In In H Hasselgård, J Ebeling & S Oksefjell Ebeling eds. (2013) *Corpus Perspectives on Patterns of Lexis*. Amsterdam: Benjamins. 13-33.

SEQUENCE AND ORDER.
THE NEO-FIRTHIAN TRADITION OF CORPUS SEMANTICS.

Michael Stubbs

Abstract

Corpus linguists often attempt to avoid assumptions imported from pre-corpus studies, by using methods which could be called "inductive", in so far as they proceed from observations about textual sequences to generalizations about order in the system. However, induction has been questioned for over 400 years (by Bacon, Hume, Popper and others), and the possibility of rigorous, theory-free induction is now generally rejected. One major phraseological model, proposed by Sinclair in the late 1990s, is certainly not a purely inductive generalization from raw corpus data. I will discuss this model using attested data on a particular construction and a distinction proposed by Firth, Halliday and Palmer between "sequence" (an observable feature of texts) and "order" (a feature of linguists' models).

1. The Neo-Firthian tradition

In the neo-Firthian tradition of linguistics, the key concern is meaning. This is obvious just from the titles of some of the main publications: "The problem of *meaning* in primitive languages" (Malinowski 1923),"The technique of *semantics*" (Firth 1935), "Learning how to *mean*"(Halliday 1975), "The search for units of *meaning*" (Sinclair 1996), "*Meaning*, discourse and society" (Teubert 2010). There could be nothing more explicit than sentences from the opening paragraph of Firth's article "Modes of *meaning*" (1951), where he says "The study of meaning is a permanent interest of scholarship ... [T]he main concern of descriptive linguistics is to make statements of meaning."

Nothing could be clearer than that. However, much linguistics – including much corpus linguistics – has avoided tackling meaning directly, and has therefore avoided the central puzzle which distinguishes the natural sciences from the social sciences. Things which natural scientists study – atoms, earthquakes, whatever – have no inherent meaning, but things which social scientists study have already been pre-interpreted by the members of society.

2. Sequence and order

Firth (1957) makes the simple but useful distinction between sequence and order to clarify the relation between observable textual data and theoretical semantic models. Firth's own writings are notorious for making difficult reading, but the distinction was quickly elaborated by his immediate students and colleagues, including Halliday (1961) and Palmer (the editor of a collection of Firth's papers). I have borrowed the title of my paper from an article by Palmer (1962). The distinction is still used by Sinclair and Mauranen (2006: 71), who say: "[I]n Firth's terms [...] sequence must be replaced by order. Order can take many forms."

The slightly different formulations in these publications can be summarized as follows. Sequence is a feature of raw data. It is concrete and linear – linear in time for spoken language and in space for written language. It is observable, and with the help of technology, we can observe the frequency of things occurring in sequence. In a rough sense, we can then make inductive generalizations about these things. However, the generalizations involve order. Sequence is one exponent of order, but order is abstract, multi-dimensional and not directly observable. It is a theoretical construct, which relies on interpretation and deduction. Firth often saw it as psychological: a case of what we expect to occur next in the linear sequence.

All one can directly observe in the linear stream of raw corpus data is co-occurrence and span, which are features of individual texts, and recurrence, which is a feature of multiple texts from independent sources. Strictly speaking, in a corpus only two things are observable: frequency and distribution. An item may occur only rarely, and therefore only in a few texts. Alternatively, it may be frequent but unevenly distributed in only a few texts, or frequent and widely distributed in many texts. Therefore, statements about frequency must always be related to statements about distribution.

If the aim is to model meaning, both similarity and variation must be taken into account, and one is dealing with order. It is clear from Firth's original definition of collocation that it is a semantic abstraction: which makes it a question of order not sequence. One of his most famous statements is:

> Meaning by collocation is an abstraction at the syntagmatic level [...] One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*. (Firth 1951: 196) [NOTE 1.]

Although Firth gives examples of what he has in mind, he is not very explicit about the possible variability in sequence, span and word-form. A few examples such as those in (1) to (6) show that the collocation can be of adjective plus noun, noun plus noun, noun plus verb, etc, and that different forms of the lemma / word family can occur in different sequences and in different spans. This immediately raises the problem of identifying linguistic units, but unfortunately, Firth does not tell us which sequences of observable word-forms count as tokens of the "same" collocation.

(1) a *dark night* in October
(2) the *dark*est of *night*s
(3) the *dark*ness of the long winter *night*
(4) the *night*s grew *dark*er and colder
(5) at *night* under cover of *dark*ness
(6) at *night*fall just as the sky *dark*ened

All my examples in this article are attested (except for a few which are used for comparison and explicitly marked as "invented"). Initial examples are from the British National Corpus (BNC), usually collected via BNCweb (Hoffmann et al 2008). Others are from the world-wide-web, usually collected via WebCorp (Renouf et al 2007). LEMMAS are represented in upper case, *word-forms* are in lower case italic.

3. An example: the *went-and-VERBed* sequence

In the rest of this paper, my examples of problems in identifying phrasal units of meaning mainly follow Sinclair's way of modelling such units (Sinclair 1996, 1999). As the work developed, various terms were used: "phrasal unit", "extended lexical unit", "semantic shift unit". A useful term is "text segment", which emphasizes that these units are not isolated speech acts, but that they have a communicative function in connected texts: this is a model which relates language and communication. The model gained immediate attention when it was proposed in the late 1990s, but there is still confusion, particularly around the concept of semantic prosody.

The sequence *went-and-VERBed* can be used in various ways. Most cases involve a literal reference to a movement:

(7) I *went and* stood in the doorway of my office
(8) she got to her feet and *went and* picked up the phone

However, even here, a compositional semantics is not quite sufficient. We can substitute *came* for *went*: *she **came** and picked up the phone* [invented example]. But both (7) and (8) would normally be interpreted as single integrated events. For example, it would be very odd to interpret (8) as meaning that she went somewhere and then, in addition, some time later, picked up the phone. In other cases, we have the same sequence of orthographic words, but the "movement" meaning of *went* seems largely irrelevant:

(9) the news got around [...] when somebody *went and* rang up a newspaper

The *went* here could be interpreted as indicating movement, but it seems at least odd to substitute *came* for *went* (? *somebody came and rang up a newspaper* [invented example]). And if we ask *why did they **go and** do that?* [invented example], it is natural to interpret the question as a way of expressing surprise or

irritation at something unexpected. In other cases again, a "movement" interpretation is most unlikely: the sequence expresses something else, such as disbelief or disapproval. (10) and (11) cannot be interpreted literally and compositionally.

(10) would you believe it, she *went and* married him
(11) they *went and* lost by six points to twenty one

The evaluative uses of the lexico-grammatical unit discussed in this article most often refer to past time, and are most often realized by the sequence *went-and-VERBed*. However, other forms occur, so I will use *GO-and-VERB* as a shorthand term for the unit.

The unit is mentioned in standard reference grammars, which sometimes call it "pseudo-coordination" (Quirk et al 1985: 978), and which list some of its characteristics. In terms of its distribution: it is rare in formal written varieties, but fairly frequent in casual conversation (Biber et al 1999: 537, 1031). In terms of its semantics: it is used when talking about two actions which are seen as a single event or "closely linked" (Sinclair et al eds 1990: 3.201). In terms of its pragmatics: it is said to be often derogatory (Quirk et al 1985: 507, 978), and to express "emotive meanings" such as "disapproval, annoyance, surprise or the like" (Huddleston & Pullum 2002: 1303).

However, the discussions in the standard grammars are very brief. There is more helpful discussion of the semantics by Wulff (2006), who studies over 5,000 occurrences of *GO-and-VERB* in the BNC, and finds mainly verbs denoting a completed action. And there is more helpful discussion of the pragmatics by Hopper (2001), who mentions examples with GO only briefly, but discusses other related double-verb constructions. He discusses both their attitudinal meanings of annoyance and frustration, and their discourse functions, arguing that they introduce a new sequence in a narrative and emphasize the speaker's current main point.

All in all, *GO-and-VERB* is the kind of unit which one would expect to be a candidate for grammaticalization: a frequent motion verb is semantically weakened and the whole sequence is pragmatically strengthened. [NOTE 2.] The grammaticalization of *going to* as a marker of the future is well known, but there seem to be no diachronic studies of the grammaticalization of *GO-and-VERB*.

In this paper I will now concentrate on cases where the co-text provides evidence for a phrasal unit which expresses disapproval or annoyance. In examples (12) to (14), which are clearly in the middle of an on-going narrative, the speakers are categorizing participants very negatively and thereby evaluating an event:

(12) but *some bastard* went and stole my pens and paper
(13) and then you went and danced with *that lout*
(14) and then *the silly little girl* went and believed the glib tales she was told by the great Australian male, didn't she?

In examples (15) to (17) each speaker is also evaluating an event within a narrative sequence. In (15) and (16) the co-text contains colloquial forms. In all three cases, the *went-and-VERBed* sequence signals that the speaker thinks that someone has done something stupid. This interpretation may depend on shared cultural assumptions, which means that we cannot analyse the language alone, but have to take into account knowledge about the world.

(15) then you *went and* forgot didn't ya?
(16) then I *went and* rode me bike into that mooring rope
(17) then you *went and* set your blanket on fire one night

There are other – even more colloquial – examples which contain formal signals of the speaker's disapproval and frustration. These are very numerous, especially in web data.

(18) the fuckin' fool *went and* sold out to Penumbra
(19) Ian *went and* bloody got married
(20) she only *went and* bloody forgot
(21) fucking *went and* arrested him
(22) have you [...] seen what [...] Harper's *went and* fucking done?

Note how different items contribute to the textual cohesion. For example, in (18), the speaker describes someone as a *fool*, and describes their action very negatively as *sold out*. The basic argument is that evidence of the meaning of different instances of the *went-and-VERBed* sequence is provided (a) by items recurring across many independent texts and (b) by items co-occurring elsewhere in the linear sequence. Meaning is not only a private psychological matter, but also a public state of affairs. The meaning is in the discourse (Teubert 2010).

Examples (9) to (22) all illustrate a conventional way for a speaker to indicate that something was going OK, and then went wrong. In some examples, the collocate SPOIL (plus a very few near synonyms) makes this explicit, as in many parallel examples such as:

(23) but then went and spoilt it all
(24) then she went and spoiled everything
(25) there the wee man goes and spoils it all
(26) of course my parents have to go and fucking spoil it
(27) then they went and fucking ruined it
(28) [they] had promising careers but then went and blew it

In these cases, there is a largely fixed unit in which only minor paradigmatic variants are possible, such as *SPOIL it all / everything / the whole thing*. Occasionally a different verb occurs: *went and spoiled it / ruined it / blew it*. But only minimal formal variation is possible. Note, however, that these statements make large assumptions about what are tokens of the "same" type, and about what is the "same" unit underlying surface variation.

In summary so far: Sinclair (in prep: chapter 12) and Teubert (2010) have emphasized that paraphrase and intertext together provide a technique of semantic analysis: they are analytic tools for studying meaning. Evidence of meaning is provided (a) by looking at the intertext and identifying recurrent segments which are formally identical or similar and (b) by looking at the immediate co-text and identifying co-occurring segments which are semantically similar. In the ideal case, the co-text will provide a paraphrase of the segment which we are interested in.

This gives us a natural way of relating text and intertext. On the horizontal syntagmatic axis of a concordance are fragments of text which often provide evidence of the meaning of an expression. On the vertical paradigmatic axis are intertextual links to similar items which have often occurred in the past. This is reminiscent of – but significantly different from – the traditional representation of syntagmatic and paradigmatic structure which derives from Saussure. In the traditional concept, on the horizontal syntagmatic axis we have what is actually co-present in the linear string, whereas on the vertical paradigmatic axis we have what could potentially have been present, but isn't. With a concordance, we have something that looks superficially similar, but the crucial difference is that corpus linguistics deals with the actual, not with the potential. On the horizontal syntagmatic axis, in individual concordance lines, we again have – as in the Saussurian model – what is actually co-present in the linear string. But on the vertical paradigmatic axis, we have what was actually present in other texts in the past. [NOTE 3.]

Intertextuality has been famously studied in literary theory, but largely neglected in linguistic theory. Paraphrase has also been largely neglected in linguistic theory, though it is central to a theory of meaning (Sinclair in prep). Semantic prosody has been much discussed (and disputed), but semantic preference has been much less discussed, although it concerns what a text is about: its topic.

4. Induction?

The examples above are all attested in corpus data. This allows the analysis to be inductive, in the rough sense that it can start with many independent observation statements, which are taken as reliable and true, and which can then be used to formulate a generalization. But the question is also: what do corpus linguists actually do? – bearing in mind that linguists often say in their published papers that they have arrived at their findings in one way, although they actually got their ideas in some rather different way.

Corpus linguists often claim that their work is either corpus-driven or corpus-based. The terms "corpus-driven" and "data-driven" were originally used by Francis (1993), in order to emphasize the fundamental difference between this approach, in which "the corpus is the main informant [... and ...] the only reliable

authority" and the approach of Quirk et al (1985) who had access to small corpora but apparently did not use them extensively (Francis 1993: 138-39). In the early days of corpus linguistics, it was certainly important to emphasize that things can be discovered in a corpus which can never be imagined by introspection. Almost twenty years later, due in particular to the use of the terms by Tognini-Bonelli (2001: 10-11), the driven/based distinction is now very well known, but its exact meaning is still being debated, for example in two recent issues of the *International Journal of Corpus Linguistics* (2010, 15 (3, 4)).

The distinction is intended to signal whether data are used (merely) to illustrate or test old categories of linguistic order, which have been taken from earlier linguistic theory (this is corpus-based analysis), or whether it is possible to induce new findings from sequences of raw textual data, and thereby avoid assumptions and self-fulfilling prophecies (this is corpus-driven analysis). The corpus-driven concept is clearly related to the concept of induction, although in all the discussion about a corpus-driven approach, there is hardly any reference to the intensive debate about induction over the past 400 years or so. The concept is usually attributed to Francis Bacon in the 1600s, though in fact it goes back much further. In the 1700s, David Hume expressed scepticism of the concept, since what has happened in the past cannot guarantee what will happen in the future. From the 1930s onwards, this scepticism was expressed even more strongly by Karl Popper (1975: e.g. 46ff), who argues that induction is simply a myth. However, by the 1960s, in a magisterial overview of the debate, Max Black (1967) is more relaxed about the concept.

Traditionally, induction is said to proceed from the particular to the general. Black (1967) introduces a sub-type of induction, which proceeds from particulars to further particulars: from the observation of particular cases to the probability of observing further similar cases in future. The traditional hope is that one can proceed from a large number of particulars to a general conclusion, but there is also the kind of case which I have illustrated, where parallel examples imply that we will find similar but variable examples in future. An argument from parallel cases is a sub-type of induction, which omits the claim to make an explicit generalization. [NOTE 4.]

Then, as I say, the question is also: what do corpus linguists actually do? One thing they do is to use parallel visual arrays as a persuasive device: concordance lines, and also frequency tables of various kinds, which are aligned and sorted to show similarities and variations. This is exactly what I did in my example of *went and spoiled*.

If we set things out in this way, we can see – at the same time – both semantic similarities, and formal variants. To the right, we have phrases meaning "all", to the left we have a discourse marker signalling an important turning point in the narrative.

```
                went and spoiled it all
                went and spoilt everything
                went and spoiled every game
                went and spoiled the whole thing
                went and spoiled the whole film
      then you went and spoiled ...
    and then I went and spoilt ...
  but then he went and spoilt ...
once again he went and spoilt ...
```

FIGURE 1 ABOUT HERE.

FIGURE 2 ABOUT HERE.

This kind of tabular presentation became common in the 1800s in illustrations in books on entomology. In illustrations such as Figure 1, we are intended to see that the insects are similar, but different, and if we see resemblances and parallels and make comparisons, then no verbal argument is necessary. (Fahnestock 2003.) It is generally accepted, I think, that people remember pictures better than verbal arguments. Seeing is often believing, but the patterns have to be perceived by the analyst. Probably the most famous example is the one shown in Figure 2: John Gould's drawings of finches which Charles Darwin had brought back from the Galapagos Islands. Gould drew them all facing in the same direction, and with the same degree of idealization. It was apparently only after Darwin had seen this visualization of the data, that he realized their significance for his theory about how to lump birds together as similar, and therefore as members of one species, but variable due to evolutionary change. (Darwin 1845.)

Similarly, corpus linguists often present findings in parallel arrays, which may be random samples, or (as in Gould's drawings) selections to illustrate a pattern. They use tabular displays as a persuasive device: concordance lines, word frequency lists, n-gram frequencies, word profiles, Zipf-type distributions, collocate clouds, dispersal plots, etc. These tabular presentations are helpful when we have patterns of similarities and variations which are visible only with the help of software. The software cannot see patterns, but it can rip texts apart, and shuffle the pieces into different formats, which allows humans to see patterns (though this still ignores the problem of what counts – intuitively – as the "same" pattern).

This way of setting out linguistic data in tables is not entirely new. We find it, for example, in 19th century historical-comparative linguistics. Saussure gives examples of paradigms in Latin, Greek and Sanskrit, and maintains that it is enough just to glance at them in order to see the relations: "Il suffit d'y jeter un coup d'oeil pour apercevoir la relation" (Saussure 1916: 15). This is of course nonsense, as Harris (2004: 102) points out: all an untrained eye can see is "a set of correspondences and lack of correspondences between three sets of spellings". To be fair, Saussure (1916: 151) also points out that there is no absolute or objective

measure of sameness: he uses the analogy of a train which leaves from Geneva for Paris at 8:45 on two on successive days. We regard it as the "same" train, although it probably consists of completely different locomotive and carriages. Some facts are best presented in tables, columns of statistics, and so on, but patterns (and their significance) depend on the point of view, interests and experience of the observer. You have to learn to see order in the sequences. This is perhaps the most basic problem in identifying phrasal units.

In summary: We can observe sequence, but we can only model order. There is widespread consensus, I think, that it is valuable to work with minimal assumptions, and to be suspicious of premature theorizing, but that it is not possible to start *tabula rasa*, that there is no neutral observation language, that there is therefore no pure induction, and that the only requirement is to formulate an idea clearly and test it. That is, we take a model and deduce its consequences. It is irrelevant how we get an idea: the important thing is to test it, and testing is deductive, not inductive.

I think it is also now fairly widely agreed that the whole distinction between corpus-driven and corpus-based has been rather exaggerated. First, the idea that empirical experience is the only guarantee of interesting theories was by and large abandoned long ago as a positivist error. Second, we can reduce the distinction between corpus-driven and corpus-based to the much simpler distinction between observable sequence and theoretical order. And third, the corpus-based position emphasizes continuity with previous work, whereas the corpus-driven position emphasizes a break. Perhaps the whole debate has more to do with academic politics than with empirical methods. These are three independent reasons for abandoning the distinction.

These points are discussed in detail in several places by Popper. For example, Popper (1975: 21-30) argues that it is a basic error to confuse the origins of knowledge with the validity of knowledge. If we doubt some claim (hypothesis, theory, etc), the thing to do is not to ask where it came from (from a corpus? from earlier theory?), but to test it. It is irrelevant to try and discover the origins (or pedigree) of an idea: this would lead, in any case, to an infinite regress. It does not matter where a theory originates, but only whether it is well-tested and ultimately whether it is correct. [NOTE 5.] Most theories are, of course, not correct, and when they are rigorously tested, errors are discovered, better theories are put in their place and then tested in turn, and knowledge develops as we learn from our errors. In other words, it is meaningless to try and distinguish between corpus-driven (ideas come from raw data) and corpus-based (ideas come from earlier theories): all ideas come from a mixture of different sources. [NOTE 6.] Empiricism does not imply that observations (sense impressions, etc) function as the true and untainted source of ideas (they don't and can't). It means that theories are tested against observations and improved when they turn out to be wrong.

5. Sinclair's model of units of meaning

Within the neo-Firthian tradition, Sinclair (e.g. 1996, 1999) makes the most sustained attempt to formulate a phrasal model of units of meaning. His model has four parameters which relate form, meaning and communicative function. Parameters 1 and 2 concern lexical and grammatical form. Parameters 3 and 4 concern topic and speech act.

If we go back to my opening examples, we find that they fit this model exactly. We have a core item and four parameters. The core is GO-*and* plus a VERB.

1. In terms of words: there are no very strong collocates, but SPOIL is prototypical, in ways that I will show below.
2. In terms of grammar: it usually follows a discourse marker, such as *and, but, so, then*; it is usually past tense.
3. In terms of the topic of the text: it is used when the speaker is talking about something which was going OK, but then went wrong.
4. In terms of the speech act: it is used when the speaker is expressing surprise and/or annoyance at someone's behaviour. It evaluates some event, and emphasizes the current main point of a narrative.

A more formal statement, using Sinclair's terminology, is as follows.

1. COLLOCATION is a sequence of co-occurring word-forms: this is the basic observational evidence. [NOTE 7.]
2. COLLIGATION concerns classes: words occurring within grammatical structures.
3. SEMANTIC PREFERENCE concerns the topic: co-ordinated choices in text.
4. SEMANTIC PROSODY concerns the speaker's evaluation: their communicative purpose in saying this now.

(3) and (4) both emphasize that these units are text segments – not isolated speech acts.

There are several logical relations between the parameters. As we move from (1) to (4), the features move from those which are objectively observable – and therefore identifiable with software – to those which require the subjective interpretation of the analyst. We move from sequence to order.

1. COLLOCATION is merely orthographic word-forms in linear sequence.
2. COLLIGATION involves syntactic categories, which can often be reliably identified, but are nevertheless abstract (e.g. negatives or modal verbs).
3. SEMANTIC PREFERENCE involves an intuitive understanding of semantic fields and of the topic of the text.
4. SEMANTIC PROSODY involves formulating generalizations about the speaker's evaluations and attitudes.

The concept which has probably attracted most interest is "semantic prosody", but there has been considerable confusion in the literature, especially about the relation between semantic preference and semantic prosody. [NOTE 8, 9.] The essential difference is as follows. Semantic preference concerns propositional content. It has to do with sense and reference: what the text is about. Semantic prosody concerns speech act force: the speaker's communicative purpose. The distinction is very similar both to Austin's (1962) distinction between locution and illocution, and also to Gazdar's (1979) distinction between semantics, which studies meaning as truth conditions, and pragmatics, which studies meaning minus truth conditions.

A further source of confusion has been whether all words have a semantic prosody, but this misses the point, namely that most words occur in longer phrasal units, and these units have predictable communicative functions. Semantic prosody is the motivation for using the text segment now. We could say that semantic prosody has two aspects: illocutionary force (e.g. making a complaint) and discourse management (e.g. emphasizing the narrative focus). Or we could make discourse management a separate fifth parameter in the model. I initially thought that this is merely a trivial question of terminology. However, I now think that splitting semantic prosody in two can usefully emphasize that any utterance is always a response to a previous utterance and that any phrasal unit is always a text segment in a longer text. This makes explicit that the phrasal model implies a dialogic view of language.

If we go back to a very simple way of putting things: Collocation and colligation have to do with how something is expressed (the form), semantic preference has to do with what is expressed (the topic), and semantic prosody has to do with why it is expressed (the speaker's motivation). The model combines form, content, speech act force and discourse function, and therefore contributes to a theory, not just of language, but of communication. This connection is ignored in much linguistics, but then it is only possible to make the connection if the analysis starts with *parole* rather than *langue*.

A more formal formulation is as follows. Parameters (1) and (2) concern the relations between signs and other signs. Parameter (3) concerns the relations between signs and the world. Parameter (4) concerns the relations between signs and speakers. This relates the model explicitly to another familiar way of looking at things, namely the famous distinctions drawn by Morris (1938), who defines syntax, semantics and pragmatics in precisely this way. Syntax concerns relations between things internal to the language, and semantics and pragmatics concern relations between the language and things external to the language.

In summary: The model concerns order not sequence. It is not an inductive generalization from observed facts. We deduce its properties. We investigate how well it fits with other things and whether it explains other things. We discover that it fits very well with other things, and that it also makes these relations more explicit and explains some new things. The major contribution of the model is suggested by this quote from the mathematician G. H. Hardy (1940/1992: 89):

> [An] idea is "significant" if it can be connected, in a natural and illuminating way, with a large complex of other [...] ideas.

One of the most important factors in the advance of systematic theory is to introduce order where there was previously disorder (Gellner 1959: 56, 221). The model meets this criterion, by relating things which were previously only poorly related: lexis, syntax, semantics and pragmatics.

6. Research problems: lexis and text

In my last main section, I set out some research problems. I have proposed a slight adaptation to the Sinclair model, as follows, in order to try and make as explicit as possible its logical structure. Parameters [1] and [2] define the form of a text segment, parameter [3] defines the content, parameter [4] defines the speaker's evaluative communicative purpose, and parameter [5] defines the textual function. [NOTE 10.]

I now have to put the phrasal units back into texts and show their textual function as part of a sequence of speech acts in a continuous text. We can go up to textual order and look at how lexis makes texts hang together: we can try and formulate a functional theory of lexis (Stubbs in prep a). And we can go up to social order and look at how cultural meanings are expressed: we can try and formulate an empirical theory of speech acts (Stubbs in prep b).

If we put the phrasal units back into the texts that the concordance software has ripped them out of, we discover that the *went and spoiled* sequence is often part of a still longer recurrent string. This example is from a student website discussing Tony Blair. [NOTE 11.] There is also a following paraphrase of the evaluation: *least effective*.

> (29) Oh, Tony. You've had ten years to write that resignation speech [...] And it *was going reasonably well* – [...] And then you *went and spoiled it all* by saying something stupid. Britain "the greatest nation on earth"? Seriously, of all the things Blair could have said [...] that had to be the least effective.

This example is from a website about a BBC soap opera [NOTE 12.] Part of the pattern is not formal but topical. We have to recognize *career going from strength to strength* as a specific instance of *going well* which fits into the topic of the text. There is also a following rough paraphrase of the evaluation: *another fine mess* (which will remind some readers of Laurel and Hardy).

> (30) With his career *going from strength to strength*, [...] it seemed like Darren might have shaken loose from his chaotic, under-achieving family. Then he *went and spoiled it all* by doing something stupid like

> sleeping with Heather. [...] Is this another fine mess that he's going to have to scheme his way out of? We'll have to see.

This example is from a film review website [NOTE 13.]. Again, there is a following paraphrase of the evaluation: *they lose the art*.

> (31) I remember watching Lady in the Water. It *was all going really well*: lots of suspense [...] when they *went and spoiled it* by bringing out the not-very-scarey wolf-type animal with grass for fur. Sometimes, [...] they lose the art of film making.

Perhaps the most famous example – a cultural icon which is the prototype of the prototype? – is in the song *Somethin' Stupid*, best known in its 1960s version by Frank and Nancy Sinatra, and also in a more recent version by Robbie Williams and Nicole Kidman.

> (32) And afterwards we drop into a quiet little place
> And have a drink or two ...
> And then I *go and spoil it all*
> By saying something stupid like "I love you".

A colleague mentioned this song when I gave an earlier version of this article as a lecture in my university. We had both presumably stored the expression *go and spoil it all*, along with previous contexts of use, but my colleague had apparently stored the text segment along with a very specific context. And – a problem for any theory of semantics – the text segment therefore meant something different to him and to me. I had either never known the song or had completely forgotten it. The song is by no means the first use of the unit, but it changed the meaning of the unit for at least some speakers. (You might think that the rest of the verse is also constructed from recurrent clichéd phrases.)

In all these cases, we have a narrative pattern. Something is going well, and then gets spoiled. The *went-and-spoiled* sequence is part of a longer phrasal unit, which is a text segment in a narrative sequence. Since it presupposes a prior sequence of events, the phrasal unit functions as an element of textual coherence, and since it occurs in independent texts, this is evidence that it is part of the system. In terms of method: We can automatically extract formal repetitions with their minor variants and argue from parallel examples in the intertext, but intuition and real world knowledge are required in order to extract semantically equivalent text segments which are examples of "things going well", as in the following examples.

> (33) it [a film] was all going really well ... when they *went and spoiled it* by bringing out the not-very-scarey wolf-type animal with grass for fur
> (34) you were having such a lovely evening ... and then I *went and spoiled it all* by doing something stupid
> (35) three of us had a rollicking good time [...] until [he] *went and spoiled everything*

(36) she kind of gets it right but then she *goes and spoils every* outfit with a random piece of clothing that just doesn't work

The variants are text-dependent and topic-dependent, but there is related vocabulary, such as *going well, lovely evening, good time, gets it right,* etc, and it might be possible to use these formal hints to identify longer recurrent sequences.

This involves a well-known, long-running, and unsolved question. How much of the perceived connectedness of text is explicit cohesion, and how much is implicit coherence which depends on knowledge from outside the text? There is really only one empirical strategy available here. Let us assume that all such relations are, in fact, explicitly signalled in some way, and then look for the signals. If you don't look for things, you won't find them.

The full canonical form looks like this:

it was going well [... example specific to text ...]
and then
someone went and spoiled it all
by doing or saying something stupid

It is canonical: first, because it expresses a familiar experience that has often been talked about in this way in the past, and second, because these actual words occur, quite frequently, in the intertext. They provide a text-independent paraphrase of the variants. The unit is semantically stable, but formally variable, due to variants which fit into the topic of the individual text. As I have now said several times, this does however depend on the problematic notion of something which stays the same, despite undergoing changes.

The phrasal unit cannot be understood apart from cultural assumptions about what constitutes "doing something stupid", such as crashing your bike, setting your blanket on fire, getting Heather pregnant when you're engaged to Libby – and presumably a large open-ended set of such things. This type of example poses problems for truth-conditional semantics, since truth now depends on the speaker's beliefs and attitudes, but it is compatible with varieties of frame semantics, or with a theory which sees representations as reproducing the social order. The crucial move from analysing linguistic units to analysing cultural units was made by Francis (1993), who proposes a cultural interpretation of phrasal units. Culture consists of recurrent representations of events which encode our attitudes and values:

> [We] can compile a grammar of the typical meanings that human communication encodes [...] the ways in which we typically evaluate situations [...] how difficult or easy life is made for us, how predictable things are, and how well we understand what is going on. (Francis 1993: 155, 141.)

You might even think that the little phrasal unit summarizes the narrative structure of the archetypal tragedy – Oedipus, Macbeth, Faust – it was all going reasonably well, until he went and spoiled it all, by marrying his mother, killing the king, making a pact with the devil, or whatever. (That was a joke.)

Sinclair himself never seemed very interested in developing such cultural interpretations of his work, which is odd, given his earlier work on discourse analysis. So, one quote is particularly interesting:

> [P]erhaps the most innovative and far-reaching development in linguistic perception in the last fifty years was the philosopher Austin's idea of illocutionary force. (Sinclair 2008: 23-24.)

The concept of semantic prosody gives considerable descriptive depth to the concepts of illocutionary force and speech act. The phrasal model is a way of building the speaker's communicative purpose into lexical items, and we are then just one step away from a theory of language as social action.

7. Sinclair and Searle

I have argued that Sinclair's model provides both a concrete strategy for describing empirical data, and also a strikingly original way of relating traditional levels of linguistic description. First, we automatically extract formal repetitions with their minor variants. Second, we manually extract semantically (and culturally) equivalent strings. Third, we include as much co-text as possible: Sinclair (in prep) calls this a "maximalist" approach.

There are certainly unanswered questions, but it is only when we make a model as explicit as possible, that we can see new research problems. Sinclair (in Sinclair et al 2004: xxiv) has proposed what he calls "a very strong hypothesis":

> For every distinct unit of meaning there is a full phrasal expression [...] the canonical form. [...] A dictionary containing all the lexical items of a language, each one in its canonical form with a list of possible variations, would be the ultimate dictionary.

This does seem over-optimistic. Phrasal units are simply too variable to be listed, and we do not know what is the optimal level of abstraction at which to describe them. For example, we do not know how to describe the relations between different units (such as different double-verb constructions).

However, there are two obvious contenders for a functional theory of speech acts: the Sinclair approach and the Austin/Searle approach. Sinclair uses attested data to show parallel cases, which are then the inductive basis of deductive reasoning about the structure of a model of language. Austin and Searle use invented data to make deductions about the structure of society. Sinclair is strong on linguistic description, but weak on social theory. Austin and Searle are weak on linguistic

description, but strong on social theory. We do not yet know how to relate the Sinclairean and the Searlean approaches to speech acts. The question here is how to relate Sinclair's bottom-up empirical description of language use (Sinclair 1996, 1999) and Searle's top-down analytic explanation of society (Searle 1995, 2010) in which he attempts to explain "the exact role of language" in the creation of social reality (Searle 2010: ix). It would require a major research programme to combine the strengths of the two approaches, but only in this way could language be integrated into a theory of social structure.

8. Concluding comments

Hanks (1997: 295) describes Sinclair's "ferocious empiricism". This is a valid description: but there is also the rationalism implicit in Sinclair's model building, which reorders the data, by reinterpreting the relations between lexis, syntax, semantics and pragmatics, and therefore reinterpreting long-standing theories of language system and use.

John Sinclair came from Edinburgh. Another – even more famous – Scottish empiricist who came from Edinburgh is David Hume (1711–1776). But Hume was an empiricist who was sceptical of induction from empirical data. [NOTE 14.] His empiricism was famously admired, but also criticised, by Immanuel Kant (1783), who confessed that David Hume had "interrupted [his] dogmatic slumber, and [given his] investigations in the field of speculative philosophy quite a new direction". Immanuel Kant's response to David Hume's scepticism was that pure empiricism is content without form, but that pure rationalism is form without content [NOTE 15.]

The neo-Firthian tradition is an attempt to develop an empirical semantics, and that requires a combination of empiricism and rationalism.

Notes

1. This can cause terminological confusion, since collocation is often defined as co-occurring word-forms, and therefore as a matter of sequence. For the abstract relation of order, Sinclair uses the term "co-selection".

2. Similar double-verb constructions also occur, with similar functions, in Swedish and Norwegian (Wiklund 1996) and Finnish (Airola 2007). Stefanowitsch (1999) discusses comparable examples from Scandinavian and other languages and argues that there are cross-linguistic semantic regularities in the use of the construction.

3. Of course, only the forms were present, and since they were different uses, they may have had different meanings. As Heraclitus argued some 2500 years ago, change is central to the universe. You cannot step into the same river twice, for other waters are continually flowing past.

4. For this, Black (1967: 169) uses the term "eduction", which he defines as follows: an "argument from sample to sample [...] a conclusion is drawn concerning approximate frequency of occurrence in a further sample obtained by the same procedure". As far as I know, he is the only source of the term, and I have found only one article (Fahnestock 2003) which quotes him. Fahnestock then provides similar examples to those in my Figures 1 and 2.

5. Popper notes an exception: the validity of sources is important to historians.

6. The same holds for the distinction between corpus-as-theory and corpus-as-method (McEnery & Hardie 2012: 147-52) which similarly confuses questions of origin and questions of validity.

7. This is a point where confusion is possible, since here collocation is seen as merely a matter of sequence. This is rather different from Firth's definition (quoted above) which involves order. See also note 1.

8. Stewart (2010), in his book on *Semantic Prosody*, is very sceptical of the concept, but, I think, makes the error of discussing semantic prosody independently of the model of which it is only one parameter. He gives references to previous work on the concept by Channell, Hunston, Louw, Partington, Sinclair, Stubbs, Tognini-Bonelli and others.

9. In other publications (Stubbs 2001: 65), I have referred to semantic prosodies as "discourse prosodies", and I argue in the present article that they are pragmatic in function. I assume that Sinclair calls them "semantic" because the evidence is observable in the text, not inferred from non-linguistic knowledge about the social context of use.

10. Hoey (2005: 13) uses the term "textual colligation" to refer to the tendency of lexis to occur at certain positions within texts, but defines this in terms of individual words.

11. Source: Gair Rhydd Cardiff's Students Weekly website. http://www.gairrhydd.com/comment/politics/842/bye-bye-blair (accessed June 2011, now unavailable).

12. Source: BBC East Enders website. http://www.bbc.co.uk/eastenders/characters/darren-miller.shtml (accessed June 2011, text now partly altered).

13. Source: New York Film Academy website. http://www.empireonline.com/news/story.asp?NID=21492 (accessed Nov 2011).

14. I discovered recently that my favourite quote from David Hume's *Enquiry Concerning Human Understanding* (1748) is also quoted by Firth (1937: 103). Hume is giving advice about reading books: "If we take in our hand any volume

[...] let us ask: Does it contain any abstract reasoning concerning quantity or number? No. Does it contain any experimental reasoning concerning matter of fact and existence? No. Commit it then to the flames: for it can contain nothing but sophistry and illusion."

15. This is the formulation in Scruton (1982: 31). What Kant (1781) said was "Gedanken ohne Inhalt sind leer, Anschauungen ohne Begriffe sind blind." (Thoughts without content are empty, perceptions without concepts are blind.)

Acknowledgements

I am very grateful to Jeanne Fahnestock (who answered my questions about the concept of "eduction"), Caty Pope, Dorothea Halbe, Naomi Hallan and two anonymous referees (who made useful comments on a previous draft), and Russell Kelly (who reminded me of the Sinatra song).

References

Airola, A. 2007. *Coordinated Verb Pairs in Texts*. Helsinki: University of Helsinki.
Austin, J.L. 1962.How to do Things with Words: Oxford: Oxford University Press.
Biber. D., Finegan, E., Johansson, S. & Conrad, S. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.
Black, M. 1967. Induction. In P. Edwards (ed) *The Encyclopedia of Philosophy*. London: Macmillan. 169-81.
Darwin, C. R. 1845. *Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world, under the Command of Capt. Fitz Roy, R.N*. 2nd ed. London: John Murray.
Fahnestock, J. 2003. Verbal and visual parallelism. *Written Communication*, 20/2: 123-52.
Firth, J. R. 1935. The technique of semantics. *Transactions of the Philological Society*. 36-72.
Firth, J. R. 1937. *The Tongues of Men*. London: Watts & Co. (Page ref to Oxford University Press edition, 1964.)
Firth, J. R. 1951/1957. Modes of meaning. In *Essays and Studies* (The English Association 1951). Also in Firth, J. R. 1957. *Papers in Linguistics 1934-1951*. London: Oxford University Press. 190-215.
Firth, J. R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*. [Special issue: *Transactions of the Philological Society*.] 1-32.
Francis, G. 1993. A corpus-driven approach to grammar: principles, methods and examples. In M. Baker, G. Francis & E. Tognini-Bonelli, E. eds *Text and Technology*. Amsterdam: Benjamins. 137-56.
Gazdar, G. 1979. *Pragmatics*. New York: Academic Press.
Gellner, E. 1959. *Words and Things*. London: Gollanz. [Page ref to Penguin edition, 1968.]
Halliday, M. A. K. 1961. Categories of the theory of grammar. *Word*, 17/3: 241-92.
Halliday, M. A. K. 1975. *Learning how to Mean: Explorations in the Development of Language*. London: Edward Arnold.
Hanks, P. 1997. Review of J. Sinclair: *On Lexis and Lexicography*. *International Journal of Corpus Linguistics*, 2, 2: 289-95.
Hardy, G. H. 1940. *A Mathematician's Apology*. Cambridge: Cambridge University Press. [Page ref to Canto edition, 1992.]
Harris, R. 2004. *The Linguistics of History*. Edinburgh: Edinburgh University Press.
Hoey, M. 2005. *Lexical Priming*. London: Routledge.
Hoffmann, S., Evert, S., Smith, N., Lee, D. & Berglund Prytz. Y. 2008. *Corpus Linguistics with BNCweb*. Frankfurt am Main: Lang.

Hopper, P. 2001. Hendiadys and auxiliation in English. In J. Bybee & M. Noonan (eds) *Complex Sentences in Grammar and Discourse*. Amsterdam: Benjamins. 145-73.
Huddleston, R. D. & Pullum, G. K. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Kant, I. 1781. *Kritik der reinen Vernunft* [*Critique of Pure Reason*]. Riga: J.F. Hartknoch.
Kant, I. 1783. *Prolegomena zu einer jeden künftigen Metaphysik* [*Prolegomena to any Future Metaphysics*]. Riga: J.F. Hartknoch.
Malinowski, B. 1923. The problem of meaning in primitive languages In C. K. Ogden & I. A. Richards (1923) *The Meaning of Meaning*. London: Paul, Trench, Trubner & Co.
McEnery, T. & Hardie, A. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
Morris, C. W. 1938. *Foundations of the Theory of Signs*. Vol. 1, No. 2, in O. Neurath et al (eds) *International Encyclopedia of Unified Science*. Chicago: Chicago UP.
Palmer, F. R. 1962. Sequence and order. *Monograph Series Language and Linguistics*, 17: 123-30.
Popper, K. R. 1975. *Conjectures and Refutations: the Growth of Scientific Knowledge*. 5th edition. London: Routledge.
Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman,.
Renouf, A., Kehoe, A. & Banerjee, J. 2007. WebCorp: an integrated system for web text search. In C. Nesselhauf, M. Hundt & C. Biewer (eds) *Corpus Linguistics and the Web*. Amsterdam: Rodopi.
Saussure, F. de. 1916. *Cours de linguistique général*. Paris: Payot.
Scruton, R. 1982. *Kant*. Oxford: Oxford University Press. [Page reference to German translation, Freiburg: Herder 1999.]
Searle, J. 1995. *The Construction of Social Reality*. London: Allen Lane.
Searle, J. 2010. *The Making of the Social World*. Oxford: Oxford University Press.
Sinclair, J. (ed) 1990. *Collins COBUILD: English Grammar*. London: HarperCollins.
Sinclair, J. McH. 1996. The search for units of meaning. *Textus*, 9/1: 75-106.
Sinclair, J. McH. 1999. The lexical item. In E. Weigand (ed). *Contrastive Lexical Semantics*. Amsterdam: Benjamins. 1-24.
Sinclair, J. McH. 2008. Borrowed ideas. In A. Gerbig & O. Mason (eds) (2008) *Language, People, Numbers*. Amsterdam: Rodopi. 21-41.
Sinclair, J. McH. in prep. *Essential Corpus Linguistics*. Ed. E. Tognini-Bonelli. London: Routledge.
Sinclair, J McH. & A. Mauranen 2006. *Linear Unit Grammar*. Amsterdam: Benjamins.
Sinclair, J. McH., Jones, S. & Daley, R. 1970/2004. *English Collocation Studies: The OSTI Report*. (ed) R. Krishnamurthy. London: Continuum.
Stefanowitsch, A. 1999. The *go-and-verb* construction in a cross-linguistic perspective. In D. Nordquist & C. Berkenfield (eds). *Proceedings of the Second Annual High Desert Linguistics Society Conference.* Albuquerque, NM: High Desert Linguistics Society.
Stewart, D. 2010. *Semantic Prosody*. London: Routledge.
Stubbs, M. 2001. *Words and Phrases*. Oxford: Blackwell.
Stubbs, M. in prep a. The textual functions of lexis. In N. Groom et al (eds). *Corpora, Grammar, Text and Discourse*. Amsterdam: Benjamins.
Stubbs, M. in prep b. Searle and Sinclair on communicative acts. In María de los Ángeles Gómez-González et al (eds). *Form and Function in Language.* Amsterdam: Benjamins.
Teubert, W. 2010. *Meaning, Discourse and Society*. Cambridge: Cambridge University Press.
Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work.* Amsterdam: Benjamins.
Wiklund, A-L. 1996. Pseudocoordination is subordination. *Working Papers in Scandinavian Linguistics*, 58: 29-53.
Wulff, S. 2006. Go-V vs. go-and-V in English: A case of constructional synonymy? In: S. Th. Gries & A. Stefanowitsch (eds). *Corpora in Cognitive Linguistics. Corpus-based Approaches to Syntax and Lexis*. Berlin: Mouton de Gruyter. 101-125.

/ FIGURES FOLLOW

Figure 1.

Plate 46, vol 1, in Dru Drury & J. O. Westwood (1837) *Illustrations of Exotic Entomology*. London: H.G. Bohn.

Source of this illustration: Wikimedia Commons.

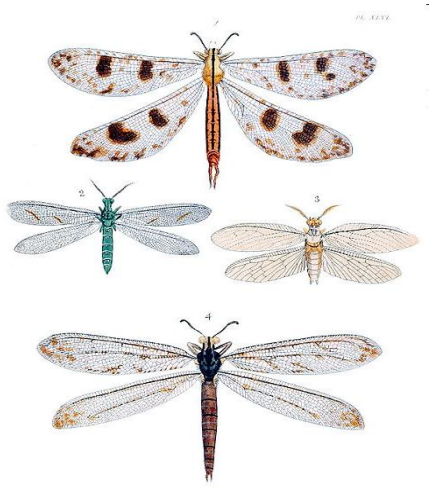

Figure 2.

From Charles Darwin (1845) *The Voyage of the Beagle.* [Various editions.]

Source of this illustration: Wikimedia Commons.