
COLLOCATIONS AND SEMANTIC PROFILES: ON THE CAUSE OF THE TROUBLE WITH QUANTITATIVE STUDIES

Michael Stubbs

FB2 Anglistik, University of Trier
D-54286 Trier, Germany

ABSTRACT

Current work on lexical collocations uses two ideas:

- words have distinctive semantic profiles or "prosodies"
- the strength of association between words can be measured in quantitative terms.

These ideas can be combined to provide comparative semantic profiles of words, which show the frequent and characteristic collocates of node words, and make explicit the semantic relations between the collocates.

Using data from corpora of up to 120 million words, it is shown that the lemma CAUSE occurs in predominantly "unpleasant" collocations, such as *cause of the trouble* and *cause of death*. A case study of this lemma is used to illustrate quantitative methods for investigating collocations. Various methods proposed in the literature are of great practical value in establishing collocational sets, but their theoretical basis is less clear. Brief comparative semantic profiles are given for related lemmas, eg REASON and CONSEQUENCE. Implications for the relation between system and use are discussed.

1. TERMINOLOGY, CONVENTIONS AND DATA

In this article I study collocations of words in corpora of data, and I start with brief definitions of some necessary terms. By collocation I mean a relationship of habitual co-occurrence between words (lemmas or word-forms). A node word may be observed to co-occur with various collocates within a certain span or window, say 4:4, ie four words to left or right. I will discuss ways of comparing, relative to the size N of corpus (in word tokens):

- f(node, collocate): joint frequency of node and collocate
- f(node), f(collocate): their independent frequencies.

I will abbreviate these to: f(n,c), f(n), f(c).

A lemma is a dictionary head-word, which is realized by various word-forms. I cite lemmas in upper case, and italicize word forms. For example, the lemma CAUSE has the forms *cause, causes, caused, causing*.

The data are from various corpora: the 1 million word LOB corpus (500 samples of 2,000 words each from written genres, eg newspapers, reports, academic articles, novels); the Longman-Lancaster corpus of written English (from which I have used only small selections of 700,000 words in 2,000 word samples from 350 different texts); the half-million word LUND corpus (87 samples of 5,000 words each from spoken genres, eg conversation, lectures and radio discussions); and a further 1.1 million words of mainly written material, including half a million words which are topically more homogeneous, on environmental issues. Some findings below are from various combinations of these corpora which total around 3.3 million words. Other findings are from the Cobuild corpus: I have used 120 million words of running text, mainly written, but some spoken, from British and American books, newspapers and magazines, BBC news broadcasts, ephemera and conversations.

Apart from a few examples of uncertain provenance quoted from dictionaries, I use only examples attested in these corpora. No examples are invented.

2. SEMANTIC PROSODIES

It is well known that some words habitually collocate with other words (Firth 1957). It is also well known that attested data are required in collocational studies, since native speaker intuitions are not a reliable source of evidence. Native speakers can often give a few examples of the collocates of a word (sometimes accurately), and they may be able to judge, very approximately, the likelihood of collocations they are presented with. But they certainly cannot document collocations with any thoroughness, and they cannot give accurate estimates of the frequency and distribution of different collocations.

In addition, it is becoming increasingly well documented that words may habitually collocate with other words from a definable semantic set. Sinclair gives several examples of words which have a "negative" semantic prosody. He shows (Sinclair 1991: 70ff) that the phrasal verb SET IN occurs primarily with subjects which refer to unpleasant states of affairs, such as rot, decay, malaise, despair, ill-will and decadence. I have corroborated this finding with the LOB corpus, which provided several "unpleasant" examples and only one "pleasant" example:

(1) before bad weather sets in; the fact that misery can set in; desperation can set in; stagnation seemed to have set in; before rigor mortis sets in.
(2) the fantastically dry and sunny spell that set in.

Sinclair (ed 1990: xi) also points out that it is bad things which BREAK OUT. Again my data gave corroborating examples:

(3) violence broke out; riots broke out; war broke out; feeling the sweat breaking out; real disagreements have broken out; a storm of protest broke out.

The noun *outbreak* collocates with diseases in examples such as *disastrous outbreaks of foot and mouth*. One of Sinclair's examples is:

(4) this caused an epidemic to break out

which collocates *caused* (see below), *break out* and *epidemic*! A third example is HAPPEN: Sinclair (1991: 112) notes that this lemma "is associated with unpleasant things - accidents and the like". Corroborating examples from LOB include:

(5) the problem of what will happen; this sort of accident can still happen; need the quarrel with Cuba ever have happened; something very untoward has happened; calm down and tell me exactly what happened.

Louw (1993) uses Sinclair's term "semantic prosody" for this collocational phenomenon. (This is prosody in its Firthian sense of a feature which stretches over several units.) And he shows, for example, that *utterly*, *bent on* and *symptomatic of* also have predominantly unpleasant collocates.

Although such negative prosodies are probably more common, positive prosodies also exist. For example *causing work* usually means bad news, whereas *providing work* is usually a good thing:

(6) when you overdraft your account, you cause extra work for the bank staff
(7) this will provide work; it will raise the standard of living

The main collocates of PROVIDE (established using the methods described below) show its positive prosody. One can provide:

(8) facilities, information, services; aid, assistance, help, support; care, food, money, nourishment, protection, security.

3. CAUSE: AN EXAMPLE OF A NEGATIVE SEMANTIC PROSODY

As a detailed example, I studied the lemma CAUSE: by looking at its entries in several dictionaries; by studying the 250 occurrences in a 1 million word corpus; and by using software (described below) to analyse further examples from small corpora and also the 38,000 occurrences in a 120 million running words. CAUSE is overwhelmingly used in contexts where cause and effect are unpleasant. The main collocates concern problems, trouble and damage, death, pain and disease.

3.1. Dictionaries

Some native speakers (but not all) that I have informally tested do produce one or two examples of such unpleasant collocations, but such native speaker data are very sparse and unreliable indeed. Neither do widely used dictionaries explicitly draw attention to such negative cases. For example, the Oxford Advanced Learners' Dictionary, the Longman Dictionary of Contemporary English, and Collins English Dictionary all give a neutral definition such as "a cause is something which produces an effect". However, the examples then include:

(9) the cause of the fire was carelessness; causes of war; cause for anxiety; cause of the accident; cause of the crime problem; cause of all my unhappiness; cause for concern; cause cancer; her rudeness was a cause for complaint.

The Cobuild dictionaries (Sinclair et al 1987, 1988) also give a neutral definition. Here, the examples are attested in a 20 million word corpus:

(10) the cause of the explosion; died of natural causes; does smoking cause cancer; difficulties caused by price increases; that's going to cause me a lot of trouble.

A minority of examples are positive or neutral (though there is no indication of how much less likely such positive examples are):

(11) every cause for confidence.

The complete Oxford English Dictionary gives both negative and positive citations from the 1300s onwards. In only one sense for CAUSE as a noun is it noted that the word may mean a person or agent who brings about something, "often in a bad sense: one who occasions, or is to blame for mischief, misfortune, etc".

In summary, dictionaries give negative examples, but do not say explicitly whether these examples are frequent or typical. They therefore give no explicit account of the lexico-semantic set which co-occurs with CAUSE, and they miss an opportunity for relating dictionary and thesaurus entries.

3.2. Corpus data: 1 million words

I studied CAUSE in a 1 million word corpus (LOB), and began to document the lexical set of its collocates by constructing a concordance of all occurrences (c. 250) of cause and causes (noun and verb), caused and causing. Nearly 80 per cent of occurrences have clearly negative collocates, usually within a span of 3:3. Conversely, a very small number of occurrences have positive collocates. The distribution is: negative 80%, neutral 18%, positive 2%.

One clear advantage of corpus data over intuitive data is that such collocates can be extensively documented. Collocates which occur as subject or object of the verb CAUSE or as prepositional object of the noun CAUSE include:

(12) abandonment, accident, alarm, anger, annoyance, antagonism, anxiety, apathy, apprehension, breakage, burning, catastrophe, chaos, clash, commotion, complaint, concern, confusion, consternation, corrosion, crisis, crowding, damage, danger, death, deficiency, delay, despondency, destruction, deterioration, difficulty, disaster, disease, disorganization, disruption, disturbance, disunity, doubt, errors, frustration, habituation (to a drug), harm, hostility, hurt, inconvenience, interference, injury, interruption, mistake, nuisance, pain, pandemonium, quarrel, rejection, ruckus, rupture, sorrows, split, suffering, suspicion, trouble, uneasiness, upset, wholesale slaughter.

Other collocations after CAUSE TO include:

(13) complain, crumble to oblivion, disintegrate, lag behind, repent, shiver, undervalue.

The phrase *cause of death* occurs several times, and medical uses are common: eg *eczema is caused by* .... A few occurrences are neutral:

(14) caused the removal of the pulpit to the side; causes the fantail to start revolving.

And the few apparently positive collocations include:

(15) caused such widespread interest; caused a pleasurable mental state; caused him to smile; caused the little boy to roll about with laughter; a cause to display such amiability.

A larger context reveals even some of these as questionably positive.

The lemma occurs in all genres represented in LOB, and in all genres, collocations are predominantly negative. It is not that positive examples in some genres are averaged away by negative examples in others. However, the newspaper press reports have only negative collocations: presumably because newspapers report predominantly crises and disasters!

These figures include the separate sense of CAUSE as "aim or principle", which occurs in different collocations, including:

(16) devoted service to this cause; conviction that your cause is right; plead a cause; take up causes.

Causes in this sense are *good, glorious, just* and *worthy*, but also *lost* and *foolish*. This sense is not frequent in LOB, but if it is omitted, then the percentage of negative instances rises. So, the 80 per cent negative count is on the low side.

4. QUANTITATIVE METHODS OF SUMMARIZING CONCORDANCE DATA

It is easy to inspect 250 concordance lines. However, to study larger corpora and to compare words, we require a method of summarizing concordance data and of estimating the frequency of association between words. Such a method requires software which

- reads in a corpus and identifies all occurrences of a node word (or pattern, eg lemma) and its frequency f(n)
- keeps a record of collocates of this node which occur in a window of defined size (eg 4 words to left and right)
- counts the frequency of joint occurrence of node and each collocate, f(n,c)
- counts the absolute frequency of each collocate in the whole corpus, f(c)
- and performs further calculations, as discussed below.

Such software has been written for corpus work at the University of Trier by Oliver Jakobs. Software in use at Cobuild is described by Clear (1993).

4.1. Raw frequencies of collocations

Often, with quantitative linguistic data, no complex statistical procedures at all are necessary. It may be sufficient simply to count and list items. For example, in a corpus of 1.5 million

words (LOB plus LUND), the following were noun collocates of CAUSE, where f(n,c) is greater than or equal to 3:

(17) accident, alarm, concern, confusion, damage, death, delay, fire, harm, trouble.

It is obvious to the human analyst that these words are semantically related. There were only a few examples (up to 8) of each individual collocation, but these negative examples are most unlikely to be a coincidence. There were no positive examples at all amongst the most frequent collocates, and only a few neutral examples (eg *effect, place, present*). Such raw frequencies require no further statistical manipulation to show a semantic pattern.

One has to distinguish, of course, between a procedure and the sample to which it is applied. I had no reason to suspect that this corpus was biased by containing lots of texts about particularly gloomy things: it contained almost 600 text samples, from many different text types, both written and spoken. But I carried out a simple check by doing the same analysis on three other small corpora. The results were the same: predominantly negative, no clearly positive, and a few neutral collocates.

> A corpus of 700,000 words, 350 samples of written English [f(n,c) is greater than or equal to 3]:

(18) accidents, damage, death, difficulty, embarrassment, headaches, pain, trouble.

> A corpus of 725,000 words, 30 samples of written and spoken English [f(n,c) is greater than 3]:

(19) concern, damage, danger, depletion, fever, headaches, migraine, pollution, poverty, problems, symptoms, trouble, unemployment, vasoconstrictions.

> A corpus of 425,000 words, comprising topically more restricted texts about environmental issues [f(n,c) is greater than 3]:

(20) blindness, cancer, concern, damage, depletion, harm, loss, ozone, problems, radiation, warming.

The frequency of even relatively common words differs considerably in different corpora and merely reflects the topics of the constituent texts. Even so, there is a large overlap in the collocates from the four corpora and this gives us even more confidence that we have identified a genuine set of recurrent collocations.

However, raw frequencies give no information on other aspects of the pattern. Raw joint frequencies f(n,c) could be more reliably interpreted if we had comparative information about how frequently words such as trouble or accident occur independently in the corpus. And positive collocates do also occur, but we do not know how much more likely is cause for concern than cause for confidence.

It is clear from these simple statistics that negative collocates of CAUSE are the most frequent, and in that sense typical. What is more difficult to provide is typical individual instances, since any specific instance is just that: specific. It will have features of its specific co-text.

Corpus linguistics has, as yet, no theory of typicality (though see Sinclair 1991: 103). However, the following examples from LOB illustrate some of these frequent collocations:

(21) East German restriction which caused today's trouble
(22) dryness can cause trouble if plants are neglected
(23) considerable damage has been caused to buildings
(24) damage caused by slugs, woodlice or mice
(25) a level that gives cause for concern
(26) I didn't see anything to cause immediate concern
(27) a certificate showing the cause of death
(28) asphyxia as the most common cause for death in drowning

4.2. Observed and expected frequencies

Many statistical calculations compare (a) how often something is actually observed and (b) how often it might be expected merely by chance. And, given how experiments are designed, one is usually hoping that (a) is much bigger than (b). A comparison of observed and expected frequencies of pairs of words usually starts from the assumption that it is meaningful to compare (a) a real corpus and (b) a hypothetical corpus consisting of the same words in random order. We then look for statistically significant deviations from this hypothetical randomness. The conventional null hypothesis would be that there is no difference between the real and hypothetical corpora.

The strangeness of this starting assumption is regularly discussed (eg Woods et al 1986: 104ff, 147ff). Standard statistical procedures assume proper random samples in which values are independent observations, but since textual data are never in this form, this calls into question whether such statistics can reasonably be used on language data: see further below.

Accepting, for the moment, this starting assumption, the probability (expected frequency) of node and collocate co-occurring can be calculated as follows. Suppose the collocation *cause - trouble* occurs 10 times in a corpus of 1 million words: then its frequency is one in 100,000 (0.00001). In general, the observed frequency O, relative to corpus size N, is:

(a): $O = f(n,c) / N$

Whether this frequency is significantly high depends on how often cause and trouble occur independently in the corpus. If cause occurs 100 times, then its probability of being the next word at some random point in the corpus is 0.0001. And, if trouble occurs 50 times, then its probability is 0.00005. If p1 and p2 are the probabilities of independent (NB!) events (such as a coin landing heads or tails on successive tosses), then the product p1p2 is the probability of their co-occurrence. So, the probability of cause and trouble co-occurring by chance is: 0.0001×0.00005, ie 5 in 1,000 million, or one in 200 million. (In a corpus of only one million words, we would have only one chance in 200 of observing this collocation at all!) In general, the expected frequency E of co-occurrence, relative to corpus size N, is:

(b): $E = f(n) / N \times f(c)/N = f(n)f(c) / N^2$

If this collocation is observed once in 100,000 words, rather than (as expected) once in 200 million, then it occurs 2,000 times more frequently than might be expected solely by chance.

In general, to calculate how much higher than chance the frequency of a collocation is, we calculate O/E.

(c): $O/E = [f(n,c) / N] / [f(n) / N \times f(c) / N]$
$= f(n,c) / [f(n)f(c) / N]$
$= [f(n,c) \times N] / f(n)f(c)$

This assumes that *cause* is adjacent to *trouble*. If we look at a window of 3:3, then we increase the probability of c occurring by chance within this window. In formulae discussed below, window size is not taken explicitly into account, though it clearly affects our results. A window of 3:3 catches examples such as the *cause of the trouble*, but not *the cause of all the trouble*. To make meaningful comparisons between different pairs of node and collocate, we have to keep window size constant. There is no real agreement in the literature about appropriate window size. Spans of 2:2 or 3:3 are often used, and Sinclair (1991) claims that little of collocational interest is found outside a span of 4:4. Some scholars claim meaningful association effects over window sizes such as 50:50, but this seems to alter the meaning of collocation, since the same content words are bound to occur at various points in a cohesive text.

A problem with this calculation of O/E is that almost any observed co-occurrence is hundreds of times more likely than by chance. Suppose, in a corpus of 1 million words, two words each occur 100 times, and co-occur just once. Then

$O/E = [f(n,c) \times N] / f(n)f(c)$
$= 1{,}000{,}000 / (100 \times 100) = 100.$

This single co-occurrence is 100 times more likely than chance! But, by definition, a single occurrence could just be due to chance. Such probability figures are artificially low, given that the data cannot be random. For example, noun-verb constraints greatly reduce the number of possible combinations. We do not have to reject the O/E ratio. We can ignore the actual values of O/E, and consider only their rank order: ie which co-occurrences are more or less unexpected, purely by chance. But our starting assumption means that conventional probability levels make little sense.

So, O/E is a (rough) indication of the strength of association between two words (Sinclair 1991: 70).

4.3. Other formulae for studying collocations

Two related statistics have been proposed in the literature, and have been referred to as "mutual information" and "t-score" (Church & Hanks 1990, Church et al 1991, Clear 1993). This literature provides very useful empirical data on collocations gathered from large corpora. However, the statistics proposed are confusing. First, the term "mutual information" comes from work in information theory, where "information" has the restricted meaning of an event which occurs in inverse proportion to its probability. Second, Church et al (1991) imply that their "t-score" is derived from a standard statistical procedure known as the "t-test": but they do not say how their variant is derived from the standard formula, and do not point out that the t-test is arguably not valid for the kind of linguistic data they discuss. (Hallan 1994.) Clear (1993) publishes no actual formulae for calculating the statistics: he refers to algorithms in work by Church, Hanks et al.

I will use the terms I-value and T-value to indicate that I am calculating things in ways proposed in this literature. But given that one cannot have "a random corpus", I will not attribute levels of statistical significance to these values. In addition, I will show that the formulae proposed by Church, Hanks et al are simple arithmetic manipulations of O and E. However, I-and T-values also have built-in corrections for the size of the corpus, and (for T) cases where node and/or collocate are themselves very frequent. Briefly:

I is a measure of the relative frequency with which words occur in collocation and independently.
T is a measure of the absolute frequency of collocations.

4.4. I-value

The formula for I is a simple variant of O/E. It compares the frequency of co-occurrence of node and collocate with the frequency of their independent occurrence. Church et al (1991: 4, 7) propose this calculation:

(d): $I(n,c) = \log_2 \{[f(n,c) \times N] / f(n)f(c)\}$

As we have seen, this is simply equivalent to:

$I(n,c) = \log_2 O/E,$ from (c)

The logarithmic value is used for historical reasons, which are hardly relevant here. Its only effect is to reduce, and therefore possibly to disguise, the differences between scores on different collocates. [NOTE 1]. I has the following characteristics. First, it compares the frequency of words within a given span of the node with their overall frequency in the whole corpus. It is sensitive to the relative proportion of mutual and independent occurrences of node and collocate. Second, it also takes into account the size of the corpus. As N increases, so I increases: I has a higher value in cases where we look at a larger corpus.

Suppose, as above, that, in a corpus of 1 million words, cause occurs 100 times, trouble occurs 50 times, and they collocate 10 times, then:

$I = \log_2 (10 \times 1{,}000{,}000) / (100 \times 50) = \log_2 2{,}000 =$ approx 11.

Genuine frequencies for CAUSE, *trouble* and their collocations across 1.5 million words were:

f(CAUSE) = 308; f(trouble) = 245; f(CAUSE, trouble) = 8. I = 7.4.

Church & Hanks (1990) and Clear (1993) show that I-values above 3 are likely to be linguistically interesting, with a window of 2:2 or 3:3. There is no strong theoretical reason for picking on this value of I, but in empirical analyses of corpus data it has been found to generate sets of semantically related words. The phrase "linguistically interesting" is admittedly undefined, but it represents an empirical claim. For several lemmas, the present article contains detailed lists of the items which the method produces. The reader can evaluate my view that these lists could not have been produced by intuition, but are, in retrospect, intuitively interesting.

I has to be interpreted with care. First, it is non-directional: it has the same value no matter which word of a pair is node or collocate, since f(n,c) has the same value as f(c,n). Clear (1993) gives the example f(kith, kin), which has the same value as f(kin, kith), although *kith* probably predicts *kin* with 100 per cent certainty, whereas *kin* can occur on its own. Or another example, from the 120 million Cobuild corpus: with BOTCH there is fairly high probability (about 1 in 7) of finding the word form *job*; but with *job* there is only a small chance (less than 1 in 1,000) of finding BOTCH. This asymmetry is lost in the calculation of I. (T is also non-directional: see below.)

Second, the behaviour of I relative to the absolute frequency of collocations is counter-intuitive, if one forgets the origins of the statistic in information theory. If the relative proportion of joint to independent occurrences remains the same, then I decreases as the absolute number of collocations increases. As Clear (1993: 279) notes: collocates can appear prominent in lists of I-values "because they are themselves rare occurrences". [NOTE 2.]

These limitations on I require that we have an alternative measure which takes into account the absolute value of f(n,c). This is the T-value.

4.5. T-value

Whereas I is a simple arithmetic variant of O/E, T is a simple variant of f(n,c). That is, T takes into account mainly (in some cases only) the absolute frequency of joint occurrence of node and collocate. Church et al (1991:8) propose the calculation:

(e): $T = \{[f(n,c) / N] - [f(n)f(c) / N^2]\} / \{[\sqrt{f(n,c)}] / N\}$

This is equivalent to, from (a) and (b):

(f): $T = (O - E) / \{[\sqrt{f(n,c)}] / N\}$

Removing N from numerator and denominator in (e):

(g): $T = [f(n,c) - f(n)f(c)/N] / \sqrt{f(n,c)}$

This still looks rather complex, but it simplifies again, since the main factor in reaching a high value for T is simply that f(n,c) should have a high absolute value. Or, in (f), E, as we have seen, will often be a very small number. Therefore the main factor in the value of T is O.

Using the invented *cause* - *trouble* figures again:

$T = [10 - (100 \times 50) / 1{,}000{,}000] / \sqrt{10}$
$= [10 - 0.005] / 3.1623 = 3.1607$

In such cases, where n and c are not very frequent (and most words are infrequent), and where the corpus is large, then f(n)f(c)/N, will be a very small number (here 0.005). In such cases, subtracting this number from f(n,c) will make only a small difference. It follows [NOTE 3] that

(h): $T = \text{approx } f(n,c) / \sqrt{f(n,c)} = \sqrt{f(n,c)}$

Therefore the main factor in the value of T is simply the absolute frequency of joint occurrences. The T-value picks out cases where there are many joint occurrences, and therefore provides confidence that the association between n and c is genuine. The statistic f(n,c) undergoes the minor arithmetic transformation to its square root. (In the above case, T = 3.1607, ie T = approx √10 = 3.1646.) Thus T rises more slowly than f(n,c), since it rises only as the square root of f(n,c). But, clearly, whether we rank order things by raw frequency or by its square root makes no difference. (Both the logarithmic value for I and the square root value for T reduce the apparent differences between collocates.)

However, T is sensitive to an increase in the product f(n)f(c). In such cases

$$T = [f(n,c) - X] / \sqrt{f(n,c)}, \text{ from (g)}$$

where X is significantly large relative to f(n,c). Since T decreases if f(n)f(c) becomes very large, the formula has a built-in correction for cases involving very common words. In practice, this correction has a large effect only with a small number of common grammatical words, especially if they are in combination with a second relatively common word. Suppose, for a corpus of 1 million words, that

n is *cause*, and f(n) = 100
c is *the*, and f(c) = 100,000
then, f(n)f(c)/N = 10,000,000 / 1,000,000 = 10

And suppose that f(n,c) = 50. Then

$$T = (50 - 10) / \sqrt{50} = 5.66.$$

Here, the simplified formula, without the correction, gives a higher result: T = approx √f(n,c) = √50 = 7.07.

If the corpus gets larger, but f(n)f(c) stays the same, then f(n)f(c)/N decreases again, and T correspondingly increases. Thus T is larger when we have looked at a larger corpus, and can be correspondingly more confident of our results. Again, this effect is noticeable only in cases where node and/or collocate are frequent.

Some real figures from the 120 million word Cobuild corpus illustrate these points. The node word form is cause.

collocate	f(c)	f(n,c)	T		√f(n,c)
the	7,033,331	9,875	25.54	is much less than	99.37
of	3,331,470	5,605	28.66	is much less than	74.87
can	235,223	1,605	33.94	is less than	40.06
problems	32,738	602	23.14	approx equals	24.54
disease	12,577	237	14.54	approx equals	15.39
nausea	794	40	6.19	approx equals	6.32

For the grammatical words, T is much lower than √f(n,c). But for the lexical words, √f(n,c) provides an approximation to T. In addition, in the top 100 collocates (by T), the rank order of nouns by T is almost the same as their rank order by raw frequency of joint occurrence. The frequent adjectives major, real, good, and single (where f(c) is less than 20,000) are rather

more out of sequence. However, for lexical words, use of raw frequency of joint occurrence as a statistic is unlikely to lead to any significant collocates being missed.

Depending on the contents of the corpus, the calculation of T may violate the assumption of independent observations in other ways. The statistics for calculating T are derived from concordance lines, but these lines are not selected at random from the language. If the corpus consists of long texts (eg whole books) or of millions of words from a single source (eg editions of a single newspaper), then one can expect many lexical repetitions across the corpus. For some purposes, it may be better to have a smaller corpus consisting of smaller varied samples (on the model of LOB's 500 samples of 2,000 words each) than a larger corpus which consists of larger homogeneous samples. Currently, such questions of corpus design remain completely unresolved (though see Biber 1990 for a detailed defence of small varied corpora).

4.6. Comparison of I- and T-values

Two sets of genuine calculations across 1.5 million words are:

CAUSE - trouble: I = 7.38, T = 2.81.
f(CAUSE) = 308; f(trouble) = 245; f(CAUSE, trouble) = 8.

the - happy: I = 1.68, T = 3.95.
f(the) = 80,833; f(happy) = 201; f(the, happy) = 33.

For CAUSE-*trouble*, √f(n,c) = √8 = 2.83 = approx T. For *the-happy*, √f(n,c) = √33 = 5.74 > T. But despite the correction, T gives weight to the absolute value 33 of f(n,c), and T is still larger than with CAUSE - *trouble*.)

These calculations usefully distinguish two different cases. First, CAUSE - *trouble* is, intuitively, a semantically significant collocation, and it gets a highish I-value. But it gets a lowish T-value, since the collocation is observed only 6 times. Second, the collocation *the - happy* is not very interesting, lexically or semantically, and I is low. But it is observed much more frequently (33 times), and we can therefore be confident that it is not a coincidence. Even against the much larger separate frequencies of n and c, it acquires a higher T-value. We are confident that there is an association (this is what T shows), though it is rather weak (this is what I shows). For these reasons, I picks out lexical collocates (which are relatively infrequent). Whereas T picks out both lexical and grammatical collocates.

4.7. Back to the beginning ...

Given the apparent complexity of some of the literature on the topic(!), and the unclear relation of proposed formulae to standard statistical tests, it is important to stress again that:

I is a simple variant of O/E
and T is a simple variant of f(n,c).

In the literature (eg Church et al 1991), different formulae are proposed, especially for T. Precisely what these variants are is less important than an understanding of the effect of the chosen variant. I and T are constructed so as to give more weight to different cases. But for

many purposes, raw figures, which present observed versus expected frequencies of collocations, may be sufficient.

Statisticians often transform data (eg to their square root or logarithm) in order to fit results to some other familiar set of values. But this can hide the original values and make them more difficult to interpret. Both square roots and logarithmic scales decrease, and possibly disguise, differences between results. Linguists should certainly keep an eye on the original raw frequencies of collocations. In addition, any statistic is merely a way of summarizing data, and any clear way of summarizing data may be useful: eg listing raw frequencies. Conventional significance tests merely indicate the strength of evidence from a single experiment: they indicate the likelihood that this evidence has been distorted by sampling errors. A result may not reach "significance", as defined by such a test, due to a bias or to natural variability in the data: and it is obvious to corpus linguists that language is highly variable. This does not mean that such results are of no interest unless they have significance levels attached to them. (Woods et al 1986: 127ff, 246ff.)

4.8. Interpreting quantitative findings

The statistics are produced mechanically by the software and can be calculated for any new corpus, but their interpretation clearly involves subjective judgements. The raw frequency of joint occurrence must, of course, be reasonably large for either statistic to make sense. What is "reasonably large" is a moot point. Single occurrences are clearly meaningless for any statistical argument, and the literature suggests ignoring cases where f(n,c) is less than 3 or 5. But the cut-off point is a matter of judgement.

It is clear that I and T-values work in practice, in that they can be combined in a semi-automatic procedure which identifies lexical sets. They can be used as filters which catch collocates likely to be of linguistic interest. Thus the output from the software can be rank ordered according to the various values, and all cases which fall below some threshold can be discarded. For example:

> Rank order output by joint frequency: discard singletons, ie all cases where f(n,c) = 1.
> Re-order by I: discard all cases where I is less than 3.
> Re-order again by T: discard all cases where T is less than 2.

These thresholds could of course be set at a higher level. The important thing is that we have a replicable procedure for filtering out cases which might be entirely due to chance. The cases which survive the filters provide a set of words, based on solid quantitative evidence, for further human interpretation. These are the cases where we can be confident that there is a strong association between node and collocate.

4.9. Documenting lexical sets

Once the main patterns are clear from these filtered results, it is worth reconsidering cases discarded by these automatic procedures. For example, in the CAUSE analysis on 1.5 million words, for the collocates in (29), it was the case that

> f(n,c) is greater than or equal to 5; I is less than 6; T is less than 2.4.

(29) accident, concern, damage, death, trouble.

These words survive the filters, and we can be confident that they are "typical" collocates. For the collocates in (30) it was the case that

> f(n,c) = 2, I is greater than 7.6 (high, but dubious given that f(n,c) is low), and t = approx 1.4 (low).

(30) chaos, complaint, disease, deficiency, mistake, nuisance.

And collocates where f(n,c) = 1, (and where I and T are therefore meaningless) included:

(31) alcoholism, commotion, drought, epilepsy, pandemonium.

Sets (30) and (31) involve small numbers, and give no confidence that the association is statistically significant. But statistics are not everything. To the human analyst, (30) and (31) provide further examples of the semantic pattern already identified with solid quantitative evidence in (29). They therefore help to build up more complete semantic sets, on the basis of core collocates for which we have good quantitative evidence, plus less frequent - but still attested - collocates.

Indeed, such a procedure is essential since, beyond highly frequent words, the relative frequencies of words are very variable in different corpora. Firth (1957) pointed out that collocations vary with genre, since they depend on the content of the texts in the corpora. This is a powerful objection to basing linguistic description purely on a mechanical use of corpus data. However, in combination with native speaker intuition (Fillmore 1992), a corpus allows us to get the facts right, amass examples and document things thoroughly, and document types of facts (eg about frequency and typicality) which are not open to introspection and which are not well described in current dictionaries and grammars.

It may also be worth grouping the data. For example, the human analyst might group all unpleasant nouns amongst the collocates: this involves subjective, but in this case simple, judgements. One can then sum their occurrences, take this value as f(n,c), and re-calculate I and T. If we total the 5 negative collocates in (29), then:

> f(n,c) = 33, f(n) = 308, f(c) = 739; and I is greater than 7, T = approx 5.7.

The high T-value for set (29) gives us more confidence that the association is genuine. I is not sensitive to absolute numbers of occurrence, but remains respectably high.

The I and T statistics can help to identify not just individual collocates, but also semantic and lexical sets. They are a step in the automatic design of a thesaurus, and when such semantic sets have been discovered, with initial human intervention, they might then be used to automate, and refine, such grouping.

5. COMPARATIVE SEMANTIC PROFILES: 120 MILLION WORD CORPUS

Using these methods (and the software described in Clear 1993), I extracted from the 120 million word Cobuild corpus comparative data on semantically related lemmas: CAUSE, AFFECT, CONSEQUENCE, CREATE, EFFECT, HAPPEN, REASON. Precisely how

comparative profiles are best presented, and in how much detail, will depend on the purpose: for example, a learners' dictionary or normative data for a stylistic analysis. But the basic idea is simple: each word is represented by a set of values which comprise a list of the most significant collocates with associated statistics.

All collocates cited below are from the top 50 T-and I-values, using a window of 4:4. For reasons discussed above, it makes little sense to estimate probability levels for such findings. However, in all cases, I is greater than 3, T is greater than 4, and usually T is greater than 10. The precise rank-order amongst collocates is at least partly dependent on the corpus, and I therefore list groups of collocates in alphabetic order. As collocates, I list mainly lexical (content) words, not grammatical (function) words.

5.1. CAUSE

The most frequent collocations (180 to 1,500 occurrences) confirm the findings already presented from smaller corpora. There are no positive collocates in the top 50 T-values. Most collocates are abstract nouns, eg:

(32) anxiety, concern, crisis, damage, distress, embarrassment, explosion, harm, loss, problem, problems, trouble.

Many examples are medical, eg:

(33) aids, blood, cancer, death, deaths, disease, heart, illness, injury, pain, suffering, symptoms, stress, virus.

Collocating adjectives include:

(34) common, considerable, great, major, root, serious, severe.

(The highest T-value of all is for by, showing that caused frequently occurs in a passive with agent.) The top I-values include words which are themselves not frequent, but, when they do occur, often co-occur with CAUSE:

(35) consternation, grievous, uproar [f(n,c) is greater than 50]; célébre, irreparable [f(n,c) is greater than 20].

Further absolute figures put things in perspective. In these 120 million words, the word form cause occurs over 16,000 times, and the lemma CAUSE about 38,000 times. If the word form cause is studied in the separate columns of a 3:3 span, then the word celebration occurs as the only positive collocate in the top 50 collocations. But this single positive instance vanishes again when the collocations for the lemma are collapsed together. For the lemma CAUSE, the dozen most frequent collocates, with raw frequencies of co-occurrence, summed for a window of 3:3 are:

(36) problem(s) 1806, damage 1519, death(s) 1109, disease 591, concern 587, cancer 572, pain 514, trouble 471, great 391, major 365, common 355, serious 351.

Closer inspection of these findings reveals one problem. The frequent collocation with *great* is partly due to phrases such as *cause for great concern*. Similarly, a frequent collocate of

CAUSE is *driving*, not because the words directly collocate, but because of the phrase *reckless driving*, which in turn occurs in phrases such as *death caused by reckless driving*. Another collocate is *natural*: due to occurrences of *death from natural causes*. Another is *grievous*: due to *cause grievous bodily harm*. Another is *irreparable*: due to *cause irreparable damage*. Another is *untold*, due to phrases such as *cause untold damage / death and destruction / heartache / misery / pain*. Such inter-collocations are beyond the scope of the methods discussed here.

5.2. CREATE

Not all lemmas have such clear prosodies as CAUSE. CREATE is "prosodically mixed or incomplete" (Louw 1993). Again, the commonest collocates are abstract nouns. But negative, neutral and positive examples are mixed in sequence amongst the top 50 T-values, as respectively:

(37) illusion, problems; atmosphere, conditions, environment, image, impression, situation, space; new, jobs, opportunities, order, wealth [f(n,c) is greater than 180].

The highest I-values include:

(38) havoc, illusion, newly [f(n,c) is greater than 50].

A larger context reveals that apparently neutral examples are both positive and negative:

(39) create a bad / false impression; create order from chaos; create the right atmosphere; create a sense of security.

The historical citations in the OED are also mixed. Many citations are positive. A separate sense is given as "Of the divine agent: to bring into being", as in:

(40) In the begynnynge God created heauen and earth. [1535, Coverdale Bible.]

(God also occurs in the top 20 in the 3.3 million word corpus, and in the top 50 in the 120 million word corpus!) But the OED also gives a few early negative collocations:

(41) Creating awe and fear in other men. [1599, Shakespeare.]
(42) Difficulties of their own creating. [1667.]

A hypothesis worth investigating might be that, since the 1500s, negative collocations have been increasing at the expense of predominantly positive collocations. The OED citations provide a hint of this, but not firm evidence. (See below on system and use.)

5.3. REASON

REASON is largely neutral. Highest T- and I-values include, respectively:

(43) apparent, different, good, main, obvious, political, real, same, simple, variety [f(n,c) is greater than 125];

(44) altruistic, cogent, compelling, discernible, earthly, extrinsic, humanitarian, obvious, ostensible, rhyme, selfish, unexplained, unfathomable, unstated, valid, [f(n,c) is greater than 10].

The collocate *good* is very common [1342], presumably due to the fixed phrase *with good reason*.

5.4. RESULT

RESULT is mixed. High T- and I-values include, respectively:

(45) disappointing, election, end, expected, final, interim, latest, losses, official, positive, preliminary, test [f(n,c) > 230]
(46) disappointing, inconclusive, preliminary, unintended, unofficial [f(n,c) is greater than 10].

5.5. AFFECT

AFFECT has a clearly negative prosody. Things are usually badly or adversely affected. Collocates with the highest T-and I-values include, respectively:

(47) adversely, badly, directly, disease, seriously, severely, worst [f(n,c) is greater than 100].
(48) adversely, drought, floods, f(o)etus, negatively, severely, worst [f(n,c) is greater than 20].

The sense of affected as in affected and conceited is clearly critical and negative. There are several medical collocates:

(49) brain, disease, f(o)etus, health.

Medical examples include:

(50) a stroke affected the brain; his face was affected, the pain extending from ...; haemophilia, one younger brother being affected; seriously affecting his whole nervous system; malaise adversely affecting his physical health.

There are also neutral collocates (eg *areas, changes, countries, people, factors, lives*). But the clearest fact is the lack of positive collocates. And the negative prosody on AFFECT can (even with no explicitly negative word in the collocational span) make it difficult to interpret utterances positively. For example, if something *affects* the accuracy of the solution, or if interest rates *affect* the cost of land, such examples are almost inevitably given a negative reading.

5.6. EFFECT

Effects are usually adverse, but can also be beneficial and positive. Highest T- and I-values include, respectively:

(51) adverse, devastating, dramatic, harmful, ill, negative, profound, toxic [f(n,c) is greater than 100]
(52) adverse [f(n,c) = 352], deleterious [f(n,c) = 19].

Several collocates are medical: *drugs, placebo, psychological, vasodilatory*. And I is high for the related: *cumulative, multiplier, snowball*.

5.7. HAPPEN

Things which happen are usually bad and unexpected (eg *accidents*), and occasionally good and unexpected (eg *miracles*). The highest I-values include:

(53) untoward, unthinkable; accident(s), tragedies; miracles [f(n,c) is greater than 10]

The top 50 T-values include no lexical words at all, only grammatical words such as:

(54) what, something, thing(s), nothing, whatever, anything.

Here, the T-value picks out characteristic syntactic constructions. Concordance lines include:

(55) not about the war but about what would happen afterwards
(56) a crisis that never should have happened
(57) puzzling as to what could have happened to his fiancee

Here, the negative collocates are missed with a window of 4:4. I therefore studied in more detail in the smaller corpora "what kinds of things happen" in English by looking at collocates within a window of 8:0. The highest T-value was accident, and the other most frequent collocates were:

(58) untoward, unthinkable; accident(s), dreadful, prevent, problems, tragedies; funny, laugh; miracles.

5.8. CONSEQUENCE

Consequences are usually bad and serious, and often unexpected. Both plural and singular are largely negative. Highest T- and I-values for the plural include, respectively:

(59) devastating, dire, disastrous, fear, grave, negative, serious, severe, suffer(ing), terrible, tragic, war [f(n,c) is greater than 30];
(60) catastrophic, devastating, dire, disastrous, grave, tragic, unintended [f(n,c) is greater than 30].

Consequences are also:

(61) incalculable, unforeseeable/seen, unpredictable.

As the Cobuild dictionary points out, *to take the consequences*, without any further qualification, implies something unpleasant.

6. DISCUSSION: ON INSTANCE AND SYSTEM; AND ON INDUCTION

The most general implication of the arguments presented here is that meaning can be analysed empirically by methods of text and corpus analysis (Sinclair 1991, Phillips 1989), as well as conceptually by more traditional analysis of semantic or lexical fields. This raises the question of inductive methods in linguistics.

6.1. Discovery procedures

Chomsky's (1957, 1965) rejection of induction, by machines or humans, is still widely assumed to be valid. In his attack on American structuralism, he rejects the concept of "discovery procedures". But he provides no real arguments against such methods, merely stating that linguistic theory is not "a manual of procedures", and asserting that there are simply no practical and mechanical ways of extracting a grammar from a corpus of utterances (1957: 50ff, 1965: 18), and that indeed no other academic discipline demands that a theory be extractable from the primary data.

I have emphasized throughout that no procedures can ever be entirely automatic. We always start with intuitions about what is interesting to study, and intuition re-enters, in designing procedures and in interpreting findings. But, given such caveats (which apply to any study of anything), quantitative procedures can identify lexical sets largely on the basis of the frequency and distribution of lexical items in a corpus, leaving the human analyst to discard a few irrelevant collocates which the procedure throws up (due to the idiosyncratic content of corpora), and to interpret the resulting lexical sets.

The computational power and the size of corpora now available are so much greater than anything Chomsky could have conceived in 1957 or 1965, that it is worth re-considering the question of how automatic discovery procedures and intuition can be combined.

6.2. Notes on corpus size

Chomsky also rejects the concept of induction in human language acquisition. But again, the computational methods now available make it worth while reconsidering this question. Patterns in an individual text are interpreted against the background of patterns in the language as we have experienced it. As Sinclair (1965) puts it:

> "Any stretch of language has meaning only as a sample of an enormously large body of text; it represents the results of a complicated selection process, and each selection has meaning by virtue of all the other selections which might have been made, but have been rejected."

But how large is "enormously large"? Our experience of "a language" comes over years via millions of words. The size of this input differs for different people: hermits, socialites, avaricious readers or bilinguals. And different numbers of words (with different type-token ratios) enter the brain, when one is chatting to friends, listening to the radio, skimming a newspaper (or doing all three at once). But even a very rough calculation indicates the order of magnitude. People are often very poor at estimating large numbers, and often linguists have no idea what kind of number might be at issue: estimates such as "billions and billions" are proposed (and are clearly incorrect).

So, very roughly, suppose a person hears/reads/produces 200 words a minute for 5 hours a day. That would be 60,000 words per day (the size of a shortish book), over 20 million words per year, or over

600 million words in 30 years.

Calculating things from the other end, 1,000 million seconds is about 32 years. If a person averaged one word every two seconds, that would be about

500 million words in 30 years.

Church and Liberman (1991: 88) engage in a similar rough calculation, and conclude that human linguistic experience is over 10 million words per year, ie

over 300 million words in 30 years.

These three estimates are extremely rough, but of the same order of magnitude, and they help to put into perspective the size of various corpora. The following estimates will not be too far out: 1 million words in a month or so; 12 million words in a year or so; and 120 million words in a decade or so. So, corpora of tens to hundreds of millions of words are within the right range for certain kinds of cognitive modelling. (These calculations refer only to the size of corpora: the content of available corpora is not what a normal speaker would be exposed to!)

Corpora of the size currently available (eg Cobuild) should contain the kinds of regularities which allow speakers to induce their implicit knowledge of collocations. Thus, continuing with our very rough calculations (given the figures above for CAUSE), a speaker/reader might come across some 40,000 examples of the lemma CAUSE in 10 years or so, with a corresponding few hundred examples of each of the most frequent collocations. Such calculations provide a glimpse of the type of patterns which are instantiated only across millions of words of text and which remain largely unconscious for speakers. It is evident that no corpus can represent the language as a whole, but I can think of no reason why the corpora I have used should be untypical with respect to semantic prosodies on relatively frequent words.

Furthermore, the findings I have presented can be checked on other data. Representativeness is often discussed in purely theoretical terms, but can be formulated as an empirical question. Findings make predictions which can be tested on other texts, text types and corpora. In practice, after a certain number of concordance lines has been examined, the same collocations start to recur, and more data provide rapidly diminishing returns in the form of new collocates. Experience will rapidly accumulate on how many examples (200 to 300?) must be examined before the core of such patterns becomes clear.

The Chomskyan position on induction is closely related to the langue-parole and competence-performance distinctions. But what such frequency data make very clear is the ultimate inseparability of system and use (Halliday 1993). CAUSE is near the stage where the word itself, out of context, has negative connotations. (AFFECT is already at this point.) The selection restrictions on CAUSE are not (yet?) categorial: it is not (yet?) ungrammatical to collocate CAUSE with explicitly positive words. But it is easy to see how an increase in frequency of use can tip the balance and change the system. More systematic diachronic data

on CAUSE and CREATE might be able to show this happening. (Such diachronic findings are commonplace for phonology.)

6.3. And a note on stylistics

Such studies allow texts to be matched against corpora. The concept of "foregrounding" a textual feature against background linguistic patterns is commonplace in much stylistics, but precise evidence on the background patterns is only now becoming available through corpus study (Louw 1993). It has doubtless often been noted that a series of words which are negative, unpleasant or pejorative (or positive, pleasant, etc) will contribute to cohesive texture. But I have here discussed a more specific cohesive mechanism. A "semantic prosody" stretches across a span of words, and therefore contributes to the cohesion of a text. An occurrence of CAUSE sets up an expectation of some unpleasant, probably abstract, word(s). If/when this occurs, then a little bit of textual cohesion results.

7. CONCLUSIONS: CAUSE FOR CONFIDENCE

These facts about CAUSE may now seem obvious. They certainly now seem "normal" to me: though they didn't before I did this small corpus study. In retrospect, statistics often seem merely to confirm what is blindingly obvious. Alternatively, if you don't like what they tell you, you can always say that statistics can prove anything. Some colleagues with whom I have discussed this analysis still argue that the word CAUSE is neutral. It is just that people talk about gloomy things: crises, problems, troubles, and the like. I have argued that CAUSE acquires guilt by association. At some point the word itself acquires unpleasant connotations, and parole affects langue.

Such methods allow semantic profiles of words to be securely based on millions of words of data. Corpus study and computational techniques are a cause for confidence that lexical descriptions in the future will provide more accurate and exhaustive documentation about words, and will give access to patterns in the language which are not accessible to unaided human observation. These patterns are probabilistic (although how probability levels might be calculated is at present unclear).

In this article, I have defined a method for identifying collocations, and shown that related methods discussed in the literature can be simplified in various ways. I have also shown that the results obtained with this method have implications for a general model of language, particularly with reference to the nature of lexico-semantic categories. The results reveal a type of relationship, between lemmas and semantic categories, which is currently captured in neither dictionaries nor grammars.

NOTES

1. The use of base 2 logarithms reflects the origin of the concept in information theory, and has no real significance here. Indeed, the logarithmic function can disguise real differences in the data, since it means that values of I do not increase linearly in proportion to N. If other values remain constant, N (or O/E) has to double, for I to increase by 1. If $I = \log_2 O/E$, then $2^I = O/E$. For example:

if O/E = 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, etc
then I = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, etc.

2. Consider the case where the number of collocations is exactly half the number of independent occurrences of n and c. For example, for increasing numbers of collocations:

f(n,c)	f(n)	f(c)	f(n,c) / f(n)f(c)
2	4	4	2/16 = 0.13
4	8	8	4/64 = 0.06
8	16	16	8/256 = 0.03

Here, I decreases in cases where there are more independent occurrences of n and c, even if there are more joint occurrences of n and c. I decreases when the probability of observing n and c independently increases - irrespective of the number of collocations observed. That seems intuitively reasonable. But it can lead to strange results. Consider the limiting case of a "perfect collocation" where:

$f(n,c) = f(n) = f(c) = y.$

Suppose *nuts* and *bolts* each occur 10 times in a corpus, and the phrase *nuts and bolts* also occurs 10 times. Every time *nuts* occurs it collocates with *bolts*, and vice versa. Then,

$O/E = (y \times N) / y^2 = N/y.$

As y increases, N/y (= O/E) decreases, and therefore I decreases: even though the association between n and c is observed in every case. Indeed the highest score would be where y = 1, and where O/E would just equal N. But if we have only one collocation of words which occur only once each, then nothing can be concluded. Given the logarithmic scale, the value of I changes slowly. Eg for increasing values of y, where N = 1 million:

y	O/E = N/y	$\log_2$ O/E = I
1	1,000,000	equals approx 20
2	500,000	equals approx 19
3	333,333	is greater than 18
5	200,000	is greater than 17
10	100,000	is greater than 16
20	50,000	is greater than 15
50	20,000	is greater than 14
100	10,000	is greater than 13
1,000	1,000	= approx 10

It is counter-intuitive that "mutual information" is lower with 1,000 cases rather than one, if we forget the origins of the concepts in "information theory": the unique collocation conveys more "information" in the sense that it is unexpected and unpredictable. (Cf Hallan 1994.)

3. Alternatively, using (f) again:

$T = [O - E] / [(\sqrt{O}) / N] = [f(n,c) / N - E] / [(\sqrt{f(n,c)}) / N]$

Disregarding E in cases where it is very small, and removing N from numerator and denominator, again:

$$T = \text{approx } f(n,c) / \sqrt{f(n,c)} = \sqrt{f(n,c)}$$

ACKNOWLEDGEMENTS

I am grateful to: the Norwegian Computing Centre for the Humanities and Longman publishers for permission to use various corpus data; Gwynneth Fox, Malcolm Coulthard and John Sinclair for arranging access to the Birmingham Cobuild corpus, the Bank of English; Andrea Gerbig for providing other corpus materials; Tim Lane for help with data extraction from Cobuild; Brigitte Grote and Oliver Jakobs who wrote software for studying corpora in Trier; Wolfram Bublitz, Andrea Gerbig and Gabi Keck for comments on previous drafts; Jeremy Clear, Naomi Hallan, Oliver Jakobs and Zoe James for trying to explain to me (several times) what I- and T-values mean (it's not their fault if I still haven't understood it, and Naomi Hallan certainly wants to reserve her position); Susan Hunston for providing me with good examples of positive prosodies (eg on PROVIDE).

REFERENCES

Baker, M., Francis, G. and Tognini-Bonelli, E. eds (1993) *Text and Technology: In Honour of John Sinclair*. Amsterdam: Benjamins.

Biber, D. (1990) Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, 5, 4: 257-69.

Chomsky, N. (1957) *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1957) *Aspects of the Theory of Syntax*. Cambridge, Mass: MIT Press.

Church, K. and P. Hanks (1990) Word association norms, mutual information and lexicography. *Computational Linguistics*, 16, 1: 22-29.

Church, K. and M. Liberman (1991) A status report on ACL/DCI. *Using Corpora*. Proceedings of 7th Annual Conference of UW centre for the New OED and Text Research.

Church, K., W. Gale, P. Hanks and D. Hindle (1991) Using statistics in lexical analysis. In U. Zernik ed *Lexical Acquisition*. Englewood Cliff, NJ: Erlbaum. 115-64.

Clear, J. (1993) From Firth principles: computational tools for the study of collocation. In Baker et al: 271-92.

Fillmore, C. J. (1992) Corpus linguistics or computer-aided armchair linguistics. In J. Svartvik ed *Directions in Corpus Linguistics*. Berlin: Mouton. 35-60.

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

Hallan, N. (1994) On choosing a collocation filter. Unpublished. University of Trier.

Halliday, M. A. K. (1993) Quantitative studies and probabilities in grammar. In M. Hoey ed *Data, Description, Discourse. Papers on the English Language in Honour of John Sinclair*. London: HarperCollins. 1-25.

Louw, B. (1993) Irony in the text or insincerity in the writer: the diagnostic potential of semantic prosodies. In Baker et al: 157-76.

Phillips, M. (1989) *Lexical Structure of Text*. University of Birmingham: English Language Research.

Sinclair, J. M. (1965) When is a poem like a sunset? *A Review of English Literature*, 6, 2: 76-91.
Sinclair, J. M. et al (1987) *Collins Cobuild English Language Dictionary*. London: HarperCollins.
Sinclair, J. M. et al (1988) *Collins Cobuild Essential English Dictionary*. London: HarperCollins.
Sinclair, J. M. et al (1990) *Collins Cobuild English Grammar*. London: HarperCollins.
Sinclair, J. M. (1991) *Corpus, Concordance, Collocation*. Oxford: OUP.
Woods, A., P. Fletcher and A. Hughes (1986) *Statistics in Language Studies*. Cambridge: CUP.