Corpus linguistics

A guide to the methodology

Anatol Stefanowitsch

Textbooks in Language Sciences 7

language science press

7 Collocation

The (orthographic) word plays a central role in corpus linguistics. As suggested in Chapter 4, this is in no small part due to the fact that all corpora, whatever additional annotations may have been added, consist of orthographically represented language. This makes it easy to retrieve word forms. Every concordancing program offers the possibility to search for a string of characters – in fact, some are limited to this kind of query.

However, the focus on words is also due to the fact that the results of corpus linguistic research quickly showed that words (individually and in groups) are more interesting and show a more complex behavior than traditional, grammar-focused theories of language assumed. An area in which this is very obvious, and which has therefore become one of the most heavily researched areas in corpus linguistics, is the way in which words combine to form so-called *collocations*.

This chapter is dedicated entirely to the discussion of collocation. At first, this will seem like a somewhat abrupt shift from the topics and phenomena we have discussed so far – it may not even be immediately obvious how they fit into the definition of corpus linguistics as "the investigation of linguistic research questions that have been framed in terms of the conditional distribution of linguistic phenomena in a linguistic corpus", which was presented at the end of Chapter 2. However, a closer look will show that studying the co-occurrence of words and/or word forms is simply a special case of precisely this kind of research program.

7.1 Collocates

Trivially, texts are not random sequences of words. There are several factors influencing the likelihood of two (or more) words occurring next to each other.

First, the co-occurrence of words in a sequence is restricted by grammatical considerations. For example, a definite article cannot be followed by another definite article or a verb, but only by a noun, by an adjective modifying a noun, by an adverb modifying such an adjective or by a post-determiner. Likewise, a transitive verb requires a direct object in the form of a noun phrase, so – barring cases where the direct object is pre- or post-posed – it will be followed by a word that can occur at the beginning of a noun phrase (such as a pronoun, a determiner, an adjective or a noun).

Second, the co-occurrence of words is restricted by semantic considerations. For example, the transitive verb *drink* requires a direct object referring to a liquid, so it is probable that it will be followed by words like *water*, *beer*, *coffee*, *poison*, etc., and improbable that it will be followed by words like *bread*, *guitar*, *stone*, *democracy*, etc. Such restrictions are treated as a grammatical property of words (called *selection restrictions*) in some theories, but they may also be an expression of our world knowledge concerning the activity of drinking.

Finally, and related to the issue of world knowledge, the co-occurrence of words is restricted by topical considerations. Words will occur in sequences that correspond to the contents we are attempting to express, so it is probable that co-occurring content words will come from the same discourse domain.

However, it has long been noted that words are not distributed randomly even within the confines of grammar, lexical semantics, world knowledge, and communicative intent. Instead, a given word will have affinities to some words, and disaffinities to others, which we could not predict given a set of grammatical rules, a dictionary and a thought that needs to be expressed. One of the first principled discussions of this phenomenon is found in Firth (1957). Using the example of the word *ass* (in the sense of 'donkey'), he discusses the way in which what he calls *habitual collocations* contribute to the meaning of words:

> One of the meanings of *ass* is its habitual collocation with an immediately preceding *you silly*, and with other phrases of address or of personal reference. ... There are only limited possibilities of collocation with preceding adjectives, among which the commonest are *silly*, *obstinate*, *stupid*, *awful*, occasionally *egregious*. *Young* is much more frequently found than *old*. (Firth 1957: 194f).

Note that Firth, although writing well before the advent of corpus linguistics, refers explicitly to *frequency* as a characteristic of collocations. The possibility of using frequency as part of the definition of collocates, and thus as a way of identifying them, was quickly taken up. Halliday (1961) provides what is probably the first strictly quantitative definition (cf. also Church & Hanks (1990) for a more recent comprehensive quantitative discussion):

> Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n lexical items) from an item x, the items a, b, c... Any given item thus enters into a range of collocation, the items with which it is collocated being ranged from more to less probable... (Halliday 1961: 276)

7.1.1 Collocation as a quantitative phenomenon

Essentially, then, collocation is just a special case of the quantitative corpus linguistic research design adopted in this book: to ask whether two words form a collocation (or: are collocates of each other) is to ask whether one of these words occurs in a given position more frequently than expected by chance under the condition that the other word occurs in a structurally or sequentially related position. In other words, we can decide whether two words *a* and *b* can be regarded as collocates on the basis of a contingency table like that in Table 7.1. The First Position in the sequence is treated as the dependent variable, with two values: the word we are interested in (here: word a), and all other words. The Second Position is treated as the independent variable, again, with two values: the word we are interested in (here: word b), and all other words (of course, it does not matter which word we treat as the dependent and which as the independent variable, unless our research design suggests a particular reason).[1]

Table 7.1: Collocation

		Second Position		
		word b	other words	Total
First Position	word a	a & b	a & other	a
	other words	other & b	other & other	other
	Total	b	other	corpus size

On the basis of such a table, we can determine the collocation status of a given word pair. For example, we can ask whether Firth was right with respect to the claim that *silly ass* is a collocation. The necessary data are shown in Table 7.2: As discussed above, the dependent variable is the First Position in the sequence, with the values silly and ¬silly (i.e., all words that are not *ass*); the independent variable is the Second Position in the sequence, with the values ass and ¬ass.

[1]Note that we are using the corpus size as the table total – strictly speaking, we should be using the total number of two-word sequences (bigrams) in the corpus, which will be lower: The last word in each file of our corpus will not have a word following it, so we would have to subtract the last word of each file – i.e., the number of files in our corpus – from the total. This is unlikely to make much of a difference in most cases, but the shorter the texts in our corpus are, the larger the difference will be. For example, in a corpus of tweets, which, at the time of writing, are limited to 280 characters, it might be better to correct the total number of bigrams in the way described.

Table 7.2: Co-occurrence of *silly* and *ass* in the BNC

		Second Position		
		ass	¬ass	Total
First Position	silly	7	2632	2639
		(0.01)	(2638.99)	
	¬silly	295	98 360 849	98 361 144
		(301.99)	(98 360 842.01)	
	Total	302	98 363 481	98 363 783

The combination *silly ass* is very rare in English, occurring just seven times in the 98 363 783 word BNC, but the expected frequencies in Table 7.2 show that this is vastly more frequent than should be the case if the words co-occurred randomly – in the latter case, the combination should have occurred just 0.01 times (i.e., not at all). The difference between the observed and the expected frequencies is highly significant ($\chi^2 = 6033.8, \text{df} = 1, p < 0.001$). Note that we are using the χ^2 test here because we are already familiar with it. However, this is not the most useful test for the purpose of identifying collocations, so we will discuss better options below.

Generally speaking, the goal of a quantitative collocation analysis is to identify, for a given word, those other words that are characteristic for its context of usage. Tables 7.1 and 7.2 present the most straightforward way of doing so: we simply compare the frequency with which two words co-occur to the frequencies with which they occur in the corpus in general. In other words, the two conditions across which we are investigating the distribution of a word are "next to a given other word" and "everywhere else". This means that the corpus itself functions as a kind of neutral control condition, albeit a somewhat indiscriminate one: comparing the frequency of a word next to some other word to its frequency in the entire rest of the corpus is a bit like comparing an experimental group of subjects that have been given a particular treatment to a control group consisting of all other people who happen to live in the same city.

Often, we will be interested in the distribution of a word across two specific conditions – in the case of collocation, the distribution across the immediate contexts of two semantically related words. It may be more insightful to compare adjectives occurring next to *ass* with those occurring next to the rough synonym *donkey* or the superordinate term *animal*. Obviously, the fact that *silly* occurs

more frequently with *ass* than with *donkey* or *animal* is more interesting than the fact that *silly* occurs more frequently with *ass* than with *stone* or *democracy*. Likewise, the fact that *silly* occurs with *ass* more frequently than *childish* is more interesting than the fact that *silly* occurs with *ass* more frequently than *precious* or *parliamentary*.

In such cases, we can modify Table 7.1 as shown in Table 7.3 to identify the collocates that *differ* significantly between two words. There is no established term for such collocates, so we we will call them *differential collocates* here[2] (the method is based on Church et al. 1991).

Table 7.3: Identifying differential collocates

		Second Position		
		word b	word c	Total
First Position	word a	a & b	a & c	a
	other	other & b	other & c	other
	Total	b	c	sample size

Since the collocation *silly ass* and the word *ass* in general are so infrequent in the BNC, let us use a different noun to demonstrate the usefulness of this method, the word *game*. We can speak of *silly game(s)* or *childish game(s)*, but we may feel that the latter is more typical than the former. The relevant lemma frequencies to put this feeling to the test are shown in Table 7.4.

Table 7.4: *Childish game* vs. *silly game* (lemmas) in the BNC

		First Position		
		childish	silly	Total
Second Position	game	12	31	43
		(6.18)	(36.82)	
	¬game	431	2608	3039
		(436.82)	(2602.18)	
	Total	443	2639	3082

[2]Gries (2003b) and Gries & Stefanowitsch (2004) use the term *distinctive collocate*, which has been taken up by some authors; however, many other authors use the term *distinctive collocate* much more broadly to refer to *characteristic* collocates of a word.

The sequences *childish game(s)* and *silly game(s)* both occur in the BNC. Both combinations taken individually are significantly more frequent than expected (you may check this yourself using the frequencies from Table 7.4, the total lemma frequency of *game* in the BNC (20 627), and the total number of words in the BNC given in Table 7.2 above). The lemma sequence *silly game* is more frequent, which might lead us to assume that it is the stronger collocation. However, the direct comparison shows that this is due to the fact that *silly* is more frequent in general than *childish*, making the combination *silly game* more probable than the combination *childish game* even if the three words were distributed randomly. The difference between the observed and the expected frequencies suggests that *childish* is more strongly associated with *game(s)* than *silly*. The difference is significant ($\chi^2 = 6.49, \text{df} = 1, p < 0.05$).

Researchers differ with respect to what types of co-occurrence they focus on when identifying collocations. Some treat co-occurrence as a purely sequential phenomenon defining collocates as words that co-occur more frequently than expected within a given span. Some researchers require a span of 1 (i.e., the words must occur directly next to each other), but many allow larger spans (five words being a relatively typical span size).

Other researchers treat co-occurrence as a structural phenomenon, i.e., they define collocates as words that co-occur more frequently than expected in two related positions in a particular grammatical structure, for example, the adjective and noun positions in noun phrases of the form [Det Adj N] or the verb and noun position in transitive verb phrases of the form [V [$_{\text{NP}}$ (Det) (Adj) N]].[3] However, instead of limiting the definition to one of these possibilities, it seems more plausible to define the term appropriately in the context of a specific research question. In the examples above, we used a purely sequential definition that simply required words to occur next to each other, paying no attention to their word-class or structural relationship; given that we were looking at adjective-noun combinations, it would certainly have been reasonable to restrict our search parameters to adjectives modifying the noun *ass*, regardless of whether other adjectives intervened, for example in expressions like *silly old ass*, which our query would have missed if they occurred in the BNC (they do not).

It should have become clear that the designs in Tables 7.1 and 7.3 are essentially variants of the general research design introduced in previous chapters and used as the foundation of defining corpus linguistics: it has two variables,

[3]Note that such word-class specific collocations are sometimes referred to as *colligations*, although the term colligation usually refers to the co-occurrence of a word in the context of particular word classes, which is not the same.

Position 1 and Position 2, both of which have two values, namely word x vs. other words (or, in the case of differential collocates, word x vs. word y). The aim is to determine whether the value word a is more frequent for Position 1 under the condition that word b occurs in Position 2 than under the condition that other words (or a particular other word) occur in Position 2.

7.1.2 Methodological issues in collocation research

We may occasionally be interested in an individual pair of collocates, such as *silly ass*, or in a small set of such pairs, such as all adjective-noun pairs with *ass* as the noun. However, it is much more likely that we will be interested in large sets of collocate pairs, such as all adjective-noun pairs or even all word pairs in a given corpus. This has a number of methodological consequences concerning the practicability, the statistical evaluation and the epistemological status of collocation research.

a. Practicability. In practical terms, the analysis of large numbers of potential collocations requires creating a large number of contingency tables and subjecting them to the χ^2 test or some other appropriate statistical test. This becomes implausibly time-consuming very quickly and thus needs to be automated in some way.

There are concordancing programs that offer some built-in statistical tests, but they typically restrict our options quite severely, both in terms of the tests they allow us to perform and in terms of the data on which the tests are performed. Anyone who decides to become involved in collocation research (or some of the large-scale lexical research areas described in the next chapter), should get acquainted at least with the simple options of automatizing statistical testing offered by spreadsheet applications. Better yet, they should invest a few weeks (or, in the worst case, months) to learn a scripting language like Perl, Python or R (the latter being a combination of statistical software and programming environment that is ideal for almost any task that we are likely to come across as corpus linguists).

b. Statistical evaluation. In statistical terms, the analysis of large numbers of potential collocations requires us to keep in mind that we are now performing multiple significance tests on the same set of data. This means that we must adjust our significance levels. Think back to the example of coin-flipping: the probability of getting a series of one head and nine tails is 0.009765. If we flip a coin ten times and get this result, we could thus reject the null hypothesis with a probability of error of 0.010744, i.e., around 1 percent (because we would have to add the probability of getting ten tails, 0.000976). This is well below the level

required to claim statistical significance. However, if we perform one hundred series of ten coin-flips and one of these series consists of one head and nine tails (or ten tails), we could not reject the null hypothesis with the same confidence, as a probability of 0.010744 means that we would expect one such series to occur by chance. This is not a problem as long as we do not accord this one result out of a hundred any special importance. However, if we were to identify a set of 100 collocations with p-values of 0.001 in a corpus, we *are* potentially treating all of them as important, even though it is very probable that at least one of them reached this level of significance by chance.

To avoid this, we have to correct our levels of significance when performing multiple tests on the same set of data. As discussed in Section 6.6.1 above, the simplest way to do this is the Bonferroni correction, which consists in dividing the conventionally agreed-upon significance levels by the number of tests we are performing. As noted in Section 6.6.1, this is an extremely conservative correction that might make it quite difficult for any given collocation to reach significance.

Of course, the question is how important the role of p-values is in a design where our main aim is to identify collocates and order them in terms of their collocation strength. I will turn to this point presently, but before I do so, let us discuss the third of the three consequences of large-scale testing for collocation, the methodological one.

c. Epistemological considerations. We have, up to this point, presented a very narrow view of the scientific process based (in a general way) on the Popperian research cycle where we formulate a research hypothesis and then test it (either directly, by looking for counterexamples, or, more commonly, by attempting to reject the corresponding null hypothesis). This is called the *deductive* method. However, as briefly discussed in Chapter 3, there is an alternative approach to scientific research that does not start with a hypothesis, but rather with general questions like "Do relationships exist between the constructs in my data?" and "If so, what are those relationships?". The research then consists in applying statistical procedures to large amounts of data and examining the results for interesting patterns. As electronic storage and computing power have become cheaper and more widely accessible, this approach – the *exploratory* or *inductive* approach – has become increasingly popular in all branches of science, particularly the social sciences. It would be surprising if corpus linguistics was an exception, and indeed, it is not. Especially the area of collocational research is typically exploratory.

In principle, there is nothing wrong with exploratory research – on the contrary, it would be unreasonable not to make use of the large amounts of language data and the vast computing power that has become available and accessible over

the last thirty years. In fact, it is sometimes difficult to imagine a plausible hypothesis for collocational research projects. What hypothesis would we formulate before identifying all collocations in the LOB or some specialized corpus (e.g., a corpus of business correspondence, a corpus of flight-control communication or a corpus of learner language)?[4] Despite this, it is clear that the results of such a collocation analysis yield interesting data, both for practical purposes (building dictionaries or teaching materials for business English or aviation English, extracting terminology for the purpose of standardization, training natural-language processing systems) and for theoretical purposes (insights into the nature of situational language variation or even the nature of language in general).

But there is a danger, too: Most statistical procedures will produce *some* statistically significant result if we apply them to a large enough data set, and collocational methods certainly will. Unless we are interested exclusively in description, the crucial question is whether these results are meaningful. If we start with a hypothesis, we are restricted in our interpretation of the data by the need to relate our data to this hypothesis. If we do not start with a hypothesis, we can interpret our results without any restrictions, which, given the human propensity to see patterns everywhere, may lead to somewhat arbitrary post-hoc interpretations that could easily be changed, even reversed, if the results had been different and that therefore tell us very little about the phenomenon under investigation or language in general. Thus, it is probably a good idea to formulate at least some general expectations before doing a large-scale collocation analysis.

Even if we do start out with general expectations or even with a specific hypothesis, we will often discover additional facts about our phenomenon that go beyond what is relevant in the context of our original research question. For example, checking in the BNC Firth's claim that the most frequent collocates of *ass* are *silly*, *obstinate*, *stupid*, *awful* and *egregious* and that *young* is "much more frequent" than *old*, we find that *silly* is indeed the most frequent adjectival collocate, but that *obstinate*, *stupid* and *egregious* do not occur at all, that *awful* occurs only once, and that *young* and *old* both occur twice. Instead, frequent adjectival collocates (ignoring second-placed *wild*, which exclusively refers to actual donkeys), are *pompous* and *bad*. *Pompous* does not really fit with the semantics that Firth's adjectives suggest and could indicate that a semantic shift from 'stupidity' to 'self-importance' may have taken place between 1957 and 1991 (when the BNC was assembled).

[4] Of course we are making the implicit assumption that there *will* be collocates – in a sense, this is a hypothesis, since we could conceive of models of language that would not predict their existence (we might argue, for example, that at least some versions of generative grammar constitute such models). However, even if we accept this as a hypothesis, it is typically not the one we are interested in this kind of study.

This is, of course, a new hypothesis that can (and must) be investigated by comparing data from the 1950s and the 1990s. It has some initial plausibility in that the adjectives *blithering*, *hypocritical*, *monocled* and *opinionated* also co-occur with *ass* in the BNC but are not mentioned by Firth. However, it is crucial to treat this as a hypothesis rather than a result. The same goes for *bad ass* which suggests that the American sense of *ass* ('bottom') and/or the American adjective *badass* (which is often spelled as two separate words) may have begun to enter British English. In order to be tested, these ideas – and any ideas derived from an exploratory data analysis – have to be turned into testable hypotheses and the constructs involved have to be operationalized. Crucially, they must be tested on a new data set – if we were to circularly test them on the same data that they were derived from, we would obviously find them confirmed.

7.1.3 Effect sizes for collocations

As mentioned above, significance testing (while not without its uses) is not necessarily our primary concern when investigating collocations. Instead, researchers frequently need a way of assessing the *strength* of the association between two (or more) words, or, put differently, the effect size of their co-occurrence (recall from Chapter 6 that significance and effect size are not the same). A wide range of such association measures has been proposed and investigated. They are typically calculated on the basis of (some or all) the information contained in contingency tables like those in Tables 7.1 and 7.3 above.

Let us look at some of the most popular and/or most useful of these measures. I will represent the formulas with reference to the table in Table 7.5, i.e, O_{11} means the observed frequency of the top left cell, E_{11} its expected frequency, R_1 the first row total, C_2 the second column total, and so on. Note that second column would be labeled OTHER WORDS in the case of normal collocations, and WORD C in the case of differential collocations. The association measures can be applied to both kinds of design.

Table 7.5: A generic 2-by-2 table for collocation research

		SECOND POSITION		
		WORD B	OTHER/WORD C	Total
FIRST POSITION	WORD A	O_{11}	O_{12}	R_1
	OTHER WORDS	O_{21}	O_{22}	R_2
	Total	C_1	C_2	N

Now all we need is a good example to demonstrate the calculations. Let us use the adjective-noun sequence *good example* from the LOB corpus (but horse lovers need not fear, we will return to equine animals and their properties below).

Table 7.6: Co-occurrence of *good* and *example* in the LOB

		Second Position		
		example	¬example	Total
First Position	good	9	836	845
		(0.2044)	(844.7956)	
	¬good	236	1 011 904	1 012 140
		(244.7956)	(1 011 895.2044)	
	Total	245	1 012 740	1 012 985

Measures of collocation strength differ with respect to the data needed to calcuate them, their computational intensiveness and, crucially, the quality of their results. In particular, many measures, notably the ones easy to calculate, have a problem with rare collocations, especially if the individual words of which they consist are also rare. After we have introduced the measures, we will therefore compare their performance with a particular focus on the way in which they deal (or fail to deal) with such rare events.

7.1.3.1 Chi-square

The first association measure is an old acquaintance: the chi-square statistic, which we used extensively in Chapter 6 and in Section 7.1.1 above. I will not demonstrate it again, but the chi-square value for Table 7.6 would be 378.95 (at 1 degree of freedom this means that $p < 0.001$, but we are not concerned with p-values here).

Recall that the chi-square test statistic is not an effect size, but that it needs to be divided by the table total to turn it into one. As long as we are deriving all our collocation data from the same corpus, this will not make a difference, since the table total will always be the same. However, this is not always the case. Where table sizes differ, we might consider using the phi value instead. I am not aware of any research using phi as an association measure, and in fact the chi-square statistic itself is not used widely either. This is because it has a serious problem: recall that it cannot be applied if more than 20 percent of the

cells of the contingency table contain expected frequencies smaller than 5 (in the case of collocates, this means not even one out of the four cells of the 2-by-2 table). One reason for this is that it dramatically overestimates the effect size and significance of such events, and of rare events in general. Since collocations are often relatively rare events, this makes the chi-square statistic a bad choice as an association measure.

7.1.3.2 Mutual Information

Mutual information is one of the oldest collocation measures, frequently used in computational linguistics and often implemented in collocation software. It is given in (1) in a version based on Church & Hanks (1990):[5]

$$(1) \quad MI = \log_2 \left(\frac{O_{11}}{E_{11}} \right)$$

Applying the formula to our table, we get the following:

$$\text{MI} = \log_2 \left(\frac{9}{0.2044} \right) = \log_2 (44.03) = 5.46$$

In our case, we are looking at cases where WORD A and WORD B occur directly next to each other, i.e., the span size is 1. When looking at a larger span (which is often done in collocation research), the probability of encountering a particular collocate increases, because there are more slots that it could potentially occur in. The MI statistic can be adjusted for larger span sizes as follows (where S is the span size):

$$(2) \quad \text{MI} = \log_2 \left(\frac{O_{11}}{E_{11} \times S} \right)$$

[5] A logarithm with a base b of a given number x is the power to which b must be raised to produce x, so, for example, $\log_{10}(2) = 0.30103$, because $10^{0.30103} = 2$. Most calculators offer at the very least a choice between the natural logarithm, where the base is the number e (approx. 2.7183) and the common logarithm, where the base is the number 10; many calculators and all major spreadsheet programs offer logarithms with any base. In the formula in (1), we need the logarithm with base 2; if this is not available, we can use the natural logarithm and divide the result by the natural logarithm of 2:

$$\text{MI} = \frac{\log_e \left(\frac{O_{11}}{E_{11}} \right)}{\log_e (2)}$$

The mutual information measure suffers from the same problem as the χ^2 statistic: it overestimates the importance of rare events. Since it is still fairly widespread in collocational research, we may nevertheless need it in situations where we want to compare our own data to the results of published studies. However, note that there are versions of the MI measure that will give different results, so we need to make sure we are using the same version as the study we are comparing our results to. But unless there is a pressing reason, we should not use mutual information at all.

7.1.3.3 The log-likelihood ratio test

The G value of the log-likelihood ratio test is one of the most popular – perhaps *the* most popular – association measure in collocational research, found in many of the central studies in the field and often implemented in collocation software. The following is a frequently found form (Read & Cressie 1988: 134):

$$G = 2 \sum_{i=1}^{n} O_i \log_e \left(\frac{O_i}{E_i} \right) \tag{3}$$

In order to calculate the G measure, we calculate for each cell the natural logarithm of the observed frequency divided by the expected frequency and multiply it by the observed frequency. We then add up the results for all four cells and multiply the result by two. Note that if the observed frequency of a given cell is zero, the expression O_i/E_i will, of course, also be zero. Since the logarithm of zero is undefined, this would result in an error in the calculation. Thus, log(0) is simply defined as zero when applying the formula in (3).

Applying the formula in (3) to the data in Table 7.6, we get the following:

$$\begin{aligned} G &= 2 \times \left(9 \times \log_e \left(\frac{9}{0.2044} \right) \right) + \left(836 \times \log_e \left(\frac{836}{844.7956} \right) \right) \\ &\quad + \left(236 \times \log_e \left(\frac{236}{244.7956} \right) \right) + \left(1011904 \times \log_e \left(\frac{1011904}{1011895.2044} \right) \right) \\ &= 2 \times ((34.0641) + (-8.7497) + (-8.6357) + (8.7956)) = 50.9489 \end{aligned}$$

The G value has long been known to be more reliable than the χ^2 test when dealing with small samples and small expected frequencies (Read & Cressie 1988: 134ff). This led Dunning (1993) to propose it as an association measure specifically to avoid the overestimation of rare events that plagues the χ^2 test, mutual information and other measures.

7.1.3.4 Minimum Sensitivity

Minimum sensitivity was proposed by Pedersen (1998) as potentially useful measure especially for the identification of associations between content words:

$$\text{MS} = \min\left(\frac{O_{11}}{R_1}, \frac{O_{11}}{C_1}\right) \quad (4)$$

We simply divide the observed frequency of a collocation by the frequency of the first word (R_1) and of the second word (C_1) and use the smaller of the two as the association measure. For the data in Table 7.6, this gives us the following:

$$\text{MS} = \min\left(\frac{9}{836}, \frac{9}{236}\right) = \min\left(0.0108, 0.0381\right) = 0.0108$$

In addition to being extremely simple to calculate, it has the advantage of ranging from zero (words never occur together) to 1 (words always occur together); it was also argued by Wiechmann (2008) to correlate best with reading time data when applied to combinations of words and grammatical constructions (see Chapter 8). However, it also tends to overestimate the importance of rare collocations.

7.1.3.5 Fisher's exact test

Fisher's exact test was already mentioned in passing in Chapter 6 as an alternative to the χ^2 test that calculates the probability of error directly by adding up the probability of the observed distribution and all distributions that deviate from the null hypothesis further in the same direction. Pedersen (1996) suggests using this p-value as a measure of association because it does not make any assumptions about normality and is even better at dealing with rare events than G. Stefanowitsch & Gries (2003: 238–239) add that it has the advantage of taking into account both the magnitude of the deviation from the expected frequencies and the sample size.

There are some practical disadvantages to Fisher's exact test. First, it is computationally expensive – it cannot be calculated manually, except for very small tables, because it involves computing factorials, which become very large very quickly. For completeness' sake, here is (one version of) the formula:

$$p_{\text{exact}} = \frac{R_1! \times R_2! \times C_1! \times C_2!}{O_{11}! \times O_{12}! \times O_{21}! \times O_{22}! \times N!} \quad (5)$$

Obviously, it is not feasible to apply this formula directly to the data in Table 7.6, because we cannot realistically calculate the factorials for 236 or 836, let

alone 1 011 904. But if we could, we would find that the p-value for Table 7.6 is 0.000000000001188.

Spreadsheet applications do not usually offer Fisher's exact test, but all major statistics applications do. However, typically, the exact p-value is not reported beyond the limit of a certain number of decimal places. This means that there is often no way of ranking the most strongly associated collocates, because their p-values are smaller than this limit. For example, there are more than 100 collocates in the LOB corpus with a Fisher's exact p-value that is smaller than the smallest value that a standard-issue computer chip is capable of calculating, and more than 5000 collocates that have p-values that are smaller than what the standard implementation of Fisher's exact test in the statistical software package R will deliver. Since in research on collocations we often need to rank collocations in terms of their strength, this may become a problem.

7.1.3.6 A comparison of association measures

Let us see how the association measures compare using a data set of 20 potential collocations. Inspired by Firth's *silly ass*, they are all combinations of adjectives with equine animals. Table 7.7 shows the combinations and their frequencies in the BNC sorted by their raw frequency of occurrence (adjectives and nouns are shown in small caps here to stress that they are values of the variables Word A and Word B, but I will generally show them in italics in the remainder of the book in line with linguistic tradition).

All combinations are perfectly normal, grammatical adjective-noun pairs, meaningful not only in the specific context of their actual occurrence. However, I have selected them in such a way that they differ with respect to their status as potential collocations (in the sense of typical combinations of words). Some are compounds or compound like combinations (*rocking horse*, *Trojan horse*, and, in specialist discourse, *common zebra*). Some are the kind of semi-idiomatic combinations that Firth had in mind (*silly ass*, *pompous ass*). Some are very conventional combinations of nouns with an adjective denoting a property specific to that noun (*prancing horse*, *braying donkey*, *galloping horse* – the first of these being a conventional way of referring to the Ferrari brand mark logo). Some only give the appearance of semi-idiomatic combinations (*jumped-up jackass*, actually an unconventional variant of *jumped-up jack-in-office*; *dumb-fuck donkey*, actually an extremely rare phrase that occurs only once in the documented history of English, namely in the book *Trail of the Octopus: From Beirut to Lockerbie – Inside the DIA* and that probably sounds like an idiom because of the alliteration and the semantic relationship to *silly ass*; and *monocled ass*, which brings to mind *pompous ass* but is actually not a very conventional combination). Finally, there

Table 7.7: Some collocates of the form [ADJ N_{equine}] (BNC)

Word a	Word b	a with b	a without b	b without a	neither
Trojan	horse(s)	37	73	12 198	98 351 475
rocking	horse(s)	34	168	12 201	98 351 380
new	horse(s)	21	113 540	12 214	98 238 008
galloping	horse(s)	17	110	12 218	98 351 438
silly	ass(es)	9	2630	340	98 360 804
prancing	horse(s)	6	17	12 229	98 351 531
pompous	ass(es)	5	250	344	98 363 184
common	zebra(s)	4	18 965	253	98 344 561
old	donkey(s)	3	52 433	643	98 310 704
old	mule(s)	3	52 433	316	98 311 031
young	zebra(s)	2	30 210	255	98 333 316
old	ass(es)	2	52 434	347	98 311 000
female	hinny(/-ies)	2	6620	17	98 357 144
braying	donkey(s)	2	9	644	98 363 128
monocled	ass(es)	1	5	348	98 363 429
large	mule(s)	1	34 228	318	98 329 236
jumped-up	jackass(es)	1	21	7	98 363 754
extinct	quagga(s)	1	428	4	98 363 350
dumb-fuck	donkey(s)	1	0	645	98 363 137
caparisoned	mule(s)	1	8	318	98 363 456

Supplementary Online Material: MNH4

are a number of fully compositional combinations that make sense but do not have any special status (*caparisoned mule*, *new horse*, *old donkey*, *young zebra*, *large mule*, *female hinny*, *extinct quagga*).

In addition, I have selected them to represent different types of frequency relations: some of them are (relatively) frequent, some of them very rare, for some of them the either the adjective or the noun is generally quite frequent, and for some of them neither of the two is frequent.

Table 7.8 shows the ranking of these twenty collocations by the five association measures discussed above. Simplifying somewhat, a good association measure should rank the conventionalized combinations highest (*rocking horse*, *Trojan horse*, *silly ass*, *pompous ass*, *prancing horse*, *braying donkey*, *galloping horse*), the distinctive sounding but non-conventionalized combinations somewhere in the middle (*jumped-up jackass*, *dumb-fuck donkey*, *old ass*, *monocled ass*) and the compositional combinations lowest (*common zebra*, *jumped-up jackass*, *dumb-fuck donkey*, *old ass*, *monocled ass*). *Common zebra* is difficult to predict – it is a conventionalized expression, but not in the general language.

Table 7.8: Comparison of selected association measures for collocates of the form [ADJ N_{equine}] (BNC)

Collocation	χ^2	Collocation	MI	Collocation	MS	Collocation	G	Collocation	Exact test
jumped-up jackass	558 883.30	*jumped-up jackass*	19.09	*jumped-up jackass*	0.045 455	*Trojan horse*	525.06	*Trojan horse*	7.78×10^{-116}
dumb-fuck donkey	152 264.90	*dumb-fuck donkey*	17.22	*pompous ass*	0.014 327	*rocking horse*	428.51	*rocking horse*	6.70×10^{-95}
Trojan horse	99 994.30	*monocled ass*	15.52	*silly ass*	0.003 410	*galloping horse*	205.79	*galloping horse*	2.13×10^{-46}
braying donkey	55 365.79	*extinct quagga*	15.48	*caparisoned mule*	0.003 135	*silly ass*	105.91	*silly ass*	1.34×10^{-24}
monocled ass	46 972.28	*caparisoned mule*	15.06	*braying donkey*	0.003 096	*prancing horse*	81.51	*prancing horse*	3.73×10^{-19}
rocking horse	45 946.30	*braying donkey*	14.76	*Trojan horse*	0.003 024	*pompous ass*	76.35	*pompous ass*	4.72×10^{-18}
extinct quagga	45 855.44	*pompous ass*	12.43	*monocled ass*	0.002 865	*braying donkey*	37.31	*braying donkey*	2.37×10^{-9}
caparisoned mule	34 259.27	*Trojan horse*	11.40	*rocking horse*	0.002 779	*common zebra*	27.29	*common zebra*	2.36×10^{-7}
pompous ass	27 622	*prancing horse*	11.03	*extinct quagga*	0.002 331	*female hinny*	25.64	*female hinny*	7.74×10^{-7}
galloping horse	18 263.01	*female hinny*	10.61	*dumb-fuck donkey*	0.001 548	*jumped-up jackass*	24.64	*jumped-up jackass*	1.79×10^{-6}
prancing horse	12 573.20	*rocking horse*	10.40	*galloping horse*	0.001 389	*dumb-fuck donkey*	23.87	*dumb-fuck donkey*	6.57×10^{-6}
silly ass	8633.06	*galloping horse*	10.07	*prancing horse*	0.000 490	*monocled ass*	19.69	*monocled ass*	2.13×10^{-5}
female hinny	3123.39	*silly ass*	9.91	*female hinny*	0.000 302	*extinct quagga*	19.68	*extinct quagga*	2.18×10^{-5}
common zebra	314.94	*common zebra*	6.33	*common zebra*	0.000 211	*caparisoned mule*	19.00	*caparisoned mule*	2.92×10^{-5}
old mule	47.12	*young zebra*	4.66	*new horse*	0.000 185	*old mule*	11.59	*old mule*	7.16×10^{-4}
young zebra	46.77	*old mule*	4.14	*young zebra*	0.000 066	*young zebra*	9.10	*young zebra*	2.95×10^{-3}
old donkey	20.49	*old ass*	3.43	*old donkey*	0.000 057	*old donkey*	7.69	*old donkey*	5.25×10^{-3}
old ass	17.70	*large mule*	3.17	*old mule*	0.000 057	*old ass*	5.88	*old ass*	1.53×10^{-2}
large mule	7.12	*old donkey*	3.12	*old ass*	0.000 038	*new horse*	2.91	*new horse*	5.15×10^{-2}
new horse	3.35	*new horse*	0.57	*large mule*	0.000 029	*large mule*	2.62	*large mule*	1.05×10^{-1}

All association measures fare quite well, generally speaking, with respect to the compositional expressions – these tend to occur in the lower third of all lists. Where there are exceptions, the χ^2 statistic, mutual information and minimum sensitivity rank rare cases higher than they should (e.g. *caparisoned mule*, *extinct quagga*), while the G and the p-value of Fisher's exact test rank frequent cases higher (e.g.*galloping horse*).

With respect to the non-compositional cases, χ^2 and mutual information are quite bad, overestimating rare combinations like *jumped-up jackass*, *dumb-fuck donkey* and *monocled ass*, while listing some of the clear cases of collocations much further down the list (*silly ass*, and, in the case of MI, *rocking horse*). Minimum sensitivity is much better, ranking most of the conventionalized cases in the top half of the list and the non-conventionalized ones further down (with the exception of *jumped-up jackass*, where both the individual words and their combination are very rare). The G and the Fisher p-value fare best (with no differences in their ranking of the expressions), listing the conventionalized cases at the top and the distinctive but non-conventionalized cases in the middle.

To demonstrate the problems that very rare events can cause (especially those where both the combination and each of the two words in isolation are very rare), imagine someone had used the phrase *tomfool onager* once in the BNC. Since neither the adjective *tomfool* (a synonym of *silly*) nor the noun *onager* (the name of the donkey sub-genus Equus hemionus, also known as *Asiatic* or *Asian wild ass*) occur in the BNC anywhere else, this would give us the distribution in Table 7.9.

Table 7.9: Fictive occurrence of *tomfool onager* in the BNC

		Second Position		
		onager	¬onager	Total
First Position	tomfool	1	0	1
		(0.00)	(1.00)	
	¬tomfool	0	98 363 782	98 363 782
		(1.00)	(98 363 781.00)	
	Total	1	98 363 782	98 363 783

Applying the formulas discussed above to this table gives us a χ^2 value of 98 364 000, an MI value of 26.55 and a minimum sensitivity value of 1, placing this (hypothetical) one-off combination at the top of the respective rankings by a

wide margin. Again, the log-likelihood ratio test and Fisher's exact test are much better, putting in eighth place on both lists ($G = 36.81$, $p_{\text{exact}} = 1.02 \times 10^{-8}$).

Although the example is hypothetical, the problem is not. It uncovers a mathematical weakness of many commonly used association measures. From an empirical perspective, this would not necessarily be a problem, if cases like that in Table 7.9 were rare in linguistic corpora. However, they are not. The LOB corpus, for example, contains almost one thousand such cases, including some legitimate collocation candidates (like *herbal brews*, *casus belli* or *sub-tropical climates*), but mostly compositional combinations (*ungraceful typography*, *turbaned headdress*, *songs-of-Britain medley*), snippets of foreign languages (*freie Blicke*, *l'arbre rouge*, *palomita blanca*) and other things that are quite clearly not what we are looking for in collocation research. All of these will occur at the top of any collocate list created using statistics like χ^2, mutual information and minimum sensitivity. In large corpora, which are impossible to check for orthographical errors and/or errors introduced by tokenization, this list will also include hundreds of such errors (whose frequency of occurrence is low precisely because they are errors).

To sum up, when doing collocational research, we should use the best association measures available. For the time being, this is the p value of Fisher's exact test (if we have the means to calculate it), or G (if we don't, or if we prefer using a widely-accepted association measure). We will use G through much of the remainder of this book whenever dealing with collocations or collocation-like phenomena.

7.2 Case studies

In the following, we will look at some typical examples of collocation research, i.e. cases where both variables consist of (some part of) the lexicon and the values are individual words.

7.2.1 Collocation for its own sake

Research that is concerned exclusively with the collocates of individual words or the extraction of all collocations from a corpus falls into three broad types. First, there is a large body of research on the explorative extraction of collocations from corpora. This research is not usually interested in any particular collocation (or set of collocations), or in genuinely linguistic research questions; instead, the focus is on methods (ways of preprocessing corpora, which association measures to use, etc.). Second, there is an equally large body of applied research that results

in lexical resources (dictionaries, teaching materials, etc.) rather than scientific studies on specific research questions. Third, there is a much smaller body of research that simply investigates the collocates of individual words or small sets of words. The perspective of these studies tends to be descriptive, often with the aim of showing the usefulness of collocation research for some application area.

The (relative) absence of theoretically more ambitious studies of the collocates of individual words may partly be due to the fact that words tend to be too idiosyncratic in their behavior to make their study theoretically attractive. However, this idiosyncrasy itself is, of course, theoretically interesting and so such studies hold an unrealized potential at least for areas like lexical semantics.

7.2.1.1 Case study: Degree adverbs

A typical example of a thorough descriptive study of the collocates of individual words is Kennedy (2003), which investigates the adjectival collocates of degree adverbs like *very*, *considerably*, *absolutely*, *heavily* and *terribly*. Noting that some of these adverbs appear to be relatively interchangeable with respect to the adjectives and verbs they modify, others are highly idiosyncratic, Kennedy identifies the adjectival and verbal collocates of 24 frequent degree adverbs in the BNC, extracting all words occurring in a span of two words to their left or right, and using mutual information to determine which of them are associated with each degree adverb.

Thus, as is typical for this kind of study, Kennedy adopts an exploratory perspective. The study involves two nominal variables: first, Degree Adverb, with 24 values corresponding to the 24 specific adverbs he selects; second, Adjective, with as many different potential values as there are different adjectives in the BNC (in exploratory studies, it is often the case that we do not know the values of at least one of the two variables in advance, but have to extract them from the data). As pointed out above, which of the two variables is the dependent one and which the independent one in studies like this depends on your research question: if you are interested in degree adverbs and want to explore which adjectives they co-occur with, it makes sense to treat Degree Adverb as the independent and Adjective as the dependent variable; if you are interested in adjectives and want to expore which degree adverbs they co-occur with, it makes sense to do it the other way around. Statistically, it does not make a difference, since our statistical tests for nominal data do not distinguish between dependent and independent variables.

Kennedy finds, first, that there are some degree adverbs that do not appear to have restrictions concerning the adjectives they occur with (for example, *very*, *re-*

ally and *particularly*). However, most degree adverbs are clearly associated with semantically restricted sets of adjectives. The restrictions are of three broad types. First, there are connotational restrictions (some adverbs are associated primarily with positive words (e.g. *perfectly*) or negative words (e.g. *utterly*, *totally*; on connotation cf. also Section 7.2.3). Second, there are specific semantic restrictions (for example, *incredibly*, which is associated with subjective judgments), sometimes relating transparently to the meaning of the adverb (for example, *badly*, which is associated with words denoting damage or *clearly*, which is associated with words denoting sensory perception). Finally, there are morphological restrictions (some adverbs are used frequently with words derived by particular suffixes, for example, *perfectly*, which is frequently found with words derived by *-able*/*-ible*, or *totally*, whose collocates often contain the prefix *un-*). Table 7.10 illustrates these findings for 5 of the 24 degree adverbs and their top 15 collocates.

Unlike Kennedy, I have used the G statistic of the log-likelihood ratio test,[6] and so the specific collocates differ from the ones he finds (generally, his lists include more low-frequency combinations, as expected given that he uses mutual information), but his observations concerning the semantic and morphological sets are generally confirmed.

This case study illustrates the exploratory design typical of collocational research as well as the kind of result that such studies yield and the observations possible on the basis of these results. By comparing the results reported here to Kennedy's, you may also gain a better understanding as to how different association measures may lead to different results.

7.2.2 Lexical relations

One area of lexical semantics where collocation data is used quite intensively is the study of lexical relations – most notably, (near) synonymy (Taylor 2003, cf. below), but also polysemy (e.g. Yarowsky 1993, investigating the idea that associations exist not between words but between particular senses of words) and antonymy (Justeson & Katz 1991, see below).

7.2.2.1 Case study: Near synonyms

Natural languages typically contain pairs (or larger sets) of words with very similar meanings, such as *big* and *large*, *begin* and *start* or *high* and *tall*. In isolation,

[6] Note that I will usually provide the frequencies for the cells O_{11}, O_{12}, O_{21} and O_{22} in tables like this, to allow you to check the calculations or to try out different association measures, but in this case lack of space prevents this. The complete dataset is part of the Supplementary Online Material, however).

Table 7.10: Selected degree adverbs and their collocates

INCREDIBLY		PERFECTLY		TOTALLY		COMPLETELY		BADLY	
Collocation	G	Collocation	G	Collocation	G	Collocation	G	Collocation	G
difficult	113.87	*normal*	989.49	*different*	3190.97	*different*	3965.12	*bruised*	573.27
lucky	95.58	*acceptable*	928.21	*dependent*	718.13	*new*	1242.61	*wrong*	410.49
fast	87.97	*clear*	880.93	*unacceptable*	706.93	*free*	404.43	*damaged*	217.62
beautiful	80.56	*happy*	822.98	*inadequate*	604.42	*wrong*	362.33	*beaten*	188.40
dangerous	77.75	*possible*	743.65	*wrong*	478.49	*unaware*	240.24	*hurt*	170.65
strong	68.13	*reasonable*	674.85	*unexpected*	459.52	*mad*	218.55	*injured*	141.80
stupid	65.32	*capable*	663.38	*unsuitable*	420.13	*refurbished*	184.58	*wounded*	138.48
efficient	61.84	*good*	545.89	*unaware*	345.90	*irrelevant*	178.25	*swollen*	113.92
simple	61.84	*adequate*	537.98	*opposed*	333.21	*separate*	172.60	*fitting*	106.13
low	59.14	*safe*	512.69	*new*	316.11	*independent*	149.81	*affected*	93.24
sexy	59.12	*natural*	469.62	*unnecessary*	303.03	*satisfied*	142.60	*broken*	71.49
naive	57.34	*competitive*	418.44	*irrelevant*	251.56	*innocent*	141.11	*mutilated*	70.78
expensive	56.66	*honest*	406.00	*alien*	251.06	*empty*	138.70	*burned*	62.07
hard	55.00	*balanced*	388.13	*confused*	249.15	*unknown*	137.75	*unstuck*	61.30
complicated	54.98	*well*	370.56	*blind*	247.58	*dry*	136.61	*lacerated*	58.56

Supplementary Online Material: LKTH

it is often difficult to tell what the difference in meaning is, especially since they are often interchangeable at least in some contexts. Obviously, the distribution of such pairs or sets with respect to other words in a corpus can provide insights into their similarities and differences.

One example of such a study is Taylor (2003), which investigates the synonym pair *high* and *tall* by identifying all instances of the two words in their subsense 'large vertical extent' in the LOB corpus and categorizing the words they modify into eleven semantic categories. These categories are based on semantic distinctions such as human vs. inanimate, buildings vs. other artifacts vs. natural entities, etc., which are expected *a priori* to play a role.

The study, while not strictly hypothesis-testing, is thus somewhat deductive. It involves two nominal variables; the independent variable Type of Entity with eleven values shown in Table 7.11 and the dependent variable Vertical Extent Adjective with the values high and tall (assuming that people first choose something to talk about and then choose the appropriate adjective to describe it). Table 7.11 shows Taylor's results (he reports absolute and relative frequencies, which I have used to calculate expected frequencies and χ^2 components).

As we can see, there is little we can learn from this table, since the frequencies in the individual cells are simply too small to apply the χ^2 test to the table as a whole. The only χ^2 components that reach significance individually are those for the category human, which show that *tall* is preferred and *high* avoided with human referents. The sparsity of the data in the table is due to the fact that the analyzed sample is very small, and this problem is exacerbated by the fact that the little data available is spread across too many categories. The category labels are not well chosen either: they overlap substantially in several places (e.g., towers and walls are buildings, pieces of clothing are artifacts, etc.) and not all of them seem relevant to any expectation we might have about the words *high* and *tall*.

Taylor later cites earlier psycholinguistic research indicating that *tall* is used when the vertical dimension is prominent, is an acquired property and is a property of an individuated entity. It would thus have been better to categorize the corpus data according to these properties – in other words, a more strictly deductive approach would have been more promising given the small data set.

Alternatively, we can take a truly exploratory approach and look for differential collocates as described in Section 7.1.1 above – in this case, for differential noun collocates of the adjectives *high* and *tall*. This allows us to base our analysis on a much larger data set, as the nouns do not have to be categorized in advance.

Table 7.12 shows the top 15 differential collocates of the two words in the BNC.

Table 7.11: Objects described as *tall* or *high* in the LOB corpus (adapted from Taylor 2003)

	Adjective				
Noun Category	tall		high		Total
humans	*Obs.:*	45	*Obs.:*	2	47
	Exp.:	22.91	*Exp.:*	24.09	
	χ^2*:*	21.31	χ^2*:*	20.26	
animals	*Obs.:*	0	*Obs.:*	1	1
	Exp.:	0.49	*Exp.:*	0.51	
	χ^2*:*	0.49	χ^2*:*	0.46	
plants, trees	*Obs.:*	7	*Obs.:*	3	10
	Exp.:	4.87	*Exp.:*	5.13	
	χ^2*:*	0.93	χ^2*:*	0.88	
buildings	*Obs.:*	3	*Obs.:*	10	13
	Exp.:	6.34	*Exp.:*	6.66	
	χ^2*:*	1.76	χ^2*:*	1.67	
walls, fences, etc	*Obs.:*	0	*Obs.:*	5	5
	Exp.:	2.44	*Exp.:*	2.56	
	χ^2*:*	2.44	χ^2*:*	2.32	
towers, statues, pillars, sticks	*Obs.:*	0	*Obs.:*	7	7
	Exp.:	3.41	*Exp.:*	3.59	
	χ^2*:*	3.41	χ^2*:*	3.24	
articles of clothing	*Obs.:*	0	*Obs.:*	7	7
	Exp.:	3.41	*Exp.:*	3.59	
	χ^2*:*	3.41	χ^2*:*	3.24	
miscellaneous artifacts	*Obs.:*	2	*Obs.:*	13	15
	Exp.:	7.31	*Exp.:*	7.69	
	χ^2*:*	3.86	χ^2*:*	3.67	
topographical features	*Obs.:*	0	*Obs.:*	5	5
	Exp.:	2.44	*Exp.:*	2.56	
	χ^2*:*	2.44	χ^2*:*	2.32	
other natural phenomena	*Obs.:*	0	*Obs.:*	5	5
	Exp.:	2.44	*Exp.:*	2.56	
	χ^2*:*	2.44	χ^2*:*	2.32	
uncertain reference	*Obs.:*	1	*Obs.:*	3	4
	Exp.:	1.95	*Exp.:*	2.05	
	χ^2*:*	0.46	χ^2*:*	0.44	
Total	58		61		119

Table 7.12: Differential collocates for *tall* and *high* in the BNC

Collocate	Collocate with TALL	Collocate with HIGH	Other words with TALL	Other words with HIGH	G
Most strongly associated with *high*					
level	0	2741	1720	36 933	240.90
education	0	2499	1720	37 175	218.94
court	0	1863	1720	37 811	161.88
quality	0	1079	1720	38 595	92.83
standard	1	1163	1719	38 511	90.35
rate	0	922	1720	38 752	79.16
proportion	0	875	1720	38 799	75.08
street	1	810	1719	38 864	60.38
school	0	676	1720	38 998	57.86
price	0	642	1720	39 032	54.93
degree	0	638	1720	39 036	54.58
speed	0	547	1720	39 127	46.75
interest	0	493	1720	39 181	42.10
risk	0	431	1720	39 243	36.78
cost	0	387	1720	39 287	33.01
priority	0	374	1720	39 300	31.89
point	0	352	1720	39 322	30.01
unemployment	0	318	1720	39 356	27.10
temperature	0	305	1720	39 369	25.99
Most strongly associated with *tall*					
man	182	3	1538	39 671	1146.54
building	82	26	1638	39 648	408.35
tree	73	26	1647	39 648	355.52
boy	40	0	1680	39 674	255.36
glass	39	2	1681	39 672	233.14
woman	38	3	1682	39 671	221.34
ship	33	0	1687	39 674	210.54
girl	32	0	1688	39 674	204.15
figure	62	93	1658	39 581	195.58
chimney	28	8	1692	39 666	141.09
order	62	176	1658	39 498	138.01
dark	23	3	1697	39 671	128.27
grass	24	5	1696	39 669	126.76
tale	20	1	1700	39 673	119.50
window	34	41	1686	39 633	117.04
story	18	0	1702	39 674	114.69
tower	24	24	1696	39 650	88.47
plant	24	28	1696	39 646	83.57
person	13	0	1707	39 674	82.80
nave	9	0	1711	39 674	57.30

Supplementary Online Material: D9P4

The results for *tall* clearly support Taylor's ideas about the salience of the vertical dimension. The results for *high* show something Taylor could not have found, since he restricted his analysis to the subsense 'vertical dimension': when compared with tall, *high* is most strongly associated with quantities or positions in hierarchies and rankings. There are no spatial uses at all among its top differential collocates. This does not answer the question why we can use it spatially and in competition with *tall*, but it shows what general sense we would have to assume: one concerned not with the vertical extent as such, but with the magnitude of that extent (which, incidentally, Taylor notes in his conclusion).

This case study shows how the same question can be approached by a deductive or an inductive (exploratory) approach. The deductive approach can be more precise, but this depends on the appropriateness of the categories chosen *a priori* for annotating the data; it is also time consuming and therefore limited to relatively small data sets. In contrast, the inductive approach can be applied to a large data set because it requires no *a priori* annotation. It also does not require any choices concerning annotation categories; however, there may be a danger to project patterns into the data *post hoc*.

7.2.2.2 Case study: Antonymy

At first glance, we expect the relationship between antonyms to be a paradigmatic one, where only one or the other will occur in a given utterance. However, Charles & Miller (1989) suggest, based on the results of sorting tasks and on theoretical considerations, that, on the contrary, antonym pairs are frequently found in syntagmatic relationships, occurring together in the same clause or sentence. A number of corpus-linguistic studies have shown this to be the case (e.g. Justeson & Katz 1991, Justeson & Katz 1992, Fellbaum 1995; cf. also Gries & Otani 2010 for a study identifying antonym pairs based on their similarity in lexico-syntactic behavior).

There are differences in detail in these studies, but broadly speaking, they take a deductive approach: they choose a set of test words for which there is agreement as to what their antonyms are, search for these words in a corpus, and check whether their antonyms occur in the same sentence significantly more frequently than expected. The studies thus involve two nominal variables: Sentence (with the values contains test word and does not contain test word) and Antonym of Test Word (with the values occurs in sentence and does not occur in sentence). This seems like an unnecessarily complicated way of representing the kind of co-occurrence design used in the examples above, but I have chosen it to show that in this case sentences containing a particular word are used as the condition under which the occurrence of another word is

investigated – a straightforward application of the general research design that defines quantitative corpus linguistics. Table 7.13 demonstrates the design using the adjectives *good* and *bad* (the numbers are, as always in this book, based on the tagged version of BROWN included with the ICAME collection and differ slightly from the ones reported in the studies discussed below).

Table 7.13: Sentential co-occurrence of *good* and *bad* in the BROWN corpus

		Bad		
		occurs	¬occurs	Total
Good	occurs	16	687	703
		(1.57)	(701.43)	
	¬occurs	110	55 769	55 879
		(124.43)	(55 754.57)	
	Total	126	56 456	56 582

Good occurs significantly more frequently in sentences also containing *bad* than in sentences not containing *bad*, and vice versa ($\chi^2 = 135.07, \text{df} = 1, p < 0.001$). Justeson & Katz (1991) apply this procedure to 36 adjectives and get significant results for 25 of them (19 of which remain significant after a Bonferroni correction for multiple tests). They also report that in a larger corpus, the frequency of co-occurrence for all adjective pairs is significantly higher than expected (but do not give any figures). Fellbaum (1995) uses a very similar procedure with words from other word classes, with very similar results.

These studies only look at the co-occurrence of antonyms; they do not apply the same method to word pairs related by other lexical relations (synonymy, taxonomy, etc.). Thus, there is no way of telling whether co-occurrence within the same sentence is something that is typical specifically of antonyms, or whether it is something that characterizes word pairs in other lexical relations, too.

An obvious approach to testing this would be to repeat the study with other types of lexical relations. Alternatively, we can take an exploratory approach that does not start out from specific word pairs at all. Justeson & Katz (1991) investigate the specific grammatical contexts which antonyms tend to co-occur, identifying, among others, coordination of the type [ADJ *and* ADJ] or [ADJ *or* ADJ]. We can use such specific contexts to determine the role of co-occurrence for different types of lexical relations by simply extracting *all* word pairs occurring in the adjective slots of these patterns, calculating their association strength

within this pattern as shown in Table 7.14 for the adjectives *good* and *bad* in the BNC, and then categorizing the most strongly associated collocates in terms of the lexical relationships between them.

Table 7.14: Co-occurrence of *good* and *bad* in the first and second slot of [ADJ_1 *and* ADJ_2]

		Second Slot		
		BAD	¬BAD	Total
First Slot	GOOD	158	476	634
		(0.89)	(633.11)	
	¬GOOD	35	136 893	136 928
		(192.11)	(136 735.89)	
	Total	193	137 369	137 562

Note that this is a slightly different procedure from what we have seen before: instead of comparing the frequency of co-occurrence of two words with their individual occurrence in the rest of the corpus, we are comparing it to their individual occurrence *in a given position of a given structure* – in this case [ADJ *and* ADJ] (Stefanowitsch & Gries (2005) call this kind of design *covarying collexeme analysis*).

Table 7.15 shows the thirty most strongly associated adjective pairs coordinated with *and* in the BNC.

Clearly, antonymy is the dominant relation among these word pairs, which are mostly opposites (*black*/*white*, *male*/*female*, *public*/*private*, etc.), and sometimes relational antonyms (*primary*/*secondary*, *economic*/*social*, *economic*/*political*, *social*/*political*, *lesbian*/*gay*, etc.). The only cases of non-antonymic pairs are *economic*/*monetary*, which is more like a synonym than an antonym and the fixed expressions *deaf*/*dumb* and *hon(ourable)*/*learned* (as in *honourable and learned gentleman/member/friend*). The pattern does not just hold for the top 30 collocates but continues as we go down the list. There are additional cases of relational antonyms, like *British*/*American* and *Czech*/*Slovak* and additional examples of fixed expressions (*alive and well*, *far and wide*, *true and fair*, *null and void*, *noble and learned*), but most cases are clear antonyms (for example, *syntactic*/*semantic*, *spoken*/*written*, *mental*/*physical*, *right*/*left*, *rich*/*poor*, *young*/*old*, *good*/*evil*, etc.). The one systematic exceptions are cases like *worse and worse* (a special construction with comparatives indicating incremental change, cf. Stefanowitsch 2007b).

Table 7.15: Co-occurrence of adjectives in the first and second slot of [ADJ_1 *and* ADJ_2] (BNC)

ADJ_1 AND ADJ_2	ADJ_1 with ADJ_2	ADJ_1 with ADJ_{other}	ADJ_{other} with ADJ_2	ADJ_{other} with ADJ_{other}	*G*
black and white	959	507	667	135 429	7348.90
economic and social	1049	1285	1286	133 942	5920.16
male and female	414	25	26	137 097	5244.75
social and economic	755	1705	862	134 240	4119.00
public and private	369	135	158	136 900	3877.60
deaf and dumb	276	43	8	137 235	3655.01
primary and secondary	262	58	25	137 217	3332.90
lesbian and gay	183	6	22	137 351	2596.57
internal and external	191	28	20	137 323	2595.41
hon. and learned	232	91	118	137 121	2594.96
political and economic	466	1166	1151	134 779	2356.74
social and political	502	1958	1139	133 963	2160.29
national and international	251	443	243	136 625	2075.94
left and right	149	37	33	137 343	1974.66
upper and lower	156	30	105	137 271	1911.70
old and new	214	462	164	136 722	1834.78
economic and monetary	266	2068	89	135 139	1802.61
physical and mental	186	467	54	136 855	1793.37
top and bottom	123	26	6	137 407	1786.23
economic and political	420	1914	1221	134 007	1671.32
local and national	186	309	180	136 887	1667.41
positive and negative	147	179	43	137 193	1653.32
good and bad	158	476	35	136 893	1560.46
private and public	161	236	160	137 005	1514.90
industrial and commercial	174	277	236	136 875	1510.40
past and present	114	60	23	137 365	1497.56
formal and informal	131	116	65	137 250	1494.07
alive and well	111	78	20	137 353	1434.91
central and eastern	155	380	97	136 930	1434.86
present and future	130	95	124	137 213	1412.34

Supplementary Online Material: CWPX

This case study shows how deductive and inductive domains may complement each other: while the deductive studies cited show that antonyms tend to co-occur syntagmatically, the inductive study presented here shows that words that co-occur syntagmatically (at least in certain syntactic contexts) tend to be antonyms. These two findings are not equivalent; the second finding shows that the first finding may indeed be typical for antonymy as opposed to other lexical relations.

The exploratory study was limited to a particular syntactic/semantic context, chosen because it seems semantically and pragmatically neutral enough to allow all kinds of lexical relations to occur in it. There are contexts which might be expected to be particularly suitable to particular kinds of lexical relations and which could be used, given a large enough corpus, to identify word pairs in such relations. For example, the pattern [ADJ *rather than* ADJ] seems semantically predisposed for identifying antonyms, and indeed, it yields pairs like *implicit*/*explicit*, *worse*/*better*, *negative*/*positive*, *qualitative*/*quantitative*, *active*/*passive*, *real*/*apparent*, *local*/*national*, *political*/*economical*, etc. Other patterns are semantically more complex, identifying pairs in more context-dependent oppositions; for example, [ADJ *but not* ADJ] identifies pairs like *desirable*/*essential*, *necessary*/*sufficient*, *similar*/*identical*, *small*/*insignificant*, *useful*/*essential*, *difficult*/*impossible*. The relation between the adjectives in these pairs is best described as pragmatic – the first one conventionally implies the second.

7.2.3 Semantic prosody

Sometimes, the collocates of a node word (or larger expressions) fall into a more or less clearly recognizable semantic class that is difficult to characterize in terms of denotational properties of the node word. Louw (1993: 157) refers to this phenomenon as "semantic prosody", defined, somewhat impressionistically, as the "consistent aura of meaning with which a form is imbued by its collocates".

This definition has been understood by collocation researchers in two different (but related) ways. Much of the subsequent research on semantic prosody is based on the understanding that this "aura" consists of connotational meaning (cf. e.g. Partington 1998: 68), so that words can have a "positive", "neutral" or "negative" semantic prosody. However, Sinclair, who according to Louw invented the term,[7] seems to have in mind "attitudinal or pragmatic" meanings that are much more specific than "positive", "neutral" or "negative". There are insightful terminological discussions concerning this issue (cf. e.g. Hunston 2007), but since

[7] Louw attributes the term to John Sinclair, but Louw (1993) is the earliest appearance of the term in writing. However, Sinclair is clearly the first to discuss the phenomenon itself systematically, without giving it a label (e.g. Sinclair 1991: 74–75).

the term is widely-used in (at least) these two different ways, and since "positive" and "negative" connotations are very general kinds of attitudinal meaning, it seems more realistic to accept a certain vagueness of the term. If necessary, we could differentiate between the general semantic prosody of a word (its "positive" or "negative" connotation as reflected in its collocates) and its specific semantic prosody (the word-specific attitudinal meaning reflected in its collocates).

7.2.3.1 Case study: True feelings

A typical example of Sinclair's approach to semantic prosody, both methodologically and theoretically, is his short case study of the expression *true feelings*. Sinclair (1996b) presents a selection of concordance lines from the COBUILD corpus – Figure 7.1 shows a random sample from the BNC instead, as the COBUILD corpus is not accessible, but Sinclair's findings are well replicated by this sample.

```
 1 f unless you 're absolutely sure of your [true feelings] . I had a similar experience several ye
 2 nces may well not reflect my employer 's [true feelings] on the matter , but once having sustain
 3  and realize it is all right to show our [true feelings] and that it is all right to be rejected
 4 wing right action : acting only from our [true feelings] , not governed by the distortions of em
 5 der , but the problem of ` reading ' the [true feelings] of the individual can be made easier by
 6  other . Having declared to Roderigo his [true feelings] about Othello , Iago later explains why
 7 ell studied in the art of disguising his [true feelings] . Let him not be frightened of me ; let
 8 rised that the TV presenter revealed her [true feelings] towards Nicola so quickly : most people
 9 embers are helpful to show each side the [true feelings] of the other , the need to accept and w
10 good husband , but you like to hide your [true feelings] . ' ` Oh , do n't be so serious , B
11  er , he has n't actually dealt with the [true feelings] that he had towards his father , and wh
12 g as ` friends ' , without revealing her [true feelings] for him . It was still light when he pi
13 t the parents will often not admit their [true feelings] about the child and the incident , acti
14 t a matter of time before she showed her [true feelings] , I was sure of that . Females -- hone
15 m for so long at last gave vent to their [true feelings] . The match had been billed in the Amer
16 eople . And got him plenty sex . Rory 's [true feelings] about the matter were complex but red-b
17 t had finally forced her to confront her [true feelings] for Arnie . Or rather , her lack of fee
18 rage in both hands , and told him of her [true feelings] , they might have had a chance to work
19 andmother finds it difficult to show her [true feelings] . ' said David . ` I think it 's a
20 er heart did more to convince her of her [true feelings] than any rational thinking . She wanted
```

Figure 7.1: Concordance of *true feelings* (BNC, Sample)

On the basis of his concordance, Sinclair then makes a number of observations concerning the use of the phrase *true feelings*, quantifying them informally. He notes three things: first, the phrase is almost always part of a possessive (realized by pronoun, possessive noun phrase or *of*-construction). This is also true of the sample in Figure 7.1, with the exception of line 11 (where there is a possessive relation, but it is realized by the verb *have*).

Second, the expression collocates with verbs of expression (perhaps unsurprising for an expression relating to emotions); this, too, is true for our sample, where such verbs are found in 14 lines: *reflect* (line 2), *show* (lines 3, 9, 14, and 19), *read* (line 5), *declare* (line 6), *disguise* (line 7), *reveal* (line 8), *hide* (line 10), *reveal* (line 12), *admit* (line 13), *give vent to* (line 15), and *tell* (line 18).

Third, and most interesting, Sinclair finds that a majority of his examples express a *reluctance* to express emotions. In our sample, such cases are also noticeably frequent: I would argue that lines 2, 3, 5, 7, 8, 10, 12, 13, 14, 15, and 19 can be interpreted in this way, which would give us a slight majority of 11/20. (Your analysis may differ, as I have made my assessment rather intuitively, instead of coming up with an annotation scheme). In many cases, the reluctance or inability is communicated as part of the verb (like *disguise*, *conceal* and *hide*), in other cases it is communicated by negation of a verb of expression (like *not admit* in line 13) or by adjectives (like *difficult to show* in line 19).

Sinclair assumes that the denotational meaning of the phrase *true feelings* is "genuine emotions". Based on his observations, he posits that, in addition, it has the semantic prosody "reluctance/inability to express emotions" – an attitudinal meaning much more specific than a general "positive" or "negative" connotation.

The methodological approach taken by Sinclair (and many others in his tradition) can yield interesting observations (at least, if applied very carefully): descriptively, there is little to criticize. However, under the definition of corpus linguistics adopted in this book, Sinclair's observations would be just the first step towards a full analysis. First, note that Sinclair's approach is quantitative only in a very informal sense – he rarely reports exact frequencies for a given semantic feature in his sample, relying instead on general statements about the frequency or rarity of particular phenomena. As we saw above, this is easy to remedy by simply determining the exact number of times that the phenomenon in question occurs in a given sample. However, such exact frequencies do not advance the analysis meaningfully: as long as we do not know how frequent a particular phenomenon is in the corpus as a whole, we cannot determine whether it is a characteristic property of the expression under investigation, or just an accidental one.

Specifically, as long as we do not know how frequent the semantic prosody 'reluctance or inability to express' is in general, we do not know whether it is particularly characteristic of the phrase *true feelings*. It may be characteristic, among other things, (a) of utterances concerning emotions in general, (b) of utterances containing the plural noun *feelings*, (c) of utterances containing the adjective *true*, etc.

In order to determine this, we have to compare our sample of the expression *true feelings* to related expressions that differ with respect to each property potentially responsible for the semantic prosody. For example, we might compare it to the noun *feelings* in order to investigate possibility (b). Figure 7.2 shows a sample of the expression [POSS *feelings*] (the possessive pronoun was included as it, too, may have an influence on the prosody and almost all examples of *true feelings* are preceded by a possessive pronoun).

```
 1  by the rest of the board ? Re-programme [your feelings] , in that case . The annual BW accounts
 2 the Asian women I spoke to told me about [their feelings] and situations . Here I shall try to d
 3 ractive , but I think you might consider [my feelings] as well as your own. , Another pause .
 4 o trust her more , dared to feel more of [my feelings] , instead of eating them away . It woul
 5 all was in order . It is hard to explain [my feelings] once I did finally set off . For the fi
 6 e family and the old person work through [their feelings] about any restrictions . This contract
 7  say . ` Nothing is ever going to change [their feelings] towards me . ` I 've tried everything
 8 han rights . It is about men reconciling [their feelings] towards their fathers and learning how
 9 l family . It is as if to let people see [your feelings] takes away some of your power . But at
10  eyelids defensively lowered to disguise [her feelings] . Crossing her legs discreetly , she du
11 nxiety ? Should n't she just accept that [her feelings] about her mother 's lifestyle were irra
12 o stop things before they went too far . [His feelings] had gone no deeper than the surface . N
13 resentment , because you do n't care for [my feelings] at all . You always think the worst of
14 etence , could n't face having to stifle [her feelings] , her crazy and immature hopes -- hope
15 Remember ? ' ` I thought I could control [my feelings] , have an exciting affair with you and
16  her and kissing her softly , she voiced [her feelings] by saying , ` I love you , Gran . '
17 our lack of understanding with regard to [his feelings] as a father . ' ` Oh , Great-gran ,
18  right , then , the doubts you had about [your feelings] . ' ` You mean my feelings towards
19 y North-West 's Billy Anderson who vents [his feelings] about the lack of North-West representa
20  that is by giving them a copy . That 's [my feelings] erm . I move . Thanks very much indeed
```

Figure 7.2: Concordance of [POSS *feelings*] (BNC, Sample)

The concordance shows that contexts concerning a reluctance or inability to express emotions are not untypical of the expression [POSS *feelings*] – it is found in four out of twenty lines in our sample, i.e. in 20 percent of all cases (lines 5, 10, 14, 15). However, it is nowhere near as frequent as with the expression *true feelings*. We can compare the two samples using the χ^2 test. As Table 7.16 shows, the difference is, indeed, significant ($\chi^2 = 5.23, \mathrm{df} = 1, p < 0.05$).

The semantic prosody is not characteristic of the noun *feelings*, even in possessive contexts. We can thus assume that it is not characteristic of utterances concerned with emotions generally. But is it characteristic of the specific expression *true feelings*, or would we find it in other contexts where a distinction between genuine and non-genuine emotions is made?

Table 7.16: Semantic prosody of *true feelings* and [POSS *feelings*]

		Prosody		
		RELUCTANCE	¬RELUCTANCE	Total
Expression	TRUE FEELINGS	11	9	20
		(7.50)	(12.50)	
	[POSS *FEELINGS*]	4	16	20
		(7.50)	(12.50)	
	Total	15	25	40

In order to answer this question, we have to compare the phrase to denotationally synonymous expressions, such as *genuine emotions* (which Sinclair uses to paraphrase the denotational meaning), *genuine feelings*, *real emotions* and *real feelings*. The only one of these expressions that occurs in the BNC more than a handful of times is *real feelings*. A sample concordance is shown in Figure 7.3.

```
 1 r-head wolf-whistles . Real situations , [real feelings] , real people , real love . The album s
 2 onal Checklist : I do my best to hide my [real feelings] from others I always try to please othe
 3  , how to manipulate , how to hide their [real feelings] and how to convince those that love the
 4 f the death of a cousin . Disguising his [real feelins] he wrote cheerfully , telling them that
 5 her words , the counsellor must seek the [real feelings] of the counsellee through careful liste
 6 tant issues are fully discussed and that [real feelings] are expressed rather than avoided . An
 7 at prevented him from ever revealing his [real feelings] to any woman . How she regretted those
 8 ing process of mystification that denies [real feelings] and experiences is a necessary prop to
 9  the play to whom he reveals some of his [real feelings] is Roderigo , but only while using him
10 sked her much sooner if he had known her [real feelings] towards him , but she had been so forma
11  of situation neither can say what their [real feelings] are . A true conversation might be ,
12  clerks are not allowed to express their [real feelings] at work , it is not surprising that the
13 k foolish in public in order to hide his [real feelings] . Men were strange creatures at times .
14 t she could smother the awakening of her [real feelings] for him ? He 'd been important enough t
15 but she hoped she managed to conceal her [real feelings] . Guessing what might greet her in the
16 ight of their honeymoon ? If Ace had any [real feelings] for her he would have taken her prohibi
17  used deliberately as a mask to hide his [real feelings] , she could only guess . ` Let me tak
18 had left him -- but his control over his [real feelings] had remained even then . But what had c
19 ' Relieved that she had not betrayed her [real feelings] , Sophie concentrated on the morning su
20 der has an insight into the Mr. Darcy 's [real feelings] during particular parts of the book . E
```

Figure 7.3: Concordance of *real feelings* (BNC, Sample)

Here, the semantic prosody in question is quite dominant – by my count, it is present in lines 2, 3, 4, 6, 7, 12, 13, 15, 17, 18 and 19, i.e., in 11 of 20 lines. This

is the exact proportion also observed with *true feelings*, so even if you disagree with one or two of my categorization decisions, there is no significant difference between the two expressions.

It seems, then, that the semantic prosody Sinclair observes is not attached to the expression *true feelings* in particular, but that it is an epiphenomenon the fact that we typically distinguish between "genuine" (*true*, *real*, etc.) emotions and other emotions in a particular context, namely one where someone is reluctant on unable to express their genuine emotions. Of course, studies of additional expressions with adjectives meaning "genuine" modifying nouns meaning "emotion" might give us a more detailed and differentiated picture, as might studies of other nouns modified by adjectives like *true* (such as *true nature*, *true beliefs*, *true intentions*, etc.). Such studies are left as an exercise to the reader – this case study was mainly meant to demonstrate how informal analyses based on the inspection of concordances can be integrated into a more rigorous research design involving quantification and comparison to a set of control data.

7.2.3.2 Case study: The verb *cause*

A second way in which semantic prosody can be studied quantitatively is implicit in Kennedy's study of collocates of degree adverbs discussed in Section 7.2.1 above. Recall that Kennedy discusses for each degree adverb whether a majority of its collocates has a positive or a negative connotation. This, of course, is a statement about the (broad) semantic prosody of the respective adverb, based not on an inspection and categorization of usage contexts, but on inductively discovered strongly associated collocates.

One of the earliest applications of this procedure is found in Stubbs (1995a). Stubbs studies, among other things, the noun and verb *cause*. He first presents the result of a manual extraction of all nouns (sometimes with adjectives qualifying them, as in the case of *wholesale slaughter*) that occur as subject or object of the verb *cause* or as a prepositional object of the noun *cause* in the LOB. He annotates them in their context of occurrence for their connotation, finding that approximately 80 percent are negative, 18 percent are neutral and 2 percent are positive. This procedure is still very close to Sinclairs approach of inspecting concordances, although is is stricter in terms of categorizing and quantifying the data.

Stubbs then notes that manual inspection and extraction becomes unfeasible as the number of corpus hits grows and suggests that, instead, we should first identify significant collocates of the word or expression we are interested in, and then categorize these significant collocates according to our criteria – note that

this is the strategy we also used in Case Study 7.2.2.1 above in order to determine semantic differences between *high* and *tall*.

We will not follow Stubbs' discussion in detail here – his focus is on methodological issues regarding the best way to identify collocates. Since we decided in Section 7.1.3 above to stick with the *G* statistic, this discussion is not central for us. Stubbs does not present the results of his procedure in detail and the corpus he uses is not accessible anyway, so let us use the BNC again and extract our own data.

Table 7.17 shows the result of an attempt to extract direct objects of the verb *cause* from the BNC. I searched for the lemma *cause* where it is tagged as a verb, followed by zero to three words that are not nouns (to take into account the occurrence of determiners, adjectives, etc.) and that are not the word *by* (in order to exclude passives like *caused by negligence, fire, exposure*, etc.), followed by a noun or sequence of nouns, not followed by *to* (in order to exclude causative constructions of the form *caused the glass to break*). This noun, or the last noun in this sequence, is assumed to be the direct object of *cause*. The twenty most frequent nouns are shown in Table 7.17, Column (a).

These collocates clearly corroborate Stubbs' observation about the negative semantic prosody of *cause*. We could now calculate the association strength between the verb and each of these nouns to get a better idea of which of them are significant collocates and which just happen to be frequent in the corpus overall. It should be obvious, however, that the nouns in Table 7.17, Column (a) are not generally frequent in the English language, so we can assume here that they are, for the most part, significant collocates.

But even so, what does this tell us about the semantic prosody of the verb *cause*? It has variously been pointed out (for example, by Louw & Chateau 2010) that other verbs of causation also tend to have a negative semantic prosody – the direct object nouns of *bring about* in Table 7.17, Column (b) and *lead to* in Table 7.17, Column (c) corroborate this. The real question is, again, whether it is the specific expression [*cause* NP] that has the semantic prosody in question, or whether this prosody is found in an entire semantic domain – perhaps speakers of English have a generally negative view of causation.

In order to determine this, it might be useful to compare different expressions of causation to each other rather than to the corpus as a whole – to perform a *differentiating collocate analysis*: just by inspecting the frequencies in Table 7.17, it seems that the negative prosody is much weaker for *bring about* and *lead to* than for *cause*, so, individually or taken together, they could serve as a baseline against which to compare *cause*.

Table 7.17: Noun collocates of three expressions of causation

(a)		(b)		(c)	
[*CAUSE* NP]	Freq.	[*BRING ABOUT* NP]	Freq.	[*LEAD TO* NP]	Freq.
problem	836	*change*	247	*increase*	219
death	358	*improvement*	43	*change*	154
damage	334	*end*	30	*conclusion*	152
concern	284	*death*	22	*development*	133
trouble	269	*downfall*	21	*loss*	123
harm	203	*result*	21	*problem*	122
difficulty	185	*reduction*	19	*death*	114
injury	139	*revolution*	19	*formation*	110
change	128	*increase*	18	*reduction*	105
pain	122	*peace*	17	*improvement*	89
confusion	113	*collapse*	14	*confusion*	80
loss	113	*transformation*	13	*creation*	76
lot	95	*development*	12	*number*	66
increase	93	*shift*	11	*award*	64
delay	90	*decline*	10	*rise*	63
distress	84	*destruction*	10	*discovery*	62
disease	81	*state*	10	*fall*	61
controversy	78	*unity*	10	*result*	61
accident	76	*effect*	9	*decline*	60
cancer	72	*event*	9	*growth*	60
		situation	9		

Supplementary Online Material: BYHW

Table 7.18 shows the results of a differential collocate analysis between *cause* on the one hand and the combined collocates of *bring about* and *lead to* on the other.

Table 7.18: Differential collocates for *cause* compared to *bring about/lead to* in the BNC

COLLOCATE	Collocate with CAUSE	Collocate with OTHER	Other words with CAUSE	Other words with OTHER	*G*
problem	836	126	11 566	15 311	778.63
damage	334	15	12 068	15 422	438.76
concern	284	10	12 118	15 427	387.24
trouble	269	9	12 133	15 428	369.27
harm	203	1	12 199	15 436	318.67
pain	122	4	12 280	15 433	167.17
death	358	136	12 044	15 301	160.75
injury	139	14	12 263	15 423	148.39
difficulty	185	51	12 217	15 386	113.91
stir	70	0	12 332	15 437	113.42
distress	84	5	12 318	15 432	103.52
havoc	62	0	12 340	15 437	100.44
alarm	57	0	12 345	15 437	92.32
delay	90	14	12 312	15 423	80.16
controversy	78	9	12 324	15 428	79.11
sensation	48	0	12 354	15 437	77.73
lot	95	18	12 307	15 419	76.05
cancer	72	9	12 330	15 428	70.73
disease	81	14	12 321	15 423	68.28
offence	55	4	12 347	15 433	64.53

The negative prosody of the verb *cause* is even more pronounced than in the frequency list in Table 7.17: Even the two neutral words *change* and *increase* have disappeared. In contrast, the combined differential collocates of *bring about* and *lead to* as compared to *cause*, shown in Table 7.19 are neutral or even positive.

We can thus conclude, first, that all three verbal expressions of causation are likely to be used to some extent with direct object nouns with a negative connotation. However, it is only the verb *cause* that has a negative semantic prosody. Even the raw frequencies of nouns occurring in the object position of the three expressions suggest this: while *cause* occurs almost exclusively with negatively

Table 7.19: Differential collocates for *bring about/lead to* compared to *cause* in the BNC

COLLOCATE	Collocate with CAUSE	Collocate with OTHER	Other words with CAUSE	Other words with OTHER	*G*
conclusion	0	155	12 402	15 282	183.49
improvement	4	132	12 398	15 305	126.52
development	11	145	12 391	15 292	109.74
change	128	401	12 274	15 036	96.20
formation	9	111	12 393	15 326	81.82
award	0	64	12 402	15 373	75.60
creation	2	77	12 400	15 360	75.55
discovery	0	62	12 402	15 375	73.23
situation	1	60	12 401	15 377	62.27
understanding	0	52	12 402	15 385	61.40
decision	0	49	12 402	15 388	57.86
qualification	0	49	12 402	15 388	57.86
establishment	1	55	12 401	15 382	56.53
arrest	1	45	12 401	15 392	45.11
speculation	4	55	12 398	15 382	42.15
suggestion	0	34	12 402	15 403	40.13
result	14	82	12 388	15 355	39.71
introduction	0	33	12 402	15 404	38.95
increase	93	237	12 309	15 200	37.85
conviction	0	32	12 402	15 405	37.77

connoted nouns, *bring about* and *lead to* are much more varied. The differential collocate analysis then confirms that within the domain of causation, the verb *cause* specializes in encoding negative caused events, while the other two expressions encode neutral or positive events. Previous research (Louw & Chateau 2010) misses this difference as it is based exclusively on the qualitative inspection of concordances.

Thus, the case study shows, once again, the need for strict quantification and for research designs comparing the occurrence of a linguistic feature under different conditions. There is one caveat of the procedure presented here, however: while it is a very effective strategy to identify collocates first and categorize them according to their connotation afterwards, this categorization is then limited to an assessment of the lexically encoded meaning of the collocates. For example, *problem* and *damage* will be categorized as negative, but a problem does not have

to be negative – it can be interesting if it is the right problem and you are in the right mood (e.g. *[O]ne of these excercises caused an interesting problem for several members of the class* [Aiden Thompson, *Who's afraid of the Old Testament God?*]). Even *damage* can be a good thing in particular contexts from particular perspectives (e.g. *[A] high yield of intact PTX [...] caused damage to cancer cells in addition to the immediate effects of PDT* [10.1021/acs.jmedchem.5b01971]). Even more likely, neutral words like *change* will have positive or negative connotations in particular contexts, which are lost in the proces of identifying collocates quantitatively.

Keeping this caveat in mind, however, the method presented in this case study can be applied fruitfully in more complex designs than the one presented here. For example, we have treated the direct object position as a simple category here, but Stefanowitsch & Gries (2003) present data for nominal collocates of the verb *cause* in the object position of different subcategorization patterns. While their results corroborate the negative connotation of *cause* also found by Stubbs (1995a), their results add an interesting dimension: while objects of *cause* in the transitive construction (*cause a problem*) and the prepositional dative (*cause a problem to someone*) refer to negatively perceived external and objective states, the objects of *cause* in the ditransitive refer to negatively experienced internal and/or subjective states. Studies on semantic prosody can also take into account dimensions beyond the immediate structural context – for example, Louw & Chateau (2010) observe that the semantic prosody of *cause* is to some extent specific to particular language varieties, and present interesting data suggesting that in scientific writing it is generally used with a neutral connotation.

7.2.4 Cultural analysis

In collocation research, a word (or other element of linguistic structure) typically stands for itself – the aim of the researcher is to uncover the linguistic properties of a word (or set of words). However, texts are not just manifestations of a language system, but also of the cultural conditions under which they were produced. This allows corpus linguistic methods to be used in uncovering at least some properties of that culture. Specifically, we can take lexical items to represent culturally defined concepts and investigate their distribution in linguistic corpora in order to uncover these cultural definitions. Of course, this adds complexity to the question of operationalization: we must ensure that the words we choose are indeed valid representatives of the cultural concept in question.

7.2.4.1 Case study: Small boys, little girls

Obviously, lexical items used conventionally to refer to some culturally relevant group of people are plausible representatives of the cultural concept of that group. For example, some very general lexical items referring to people (or higher animals) exist in male and female versions – *man*/*woman*, *boy*/*girl*, *lad*/*lass*, *husband*/*wife*, *father*/*mother*, *king*/*queen*, etc. If such word pairs differ in their collocates, this could tell us something about the cultural concepts behind them. For example, Stubbs (1995b) cites a finding by Baker & Freebody (1989), that in children's literature, the word *girl* collocates with *little* much more strongly than the word *boy*, and vice versa for *small*. Stubbs shows that this is also true for balanced corpora (see Table 7.20; again, since Stubbs' corpora are not available, I show frequencies from the BNC instead but the proportions are within a few percent points of his). The difference in associations is highly significant ($\chi^2 = 217.66, \mathrm{df} = 1, p < 0.001$).

Table 7.20: *Small* and *little girls* and *boys* (BNC)

		Second Position		
		boy	girl	Total
First Position	little	791	1148	1939
		(927.53)	(1011.47)	
	small	336	81	417
		(199.47)	(217.53)	
	Total	1127	1229	2356

This part of Stubbs' study is clearly deductive: He starts with a hypothesis taken from the literature and tests it against a larger, more representative corpus. The variables involved are, as is typical for collocation studies, nominal variables whose values are words.

Stubbs argues that this difference is due to different connotations of *small* and *little* which he investigates on the basis of the noun collocates to their right and the adjectival and adverbial collocates to the left. Again, instead of Stubbs' original data (which he identifies on the basis of raw frequency of occurrence and only cites selectively), I use data from the BNC and the G test statistic. Table 7.21 shows the ten most strongly associated noun collocates to the right of the node word and Table 7.22 shows the ten most strongly associated adjectival collocates to the left.

Table 7.21: Nominal collocates of *little* and *small* at R1 (BNC)

COLLOCATE	Collocate with LITTLE	Collocate with SMALL	Other words with LITTLE	Other words with SMALL	*G*
Most strongly associated with *little*					
bit	2838	30	33 606	31 214	3331.01
girl	1148	70	35 296	31 174	1008.61
doubt	546	3	35 898	31 241	647.25
time	595	23	35 849	31 221	579.93
while	435	0	36 009	31 244	541.06
evidence	324	0	36 120	31 244	402.53
attention	253	1	36 191	31 243	302.56
chance	273	21	36 171	31 223	220.00
money	194	4	36 250	31 240	207.73
interest	213	12	36 231	31 232	189.11
Most strongly associated with *small*					
number	23	1118	36 421	30 126	1553.13
group	123	1089	36 321	30 155	1057.19
amount	7	670	36 437	30 574	974.34
business	36	784	36 408	30 460	971.22
firm	15	456	36 429	30 788	594.11
proportion	0	332	36 444	30 912	515.24
scale	1	265	36 443	30 979	399.02
company	15	316	36 429	30 928	386.62
area	15	302	36 429	30 942	366.16
mammal	0	203	36 444	31 041	314.58

Table 7.22: Adjectival collocates of *little* and *small* at L1 (BNC)

Collocate	Collocate with LITTLE	Collocate with SMALL	Other words with LITTLE	Other words with SMALL	*G*
Most strongly associated with *little*					
nice	356	4	4719	1083	112.25
poor	248	0	4827	1087	98.46
pretty	119	0	4956	1087	46.69
tiny	95	0	4980	1087	37.19
nasty	60	0	5015	1087	23.42
funny	67	1	5008	1086	19.19
dear	47	0	5028	1087	18.32
sweet	42	0	5033	1087	16.36
silly	59	1	5016	1086	16.30
lovely	92	5	4983	1082	13.84
Most strongly associated with *small*					
other	59	141	5016	946	282.80
only	36	119	5039	968	268.74
proximal	0	28	5075	1059	97.76
numerous	4	30	5071	1057	81.67
far	3	19	5072	1068	49.83
wee	2	15	5073	1072	40.67
existing	0	11	5075	1076	38.26
various	6	18	5069	1069	38.01
occasional	1	12	5074	1075	35.08
new	25	28	5050	1059	33.95

This part of the study is more inductive. Stubbs may have expectations about what he will find, but he essentially identifies collocates exploratively and then interprets the findings. The nominal collocates show, according to Stubbs, that *small* tends to mean 'small in physical size' or 'low in quantity', while *little* is more clearly restricted to quantities, including informal quantifying phrases like *little bit*. This is generally true for the BNC data, too (note, however, the one exception among the top ten collocates – *girl*).

The connotational difference between the two adjectives becomes clear when we look at the adjectives they combine with. The word *little* has strong associations to evaluative adjectives that may be positive or negative, and that are often patronizing. *Small*, in contrast, does not collocate with evaluative adjectives.

Stubbs sums up his analysis by pointing out that *small* is a neutral word for describing size, while *little* is sometimes used neutrally, but is more often "non-literal and convey[s] connotative and attitudinal meanings, which are often patronizing, critical, or both." (Stubbs 1995b: 386). The differences in distribution relative to the words *boy* and *girl* are evidence for him that "[c]ulture is encoded not just in words which are obviously ideologically loaded, but also in combinations of very common words" (Stubbs 1995b: 387).

Stubbs remains unspecific as to what that ideology is – presumably, one that treats boys as neutral human beings and girls as targets for patronizing evaluation. In order to be more specific, it would be necessary to turn around the perspective and study all adjectival collocates of *boy* and *girl*. Stubbs does not do this, but Caldas-Coulthard & Moon (2010) look at adjectives collocating with *man*, *woman*, *boy* and *girl* in broadsheet and yellow-press newspapers. In order to keep the results comparable with those reported above, let us stick with the BNC instead. Table 7.23 shows the top ten adjectival collocates of *boy* and *girl*.

The results are broadly similar in kind to those in Caldas-Coulthard & Moon (2010): *boy* collocates mainly with neutral descriptive terms (*small*, *lost*, *big*, *new*), or with terms with which it forms a fixed expression (*old*, *dear*, *toy*, *whipping*). There are the evaluative adjectives *rude* (which in Caldas-Coulthard and Moon's data is often applied to young men of Jamaican descent) and its positively connoted equivalent *naughty*. The collocates of *girl* are overwhelmingly evaluative, related to physical appearance. There are just two neutral adjective (*other* and *dead*, the latter tying in with a general observation that women are more often spoken of as victims of crimes and other activities than men). Finally, there is one adjective signaling marital status. These results also generally reflect Caldas-Coulthard and Moon's findings (in the yellow-press, the evaluations are often heavily sexualized in addition).

Table 7.23: Adjectival collocates of *boy* and *girl* at L1 (BNC)

Collocate	Collocate with BOY	Collocate with GIRL	Other words with BOY	Other words with GIRL	*G*
Most strongly associated with *boy*					
old	634	257	5385	7296	279.98
small	336	81	5683	7472	237.78
dear	126	45	5893	7508	61.30
lost	41	1	5978	7552	58.54
big	167	89	5852	7464	46.02
naughty	71	22	5948	7531	39.75
new	124	69	5895	7484	31.31
rude	19	0	6000	7553	30.93
toy	16	0	6003	7553	26.04
whipping	14	0	6005	7553	22.78
Most strongly associated with *girl*					
young	351	820	5668	6733	111.04
pretty	23	132	5996	7421	62.59
other	194	444	5825	7109	54.56
beautiful	13	87	6006	7466	46.13
attractive	1	35	6018	7518	33.58
blonde	1	29	6018	7524	26.89
single	1	27	6018	7526	24.68
dead	12	57	6007	7496	22.67
unmarried	0	17	6019	7536	19.94
lovely	18	66	6001	7487	19.45

This case study shows how collocation research may uncover facts that go well beyond lexical semantics or semantic prosody. In this case, the collocates of *boy* and *girl* have uncovered a general attitude that sees the latter as up for constant evaluation while the former are mainly seen as a neutral default. That the adjectives *dead* and *unmarried* are among the top ten collocates in a representative, relatively balanced corpus, hints at something darker – a patriarchal world view that sees girls as victims and sexual partners and not much else (other studies investigating gender stereotypes on the basis of collocates of *man* and *woman* are Gesuato (2003) and Pearce (2008)).