MEMORIAL ARTICLE: JOHN SINCLAIR (1933–2007)

The Search for Units of Meaning: Sinclair on Empirical Semantics

MICHAEL STUBBS

Universität Trier, Germany

John McHardy Sinclair has made major contributions to applied linguistics in three related areas: language in education, discourse analysis, and corpus-assisted lexicography. This article discusses the far-reaching implications for language description of this third area. The corpus-assisted search methodology provides empirical evidence for an original and innovative model of phraseological units of meaning. This, in turn, provides new findings about the relation between word-forms, lemmas, grammar, and phraseology. The article gives examples of these points, places Sinclair’s work briefly within a tradition of empirical text analysis, and identifies questions which are currently unanswered, but where productive lines of investigation are not difficult to see: (1) linguistic-descriptive (can we provide a comprehensive description of extended phrasal units for a given language?) and explanatory (what explains the high degree of syntagmatic organization in language in use?), and (2) socio-psychological (how can the description of phrasal units of meaning contribute to a theory of social action and to a theory of the ways in which we construe the social world?).

John McHardy Sinclair (14 June 1933–13 March 2007) contributed significantly to three central areas of applied linguistics: language in education, discourse analysis, and corpus-assisted lexicography. Throughout all this work, his method of linguistic analysis was to search for patterning in long authentic texts, and he argued consistently against the neglect and devaluation of textual study in much recent linguistics.

In the 1960s, his early corpus work followed the principle that conversation is ‘the key to a better understanding of what language really is and how it works’ (Firth 1935: 71), and argued that spoken English would provide evidence of ‘the common, frequently occurring patterns of language’ (Sinclair et al. 1970/2004: 19). In the 1970s, his work on audio-recorded spoken language in school classrooms described characteristic units of teacher–pupil dialogue, and developed structural categories for analysing long texts, as opposed to the short invented sentences which were in vogue at the time (Sinclair and Coulthard 1975). In the 1980s and 1990s, he studied patterning which is visible only across machine-readable corpora of hundreds of millions of running words (Sinclair 1991, 2004a). This led to his theory of phraseology
and to the Cobuild series of language reference materials for advanced learners of English. The first Cobuild dictionary appeared in 1987 and was followed by further dictionaries, grammars and teaching materials, as well as detailed discussions of their theoretical principles (Sinclair (ed.) 1987; Barnbrook 1996).

Accounts of Sinclair’s life and work have appeared in several places,¹ and his publications are comprehensively listed in the Supplementary Data, which accompanies the on-line version of this article. I will not repeat this material here, but will concentrate on his theory of phraseology. It is this which probably has the most far-reaching implications for language description and language teaching.

EMPIRICAL SEMANTICS: TECHNOLOGY AND OBSERVATION

Sinclair’s work draws on the British tradition of empirical text analysis developed by Firth, Halliday, and others, but it is Sinclair who develops the most radical implications, namely that multi-word units of meaning can be discovered by observing recurrent patterns across large text collections. His best known work is based on ‘a new view of language and the technology associated with it’ (Sinclair 1991: 1), and on the observation that ‘the language looks rather different when you look at a lot of it at once’ (Sinclair 1991: 100). Technology leads to a radically revised perception of the object of study for linguistics, because it becomes possible to observe patterns of language use which are otherwise invisible. Sinclair is one of the very few linguists who has discovered many things which people had simply not noticed, despite thousands of years of textual study—because they are observable only with the help of computer techniques which he helped to invent. His discoveries are due to a number of related principles and insights.

First is his plain text policy (Sinclair 2004a: 190–1). In contrast to much other corpus work, Sinclair’s principle is to rely as little as possible on annotating the data, since grouping word-forms into lemmas and part-of-speech categories can hide the patterns. In the worst case, a linguist tags data with categories from pre-corpus studies, searches the data, and finds the categories, in a vicious methodological circle. A widespread belief, that raw text is not amenable to systematic analysis, led to Saussure’s despair of studying parole and to Chomsky’s withdrawal into introspection. Sinclair has shown that we now have the technology to discover patterns in raw textual data—and that these patterns are rather different from what we previously thought. His slogan for the approach is ‘trust the text’ (Sinclair 2004a).

Second, as Sinclair (2008: 31, 38) points out, although linearity, directionality, and redundancy are all obvious features of text, they are largely ignored in linguistic theory. The role of paradigmatic choice has often been over-estimated, and syntagmatic constraints on linear sequences correspondingly underestimated. A major motivation for his approach is therefore to remedy this neglect of syntagmatic organization (Sinclair and Mauranen 2006: xviii).
In an individual text, we can observe neither repeated syntagmatic relations nor any paradigmatic relations at all, but it is precisely these two things which concordances make visible (Tognini-Bonelli 2004: 18). An individual text is designed to be read as a whole, linearly, from left to right. A concordance is designed to be read as a series of fragments, vertically, from top to bottom.

A concordance brings together utterances which have been produced at different times by different speakers, makes visible recurrent patterns, and allows us to count them. It freezes instances of language behaviour, and provides public data which can be studied at leisure (Mauranen 2004: 103). This is analogous to the effect of a camera, which can freeze a series of events in time, thereby turning a process into a product, which is publicly accessible to different observers. The development of the natural sciences was made possible by the powers of observation which were, in turn, made possible by the invention of the lens, and hence the telescope, the microscope, and the camera (Macfarlane and Martin 2002). Presumably no-one these days claims that observations can entirely avoid subjectivity. We select what to study because we have ideas about what is interesting, and different methods of observation give access to different phenomena. Nevertheless, corpus data allow us to study language ‘with a degree of objectivity…where before we could only speculate’ (Kilgarriff 1997: 137).

It is sometimes objected that corpus methods over-emphasize quantitative aspects of language use. However, no language teacher could conceivably ignore frequency in presenting vocabulary, since the distribution of units in language use is so uneven. A few words occur once every few seconds or so, but most words occur only once every few weeks or so (as Sinclair once summarized Zipf’s law to me). For example, ten of the high frequency words in English (the, of, and, to, a, in, that, is, was, it) make up over 20 per cent of most running text. Sinclair (e.g. 1999a) often points out that linguistic theory underestimates the importance of high frequency words. In any case, observational tools are designed to reveal something which we could not otherwise see, such as recurrent collocations. The notion of what counts as the same in a language system is a foundational question for linguistics, and a central task is therefore to discover the units (morphemes, phonemes, distinctive features, etc.) which regularly recur, and which are the ‘very essence of language’ (Harris 1988: 19).

In summary: The computer-assisted methods which can help in discovering patterns of language use involve new observational techniques, they are quantitative, and they provide a way of studying the relation between paradigmatic oppositions and syntagmatic constraints. Major theoretical advances have often come when linguists have realized the significance of different forms of data. Corpora are just data and quantitative methods are just methods, but their combination has led to a major shift in theory, and it is this theory which has to be evaluated.
A concordance presents parallel cases in tabular form—typically dozens or hundreds of instances—and this visual display is part of the argument (Fahnestock 2003: 142). The methods based on the use of concordance data are often thought of as ‘induction’: many observations (down the vertical axis of a concordance) lead to a generalization. However, we are always working with a sample drawn from an indefinitely large population whose parameters of variation are unknown. If we increase the size of the sample, we increase its diversity, but no matter how large the sample becomes, we will never know whether it includes all the patterns of the language in their typical proportions. Indeed, because of differences between text collections due to genre and topic, it is most unlikely that this will be the case (Sinclair 2008: 30).

It is therefore useful to distinguish between ‘induction’ and what is sometimes called ‘eduction’. Induction reasons from particulars to the general, and is notoriously susceptible to refutation by counter-example. Eduction reasons from particulars to more particulars, and the concept of a single counter-example does not apply. Many observations of a pattern lead to a prediction that similar patterns will be observed in the future. But absolutely fixed patterns are extremely rare, and a frequent conclusion is that a given pattern is ‘typical’ or ‘canonical’, but that it has variants. Fahnestock discusses these points in detail, and shows how parallelism across both textual and visual arrays constitute ‘schematic presentations of evidence in an argument’ (2003: 139).

‘THE SEARCH FOR UNITS OF MEANING’: OSTI TO COBUILD

It is simplistic to pick out just one theme in all this work, but ‘the search for units of meaning’ (the title of Sinclair 1996) is central, and signals a defining problem in linguistics. In the 1960s, Sinclair et al. (1970/2004: 3) had asked how objective data and subjective meaning are related: ‘(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’s early corpus work, informally published as the OSTI Report (UK Government Office for Scientific and Technical Information, Sinclair et al. 1970/2004), describes quantitative research on computer-readable data, carried out between 1963 and 1969, but not formally published until 2004. 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 the OSTI spoken corpus of 135,000 running words. The project was in touch with work by Francis and Kučera (1967) on their one-million-word corpus of written American English, but little other comparable work was available.

Yet the OSTI Report explicitly formulated many ideas which are still central to corpus linguistics. The unit of lexis is unlikely to be the word in all cases. 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). There is a relation ‘between statistically defined units of lexis and postulated units of meaning’ (Sinclair et al. 1970/2004: 6). As Sinclair 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’. This tradition of computer-assisted language analysis was concerned, from the beginning, with a theory of meaning, and it provides the context for Sinclair’s ambitious aim: ‘the ultimate dictionary’ would list all the lexical items in the language with their possible variants (Sinclair et al. 1970/2004: xxiv).

Because machines in the 1970s were not powerful enough to handle large quantities of data, the work was shelved, and started again in the 1980s as the COBUILD project in corpus-assisted lexicography (Sinclair (ed.) 1987; Moon 2007). Sinclair’s long-term vision of linguistics was formulated in the 1960s, and he then waited till the technology—and everyone else’s ideas—had caught up with him.

CO-SELECTION: LEMMAS AND WORD-FORMS

Sinclair proposes co-selection as a central descriptive mechanism of language in use. Examples of lexical co-selection which tend to spring to mind are cases where a relatively infrequent word has a strong tendency to co-occur with a restricted set of collocates. Boggle usually co-occurs with mind, or quaff usually co-occurs with beer or wine. More generally, collocation is often thought of as a relation between two lemmas. For example, different forms of the lemmas hard–work, play–role, strong–argue and heavy–rain can co-occur as follows:

- hard work, hard-working, works very hard, to work harder, a hard worker
- play a role, play a key role, the role played by, role-play, a new role to play
- a strong argument, strongly argued, the argument will be strengthened
- heavy overnight rain, heavy autumn rains, heavy rainfall, raining heavily

However, there are frequently restrictions on the forms of the lemma. For example, the lemmas heavy–drink co-occur as heavy drinker and drink heavily, but not as *heavily drunk.

In addition, as Greaves (2007) and Warren (2007) point out, in the collocation play–role the noun is often preceded by roughly synonymous adjectives. This provides evidence of a longer pattern, in which the adjective signals the speaker’s evaluation:

- play a(n) <central, crucial, important, key, leading, major, pivotal, vital> role
Greaves, Warren, and Cheng (Cheng et al. 2006, 2009) have developed software (illustrated below) to investigate empirically the extent to which collocational units can vary on three dimensions: the forms of the lemmas, their relative position to left and right of each other, and their distance from each other.

One of the clearest findings of corpus analysis is that different forms of a lemma often have quite different frequencies and collocates, and therefore different meanings. Sinclair (2004a: 31) notes that plural eyes and singular eye have little overlap in their top 20 collocates: blue and brown collocate only with plural eyes, and singular eye occurs in expressions to do with visualizing and evaluating.

- KEEP an eye on; TURN a blind eye to; in the public eye; with the naked eye; as far as the eye can see, in his mind’s eye; more than meets the eye; KEEP an eye out for

Similarly, the collocation seek-asylum occurs in various forms (asylum seekers, seeking asylum, etc.). However, different forms of seek co-occur with different collocates. In a 200-million-word corpus, I studied the 20 most frequent collocates of the different word-forms (Stubbs 2001: 27–8). The forms seek, seeking and sought all shared the collocates asylum, court, government, help, political, support. The forms seeks and seek shared only one collocate: professional. And the pairs seeks/sought and seeks/seeking had no shared collocates. The word-form seeks is frequent in lonely hearts ads, where its frequent collocates include attractive, black, caring, female, guy, lady, male, man, professional, similar:

- female 31, single, seeks well educated gentleman

The overlap in the collocates provides a measure of the semantic distance between the word-forms. Three word-forms (seek, seeking, sought) often occur in texts on political and legal topics, with collocates from the semantic field of “help and support”, but the word-form seeks is only distantly attached to this cluster. How lemmas and their forms should be divided between headwords in dictionaries is an empirical question.

CO-SELECTION: LEXIS AND GRAMMAR

Sinclair (1992: 14) illustrates co-selection of lexis and grammar with the example lap. In its body-part sense, the word does not occur as the subject or object of a verb, but only in prepositional phrases: her knitting lying in her lap, he lifted the cat off his lap. In other phraseology, it has different meanings (e.g. we’re on the last lap now). It therefore makes little sense to ask what the individual word means, since this depends on how it is used in different phrases and grammatical constructions.

The word ebb provides a similar example. It occurs around 250 times in the British National Corpus (BNC). Only a few uses are literal (rocky reefs exposed by
the ebb tide). Most occurrences are of the noun, most frequently in phrases such as a low ebb and ebb and flow. The descriptive problem lies in the variants:

- at (such) a low ebb; at this low ebb
- at an/a <appalling, extremely, fairly, particularly, very> low ebb
- at <her, his, one’s, its, the> lowest ebb

The most frequent form is at a low ebb, but the unit is more abstract and has indeterminate boundaries. The collocation at-Low-ebb is typically used to talk about people’s morale or spirits, which are at a lower ebb than for some time in the past. The most frequent verb is be, but a few other verbs occur:

- her spirits were at their lowest ebb
- with teachers’ morale at its lowest ebb in living memory
- at their lowest ebb for 20 years
- at its lowest ebb in history
- staff levels have reached a low ebb
- credit had sunk to its lowest ebb

This is a clear example of a word occurring in restricted patterns, typically in a unit which is ‘a single lexical choice whose realization is six or seven words long, and within which there is some variation’ (Sinclair 2004b: 290). There is no possible paradigmatic contrast between definite and indefinite articles, or between high and low. In practical terms, there is little point in knowing the word ebb, without knowing its phraseology. In theoretical terms, the problem is to establish the internal variability and external boundaries of the phrasal unit. The membership of the category (as with most linguistic categories) is a matter of degree.

Taylor (2002: 102) discusses the grammatical pattern N1 by N1. Examples in the BNC which occur 20 times or more are:

- step by step (165), day by day (117), year by year (105), bit by bit (47), week by week (32), line by line (31), case by case (30), month by month (24), stage by stage (21), inch by inch (20)

The lexis usually denotes small units of time or space. The phrase day by day is frequent, decade by decade much less frequent (3), and century by century does not occur in the BNC, although a search of the worldwide web provides examples such as

- it was possible, century by century, to follow the town’s urban development

The construction has a few conventionalized exponents, but it is the idiomatic pattern itself which carries the meaning of a gradual, steady, often deliberate and methodical process. It is not possible to give a definitive list of its lexis, since it is partly productive:

- she washed a Cos lettuce, leaf by leaf
- they worked their way up floor by floor
The observation that different words occur in different grammatical patterns leads to the idea of contrasting a global with a local grammar (Gross 1993). Traditional grammatical models have never been able to cope with everything which occurs in all varieties of text. So, what is required is descriptions of specific areas of language, where specific meanings are expressed. Hunston and Sinclair (2000) discuss the language of evaluation and show that evaluative adjectives often occur in the pattern:

- there + link verb + something + adjective + about

The BNC has over 190 occurrences of the pattern there be something ADJ about. The adjectives are evaluative, and usually unfavourable. The most frequent express the concept of “strange” (e.g. abnormal, curious, eerie, fishy, funny, paradoxical, peculiar, strange, odd, unusual) or of “familiarity versus difference” (e.g. different, familiar, new, special, unique). But it is not possible to provide a definitive list: there are also single occurrences of absurd, crude, delicious, evil, God-like, horrible, insulting, irritating, naive, primitive, right, risible, spectacular, suspicious, tragic, unsatisfactory, vital, wonderful, and others. Again, we have a strong central pattern with a long tail of variants.

**FREQUENT WORDS IN FREQUENT PHRASES**

Sinclair argues that very frequent words need to be described in their own terms: ‘their frequency makes them dominate all text’, but few of them ‘have a clear meaning independent of the context’. For example, the word way ‘appears frequently in fixed sequences’, where different patterns characterize different meanings, and where the resultant phrases ‘are frequently used metaphorically rather than literally’ (Sinclair et al. 1970/2004: 157–59, 163, 110–11). It looks like a high frequency noun, but in terms of its usage is in a class of its own (Sinclair 1999a: 166–72). The word is frequent because it occurs in frequent phrases, and its meaning depends on the phraseology. On its own it seems to have many meanings, but the phrases are unambiguous.

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

The word frequently occurs in longer conventionalized phrases which express topic-independent pragmatic meanings, as in these quasi-proverbs and clichés, speech acts, and discourse markers:

- see/know which way the wind is blowing; has become a way of life; if that’s the way you want it; laughing all the way to the bank; let me put it this way; only one way to find out; that’s one way of putting it; that’s the way I look at it; there is no way of knowing/telling; well on the way to recovery

The word point is also very frequent: at about rank 30 in a frequency list of nouns in the BNC. Under point as a head-word, Cobuild (1995a) gives about 30 senses, which seems to imply that the word is highly ambiguous, but the
Concordance 1: Illustrative examples of strong point preceded by a negative.

It was soon clear that rowing was not my strong point. At hockey there was a very strong point for arithmetic was not her strong point especially mental arithmetic. Negotiating between men is evidently not your strong point. Perhaps a few lessons might help. Zoological accuracy was not Tulip’s strong point. The animal was a chimps. Knowledge that, cooking was not Stella’s strong point, for it had turned out to be a strong point. She was on the point of saying something or argument anyway. That wasn’t her strong point. Her eyesight was her strongest point at the moment, whether she liked it or not. The two-word string strong point is still theoretically ambiguous, but in practice it is almost always used in an abstract sense. It can be used positively (My hon. friend makes a strong point). But, as in the illustrative lines in Concordance 1, it is most often the core of a speech act which has the form:

- x be NEG y’s strong point

Variable y is always a possessive pronoun or a proper name. Variable x is often something technical and/or numerical, and relates to the topic of the co-text. The whole unit is a conventional and ironic way of criticising someone by understatement. If you say that cooking is not her strong point, you mean “her cooking is terrible”. We have a canonical form, with minor variants, and a clear pragmatic force.

A MODEL OF EXTENDED LEXICAL UNITS (ELUs)

In a series of articles, Sinclair (1996, 1998, 2005) draws together his observations about co-selection and identifies semantic units of a kind which had not

phrases in which it occurs are not ambiguous. Some of the most frequent in the BNC are

- from the point of view of; it is at this point that; point you in the right direction; but that’s not the point; she was on the point of; at pains to point out that; I don’t see the point

The two-word string strong point is still theoretically ambiguous, but in practice it is almost always used in an abstract sense. It can be used positively (My hon. friend makes a strong point). But, as in the illustrative lines in Concordance 1, it is most often the core of a speech act which has the form:

- x be NEG y’s strong point

Variable y is always a possessive pronoun or a proper name. Variable x is often something technical and/or numerical, and relates to the topic of the co-text. The whole unit is a conventional and ironic way of criticising someone by understatement. If you say that cooking is not her strong point, you mean “her cooking is terrible”. We have a canonical form, with minor variants, and a clear pragmatic force.

A MODEL OF EXTENDED LEXICAL UNITS (ELUs)

In a series of articles, Sinclair (1996, 1998, 2005) draws together his observations about co-selection and identifies semantic units of a kind which had not
previously been described. He argues 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 the faulty assumption that the word is the primary
unit of meaning, and to the misleading separation of lexis and grammar. The
model is extremely productive, and many further examples have been dis-
covered by other linguists (e.g. Channell 2000; Francis 1993; Hunston 2007;

In one of his best known case studies, Sinclair (1998) discusses the
verb BUDGE.

were stuck fast. We couldn’t budge them.
the handle didn’t budge. She tried again and again
the rope would not budge. A tug-o’-war ensued
neither bribe nor threat will budge him from the truth
Mrs Thatcher refused to budge in her hostility to the EC

The meaning is not conveyed by the individual word alone. The verb
regularly occurs in a longer construction, with either a grammatical or
implied negative to the left (didn’t budge; refused to budge). A larger set of
examples would show the things that typically don’t budge (e.g. doors,
lids of jars, obstinate people): this is the ‘semantic preference’ of the
verb, which relates to the topic of the text. The whole construction is
used to tell a little narrative whose typicality we all recognize: the speaker
has tried repeatedly to do something, has failed, and is now annoyed. This
overall evaluative ‘semantic prosody’ is the communicative function of the
whole unit.

One of Sinclair’s main contributions to linguistic theory is his model of
an ELU which is held together by relations of lexis, syntax, semantics,
and pragmatics. These units (Sinclair 1996, 1998, 2005) have an obligatory
core and an obligatory semantic prosody. Their internal structure has
four parameters, which take different values and go from concrete to
abstract: from observed word-forms (1) to hypothesized communicative
functions (4).

(1) COLLOCATION is the relation of co-occurrence between an obligatory
core word or phrase (the node) and individual COLLOCATES: word-
tokens which are directly observable and countable in texts.
(2) COLLIGATION is the relation of co-occurrence between the node and
abstract grammatical categories (e.g. past participles or quantifiers).
A traditional category such as “negative” may be realized grammati-
cally (would not budge) or semantically (would hardly budge, refused to
budge).
(3) SEMANTIC PREFERENCES is the relation of co-occurrence between the phrasal unit and words from characteristic lexical fields. Recurrent collocates provide observable evidence of the characteristic topic of the surrounding text (e.g. typical subjects or objects of a verb).

(4) SEMANTIC PROSODY is the function of the whole extended unit. It is a generalization about the communicative purpose of the unit: the reason for choosing it (and is therefore related to the concept of illocutionary force).

Two further criteria in defining units are the strength of attraction between node and collocates (though there is considerable debate about which statistics are appropriate: see Evert 2005); and their distribution in text-types: whether they occur widely in general English or are restricted to broad varieties (e.g. journalism or technical and scientific English) or specialized text-types, such as recipes (add to taste), weather forecasts (warm front), or horse racing commentaries (entering the final furlong).

In summary, the model has these components:

- COLLOCATION tokens co-occurring word-forms
- COLLIGATION classes co-occurring grammatical classes
- SEMANTIC PREFERENCE topics textual coherence
- SEMANTIC PROSODY motivation communicative purpose

Relations (1) and (2) deal with how linguistic signs relate syntagmatically to one another; (3) deals with how linguistic signs relate semantically to the topic of the text; (4) deals with how linguistic signs relate pragmatically to their users. This model of meaningful human action is therefore based on two ways of grasping linguistic units: (1) and (2) are evidence of (3) and (4).

Relation (1) collocation and (2) colligation provide an objective view from the outside. They describe the observable behaviour of a social group. Collocation operationalizes the search for phrasal units: it makes only minimal assumptions about the division of text into orthographic words. Colligation assumes an analysis into grammatical classes: it is not directly observable. Collocation and colligation are part of our unconscious linguistic behaviour.

Relation (3) semantic preference and (4) semantic prosody provide a subjective view from the inside. Textual collocates can be identified automatically by the software, and provide observable evidence, but discourse topics can be identified only intuitively, and semantic prosody is a hypothesis about the conscious and intentional social action. The pragmatic function is often difficult to formulate, partly because there may be no word in the language which can serve as a descriptive label (Sinclair 1998: 20). The hypothesis is based on observable data, but improved formulations—and counter-examples—are always possible.
The model, especially the concepts of semantic preference and semantic prosody, has received substantial commentary (Partington 2004; Hunston 2007), but readers seem to have had problems in interpreting the model in three areas. (a) The model has largely been illustrated by individual case studies, and it remains unclear whether all phrases, or only some, have semantic prosodies. (b) The term ‘semantic prosody’ is often regarded as a rather vague indicator of the attitudinal connotation of a word: something good/desirable or bad/undesirable. This was perhaps the case with early examples (e.g. Sinclair 1991: 70–5; Louw 1993), but more recent analyses propose much more specific communicative functions, as in the BUDGE example (‘failed attempt and frustration’). (c) There is confusion between semantic preference and semantic prosody, and it might be helpful to use different terms here, in order to distinguish between semantic relations (to the topic of the surrounding text), and the pragmatic function (of the whole phrasal unit). These problems can largely be solved if ‘semantic prosody’ is used in Sinclair’s own sense, to refer to communicative purpose. This is also Hunston’s (2007) interpretation: she distinguishes clearly between the attitudinal connotation of an individual word and the discourse function of an extended unit of meaning.

SEARCH METHODS: AN EXAMPLE

In relatively early papers, Sinclair (1991: 110–15) distinguished between two principles: the idiom principle (that a speaker has available a large number of semi-preconstructed phrases) and the open-choice principle (to which speakers have recourse when no conventional phraseology is available). He later argued (Sinclair 1996) that this distinction is too sharp, and talked of two tendencies: the phraseological tendency (where words ‘go together and make meanings by their combinations’) and the terminological tendency (where words ‘have a fixed meaning in reference to the world’). The final move (Sinclair 1999b: 2–3) was to argue that ‘most of the so-called fixed expressions are not fixed at all’: there are just ‘variable expressions’. The ebb example above is typical: a phrasal unit has one frequent canonical form and a long tail of variants (cf. Moon 1998). This finding has consequences for search methodology (Tognini-Bonelli 2007) and for the level of abstraction at which the units must be described.

In a project on which Sinclair was a consultant, Cheng et al. (2006, 2009) have developed a new generation of concordance software (ConcGram), which can search for variants of two or more lemmas, which co-occur in different sequences, either adjacent or not. For example, the lemmas cause–problem frequently co-occur, but the lemmas vary in form (e.g. cause and causing) and in their relative position (e.g. causes a problem or a problem caused by),
Concordance 2. Illustrative examples of CAUSE–PROBLEM in a span of 6:6 in a 6-million-word corpus. Every third example.

01 dvantage of flexibility which can cause problems during installation. Fibregla
02 our behaviour. this can sometimes cause problems, for an over-night change in
03 go But money brings yu down Money causes problems anywea money is found, Food
04 t in present day Kazakhstan, were causing problems ( see below for more detail
05 under the rating system. This is causing problems for the local authority. It
06 ing the raft of legislation could cause real problems. On the single market, t
07 nsion on It’s not actually cos it caused these problems. With this personal co
08 ad confessed to the breaches that caused constant problems- he frequently anno
09 erratic nature of Patrice Lumumba
10 ol it is only the fish leech that caused any serious problems. This fastens it
11 longer wanted as first homes they cause fire or overheating problems—electri
12 cleared away and areas likely to cause relatively few problems; but when they
13 at at once trivializes systematic causes of poverty and magnifies the problems
14 ty Council to act on the problems caused by the gypsies at the end of the Abin
15 shington to confront the problems caused by its continuing non-military suppor
16 at percentage of our problems are caused by hail, rain, wind, snow, blow, fire
17 variety of other problems can be caused by echo, the two most important of wh
18 owed when the problem is directly caused by an error in the sub-system being t
19 s and problems on the borderlands caused the country’s chieftain, Khama III, t
20 cant problem on college campuses, causing faculty, staff, students, and parent

and the collocates may be adjacent or not (e.g. causes problems and causes severe problems). To generate Concordance 2, the ConcGram software was set to find all examples of CAUSE–PROBLEM within a span of 6:6 in a corpus of 6 million running words.

In this small illustrative sample, CAUSE is much more frequent as a verb, and PROBLEM is much more frequent in the plural. This could be checked in a larger corpus. Other software reveals further semantic variation. In the BNC, the top ten occurrences of CAUSE + (ADJ) + NOUN are:

- causes problems 180
- causing death 99
- cause trouble 60
- causes problems 45
- causing damage 41
- causing problems 41
- cause damage 40
- cause cancer 37
- cause difficulties 36
- cause injury 36

causing unnecessary suffering 14
causing serious injury 11
cause serious injury 10
cause mental handicap 9
cause serious damage 8
cause severe damage 8
caused extensive damage 7
causing criminal damage 7
caused considerable damage 6
caused great concern 6
The top ten occurrences of nouns before and after *caused by* are:

<table>
<thead>
<tr>
<th>Noun</th>
<th>Frequency</th>
<th>Noun</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>damage caused by</td>
<td>165</td>
<td>caused by negligence</td>
<td>20</td>
</tr>
<tr>
<td>problems caused by</td>
<td>103</td>
<td>caused by changes</td>
<td>16</td>
</tr>
<tr>
<td>injuries caused by</td>
<td>36</td>
<td>caused by fire</td>
<td>14</td>
</tr>
<tr>
<td>pollution caused by</td>
<td>29</td>
<td>caused by lack</td>
<td>13</td>
</tr>
<tr>
<td>difficulties caused by</td>
<td>21</td>
<td>caused by people</td>
<td>13</td>
</tr>
<tr>
<td>losses caused by</td>
<td>19</td>
<td>caused by differences</td>
<td>11</td>
</tr>
<tr>
<td>disruption caused by</td>
<td>18</td>
<td>caused by smoking</td>
<td>10</td>
</tr>
<tr>
<td>loss caused by</td>
<td>18</td>
<td>caused by alcohol</td>
<td>9</td>
</tr>
<tr>
<td>pain caused by</td>
<td>18</td>
<td>caused by exposure</td>
<td>8</td>
</tr>
<tr>
<td>delays caused by</td>
<td>17</td>
<td>caused by bacteria</td>
<td>7</td>
</tr>
</tbody>
</table>

As Barlow (2004: 206–8) shows very clearly, these methods transform texts, by ripping them apart into concordance lines and lists of words and phrases. When these textual fragments are further rearranged into tables, alphabetically or by frequency, the arrangement becomes part of the argument. The methods are a necessary estrangement device, which force necessary distance between the observer and the way in which we normally experience running text. Without them, we cannot see the patterns, but because different methods of observation provide multiple views of the text, which highlight some patterns and obscure others, such methods are exploratory and corpus data are ‘essentially indirect’ (Sinclair 1999b: 1; 2004c: 7).

Table 1 shows a lexical profile generated from all 3,250 occurrences of *causing* in the BNC. It shows, with their frequencies, the top 15 noun and verb collocates within a span of 2:2. Some collocates, especially to the right, are hyponyms of *problems* (e.g. *alarm, confusion, delays, explosion, havoc*). Several collocates are medical (e.g. *cancer, cells, disease, gene, injury*). Others are due to legal phrases such as

- charged with causing grievous bodily harm
- denies causing actual bodily harm

So far, this analysis of *cause* above generalizes across mixed reference corpora, looks only at the surrounding lexis, and ignores the detailed interaction between semantic field, grammar, and text-type. First, with an object such as *damage* or *trouble*, *cause* can take a double object construction, but this is not possible with words for illnesses and diseases (Klotz 1997: 102–4). Semantic preference interacts with grammar:

- it *caused* a lot of trouble   it *caused him* a lot of trouble
- smoking *causes* cancer       *smoking caused him* cancer

Second, *cause* has no necessarily undesirable connotations in scientific and technical English. Semantic prosody interacts with text-type:

- their effectiveness in causing the plastic flow of materials
But third, Hunston (2007: 263) points out that this formulation is not quite right. It is when a human being is involved that *cause* implies something undesirable. If cause and effect are inanimate (which is often the case in scientific and technical texts), then there is no necessary attitudinal connotation. In a word, there is co-selection of lexis, syntax, semantics, pragmatics, and text-type.

We now have criteria for the prototype structure of the ELU. Canonical phrases include *cause problems* and *cause serious damage*. The most frequent nouns (*problems* and *damage*) have general superordinate meanings, and can therefore occur in a wide range of texts. The most frequent adjectives (*severe* and *serious*) provide evidence of the evaluative semantic prosody of the unit. Hyponyms of *problems* and *damage* (e.g. *disruption, injury*) are topic-dependent and make cohesive links within specific texts. The paradigmatic variation in noun collocates to left and right makes the unit fit into its co-text, and are evidence of the semantic preference of the unit. The topic might be *pollution* and the cause might be *negligence*. More specialized again are technical
expressions, such as cause grievous bodily harm, which typically occur in legal and journalistic texts. In scientific and technical texts the semantic preference and the semantic prosody are likely to be cancelled.

The vertical and horizontal axes of concordance data and lexical profiles provide evidence of form and meaning respectively. The vertical axis provides evidence of recurrent forms across many instances of language use: this is often very obvious even to superficial inspection. The horizontal axis provides evidence of meaning, both in individual instances (individual concordance lines) and also in the language system. This is often less obvious, because different collocates typically occur in different instances, and it needs the experience and intuition of the analyst to group these collocates into semantic sets. In addition, we typically have a few prominent items, with a long tail of less frequent or unique exponents (Sinclair 2004b: 286), and the appropriate level of delicacy in description remains a decision for language teachers or lexicographers. The software provides objective and quantifiable evidence, but the best way to present such findings requires judgement. As Labov (1972: 258) puts it: ‘By the time the analyst knows what to count, the problem is practically solved’.

SUMMARY AND IMPLICATIONS

John Sinclair’s work shows how to use empirical evidence to tackle one of the deepest questions in the philosophy of language: the nature of units of meaning. As Kant put it (Scruton 2001/2004: 31, 41): pure empiricism is content without form, but pure rationalism is form without content. Sinclair’s model has both: corpus observations provide high empirical content, and interpretation provides the elegant phraseological model.

Sinclair questions traditional distinctions between logic, rhetoric, and grammar, which have been familiar since the medieval Trivium, and which are still often taken for granted. He is sceptical of logical approaches to language (e.g. natural language processing), he reinterprets rhetoric as discourse and builds concepts of communicative function into ELUs, and he severely questions the invented introspective data which were prominent in grammatical theory from the 1960s onwards. As he says, alluding to this tradition, ‘one does not study all of botany by making artificial flowers’ (Sinclair 1991: 6). His work supports a long and very different tradition of text and corpus analysis, derived from his own British teachers and colleagues (especially Firth and Halliday), but represented in a broader European tradition (for example, Jespersen) and in a much more restricted American tradition (for example, Fries). This tradition has always emphasized the description of meaning, but it is Sinclair who makes the most sustained attempt to develop an empirical semantics: what Hanks (1997) has called his ‘ferocious empiricism’.

However, his research programme remains to be developed in several directions. Implications and applications for teaching materials are discussed by
Willis (1990), for pattern grammar by Hunston and Francis (2000), and for lexicography by Moon (2007) and by the articles in the special issue of the *International Journal of Lexicography* (21/3, 2008) which is devoted to Sinclair’s work. I will mention here some implications of the model of phraseology.


A major descriptive task, for both theoretical and applied linguists, is a comprehensive list of the units of meaning in a given language. The general method is clear. If we discover phrases which are both frequent and widely distributed across a corpus, then they are not text-dependent, but part of the patterns of the language (Sinclair *et al*. 1970/2004: 79). The major descriptive problem is that the units are internally variable and have indeterminate boundaries, and that similar units are often related to each other in taxonomic hierarchies (Croft 2001: 25; Stubbs 2007). The general solution is to describe their canonical, prototypical forms, but deciding the appropriate level of delicacy for different purposes is a matter of interpretation. We have many convincing case studies and the beginnings of such descriptions in corpus-based dictionaries, but we still need thorough-going phraseological dictionaries.

2. A theory of textual cohesion

The question then arises as to why certain phrasal units occur in many different texts. One answer is that they do not depend on the topic of the text, but serve text-management functions, such as signalling narrative structure, topicalization, point of view, and the like. This provides the link between text and corpus analysis. Here are just two examples.

First, the double verb construction went and VERBed is a conventional way of marking the end of a segment of narrative.

- he put the phone down and went and got himself a malt whisky
- so I went and toddled off to find somebody
- and then, would you believe it, she went and married him
- then you went and got married again, you can’t help yourself
- then he went and jumped out of a plane
- then she went and spoiled everything by behaving as if pissed
- Paul Bodin then went and missed a penalty

In (a) and (b) the verb went seems redundant: it is assumed by the action which follows. In (c) to (g), went cannot be interpreted literally at all: it is a pragmatic signal of something the speaker did not expect or did not approve of. Along with the co-occurring (and) then or so, it also signals the conclusion of part of the action.²

Second, there are several words (e.g. blatant, downright, mere, outright, sheer, utter) which have purely pragmatic meanings. They do not denote anything
in the world, but signal the attitude of the speaker and a textual contrast. For example, the *x and/or outright y* construction contributes to textual coherence by contrasting two points on a scale: a middle point and an extreme point of which the speaker disapproves.

- fraught with inaccuracies and outright fallacies
- through ruthless speculation and outright fraud
- deeply ingrained suspicion and outright hostility
- signs of strain or outright contradiction
- misguided speculation or outright corruption
- imbued with deep suspicion or outright opposition

An increasing number of case studies (e.g. Hunston and Francis 2000: 185–8; Partington and Morley 2004; Hoey 2005; Mahlberg 2005) show how lexis and phraseology contribute to textual structure and organization. If they could be integrated, such case studies would contribute to a functional theory of lexis.

3. A theory of collocation

We know from many descriptive studies that the syntagmatic attraction between linguistic units is much stronger than is often realized, and this attraction can be measured in different ways. For example, the Cobuild (1995b) database gives the 10,000 most frequent word-forms along with their 20 most frequent collocates in a 200-million word corpus. In the following illustrations (from Stubbs 2006), a statement such as *node <collocate 10%>* means that a node word co-occurs in 10 per cent of cases with its top collocate within a span of four words to left or right:

- drive <car 3%>; origin <country 4%>; basis <regular 4%>
- room <dining 6%>; hate <love 7%>; ending <year 8%>
- chopped <finely 14%>; answers <questions 16%>; pose <threat 18%>
- solving <problem 37%>; conditioning <air 57%>
- eighteenth <century 77%>

Indeed, the strength of attraction between collocates is much stronger than such figures suggest, since such calculations, between two individual word-forms, would count separately *inflicts damage, inflicted a wound,* and *inflicting an injury,* although all three phrases seem part of a single semantic pattern.

Kilgarriff (2006) goes a step further and identifies those words which usually co-occur with a small set of particular collocates, rather than appearing freely with large numbers of collocates. Using data from the BNC, he identifies the 100 nouns and verbs which show the ‘most collocational’ behaviour (for nouns, with respect to the verbs they are objects of, and for verbs, with respect
to their object-nouns). Many high frequency nouns frequently occur with light support-verbs, and low frequency nouns often occur largely in idioms:

- take place, take care, take advantage, pay attention
- take the mickey, come a cropper, enter/join the fray

Corpus studies have now documented in detail that people’s language use is typically very repetitive, but it is far from trivial to explain what level of repetitiveness is optimum in a functioning system. Semantic and functional relations between collocates are studied by Mel’čuk (1998) and Jones (2002), and psycholinguistic factors are studied by Wray (2002, 2008) and Ellis et al. (in press). We need to move from description to explanation by using both semantic and psycholinguistic data.

4. A theory of social action

In his inaugural lecture, given only a few years after the publication of How to Do Things with Words (Austin 1962), Sinclair (1966) argued that ‘utterances do things rather than just mean things’ (emphasis in original). Sinclair (2008: 24) returns to this point that, before Austin, linguists were ‘embarrassingly short’ of ideas about language interaction. Both speech acts and ELUs build the speaking agent into units of language structure, and the concept of illocutionary force seems close to the concept of semantic prosody. Yet, despite his early work on classroom discourse, and despite the social orientation of work by Firth and Halliday, Sinclair remained notoriously unwilling to draw out the social-psychological implications of his work on phraseology. It is only corpora which can provide data for studying prosodies from the bottom up, and therefore show how we could do real ‘ordinary language philosophy’. Austin and Searle work deductively. They use only invented introspective data. They propose a powerful theory of speech acts, and Searle (1995) is strong on social theory. Conversely, Sinclair works inductively. He uses authentic empirical data. He also proposes a powerful theory of functional language units, but his work lacks a general social theory. We need a clearer understanding of the relation between these two traditions.

5. A theory of how people construe the world

Despite his resistance to socio-psychological approaches, Sinclair’s phraseological model does use unashamedly psychological concepts such as motivation and evaluation. ELUs, such as the wouldn’t budge construction (Sinclair 1998), are little schemas of cultural knowledge. The par for the course construction signals that things have gone wrong, yet again, in just the way that we would have predicted (Channell 2000: 47–50). The not the end of the world construction signals sympathy for someone who has suffered a disappointment (Stubbs 2007). ELUs assume typical scenarios in the everyday world, and presumably we learn how to construe the social world through the phraseology. Speakers are conscious social agents whose constant comments
on the actions of themselves and others point to phenomena which really exist (Harré and Secord 1972). To complain is part of being discontented, to express frustration is part of being frustrated. ELUs are part of the way in which actions and states of mind are mediated by language. Meaning depends on intertext: it derives from what has been said frequently and successfully by innumerable speakers in the past. But often the semantic prosodies have no conventional name (Sinclair 1998: 20) and can be discovered only by studying how the lexico-grammar expresses taken-for-granted cultural meanings. We need to compile ‘a grammar of the typical meanings that human communication encodes’ (Francis 1993: 155) and relate the linguistic and cognitive schemata which underlie how we understand social behaviour (Teubert 2008).

These five points indicate how Sinclair’s empirical work could be developed in linguistic, cognitive and social directions. Although he seems generally unwilling to speculate on such topics, Sinclair (2007: 1) points out that, whereas corpus linguists like nothing better than empirical findings supported by levels of statistical significance, people outside this narrow circle want to know how it all hangs together, and how all the empirical information contributes to solving the great intellectual puzzles of language in society. He asks: How should all this work be evaluated? How does empirical linguistics contribute to ‘wider issues’, and how can it be used ‘as a foundation for a broad range of intellectual exploration’? His approach to language may be ‘ferociously empiricist’, but it is the antithesis of naive positivism. By changing the frame of reference for our observations, Sinclair has shown ways of describing the order which underlies the apparent chaos of parole. His approach to phraseology is a major advance in describing the facts. We now have to ask how far this approach can help in explaining the facts.

SUPPLEMENTARY DATA

There is online Supplementary Data available at Applied Linguistics online.

CONVENTIONS AND CORPORA

Lemmas are represented as small caps. Word-forms are represented in lower case italic. ‘Meanings’ are in double quotes. ‘Quotes’ from other authors and technical terms are in single quotes. <Diamond brackets> enclose typical collocates of a node-word. Many of the examples are from the British National Corpus (BNC): 100 million running words of contemporary English, 90 million written and 10 million spoken, sampled from over 4,000 texts from a wide range of text-types. Many of the examples of n-grams (recurrent multi-word strings) were extracted from the BNC by using William Fletcher’s BNC-interface at http://www.phrasesinenglish.org. This database allows us to start from a node word, and extract all recurrent n-grams with the node at each position, down to a specified cut-off frequency. References to a six-million word corpus are to a collated version of the Brown, LOB, Frown, FLOB and BNC-baby
corpora: this comprises five million written and one million spoken running words.

ACKNOWLEDGEMENTS

For comments on a previous version of this paper, I am grateful to Bettina Fischer-Starcke and to three anonymous reviewers.

NOTES


2 Airola (2007) studies the function of similar double verb constructions in Finnish, and shows that they seem semantically redundant if regarded out of context, but that they can be explained by looking at their pragmatic function.

REFERENCES


Sinclair, J. 2005. ‘The phrase, the whole phrase and nothing but the phrase’ Plenary lecture, Phraseology 2005, Louvain-la-Neuve, October.


