(r.krishnamurthy@aston.ac.uk)

Abstract

This paper discusses three important aspects of John Sinclair's legacy: the corpus, lexicography, and the notion of 'corpus-driven'. The corpus represents his concern with the nature of linguistic evidence. Lexicography is for him the canonical mode of language description at the lexical level. And his belief that the corpus should 'drive' the description is reflected in his constant attempts to utilize the emergent computer technologies to automate the initial stages of analysis and defer the intuitive, interpretative contributions of linguists to increasingly later stages in the process. Sinclair's model of corpus-driven lexicography has spread far beyond its initial implementation at Cobuild, to most EFL dictionaries, to native-speaker dictionaries (e.g. the *New Oxford Dictionary of English*, and many national language dictionaries in emerging or re-emerging speech communities) and bilingual dictionaries (e.g. Collins, Oxford-Hachette).

1. Linguistic evidence: from text to corpus

1.1 *Introduction*

John Sinclair believed that natural language use constituted the best source of linguistic evidence. Such use can only be found in authentic communicative texts. He believed in the importance of language as text (not as words or sentences), and therefore urged the inclusion of whole texts (not text extracts) in the corpus. Lexicography, operating at the level of lexis, involves the least degree of abstraction away from the text and therefore incurs the least accompanying loss of meaning. A corpus-driven approach involves a bottom-up methodology, beginning by selecting unedited examples from the corpus, identifying their shared and individual features, and only then grouping them for the purpose of lexicographic presentation. The degree of 'drivenness' applies simultaneously to the computational automation of the process, and the attendant withholding of human linguistic intuition.

The consequence of these pivotal concerns is a lexicography that serves as a commentary on examples (closer in function to the early manuscript glosses which launched the entire lexicographic enterprise), examples that are selected automatically for their typicality from authentic texts; and therefore also a lexicography that re-inspects the notion of lemma (headword), and dismantles and reconstructs the elements of traditional dictionary entries, in order to report more faithfully and accurately on the relationship between forms and meanings.

1.2 *The fundamental importance of text*

One of the most consistent themes emerging from Sinclair's writing, and evident even in his early publications, is the insistence that all linguistic study must start from text: 'Every morpheme in a text must be described both grammatically and lexically' (1966: 423) and 'evidence from text which is not a spontaneously produced, continuous stretch of natural language cannot be assumed to be reliable' (1970: 28). This culminates in: 'I have placed text more and more centrally during my career and paid ever-increasing attention to it . . . Now I have very little time for any work, including my own backlist, which is not rooted in the actual patterns of occurrence of words in text' (2007: 156).

1.3 *Terminology: from 'text' to 'corpus'*

In Sinclair (1966), the references to text are frequent: 'in a given stretch of text . . . items in a text . . . examining a text', but the importance of text as a source of evidence is assumed rather than explicitly stated, and the term *corpus* itself is absent. In Sinclair (1970: 23–24) the term *corpus* is used only in connection with the Brown University collection, and even that is more usually referred to as 'text'. In his 1987 lecture *The Dictionary of the Future*, he introduces the term cautiously: 'We called this our corpus' (1987b: 2), no doubt because the term was still not widely recognized in its computational sense outside specialist circles. The term finally becomes convention in the title of his 1991 book, *Corpus Concordance Collocation*.

1.4 *Corpus size*

The need for large quantities of text is a natural concomitant of Sinclair's decision to make lexis one of the foci of his attention: 'there is no easy way of collecting a few thousand occurrences of any lexical item' (1966: 412). He establishes this by practical research: 'our results show that a comprehensive description of English lexis would require a mammoth text' (1970: 8). In early works, given the technology available, he is uncertain about the exact amount: 'Valuable information can be obtained from quite short texts, but large quantities are desirable for full lexical descriptions. At the moment it is impossible even to guess what the optimum text length will be' (1970: 16). Large corpora are required for all aspects of language study, not just for lexis: 'It is, therefore, necessary to have access to a large corpus because the normal use of language is highly specific, and good representative examples are hard to find' (1991: 101).

1.5 *Sinclair: influences and 'schools'*

Sinclair acknowledges the influence of his mentors in directing his path towards both lexis and data, for example Angus Mackintosh: 'his interest in vocabulary was infectious, and his farsightedness guided me into corpus work and computing in 1960' (Sinclair 1991: xiii), and Michael Halliday: 'it was he who taught me to trust the text... he supported the start of corpus research in Edinburgh, and he encouraged my early attempts to understand lexis' (Sinclair 2004: vii). Sinclair firmly derives his work on collocation 'from an oft-quoted remark of J.R. Firth (1957): "Meaning by collocation is an abstraction at the syntagmatic level" ' (1970: 3). He also refers to Firth for 'colligation' (2004: 32) and 'context of situation' (2004: 102–103).

Carter (in Sinclair 2004: 2) says 'Sinclair is in a distinct tradition of British linguistics. This tradition owes much to the foundations built by Professor J. R. Firth in the 1950s and extended by Professor Michael Halliday in the 1960s' and 'Sinclair is firmly in the Firthian tradition'.

However, Sinclair himself is extremely wary of subscribing to the notion of 'linguistic schools': 'The formation of "schools" of linguistics is a constant danger, and one that I have been at pains to avoid, neither instituting one nor joining any. They relieve members from thinking for themselves, passing the burden up to the guru. Worse, they corral members inside a protective coating, so that uncomfortable ideas can be either ignored, or sanitised before circulation, or percolated through a fine mesh' (2007: 157).

1.6 *Text types*

The primacy of text is a core principle in literary studies, but Sinclair criticizes 'the assumption that one seriously studies only a literary text, and it is quaint and curious to examine any old piece of language' (1968: 82). He goes on to discuss the 'power of linking literary text to other text... the exceptionally detailed common ground assumed in the minutiae of linguistic analysis' (ibid: 88), and says 'we examine a text with reference to a system of agreed signals which constitute the most complex of all our social relationships' (ibid.). This reflects the general shift in language study away from specialized texts (centuries ago, solely religious texts; then 'elite texts', as in Johnson's dictionary, in which the examples were selected from 'the best writers').

1.7 *Intuition and text*

Sinclair's focus on authentic, communicative text is in marked contrast to Chomsky's assertion that 'Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogenous speech community, who knows its language perfectly' (Chomsky 1965: 3).

Sinclair rarely criticized this viewpoint directly, but often cautioned against its consequences. For example, when discussing sentences created by informants to exemplify particular words, ' "synthetic" text has more clearly-marked lexical patterning. One's suspicions are aroused by this, for it suggests that such text is evidence of people's linguistic behaviour rather than their language behaviour' (1970: 8), and 'Starved of adequate data, linguistics languished . . . It became fashionable to look inwards to the mind rather than outwards to society . . . The communicative role of language was hardly referred to' (1991: 1) and 'Students of linguistics over many years have been urged to . . . prefer their intuitions to actual text where there was some discrepancy. Their study has, therefore, been more about intuition than about language' (ibid: 4). Even here, Sinclair mitigates his views: 'It is not the purpose of this work to denigrate intuition – far from it. The way a person conceptualizes language and expresses this conceptualization is of great importance and interest precisely because it is not in accordance with the newly observed facts of usage' (ibid.).

Other scholars view the 'discrepancy' as much more profound or divisive (e.g. Carter in Sinclair 2004: 2, Léon 2005, De Beaugrande 2000). Certainly Chomsky's recent remarks seem extreme by comparison: 'Corpus linguistics doesn't mean anything . . . I don't pay much attention to it. I don't see much in the way of results' (Andor 2004: 97).

2. Lexicography

2.1 *Lexis as an independent level*

Sinclair says 'I am not here going to defend the setting-up of lexis as an independent part of language form. That is done in a paper presenting the present approach to lexis by M. A. K. Halliday elsewhere in this volume' (1966: 1). However, he takes nothing for granted, and reminds us of the basics from time to time: 'A word is a somewhat arbitrary string of signs' (1970: 5) and 'the set of four choices a b c k, arranged in the sequence b a c k with no spaces, is an important linguistic event in its own right, long before it is ascribed a word-class or a meaning' (1991: 117).

2.2 *Lexis and collocation*

Sinclair readily points out the importance of the combinatorial and structural elements involved: 'lexis . . . describes the tendencies of items to collocate with each other... such tendencies cannot be described in terms of small sets of choices... lexical items do not contrast with each other in the same sense as grammatical classes contrast' (1966: 411). He is aware of the weak theoretical situation of lexis: 'At the present time, lexical statements look very much weaker than statements made using the precise and uncompromising machinery of grammar. There is a much less elaborate framework to lexical descriptions and much less certainty in the statements' (1966: 411–412). But he remains optimistic: 'The theory of lexis opens up exciting areas for describing language more accurately and more usefully' (1966: 429). The lexical statement becomes his basic unit of lexicography.

2.3 *Lexis and meaning*

Sinclair gives us a detailed definition of lexis 'The lexis of a language is the set of all its word-forms (q.v.). Lexical structure is the organization of these word-forms into units such as collocations (q.v.) and idioms (q.v.). The origins of lexical evidence are the word-patterns in texts. Position and frequency of occurrence are important criteria, and in the first instance priority is given to the formal arrangements of word-forms, rather than to intuitive conclusions about meaning. This distinguishes lexis from semantics, which is centrally concerned with the organization of meaning' (1991: 174). Sinclair distinguishes between *word-form* (the formal unit) and *lexical item* (the unit of meaning), a vital issue for lexicography. One consequence of giving collocation a high priority is that 'The meaning of words chosen together is different from their independent meanings. They are at least partly delexicalized' (2004: 20). Nevertheless, he concedes: 'The starting point of the description of meaning in language is the word. This is one of two primitives in language form, the other being the sentence. The sentence is the unit that aligns grammar and discourse, and the word is the unit that aligns grammar and vocabulary' (ibid: 24).

2.4 *Lexis and grammar*

Sinclair states early on that 'The two interpenetrating ways of looking at language form are **grammar** and **lexis**' (his emphasis; 1966: 411). Grammar regards language as a 'large number of separate choices, each choice being from a small list of possibilities' (ibid). Grammar can help to describe lexis, but both are necessary for language description: 'Every morpheme in a text must be described both grammatically and lexically... Each successive form in a text is a lexical item or part of one, and there are no gaps where only grammar is to be found' (1966: 423).

Sinclair emphasizes the inseparability of grammar and lexis: 'Is it wise to divide language patterning into grammar and something else (be it lexis or semantics or both) before considering the possibility of co-ordinated choice?' (1991: 3). He is critical of over-reliance on grammar 'Students of grammar are often victims of the "all or nothing" argument, which does not allow a few exceptions to a pronounced tendency. Students of lexis in the early days were made to feel that this kind of statistical evidence was somehow not as good as the wholesome, contrived rules of grammar. Now it is manifest that the nature of text is not to follow clear-cut rules, but to enjoy great flexibility and innovation' (ibid: 6). He concludes that 'grammar is part of the management of text rather than the focus of the meaning-creation' (ibid: 8), and 'adjustment of meaning and structure is a regular feature of a language. It can be used to provide valuable evidence for lexicography, suggesting sense divisions, and identifying phrase units with distinctive patterning. Then, by using the same evidence in reverse, the traditional domain of syntax will be invaded by lexical hordes' (ibid: 65).

Sinclair criticizes the traditional linguistic dichotomies, especially between grammar and lexis: 'It is, therefore, unnecessary to make a sharp distinction between abstract and actual language structure – the sort of distinction embodied in Saussure's langue and parole or Chomsky's competence and performance' (ibid: 103). He goes further: 'It is folly to decouple lexis and syntax, or either of those and semantics' (ibid: 108). The problem is that 'Virtually all grammars are constructed on the open-choice principle' (ibid: 109), whereas 'the principle of idiom is far more pervasive and elusive than we have allowed so far' (ibid: 111) and 'at least as important as grammar in the explanation of how meaning arises in text' (ibid: 112). The point is that 'open-choice is a process which goes on in principle all the time, but whose results are only intermittently called for' (ibid: 114).

The *Collins Cobuild English Grammar* (Sinclair et al. 1990) fulfilled Sinclair's vision of using corpus evidence to populate syntax with 'lexical hordes'. And the *Collins Cobuild Grammar Patterns 1: Verbs* (Sinclair et al. 1996) and *2: Nouns and Adjectives* (Sinclair et al. 1998) finally 'recoupled' lexis, syntax, and semantics.

2.5 *Lexical item*

Sinclair asserts that 'A lexical item is a unit of language representing a particular area of meaning which has a unique pattern of co-occurrence with other lexical items. It cannot always be identified with an orthographic word' (1970: 9) and lists morphemes, homographs, paradigmatically associated words (e.g. *kick, kicks, kicking, kicked*), syntagmatically associated words (*run to seed*), and multiverbal items (e.g. idioms) as candidates. Later, he says 'single words chosen on open-choice principles, that leave no trace of their use, are examples of the limiting case of lexical item' (2004: 39).

2.6 *Dictionaries*

Sinclair says ‘It might sound strange to suggest that great difficulties lie in the way of anyone who wants to study the vocabulary of a language. One could point to the great dictionaries . . .’ (1966: 410) and continues to refer frequently in his work to dictionaries, their functions, and their importance: ‘A dictionary is the only really successful reference tool for language’ (1987b: 5); ‘Mostly, utterances are working in the world to get things done. The utterances in dictionaries are concerned with the inner world of language. Insulated from the need to do anything, the dictionary statements are part of a discourse which is not directly about the world at all’ (ibid: 6); ‘A glance at any dictionary will confirm the status of the word as the primary unit of lexical meaning’ (2004: 25).

Sinclair’s criticizes traditional dictionaries for their limited functionality: ‘“Ordinary” dictionaries are designed primarily to offer support to the reader, and not the writer. Given that restricted objective, it is not necessary to state the limits and constraints on structure and usage. All that is required is recognition criteria’ (1987a: 106); and for their selection criteria: ‘sparing in citations of the inflected forms of words, and rather generous in citing the derived forms’ (1987a: 104), despite the fact that ‘Most of the derived forms are regular, and many have little prospect of ever being used’. He takes issue with ‘the classic indecision of dictionaries about transitivity, enshrined in the meaningless message *v.t. + i.*’ (1985: 14–15), and observes that ‘Frequently grammar and etymology have guided the choice of entry division, and semantics the subclassifications’ (1970: 5).

Sinclair praises ‘Those dictionaries that stand as milestones in our cultural history use real citations: Dr Johnson’s Dictionary of 1755 and the Oxford Dictionary begun by Murray in 1878. They understood that the dictionary is really just a commentary on the examples’ (1987b: 2) and hence criticizes ‘the widespread custom of made-up examples . . . when the examples are concocted by the same lexicographer, they have no value at all . . . usage cannot be thought up – it can only occur’ (1984: 3). He is also critical of many dictionary conventions (reliance on partial forms, abbreviations, and various symbols): ‘How good a sample of the language is the dictionary itself? Is it even written in the language it purports to describe?’ (1987b).

2.7 *Lexicography*

Sinclair promoted lexicography as the legitimate mode of expression of linguistic description at the lexical level. However, he acknowledged its deficiencies: ‘At present, lexicography is a group of specialised skills, a body of received practical wisdom . . . There is no overt rationale, and controversies rage in the gulf between principles and practice, with no sign of resolution’ (1984: 1). He urged that ‘Evidence of secondary sources and the evidence of introspection should be brought in at a late stage in the process of compilation . . . the initial evidence should always be . . . from the observation of language in use' (1985: 3), adding that 'the most favourable point for the operation of introspection' was 'in evaluating the evidence rather than creating it' (ibid.).

He bemoans the fact that 'the lack of external standards of evaluation narrows the range of possible work done as lexicography, causes it to be introspective and conservative. Its security lies essentially in repeating successful practice, and it is highly resistant to innovation, experiment, or even discussion outside the small group of established practitioners' (1984: 4). He sees lexicography as operating 'at the intersection of Linguistics and Information Technology' (1984: 6), then adds a third factor, Experience (incorporating principles and practice), and suggests that 'lexicography may come to be regarded as a special variety of computational linguistics' (ibid: 8). He notes that 'Lexicography is one of the places where language study meets the general public: there are few enough of these, and most of the others do not have the same high standards' (1984: 13).

3. Corpus-driven lexicography

3.1 *Computer technology*

The computer now occupies a central role in most of our lives, but as long ago as 1966, Sinclair had forecast that in the formal study of vocabulary, 'all sorts of problems lie ahead, problems which are not likely to yield to anything less imposing than a very large computer' (1966: 410), because 'the patterns perceived by a trained linguist examining a text are unreliable' (ibid: 413).

He later extols 'the ability of computers to organise abundant textual evidence. The storage and scanning of very long texts provides a close-to-objective basis on which language patterns can be observed' (1984: 13). As the technical problems in creating large electronic corpora are gradually resolved, he states that 'The selection process becomes a selection of texts for the corpus, not instances for a dictionary. Once it is decided to include a text, then all the instances of all the words constitute the evidence' (1985: 4). He expresses great satisfaction with the new technology: 'The quality of evidence about the language which can be provided by concordances is quite superior to any other method; once lexicography takes full advantage of this evidence, it will be impossible to go back to a reliance on pre-computational techniques' (1985: 7).

3.2 *Initial problems*

Sinclair is wary of merely applying technology to traditional ideas: 'those varieties of computational linguistics which used to ignore corpus evidence have quite dramatically switched in recent years in their attitude to corpora, but have retained models of language which are not justified by the evidence they now have' (1991: 22). Meanwhile, he notes other practical problems that have started to emerge: 'there are some words that occur too often, and some that do not occur enough. Consequently, there is only a central set of words for which the evidence is both comprehensive and convenient' (1985: 8) and 'it seems that no set of texts, no matter how extensive, provides enough evidence for the description of its own vocabulary. There is always a huge tail of words which have only a handful of occurrences' (1987a: 152).

Grammar is a particular problem: 'Published grammars are much too general... and words are stubbornly individual... when looking at individual words... many people found it difficult not to generalise well beyond what was attested' (1987a: 107). Sinclair concedes that 'there is no prospect of either perfection or neutrality in this work' (ibid.).

3.3 *The COBUILD project*

Looking Up documented the new lexicographic methodology utilised in the Collins-funded COBUILD project at Birmingham University, and Sinclair celebrates 'the ability to get for the first time a view of the language which is both broad and comprehensive... Cobuild has created the first wholly new dictionary for many years' (1987a: vii). The presentation represented 'a fairly sharp break with traditional lexicography' (ibid: viii). He is especially pleased with the examples: 'We concentrate on real examples drawn from the corpus, with little or no editing. This is because no-one yet knows what breathes life into English which occurs naturally, and we are not rash enough to suggest that it is better to concoct examples than to select from our rich store' (1987b: 2). He emphasizes 'how carefully the language is patterned... how the description is very sensitive to the number of instances of a form' (1987a: 150). One of the key findings is that 'the whole drift of the historical development of English has been towards the replacement of words by phrases, with word-order acquiring greater significance' (ibid.). Hanks (2004, forthcoming) has developed the analytical methodology further in his 'Corpus Pattern Analysis'.

3.4 *The future*

In *Looking Up*, Sinclair forecasts that 'As statistical methods improve, the extent and reliability of the linguistic statements will increase. As computational processing improves, the quantity and quality of the linguistic evidence will increase... A fully automatic dictionary is at the design stage' (1987a: 152). Church and Hanks (1989) led the way towards the improvement of statistical methods - initially using Mutual Information (*MI;* later, *t-score* was added – see

Church et al. 1991, 1994). *MI* was the basis for Sketch Engine (Kilgarriff et al. 2004), a user-friendly statistical tool which operates on any part-of-speech tagged corpus in any language. For English, it lemmatizes the words, selects the most common and important syntagmatic relations, and displays the most statistically salient lexical items in each relation (e.g. adjective + head noun; subject / object / prepositional object of verbs). It was used extensively for the Macmillan English Dictionary (Rundell 2002).

In *The Dictionary of the Future*, Sinclair envisages the dictionary as ‘a device through which the user will observe the living language. Not the frozen fillets of the printed citations, nor the stuffed dummies of the made-up examples, but the language as it is when it is being used . . . language *through* the dictionary . . . the next target for progressive lexicography’ (1987b: 5). In *Trust the Text*, he says ‘models that arise from corpus-driven studies . . . have a holistic quality that makes them attractive. The numerical analysis is aligned closely with meaningful analysis; lexis and grammar are hardly distinguished, surface and abstract categories are mixed without difficulty. As a result some of the problems of conventional description are much reduced – for example there will be little word-based ambiguity left when this model has been applied thoroughly’ (2004: 39).

3.5 *Conclusion*

Elena Tognini-Bonelli (2001: 84) describes corpus-driven linguistics thus: ‘The corpus, therefore, is seen as more than a repository of examples to back pre-existing theories or a probabilistic extension to an already well defined system. The theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus. Indeed, many of the statements are of a kind that are not usually accessible by any other means than the inspection of corpus evidence. Examples are normally taken verbatim . . . recurrent patterns and frequency distributions are expected to form the basic evidence for linguistic categories; the absence of a pattern is considered potentially meaningful’.

Similarly, corpus-driven lexicography does not use a corpus to find examples to fit pre-existing entries; the new entries, sense divisions, and definitions are fully consistent with, and reflect directly, the evidence of the corpus; examples are used verbatim; recurrent patterns form the basis for lexicographic categories; and the absence of an entry, or a pattern in an entry, is a meaningful lexical statement.

3.6 *Postscript*

The corpus-driven nature of Sinclair’s work was not confined to lexicography, but applied to all of his linguistic thinking: ‘The advent of the corpus has been the most thrilling development in language study during my career, and much of my work has been in celebration of this bountiful resource . . . the corpus has things to tell me, and I try to work out where it is heading. I have been surprised at the confidence of so many scholars, who seem to think that they have something to tell the corpus. While my position always runs the risk of appearing naïve, of reinventing the wheel or stating the obvious, I feel that I am on safer ground . . . The fact that the theories available to me did not alert me at all to the strongly recurrent patterns found in a corpus nor explained them when they emerged caused me to view theories with increasing suspicion' (2007: 157).

References

A. Dictionaries.

Pearsall, J. and Hanks, P. (eds.) 1998. *New Oxford Dictionary of English*. Oxford:OUP.

Rundell, M. (ed.) 2002. *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan.

Sinclair, J.M., Hanks, P., Fox, G., Moon, R., Stock, P. (eds.) 1987. *Collins COBUILD English Language Dictionary*. London: Collins ELT.

B. Other Literature.

Andor, J. 2004. 'The master and his performance: An Interview with Noam Chomsky.' *Intercultural Pragmatics* 1/1: 93–111.

Bazell, C.E., J.C. Catford, M.A.K. Halliday and R.H. Robins (eds.) 1966. *In Memory of J.R. Firth*, London: Longman.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*, Cambridge, MA: MIT Press.

Church, K. and Hanks, P. 1989. 'Word Association Norms, Mutual Information, and Lexicography', in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*: 76–83; reprinted in *Computational Linguistics*, 16:1, 1991: 22–29.

Church, K., Gale, W., Hanks, P. and Hindle, D. 1991. 'Using Statistics in Lexical Analysis' in U. Zernik (ed.), *Lexical Acquisition: Using On-line Resources to Build a Lexicon*. Mahwah, NJ: Lawrence Erlbaum Associates, 115–164.

Church, K., Gale, W., Hanks, P., Hindle, D. and Moon, R. 1994. 'Lexical Substitutability' in B. T. S. Atkins and A. Zampolli (eds.), *Computational Approaches to the Lexicon*. Oxford: Oxford University Press, 153–177.

De Beaugrande, R. 2000. *Functionalism and Corpus Linguistics in the 'Next Generation'*. Available at http://www.beaugrande.com/Functionalism%20and%20Corpus%20Linguistics.htm Accessed on 01/04/08.

Firth, J. R. 1951. 'Modes of meaning.' In Firth, J.R. 1957: 190–215.

Firth, J. R. 1957. *Papers in Linguistics 1934-1951*, London: Oxford University Press.

Hanks, P. 2004. 'Corpus Pattern Analysis', in Geoffrey Williams and Sandra Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress*, EURALEX 2004, Université de Bretagne-Sud, 87–97.

Hanks, P. (forthcoming) *Norms and Exploitations: Corpus, Computing, and Cognition in Lexical Analysis*. MIT Press.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. 2004. 'The Sketch Engine', in G. Williams and S. Vessier (eds.) *Proceedings of the Eleventh EURALEX*

International Congress, EURALEX 2004, Université de Bretagne-Sud: 105–116. Reprinted in P. Hanks (ed.) 2007. *Lexicology: Critical Concepts in Linguistics*, London: Routledge, 230–242.

Krishnamurthy, R. (ed.) 2004. *English Collocation Studies: The OSTI Report* by John Sinclair, Susan Jones and Robert Daley, London: Continuum.

Léon, J. 2005. 'Claimed and Unclaimed Sources of Corpus Linguistics', *Henry Sweet Society Bulletin*, 44: 36–50; reprinted in W. Teubert and R. Krishnamurthy (eds.) 2007: 326–341.

Sinclair, J. M. 1966. 'Beginning the study of lexis' in Bazell et al. (eds.) *In Memory of J.R. Firth*, 410–430.

Sinclair, J. M. 1968. 'English Language in English Studies', *Educational Review* 20, 82–94.

Sinclair, J. M. 1970. *English Lexical Studies*, report to the Office of Scientific and Technical Information; published as R. Krishnamurthy (ed.) 2004.

Sinclair, J. M. 1984. 'Lexicography as an academic subject.', In R.R.K. Hartmann (ed.) *LEXeter 83 Proceedings*, Lexicographica Series Maior No 2, Tubingen: Max Niemeyer Verlag, 3–12.

Sinclair, J. M. 1985. 'Lexicographic Evidence.' in Ilson, R. (ed.) *Dictionaries, Lexicography and Language Learning*. ELT Documents 120, Pergamon, 81–94.

Sinclair, J. M. (ed.) 1987a. *Looking Up. An account of the COBUILD project in lexical computing*. London: Collins ELT.

Sinclair, J. M. 1987b. *The Dictionary of the Future*. Collins English Dictionary Annual Lecture. University of Strathclyde, 6 May 1987.

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

Sinclair, J. M. 2004. *Trust the Text – Language, corpus and discourse*, London: Routledge.

Sinclair, J. M. 2007. Preface, *International Journal of Corpus Linguistics*, 12/2: 155–157.

Sinclair, J. M., Fox, G., Bullon, S., Krishnamurthy, R., Manning, E. and Todd, J. (eds.) 1990. *Collins Cobuild English Grammar*. London: Collins ELT.

Sinclair, J. M., Fox, G., Francis, G., Hunston, S., Manning, E. (eds.) 1996. *Collins Cobuild Grammar Patterns 1: Verbs*. London: HarperCollins Publishers.

Sinclair, J. M., Francis, G., Hunston, S., Manning, E. (eds.) 1998. *Collins Cobuild Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins Publishers.

Teubert, W. and Krishnamurthy, R. (eds.) 2007. *Corpus Linguistics: Critical Concepts in Linguistics. Volumes 1–6*. London & New York: Routledge.

Tognini-Bonelli, E. 2001. 'The corpus-driven approach'. Chapter 2 in *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John Benjamins pp. 84–100; eprinted in W. Teubert and R. Krishnamurthy (eds.) 2007: 74–92.