Published in M Nenonen & S Niemi eds (2007) *Collocations and Idioms*. Joensuu: Joensuun Yliopisto. 317-29.

Michael Stubbs
University of Trier, Germany

Notes on the History of Corpus Linguistics and Empirical Semantics

This is a paper on empirical semantics. One traditional view is that semantics cannot be empirical, because meaning is cognitive and conceptual, invisible, and therefore impossible to study via observable data. However, an alternative view, which also has a long tradition, bases the study of meaning on textual data. This is sometimes referred to as the view that 'meaning is use': and language use is clearly observable in texts. This text-based view also tends to emphasize that the units of meaning are not individual words, but longer multi-word units, collocations and extended phrasal units of various kinds.

1. ON EARLY WORK ON CONCORDANCES AND COLLOCATIONS

In 1737 Alexander Cruden published a concordance of the Bible (Cruden 1737). It presents the node word in eight or ten words of co-text, and sometimes lists separately concordance lines which reveal recurrent collocations (e.g. cases where *darkness* collocates with *light* or *day)*, and recurrent phrases (e.g. *Land of darkness* and *out of darkness*). Cruden spent many years in mental hospitals, but the question of cause and effect is unclear. Was he driven mad by preparing a concordance of the Bible manually and single-handedly? Or would you have to be mad to attempt this in the first place?

In 1790 Samuel Ayscough published an index to words in Shakespeare (Ayscough 1790). This presents node words in often just three or four words of co-text. However, the title is interesting: *An index to the remarkable passages and words made use of by Shakespeare; calculated to point out the different meanings to which the words are applied*. This gives a very explicit statement of the theory (which is often attributed well over a hundred years later to Wittgenstein and Austin) that 'meaning is use'.

Other work is often cited in the development of phraseology. In 1909 Charles Bally published work on the stylistics of French, in which he pointed out that 'conceptual units' are often idiomatic multi-word 'phraseological units' (Bally 1909). In 1933 Harold Palmer (1877-1949) published his *Report on Collocations* (Palmer 1933). He talks of 'many odd comings-together-of words'. In 1957 J R Firth published an article which is widely quoted for its famous statistical definition of collocation: 'You shall know a word by the company it keeps. [...] The habitual collocations in which words [...] appear are quite simply the mere word accompaniment.' (Firth 1957).

One other technical development was necessary for a systematic analysis of collocation: the KWIC (Keyword in Context) format for concordances, which (given computer technology) allows two innovations: a pattern can be aligned in the centre of the page or screen, and concordance lines can be re-ordered to the left and right. It is the possibility of rapidly re-displaying lines, alphabetically or by frequency, which makes patterns of co-occurrence visible. The term KWIC originates in work done in the late 1950s by Hans Peter Luhn (1896-1964). From 1941 he worked with others at IBM on methods of information retrieval, and invented a method of indexing books and articles. The idea was that a stop list of irrelevant words could be removed from titles, and that the main content words could then be permuted: that is separately aligned. (Stevens 1965, Soy 1998.) Key sentences from his 1960 article (Luhn 1960: 289) are as follows:

> 'Keyword-in-context indexing [...] may be applied to the title of an article, its abstract or its entire text. [...] By making the keywords assume a fixed position within the extracted portions [...] the KWIC index is generated.'

From the 1950s to the 1970s, information was often input into machines on punched cards, and early programming languages (such as Fortran) assumed that information was given on the cards in fixed columns, from 1 to 80. The idea of aligning keywords followed naturally from this. What is still not absolutely clear is when the importance of KWIC concordances was recognized by linguists. It cannot have taken long, since concordance packages were avialable from the mid-1960s. COCOA (COunt and COncordance Generation on Atlas) was developed in 1967, and the CLOC (CoLOCation) software was commissioned by Sinclair in the 1970s (Reed 1986).

2. ON CORPUS-DRIVEN STUDIES OF COLLOCATION

An early seminal text (Sinclair et al 1970/2004) is 'the OSTI report' (UK Government Office for Scientific and Technical Information). This reports quantitative research carried out between 1963 and 1969, but hardly accessible until it was formally published in 2004. The main findings are based on 135,000 words of computer-readable spoken text, and the report makes substantial progress towards a statistical theory of collocation. It formulates very clearly many of the basic questions and principles of modern corpus linguistics, in particular the concept of statistical tests for collocational attraction. Some questions, which are still at the centre of corpus studies, include: What kinds of lexical patterning can be found in text? How can collocation be objectively described? What size of span is relevant to collocation? How can collocational evidence be used to study meaning? It argues that the unit of lexis is unlikely to be the individual word in all cases, and that units of meaning can be defined via statistically defined units of lexis. In other words, the OSTI Report tackles a question which has still never had a satisfactory answer: how can the units of meaning of a language be objectively and formally identified? This tradition of corpus work was concerned, from the beginning, with a theory of meaning. In the careful phrasing of the report (p.6), there is a relation 'between statistically defined units of lexis and postulated units of meaning'.

One other early computer-assisted project (Allén et al 1975) notes the intellectual climate in the 1970s, when 'linguists tended to lay too much emphasis on introspection' (p.xxxi), and emphasizes the 'meaningful quantification' of authentic material. The authors argue that recurrence is 'the methodological foundation of the investigation' (p.xxxiii)and that phraseology (collocations and word combinations) is 'the area of intersection between

grammar and lexicon' (p.xxxii); and they provide a method of measuring how fixed words are in their use in given phrases (p.xlv).

A dramatic increase in the quantity of data (plus the ability to search and redisplay the data in different ways) leads to a significant change in the quality of analysis. The key techniques are: concordancing a large number of texts in KWIC format; permuting and redisplaying node words in concordance lines; extracting recurrent fixed word-strings (and their minor variants); and calculating the strength of attraction between co-selected words (this is not discussed here). When these techniques are combined they make recurrent patterns easy to recognize. Exactly what we can observe with concordance lines from a large general corpus is well discussed by (Tognini-Bonelli 2004):

> On the horizontal (syntagmatic) axis we have:
> a single instance of *parole* in the form a fragment of a meaningful speech act
> which has been extracted from an individual text / speaker
> (and which originally occurred as a piece of natural communicative behaviour).
>
> On the vertical (paradigmatic) axis we have:
> evidence of *langue* in the form of repeated formal patterns
> which have been extracted from language use across a speech community
> (and which have been displayed artificially by the linguist).

When they are displayed appropriately, concordance lines from many texts reveal repetitions which are evidence of units of meaning. We can see this in the small example below. To the left of the node phrase are semantically related lemmas (CLEAR, KNOW, SEE and ACCORDING TO, DEPENDING ON). To the right are forms of BLOW, most frequently *blowing*.

```
 1.  ity anywhere in the world.  It's clear which way the wind blows today.  All sig
 2. ike so many others, Baron de Tracy knew which way the wind blew and where gain w
 3. e been here before.  PLAYER: And I know which way the wind is blowing.  GUIL: Op
 4. he wind, but it mattered little to know which way the wind was blowing: wherever
 5. n: "You Don't Need A Weatherman To Know Which Way The Wind Blows."  In June 1968
 6. ered by his rage.  She had always known which way the wind was blowing, he thoug
 7. take anything for granted so when I saw which way the wind was blowing I just bu
 8. e affect on the game.  Difficult to see which way the wind is blowing at the mom
 9. lumb bob.  Hold their finger up and see which way the wind's blowing.  That's it
10  fter the acquisition is complete to see which way the wind is blowing.  "If we'r
11  r garden, as well as the ability to see which way the wind is blowing from the c
12  ws, lifting shutters, lingering  To see which way the wind blew, looking  Outwar
13   tonight," he told her, "I couldn't see which way the wind was blowing.  I decid
14. nary news-letter and then later, seeing which way the wind blew, had made an unh
15. ning.  Decisions are taken according to which way the wind blows", according to
16. enis Healey "wobbles about according to which way the wind is blowing".  And abo
17. the other end of the town, depending on which way the wind is blowing.  Occasion
18.  that would be either way, depending on which way the wind blew.  If the wind bl
19. soud, sometimes opposing him, depending which way the wind is blowing.  The batt
20. also ask the Right Honourable Gentleman which way the wind is blowing?  Before t
```

Without human intervention, the software can identify exactly repeated word-strings (n-grams). These are evidence for more abstract semantic patterns which are identifiable via subjective judgements about semantically related word-sets.

3. ON CORPUS TOOLS FOR PHRASEOLOGY EXTRACTION: *PIE*

The British National Corpus (BNC) is a widely used corpus of 100 million running words of spoken and written British English, and several interfaces to the BNC provide evidence

about recurrent phraseology. William Fletcher's PIE data-base (Phrases in English at http://pie.usna.edu) allows three types of recurrent strings to be extracted from the BNC:

- n-grams are uninterrupted strings of 1 to 8 orthographic word-forms
- p(hrase)-frames are n-grams with one variable lexical slot
- PoS-grams are strings of part of speech tags.

PIE allows the user to extract word-sequences defined by length (1 to 8) plus any combination of lexis and BNC grammatical tags. In any slot the user can specify

- nothing
- or a word-form or pattern e.g. kn?w* = *know, knows, knew, knowledge*, etc
- or a part of speech tag e.g. NOUN, LEXICAL VERB (ca 65 categories)
- or a combination of these e.g. kn?w* + VERB (excludes *knowledge*)

PIE is a major advance on static word-frequency lists, since the user can generate different lists of both single words and multi-word units: for example, the 100 most frequent nouns, the top 100 adjective-noun combinations, the top 50 4-grams (content unspecified), frequent 6-grams which end with the word *way* (e.g. *let me put it this way*), or all n-grams of length 2 to 8, which contain the word *way* in any position (see below). The total number of definable patterns is astronomically high. We transform the data into quantities. This allows patterns to be seen. We then interpret the quantitative patterns as evidence of phrasal constructions. What is frequent in the corpus is evidence of what is typical in the language use of the hundreds of speakers who are represented in the corpus.

4. ON SINCLAIR'S MODEL OF EXTENDED LEXICAL UNITS

A major theoretical proposal to come out of corpus studies is Sinclair's (1998, 2005) model of an extended lexical unit. This has the following structure.

[1] COLLOCATION is the relation between the node word and individual word-forms which co-occur frequently with it.
[2] COLLIGATION is the relation between the node word and grammatical categories which co-occur frequently with it.
[3] SEMANTIC PREFERENCE is the relation between the node word and semantically related words in a lexical field.
[4] SEMANTIC PROSODY is the discourse function of the unit: it describes the speaker's evaluative attitude.

Relations [1] to [4] are increasingly abstract.

[1] Collocates are word-tokens: individual word-forms, directly observable in texts.
[2] Colligation refers to word-classes (such as past participles or quantifiers), not directly observable, but abstractions based on the behaviour of the word.
[3] Semantic preference refers to a class of words in a lexical field which share some semantic feature. This will relate to the topic of the surrounding co-text.
[4] Semantic prosodies are the motivation for speaking, and therefore related to concepts of speech act and illocutionary force.

In summary, the structure of extended lexical items is as follows:

[1] collocation	tokens	co-occurring word-forms
[2] colligation	classes	co-occurring grammatical classes
[3] semantic preference	topics	lexical field, similarity of meaning
[4] semantic prosody	motivation	communicative purpose

This model integrates lexis fully within the traditional concerns of linguistic theory. A lexical unit consists of lexical, syntactic, semantic and pragmatic components. Relations [1] to [4] correspond to the classic distinctions drawn by Morris (1938). Syntax deals with how linguistic signs relate to one another (here collocation and colligation), semantics deals with how linguistic signs relate to the external world (here lexical sets and the phenomena they denote), and pragmatics deals with how linguistic signs relate to their users (here expression of speaker attitude).

5. FROM WORDS TO PHRASEOLOGY: *WAY*

If we extract the most frequent phraseology around frequent words, we can investigate a general hypothesis (proposed by Sinclair et al 1970, Sinclair 1999, Summers 1996): that frequent words are frequent because they occur in frequent phrasal constructions which express conventional pragmatic functions in text. For example, the word-form *way* is amongst the top ten nouns: somewhere around rank 100 in frequency lists from large corpora. Its general phraseology has been discussed within both pattern grammar (Sinclair 1999, Hunston & Francis 2000: 101-02, 110) and construction grammar (Goldberg 1996).

The *Oxford English Dictionary* documents different senses which have developed historically from concrete to abstract, to give a range of meanings, some purely idiomatic.

- position: *the other way round*
- method: *the correct way of holding it*; *the best way to do it*
- temporal: *always*; *all the way through the film*
- adverbial: *away*
- concessive: *in a way*
- discourse marker: *by a long way*, *anyway*, *by the way*

This list illustrates that the word *way* is frequent because it occurs in frequent phrases. It only rarely conveys the literal denotation "path". Its literal denotation is often weakened, and the evaluative connotation of the resultant phrase is often strengthened.

PIE provides us with a systematic method of identifying such phrases. We can start from a node word, and extract all recurrent n-grams with the node at each position, down to a cut-off frequency (say at least one per ten million in the BNC). We can identify variants of these n-grams (using the p-frame and concordance software); identify recurrent patterns around the unit (i.e. evidence of its meaning); and look for other strings which are semantically and functionally related. For example, the word *way* frequently occurs in topic-independent phrases which express pragmatic meanings. They either contribute to information management or express general evaluative meanings. Examples include:

- (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

Although native speakers cannot generate comprehensive lists of such phrases from introspection, they recognize them as idiomatic and conventional ways of expressing culturally important meanings. Several have have strong pragmatic meanings, including speech acts (e.g. threat) and discourse markers (which open or close discourse sequences).

The software allows us to search for n-grams containing the node word at each position in strings of decreasing lengths. With 8-grams this looks like this:

way xxx xxx xxx xxx xxx xxx xxx
xxx **way** xxx xxx xxx xxx xxx xxx
xxx xxx **way** xxx xxx xxx xxx xxx
xxx xxx xxx **way** xxx xxx xxx xxx
xxx xxx xxx xxx **way** xxx xxx xxx
xxx xxx xxx xxx xxx **way** xxx xxx
xxx xxx xxx xxx xxx xxx **way** xxx
xxx xxx xxx xxx xxx xxx xxx **way**

For example, here are all 7- and 8-grams with *way* in any position which occur 10 times or more in the BNC:

	gram	freq
	in such a way as to make	26
	in such a way that it is	26
>	is still a long way to go	20
>	there is a long way to go	20
>	there is still a long way to	20
>	there is still a long way to go	20
	in much the same way as the	19
	that 's the way to do it	17
>	there 's a long way to go	17
	there was no way i was going	16
	there was no way i was going to	15
	was no way i was going to	15
>	have a long way to go before	14
	one way of doing this is to	14
	in such a way that it can	13
>	still have a long way to go	13
	in such a way as to ensure	12
	it in such a way that it	12
	laughing all the way to the bank	12
	to do with the way in which	12
>	we have a long way to go	12
	will the right hon. gentleman give way	12
>	a long way to go before we	11

	as she made her way to the	11
	example of the way in which the	11
	that 's the way it should be	11
>	's still a long way to go	10
	and the way in which they are	10
	be treated in the same way as	10
	in such a way as to exclude	10
	it 's a long way from the	10
	the only way you can do it	10
	to stand in the way of the	10

Let us then take the most frequent repeated pattern (marked in the data above):

- there is (still) a long way to go
- we (still) have a long way to go (before) we

No occurrences in the BNC refer to concrete and literal journeys, such as: "we set off from Scotland at 8 o'clock this morning, but **there is still a long way to go** before we arrive in London". All occurrences have a conventional abstract metaphorical interpretation, and concordance lines show that it is part of a longer narrative sequence. The words in bold appear in different occurrences.

> The speaker is encouraging people who have responded to a **challenge**: much has been **achieved**, the **early signs are encouraging**, and things look **promising**. **However**, **in spite of improvements**, the speaker **realises** or **admits** that much **still needs** to be done: **plans are underway**, but the speaker is now **warning** that an **effort** is **needed** to **make a success** of things, and that there will be a **struggle before** the target is reached.

Across the whole corpus the collocates make explicit the pragmatic force of the construction. To the left are words from the semantic fields of "achievement" and "progress". To the right are words from the fields of "plans" and "struggle". In individual texts the specific collocates contribute to textual cohesion. Once we have this candidate for a phrasal construction, we can use the p-frame software plus concordance lines to identify related recurrent strings. This depends on subjective decisions: but these decisions are based on replicable quantitative data. A wider search produces other, also highly conventionalized, ways of expressing this complex speech act.

- I believe [this] is starting to show improvements but **we still have a long way to go**. It is my intention this approach should extend throughout the company.
- While saying that some measures the company had taken to get back on track were bearing fruit, he warned: 'We should not let ourselves become overconfident. **There is still a long and difficult road ahead of us** before we can become truly competitive again'. [*Time,* 166/17, 24 Oct 2005, p.39.]

The construction has the structure of a lexical item proposed by Sinclair (1998: 14-20):

[1] Largely fixed lexical core (*there is / we have, still, long, way / road*).
[2] Colligation: frequent syntactic frame (*but / so ... before*).
[3] Semantic preference: lexical fields of "progress", "challenge".
[4] Semantic prosody: encouragement plus warning.

A sample dictionary entry, which gives the canonical form with its main variants and pragmatic function, might look like this:

	{there BE	(still)	}
(but / so)	{PRONOUN	(still) HAVE	} a long way to go (before)
	{PRONOUN 've / 's	(still) got	}

This conventional speech act expresses encouragement about achievements in the past plus a warning not to give up in the face of opposition in the future.

A second construction has been intensively studied within both construction grammar (e.g. Goldberg 1996) and pattern grammar (Francis et al (1996: 330-38). Examples are:

• she	made her way	along the corridor
• he knew he must	find his way	through that cave
• they	worked their way	up the stream
• in a futile attempt to	claw his way	back to the surface
• they	forced their way	through the thick vegetation
• he must	grope his way	into the labyrinth
• he	picked his way	back down the ladder

PIE can identify the most frequent verbs: MAKE and FIND. Other verbs (e.g. FORCE, PUSH, WORK, SMASH, WIND, CLAW, PICK, THREAD, ELBOW, GROPE) indicate the typical semantic prosody: difficulty where force is used or where care is necessary, and/or where there is an awkward winding movement. Semantically related verbs which occur in the construction are described by Francis et al (1996: 330-38). Often there are unpleasant connotations, of force and violence (*a burglar burning his way through a safe*), of dishonesty (*people who bluff their way through music*), of illegal activities (*politicians who cheated their way to government*), of stupidity (*we muddled our way through*), or of general unpleasantness (*I unpeeled my way through the sodden address book*). However, the construction is very productive, and many other verbs occur (not all with unpleasant connotations). The pattern involves

[1] Largely fixed lexical core (*way*). MAKE is the most frequent verb.
[2] Colligation: possessive determiner + direction expression.
[3] Semantic preference: most other lexical verbs from predictable lexical fields.
[4] Ssemantic prosody: difficulty in overcoming some barrier.

The pattern is productive precisely because the meaning is associated with the pattern rather than with individual words within it (Hunston 2002: 140): so *unpeel* can be interpreted without difficulty as a verb expressing purposeful movement:

- I unpeeled my way through the sodden address book [attested].

6. CONCLUSIONS

The techniques I have discussed are very good at revealing communicative acts of specific kinds. If recurrent word-strings are both frequent and fairly evenly distributed across a corpus, then it follows that they have little to do with the content of individual texts, but rather with general communicative functions which speakers frequently express, independently of what they are talking about. The techniques can discover conventional ways of managing information and of expressing 'the typical meanings that human communication encodes' (Francis 1993: 155). There are no purely automatic inductive discovery methods for identifying phrasal constructions. However, automatic methods can find recurrent strings with limited formal variation and provide empirical quantitative data on phraseology. All methods have their limitations, but the question is not: Do they tell us everything we want to know about phraseology? (Clearly no.) But are they better than trying to discover patterns by introspection? (Clearly yes.)

ACKNOWLEDGEMENTS

For information of the history of corpus linguistics I am very grateful to John Sinclair, Bill Fletcher and Jutta Steckeweh. Bill Fletcher not only generously provided me with copies of original articles by Hans Peter Luhn and others, but he is responsible for designing and implementing the PIE data-base from which I have drawn my data.

REFERENCES

Allén, S. et al 1975. Nusvensk frekvensordbok. Stockholm: Almqvist & Wiksell.

Ayscough, S 1790. An Index to the Remarkable Passages and Words Made Use of by Shakespeare. London: Stockdale.

Bally, C. 1909. Traité de stylistique française. Geneva: Librairie Georg & Cie.

Cruden, A. 1737. A Complete Concordance to the Holy Scriptures. London: Tegg.

Firth, J. R. 1957. A synopsis of linguistic theory 1930-1955. Transactions of the Philological Society. Special Volume. Studies in Linguistic Analysis. Oxford: Blackwell. 1-32.

Fletcher, W. 2003-05. PIE Phrases in English.. http://pie.usna.edu.

Francis, G. 1993. A corpus-driven approach to grammar. In M. Baker et al (eds.) Text and Technology. Amsterdam: Benjamins. 137-56.

Francis, G., Hunston, S. & Manning, E. 1996. Grammar Patterns. 1: Verbs. London: HarperCollins.

Goldberg, A. E. 1996. Making one's way through the data. In Shibatani, M. & Thompson, S. A. (eds) Grammatical Constructions. Oxford: OUP. 29- 53.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Oxford: OUP.

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

Luhn, H. P. 1960. Keyword-in-context index for technical literature. American Documentation, xi, 4: 288-95.

Morris, C. W. (1938) Foundations of the theory of signs. In O. Neurath, R. Carnap & C. W. Morris eds International Encyclopedia of Unified Science. Chicago: Chicago University Press. 77-138.

Palmer, H. E. 1933. Second Interim Report on Collocations. Tokyo: Kaitakusha.

Reed, A. 1986. DOC: CLOC V00309. Available at http://www.decus.org/libcatalog/

document_html/v00309_1.html. Accessed Nov 2005.
Sinclair, J. 1998. The lexical item. In E. Weigand (ed) Contrastive Lexical Semantics. Amsterdam: Benjamins. 1-24.
Sinclair, J. 1999. A way with common words. In H. Hasselgård & S. Oksefjell (eds) Out of Corpora. Amsterdam: Rodopi. 157-79.
Sinclair, J. 2005. The phrase, the whole phrase and nothing but the phrase. Plenary lecture to Phraseology 2005, Louvain-la-Neuve, October 2005.
Sinclair, J. M., Jones, S. & Daley, R. 1970/2004. English Collocation Studies: The OSTI Report. (Ed.) R Krishnamurthy. London: Continuum. [Originally circulated as a mimeoed report in 1970.]
Soy, S. 1998. Class notes: H P Luhn and automatic indexing. Available at http://www.gslis.utexas.edu/~ssoy/organizing/l391d2c.htm. Accessed Nov 2005.
Stevens, M. E. 1965. Automatic indexing: a state of the art report. Available at http://www.itl.nist.gov/iaui/894.02/works/pubs/mono91/01.txt. Accessed Nov 2005.
Summers, D. 1996. Computer lexicography: the importance of representativeness in relation to frequency. In J. Thomas & M. Short (eds) Using Corpora for Language Research. London: Longman. 260-66.
Tognini-Bonelli, E. 2004. Working with corpora. In C. Coffin et al (eds) Applying English Grammar. London: Arnold. 11-24.