doi:10.1093/ijl/ecn031

THE LEXICOGRAPHICAL LEGACY OF JOHN SINCLAIR

Patrick Hanks: *Faculty of Informatics, Masaryk University, Brno, CZ (hanks@fi.muni.cz)*

Abstract

John Sinclair opened up possibilities for new kinds of dictionaries. He assigned a central role to collocations and phraseology, insisting on close attention to textual evidence coupled with a broad theoretical perspective and ruthless jettisoning of hypotheses that do not fit the facts. He aimed to create dictionaries that would help students to write and speak idiomatically. In the tradition of Dr Johnson and OED, these would be based on evidence rather than speculation, but evidence of contemporary usage, not literary citations. In this paper, I look at some possibilities inspired by this approach. I suggest that a synthesis between Sinclairian corpus linguistics and construction grammar is overdue.

1. Introduction

In an era when much of linguistics was dominated by speculative theories of syntax, focusing on the logical form of idealized sentences without empirical investigation of either usage or meaning, John Sinclair (1933 – 2007) followed Francis and Kučera (1964) in devising computational approaches to collecting evidence. In Sinclair (1966) and Sinclair et al. (1970), he introduced a focus on collocations. In Sinclair (1987) he noted that, if a language contains several hundred thousand word types, then a corpus of only one million words will not be large enough to distinguish significant collocations from random co-occurrences. He looked dispassionately at the growing evidence, which he described as 'disturbing', and he thought deeply about meaning in language. He addressed issues of form, structure, pragmatics, and content beyond the level of the sentence. He viewed the lexical item as an entry point to phraseological meaning, not necessarily as a meaningful unit in itself. He saw that the meaning of utterances is flexible and probabilistic, often arising out of a dynamic interpretation of collocations, not just a list of dictionary 'meanings'. In the course of analysing data, he found it necessary to challenge received notions of polysemy, lemmatization, dictionary structure, discourse, and grammar.

International Journal of Lexicography, Vol. 21 No. 3. Advance access publication 6 August 2008

He was unimpressed by other linguists' tendency to use corpora as 'fish ponds' in which to fish for examples supporting their theories, rather than to look and see what is going on. Above all, he showed that innovation in lexicography is possible at a nontrivial level.

It is too early to evaluate his full impact on our understanding of the phenomenon of language and the possibilities for lexicography, which have scarcely begun to be realized. In the second section of this article I give an example of post-Sinclairian corpus analysis. In the third section, I briefly compare aspects of the Sinclairian approach with parallels in construction grammar. There are similarities as well as differences, and both are of lexicographical interest.

2. Sinclair as a lexicographical influence

John Sinclair was not a lexicographer. He did not compile an index of a culture through its terminology, nor did he churn out dictionary entries in profusion on a received model. Instead, he challenged received models, leaving the churning out to others. He studied words in use in order to shed light on how language works. He was not afraid to abandon the comfortable certainties of traditional syntactic theory, nor, indeed, his own earlier hypotheses. For him, lexicography was a means to an end. His early work was in discourse analysis and grammar as well as lexis, but by getting involved with the complex business of dictionary publishing, he could:

(a) get funding for the creation of ever larger corpora (it is hard for us in the age of the Internet to remember how difficult and expensive this was in the 1980s);
(b) encourage the study of lexis as a linguistic level, looking at multiple contexts for each word in order to see what is really going on, rather than accepting speculative theories about what might be going on;
(c) mastermind the creation of dictionaries, grammars, and course books that would help learners get to grips with idiomatic and pragmatic uses of language, as opposed to teaching them word lists and grammatical abstractions.

In pursuit of the second and third of these objectives, he sometimes made—and even insisted on—suggestions that his co-workers thought grossly impractical. Some of these suggestions remain unimplemented; others turned out, after a little experimentation, to be perfectly practical after all. Cobuild's 'full-sentence explanations', addressing the user directly, are a case in point. From the outset of the project, Sinclair was determined to get rid of the apparatus of traditional dictionaries (reductionist and substitutable definitions, simplistic grammatical categorizations, overuse of brackets, etc.), an apparatus which he thought was to a large extent both theoretically unsound and meaningless to users. However, it was not clear at first what should be put in its place, and first-stage compilation on Cobuild proceeded using a more or less traditional model minus the brackets. It was only at the final editing stage that the new style was developed to the point where it could be implemented systematically. Even then, it could not have been done if it had not been for Jeremy Clear, who wrote a suite of computer programs implementing it semi-automatically. The results were checked and tidied up by the lexicographers during the final editing stage.

The rationale of full-sentence explanations is discussed in Hanks (1987). Applied systematically, it forces the lexicographer to encode each definiendum in phraseology that is most associated with a distinct sense. Subsequent English learners' dictionaries have adopted full-sentence definitions piecemeal, adding them to the inventory of defining strategies for use in emergencies, when substitutable definitions fail. The distinction between systematic and piecemeal application is an important one. By adopting the full-sentence style systematically, Cobuild pointed the way ahead to more formal approaches to lexical analysis, one of which is illustrated below.

2.1 *The vicious circle of dictionary publishing*

Dictionary publishing is characteristically caught in a vicious circle. It is a cut-throat competitive business, in which marketing is at least as important as content. Profits are potentially large, but investment budgets are large too (and development capital is tied up for an unconscionably long time), risks are high, and profit margins are thin. Dictionary publishers tend to pride themselves on being 'market-driven'. This is the root of a problem. Existing dictionaries create certain expectations among users about what dictionaries will be like. These expectations are conservative: people expect new dictionaries to be improved versions of old ones, not radical new departures. How could it be otherwise? Not being professional corpus analysts or lexicographers, the dictionary-buying public cannot propose innovations; they cannot know what the data is like—and if they did, they would be most unlikely to know what generalizations are necessary or innovations possible. So dictionary publishers are typically conservative, driven by an unthinking market and opposed to any innovation that might frighten away buyers.

Sinclair offered a way out of this vicious circle, for courageous publishers. The way out is not slavish mimicry of the practices of Cobuild, still less of any other dictionary. That would simply be to start a new vicious circle. Instead, the way out is to start painstakingly marrying observed facts with new possibilities for description—examining data with an open mind, then looking at users' needs, and so gradually working up a framework for analysis and description that will do least distortion to evidence and be most helpful to the target audience. What is the data telling us? Forget what theorists have told us—they don't look at data. Also, take what the market says it wants with a pinch of salt. Instead, study data intensively. What do users need to be told, and how should it be expressed?

2.2 *Examining and classifying data: seeing what is there*

Sinclair (1966, 1970, 1987, 1991) initiated a long and (now) thriving tradition of empirical lexical analysis. From a lexicographical point of view, the most important milestone in this tradition is the first Cobuild dictionary (1987), but there are other milestones too, notably the *Verb Patterns* of Francis et al. (1996) and the *Pattern Grammar* of Hunston and Francis (2000). The distinction between a pattern grammar and a pattern dictionary is important. A pattern grammar, like all grammars, seeks structural generalizations that will be true of relations among sets of lexical items. A pattern dictionary, on the other hand, recognizes that every lexical item is associated with a unique set of collocational and colligational relationships. A pattern dictionary must therefore examine each lexical item individually and ask what patterns it participates in. In order to do this effectively, guidelines must be established for what counts as a pattern, and a metalanguage must be developed for pattern description. These issues are discussed by Hanks and Pustejovsky (2005).

Pattern elements are preferred collocations (individual words that frequently occur in close proximity to a target word) and colligations (words in grammatical relationships with a target word). The scope of a pattern element ranges from individual lexical items—e.g. *havoc* in relation to *wreak* (see Moon, this volume)—to extremely large classes, e.g. *amass* in relation to nouns denoting a quantity of almost anything: *land, wealth, debts, information, ideas, a collection of artefacts, an army, followers*, etc. These examples represent extremes of a collocational continuum. Within that continuum, each word has its own special place and its own set of preferences. Even a very open-choice word such as *amass* turns out to have distinct preferences and dispreferences, as the Sketch Engine (Kilgarriff et al. 2004) shows. Business people *set goals* and hope to *amass fortunes*. Both these nouns are found in corpus-based lists of statistically significant direct objects of *amass*, but *amassing goals* activates a quite different meaning of *goal* (see 1) from that in *setting goals*. We find *amassing goals* only in sports reports, and then normally only in a phrase governed by a cardinal number or a quantifier such as *enough*. Business goals are not 'amassed'; fortunes are not 'set'.

(1) McCoist . . . has . . . amassed 43 goals in all competitions.

Amass and *accumulate* are verbs with similar meanings and similar sets of preferred direct objects (though there are subtle differences, which a lexicographer might want to tease out). However, only *accumulate* has an inchoative sense, in which no human agency is involved. (2) is natural, (3) is not.[1]

(2) the sediment accumulates over eons.
(3) **the sediment amasses over eons.*

Examination of multiple instances reveals that the semantic class of a collocate may have an effect on the meaning of a content word elsewhere in the clause—as also may function words such as determiners. Compare *find one's way around a location* (4) with *find a way around a problem* (5). Different meanings of *find* are activated by the different phraseology.

(4) signposts and directions which help you to find your way around the hospital easily.
(5) Borland has obviously found a way around the problem of maintaining data integrity.

2.3 *Polysemy or conventional metaphor?*

Now consider the noun *incarnation*. This is an important technical term in Christian theology. Etymologically, its literal meaning is 'into flesh', Latin *in carne*. It denotes in particular the Christian belief that God became a human being 'in the flesh' in the person of Christ. It denotes beliefs in other religions, too, about spiritual beings taking bodily form.[2] These usages are still common, as in (6) and (7), and it can be argued that knowledge of the word's etymological meaning enriches understanding of its contribution to other conventional uses, for example (8) and (9), which are classified as conventional metaphors.

(6) The incarnation is therefore an extreme instance of divine immanence within the creation.
(7) the divine baby Rama was born as Vishnu's seventh incarnation on Earth.
(8) It is some five years since I worked with Mr Edell in his previous incarnation as a lawyer, but he struck me as a man who would do the necessary.
(9) When reality catches up with such people, it is usually somewhere near the fraught crossroads where sex, power and violence meet. Kennedy, now known to have shared molls, and perhaps Marilyn, with Mafia leader Sam Giancana, is the incarnation of this syndrome.

None of this would have impressed Sinclair. Rather, he would have taken a longer, harder look at the corpus evidence and might have made observations such as the following:

- There are 319 hits for the lemma *incarnation* in BNC.
  - Only 31 of them (10%) are plural: the noun is normally singular.

  - In Christian texts it is always singular, and usually with *the* or *Christ's*.
- 49 of the 319 hits—7 of them plural—are in the pattern *incarnation(s) of* N2.
  - N2s in this pattern include *Christ, God, Trinity, Virgin, deity, Isis, Durga, devilry, evil forces, Hitler's Englishman, John Mayall.*
- 19 hits are in the pattern *incarnation as* N2. They are all singular.
  - N2s in this pattern include a wide variety of nouns, including *a professional legislature* and *a frisbee.*
- There are 31 hits for *previous incarnation* and 7 hits for *earlier incarnation*. In this pattern, *incarnation* is normally a count noun, although elsewhere it may be a mass noun.
  - Related to these are expressions such as *present incarnation*, *current incarnation*, *later incarnation*, and *future incarnation*. This relationship embodies the fact that the adjectives *present*, *current, later*, and *future* have similar patterns of distribution to those of *previous* and *earlier*. (Sinclair would defer the obvious move of invoking a semantic common feature.)
- Statistically significant collocations with *incarnation* (unstructured—i.e. not colligations) include *trinity, resurrection, passion,* and *cross.*
- *the Incarnation*—often written with a capital I—is a common expression in Christian writings, where it is often used without a modifier or qualifier.
- There are 10 hits for *the story of the Incarnation*, but they are all from the same text.

Several other observations about lexis as a linguistic level could be made, even for such an infrequent, simple, and straightforward word as *incarnation*. How might such information be encapsulated in a dictionary? What is to be selected? There is, of course, no single right answer to this question. It depends in part on what sort of dictionary is being compiled. Let us assume a dictionary for encoding use and look first at Cobuild (1987).

1. Someone who is an **incarnation** of a particular quality represents that quality in human form in a very strong way. EG *Miss Lenaut, that incarnation of feminine beauty . . . He became the incarnation of evil.* N-COUNT; N + *of*

2. An **incarnation** is an instance of being alive on earth in a particular form. Some religions believe that people have several incarnations in different forms. EG *Perhaps they were lovers in a previous incarnation . . . Christians insist upon only one incarnation.* N-COUNT

Incarnation is identified in sense 1 as a count noun with *of*—the most salient pattern—but otherwise little attention is paid to phraseology. Cobuild is still a sense-driven dictionary. This is part of what Sinclair had in mind when he described Cobuild, in its introduction, as a very traditional dictionary. It picks out only one—the most salient one—of the phraseological patterns.

By contrast, a phraseological pattern dictionary entry might look something like this:

- **an incarnation of a being or deity** = appearance on earth in bodily form:
  - *Rama was born as Vishnu's seventh incarnation.*
- **Christ's Incarnation, the Incarnation** = (in Christian belief) God being born on earth in human form:
  - *Christ is the incarnation of God.*
- **the incarnation of something** = a typical or outstanding example of it:
  - *She was the incarnation of everything that had gone amiss in Sylvie's own life*
- **someone's or something's incarnation as something** = the event of taking a particular character or form:
  - *Charlie's incarnation as a Norse explorer*
- **in a previous incarnation** = in a previous life or (humorously) a previous period of someone's life:
  - *I worked with Mr Edell in his previous incarnation as a lawyer.*
- **in one's present/current incarnation** = as one is now.

Notice (a) that this approach is driven by phraseology, not meaning, and (b) that it abandons the goal of representing all imaginable possibilities. Instead, it represents normal usage. It is, therefore, suitable to help a student who wants to use the word normally and idiomatically.

3. Corpus linguistics and construction grammar

There is little common ground between Sinclair and generative linguistics. However, in recent years the defects of generative theory have become increasingly obvious even within the speculative American tradition. One such defect was a failure to engage seriously with the nature of meaning, and one response to this was the Generative Lexicon theory of Pustejovsky (1995), which directly addresses issues such as the multiple facets of word meaning and coercion of words to have different meanings in different contexts.

Another set of problems arose out of the sharp distinction between lexicon and grammar. Chomskyan grammar was conceived as a sort of concrete mixer: a machine into which raw material (lexical items) is poured at one end in order to produce well-formed sentences (without reference to meaning) at the other end. The currently fashionable theory of Construction Grammar originated (Fillmore et al. 1988) in part in response to this. It is explained most succinctly by Michaelis (2006); also in Chapter 10 of Croft and Cruse (2004); more extensively in Goldberg (1995). Construction Grammar deserves the attention of lexicographers, for it asks, among other things, where meaning is to be found, and comes up with the answer that meaning resides in constructions. A construction is a linguistic element that cannot be broken down into smaller units without loss of meaning. A construction may be anything from a single word to a whole phrase.

> Constructions may specify, not only syntactic but also lexical, semantic, and pragmatic information. . . . lexical items . . . may be viewed . . . as constructions in themselves. (Fillmore et al. 1988: 501).
>
> There is no principled divide between 'lexicon' and 'rules'. . . . The proposal, then, is to expand the role of the traditional lexicon to include productive or semi-productive phrasal patterns that have previously been assumed to lie within the domain of syntax. (Goldberg and Jackendoff 2004: 532, 535)

There is no space here to do more than comment briefly on certain aspects of a family of constructions, namely resultatives, discussed at length by Goldberg and Jackendoff (2004). They show clearly that resultative meaning is not created solely by the application of grammatical rules to an input of lexical items.

(10) *Bill belched his way out of the room.*

There is more to the meaning of (10) than an ordered concatenation of words and word classes: such a concatenation does not yield the interpretation that Bill moved, nor that he belched protractedly while doing so, yet this is what the sentence means. None of the words or phrases in (10) necessarily implies movement: *out of the room* can also imply location, as in *Bill is out of the room at the moment*. And yet (10) expresses a movement event coupled with a belching event. It contrasts with, say, *Bill belched out of the window*, in which Bill may merely have turned his head. Goldberg and Jackendoff offer a complex and not entirely satisfactory analysis, discussion of which must await another occasion. Here, I will propose an alternative. Corpus analysis shows that, among other things, *way* participates in the following pattern:

A. [[Human]] V {[REFLDET] way} [Adv[Direction]][3]

Verbs that participate in this construction include at the least the following classes:

> **Verbs of manner of movement:** *barge, bulldoze, bumble, burrow, crash, crawl, cruise, dance, dodge, gallop, jangle, jostle, lurch, manoeuvre, move, navigate, paddle, plod, plough, ramble, sashay, scramble, scuttle, shuffle, sidle, skip,*

sleep-walk, slog, squirm, stamp, steamroller, stagger, steer, stumble, swashbuckle, swim, waddle, weave, worm, wriggle, zigzag.

Verbs of noise emission: *bang, bellow, blast, breathe, chuckle, clang, clatter, croak, crunch, gargle, giggle, growl, honk, hoot, moan, puff, roar, sing, sniff, sniffle, sob, splash, squelch, strum, tap, thunder, tootle, twang, warble, weep, wheeze, whistle, whine, whoop, yowl.*

Eating and drinking verbs: *chew, chomp, drink, eat, gnaw, gnash, gobble, munch, slurp.*

Verbs of violent physical action: *batter, battle, beat, bomb, burst, butt, claw, elbow, fight, force, grope, hack, jab, jerk, kick, poke, power, push, scorch, scratch, shoot, shoulder, shove, shovel, slash, smash, struggle, thrust, wrestle.*

Other physical action verbs: *bob, bow, bum, bump, clown, cut, dig, edge, elbow, feel, fumble, grope, hack, idle, oil, pick, squeeze, thread, thumb.*

Cognitive action verbs: *bamboozle, bluff, bribe, buy, cajole, cheat, con, dupe, earn, flaunt, gamble, joke, lie, negotiate, play, plot, pose, pray, preach, puzzle, scowl, spend, talk, trick, waffle, wangle, wheedle.*

Light verbs: *find, make, wend, work.*

The cognitive action verbs (which typically have negative semantic prosody) denote achievement of an abstract goal. Many of the other verbs here, including verbs of movement, are used for both abstract and concrete goals. If a verb is already, in its most literal sense, a verb of movement or goal orientation, then use in pattern A reinforces the notion of movement towards a goal. If the verb is not already a verb of movement or goal orientation, it is coerced into being one by the construction. Occasionally, e.g. *fuff* in (11) and *pole* in (12), a word that is not normally a verb at all is coerced into being one by the construction, with a goal-orientation meaning.

(11) Gesner had always found he couldn't fuff his way through it.
(12) The first body . . . was discovered accidentally by a bargee, who had been poling his way up the river.
(13) Mrs Thatcher stonewalled her way through Question Time.

(13) implies a goal achieved (Mrs Thatcher survived Question Time politically that day), even though *stonewalling* is not a normal member of the set of cognitive action verbs.

Pattern A contrasts with other patterns in which verb + *way* may participate. One of the closest is B.

B. [[Human]] V {[DET] way} {of –ING}

If a verb is found in both pattern A and pattern B, as in (4) and (5) above, it will have a different meaning in the two patterns.

These findings elaborate and occasionally correct Goldberg and Jackendoff's account of the *way* resultative construction. Goldberg and Jackendoff discuss other resultatives, which there is no space to discuss here.

One further point must be made about Goldberg and Jackendoff's important paper, however. This concerns their examples, which, in the tradition of speculative linguistics, are mostly invented. Some of the invented examples can be shown, by comparison with corpus evidence, to be idiomatic (e.g. *belch* in (10) is a verb of noise making). Others (e.g. 14) are unnatural.

(14) **Fred watered the plants flat.*

(14) is an example concocted by Goldberg and Jackendoff and, according to them, it is idiomatic. But *water*, unlike, say, *hammer*, is not a verb that normally participates in this construction. In the unlikely event of such a sentence being found in a real text, we should say that it is an exploitation of a norm, not an idiomatic use. Goldberg and Jackendoff do not make this distinction. When, occasionally, they cite an authentic example from Google, they are apparently unaware of the argument by corpus linguists that authenticity alone is not enough: evidence of conventionality is also needed.

This cavalier attitude to evidence both complicates and vitiates many a good study in generative linguistics. For half a century, it has left linguistics drowning in speculation about bizarre possibilities, rather than attending to the structure of normal language. At the very least, generative linguists could learn something from Sinclair about how to analyse evidence.

4. Conclusion

Probably the most important part of John Sinclair's legacy is his insistence on analysis of corpus evidence in order to establish details of normal phraseology. This manifests itself most clearly in his studies of collocational preferences. There is some similarity between Sinclair and construction grammarians in that both assign a central role to meaning and reject compositional theories of meaning, but Sinclair was and remains ahead of construction grammarians in at least two ways: (a) his wider perspective on discourse and pragmatics (beyond the sentence), and (b) his analyses of empirical evidence. Construction grammarians, however, have much to contribute by way of theoretical insights into the nature of the lexicon, and a synthesis would benefit both theoretical linguistics and practical lexicography.

Acknowledgement

This work was supported in part by grant 1ET100300419 of the Czech Academy of Sciences.

Notes

[1] Real examples (i.e. those taken from a text or corpus) are printed in roman, invented examples in italics. Invented examples that are judged to be unnatural are preceded by an asterisk.

[2] The reason for prioritizing Christianity is the central role that that religion has played in the development of the English language.

[3] The terminology is that of the Corpus Pattern Analysis project (see Hanks and Pustejovsky 2005). For the distinction between REFLDET (reflexive possessive determiners) and other uses of possessive determiners, consider the ambiguity of 'Forster took his place'. Did Forster replace someone else, or did he take the place that had been assigned to him?

References

Croft, W., and D. A. Cruse. 2004. *Cognitive Linguistics*. Cambridge University Press.

Fillmore, C. J., P. Kay, and M. C. O'Connor. 1988. 'Regularity and Idiomaticity in Grammatical Constructions: The case of *let alone'*. In *Language* 64: 501–538.

Francis, G., S. Hunston, and E. Manning. 1996. *Collins Cobuild Grammar Patterns 1: Verbs*. London: HarperCollins.

Francis, W. N., and H. Kučera. 1964. *Manual of Information to accompany a Corpus of Present-Day Edited American English*. Providence, RI: Brown University.

Goldberg, A. E. 1995. *Constructions: a Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Goldberg, A. E., and R. Jackendoff. 2004. 'The English Resultative as a Family of Constructions'. In *Languag*e 80 (3): 532–568.

Hanks, P. 1987. 'Definitions and Explanations' in J. Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Collins ELT, 116–136.

Hanks, P., and J. Pustejovsky. 2005. 'A pattern dictionary for natural language processing'. In *Revue Française de Linguistique Appliquée* 10(2): 63–82.

Hunston, S., and G. Francis. 2000. *Pattern Grammar*. Amsterdam: Benjamins.

Kilgarriff, A., P. Rychlý, P. Smrž, and D. Tugwell. 2004. 'The Sketch Engine'. In *EURALEX 2004 Proceedings*. Lorient, France.

Michaelis, L. A. 2006. 'Construction Grammar' in E. K. Brown (ed), *Encyclopedia of Language and Linguistics, 2nd edition*. Elsevier. Vol. 3: 73–84.

Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

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

Sinclair, J. M. 1987. 'The Nature of the Evidence' in J. Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Collins ELT: 150–159.

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

Sinclair, J. M. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Sinclair, J. M., S. Jones, and R. Daley. 1970 [2004, ed. R. Krishnamurthy]. *English Lexical Studies*, report to the Office of Scientific and Technical Information (OSTI).