5 *Collocations*

A COLLOCATION is an expression consisting of two or more words that correspond to some conventional way of saying things. Or in the words of Firth (1957: 181): "Collocations of a given word are statements of the habitual or customary places of that word." Collocations include noun phrases like *strong tea* and *weapons of mass destruction*, phrasal verbs like *to make up*, and other stock phrases like *the rich and powerful*. Particularly interesting are the subtle and not-easily-explainable patterns of word usage that native speakers all know: why we say *a stiff breeze* but not *??a stiff wind* (while either *a strong breeze* or *a strong wind* is okay), or why we speak of *broad daylight* (but not *?bright daylight* or *??narrow darkness*).

COMPOSITIONALITY

Collocations are characterized by limited *compositionality*. We call a natural language expression compositional if the meaning of the expression can be predicted from the meaning of the parts. Collocations are not fully compositional in that there is usually an element of meaning added to the combination. In the case of *strong tea*, *strong* has acquired the meaning *rich in some active agent* which is closely related, but slightly different from the basic sense *having great physical strength*. Idioms are the most extreme examples of non-compositionality. Idioms like *to kick the bucket* or *to hear it through the grapevine* only have an indirect historical relationship to the meanings of the parts of the expression. We are not talking about buckets or grapevines literally when we use these idioms. Most collocations exhibit milder forms of non-compositionality, like the expression *international best practice* that we used as an example earlier in this book. It is very nearly a systematic composition of its parts, but still has an element of added meaning. It usually refers to administrative efficiency and would, for example, not be used to describe a cooking technique although that meaning would be compatible with its literal meaning.

TERM
TECHNICAL TERM
TERMINOLOGICAL PHRASE

There is considerable overlap between the concept of *collocation* and notions like *term*, *technical term*, and *terminological phrase*. As these names sug-

TERMINOLOGY EXTRACTION

gest, the latter three are commonly used when collocations are extracted from technical domains (in a process called *terminology extraction*). The reader be warned, though, that the word *term* has a different meaning in information retrieval. There, it refers to both words and phrases. So it subsumes the more narrow meaning that we will use in this chapter.

Collocations are important for a number of applications: natural language generation (to make sure that the output sounds natural and mistakes like *powerful tea* or *to take a decision* are avoided), computational lexicography (to automatically identify the important collocations to be listed in a dictionary entry), parsing (so that preference can be given to parses with natural collocations), and corpus linguistic research (for instance, the study of social phenomena like the reinforcement of cultural stereotypes through language (Stubbs 1996)).

There is much interest in collocations partly because this is an area that has been neglected in structural linguistic traditions that follow Saussure and Chomsky. There is, however, a tradition in British linguistics, associated with the names of Firth, Halliday, and Sinclair, which pays close attention to phenomena like collocations. Structural linguistics concentrates on general abstractions about the properties of phrases and sentences. In contrast, Firth's *Contextual Theory of Meaning* emphasizes the importance of context: the context of the social setting (as opposed to the idealized speaker), the context of spoken and textual discourse (as opposed to the isolated sentence), and, important for collocations, the context of surrounding words (hence Firth's famous dictum that a word is characterized by the company it keeps). These contextual features easily get lost in the abstract treatment that is typical of structural linguistics.

A good example of the type of problem that is seen as important in this contextual view of language is Halliday's example of strong vs. powerful tea (Halliday 1966: 150). It is a convention in English to talk about *strong tea*, not *powerful tea*, although any speaker of English would also understand the latter unconventional expression. Arguably, there are no interesting structural properties of English that can be gleaned from this contrast. However, the contrast may tell us something interesting about attitudes towards different types of substances in our culture (why do we use *powerful* for drugs like heroin, but not for cigarettes, tea and coffee?) and it is obviously important to teach this contrast to students who want to learn idiomatically correct English. Social implications of language use and language teaching are just the type of problem that British linguists following a Firthian approach are interested in.

In this chapter, we will introduce the principal approaches to finding col-

locations: selection of collocations by frequency, selection based on mean and variance of the distance between focal word and collocating word, hypothesis testing, and mutual information. We will then return to the question of what a collocation is and discuss in more depth different definitions that have been proposed and tests for deciding whether a phrase is a collocation or not. The chapter concludes with further readings and pointers to some of the literature that we were not able to include.

The reference corpus we will use in examples in this chapter consists of four months of the New York Times newswire: from August through November of 1990. This corpus has about 115 megabytes of text and roughly 14 million words. Each approach will be applied to this corpus to make comparison easier. For most of the chapter, the New York Times examples will only be drawn from fixed two-word phrases (or bigrams). It is important to keep in mind, however, that we chose this pool for convenience only. In general, both fixed and variable word combinations can be collocations. Indeed, the section on mean and variance looks at the more loosely connected type.

5.1 Frequency

Surely the simplest method for finding collocations in a text corpus is counting. If two words occur together a lot, then that is evidence that they have a special function that is not simply explained as the function that results from their combination.

Predictably, just selecting the most frequently occurring bigrams is not very interesting as is shown in Table 5.1. The table shows the bigrams (sequences of two adjacent words) that are most frequent in the corpus and their frequency. Except for *New York*, all the bigrams are pairs of function words.

There is, however, a very simple heuristic that improves these results a lot (Justeson and Katz 1995b): pass the candidate phrases through a part-of-speech filter which only lets through those patterns that are likely to be "phrases".[1] Justeson and Katz (1995b: 17) suggest the patterns in Table 5.2. Each is followed by an example from the text that they use as a test set. In these patterns A refers to an adjective, P to a preposition, and N to a noun.

Table 5.3 shows the most highly ranked phrases after applying the filter. The results are surprisingly good. There are only 3 bigrams that we would not regard as non-compositional phrases: *last year*, *last week*, and *first time*.

1. Similar ideas can be found in (Ross and Tukey 1975) and (Kupiec et al. 1995).

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Table 5.1 Finding Collocations: Raw Frequency. $C(\cdot)$ is the frequency of something in the corpus.

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

Table 5.2 Part of speech tag patterns for collocation filtering. These patterns were used by Justeson and Katz to identify likely collocations among frequently occurring word sequences.

$C(w^1\ w^2)$	w^1	w^2	tag pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Table 5.3 Finding Collocations: Justeson and Katz' part-of-speech filter.

York City is an artefact of the way we have implemented the Justeson and Katz filter. The full implementation would search for the longest sequence that fits one of the part-of-speech patterns and would thus find the longer phrase *New York City*, which contains *York City*.

The twenty highest ranking phrases containing *strong* and *powerful* all have the form A N (where A is either *strong* or *powerful*). We have listed them in Table 5.4.

Again, given the simplicity of the method, these results are surprisingly accurate. For example, they give evidence that *strong challenge* and *powerful computers* are correct whereas *powerful challenge* and *strong computers* are not. However, we can also see the limits of a frequency-based method. The nouns *man* and *force* are used with both adjectives (*strong force* occurs further down the list with a frequency of 4). A more sophisticated analysis is necessary in such cases.

Neither *strong tea* nor *powerful tea* occurs in our New York Times corpus.

w	$C(\textit{strong}, w)$	w	$C(\textit{powerful}, w)$
support	50	force	13
safety	22	computers	10
sales	21	position	8
opposition	19	men	8
showing	18	computer	8
sense	18	man	7
message	15	symbol	6
defense	14	military	6
gains	13	machines	6
evidence	13	country	6
criticism	13	weapons	5
possibility	11	post	5
feelings	11	people	5
demand	11	nation	5
challenges	11	forces	5
challenge	11	chip	5
case	11	Germany	5
supporter	10	senators	4
signal	9	neighbor	4
man	9	magnet	4

Table 5.4 The nouns w occurring most often in the patterns "*strong* w" and "*powerful* w".

However, searching the larger corpus of the World Wide Web we find 799 examples of *strong tea* and 17 examples of *powerful tea* (the latter mostly in the computational linguistics literature on collocations), which indicates that the correct phrase is *strong tea*.[2]

Justeson and Katz' method of collocation discovery is instructive in that it demonstrates an important point. A simple quantitative technique (the frequency filter in this case) combined with a small amount of linguistic knowledge (the importance of parts of speech) goes a long way. In the rest of this chapter, we will use a stop list that excludes words whose most frequent tag is not a verb, noun or adjective.

Exercise 5-1
Add part-of-speech patterns useful for collocation discovery to Table 5.2, including patterns longer than two tags.

2. This search was performed on AltaVista on March 28, 1998.

Sentence: *Stocks crash as rescue plan teeters*

Bigrams:

stocks crash	*stocks as*	*stocks rescue*		
	crash as	*crash rescue*	*crash plan*	
		as rescue	*as plan*	*as teeters*
			rescue plan	*rescue teeters*
				plan teeters

Figure 5.1 Using a three word collocational window to capture bigrams at a distance.

Exercise 5-2
Pick a document in which your name occurs (an email, a university transcript or a letter). Does Justeson and Katz's filter identify your name as a collocation?

Exercise 5-3
We used the World Wide Web as an auxiliary corpus above because neither *stong tea* nor *powerful tea* occurred in the New York Times. Modify Justeson and Katz's method so that it uses the World Wide Web as a resource of last resort.

5.2 Mean and Variance

Frequency-based search works well for fixed phrases. But many collocations consist of two words that stand in a more flexible relationship to one another. Consider the verb *knock* and one of its most frequent arguments, *door*. Here are some examples of knocking on or at a door from our corpus:

(5.1) a. she knocked on his door

b. they knocked at the door

c. 100 women knocked on Donaldson's door

d. a man knocked on the metal front door

The words that appear between *knocked* and *door* vary and the distance between the two words is not constant so a fixed phrase approach would not work here. But there is enough regularity in the patterns to allow us to determine that *knock* is the right verb to use in English for this situation, not *hit*, *beat* or *rap*.

A short note is in order here on collocations that occur as a fixed phrase versus those that are more variable. To simplify matters we only look at fixed phrase collocations in most of this chapter, and usually at just bigrams. But it is easy to see how to extend techniques applicable to bigrams

to bigrams at a distance. We define a collocational window (usually a window of 3 to 4 words on each side of a word), and we enter *every* word pair in there as a collocational bigram, as in Figure 5.1. We then proceed to do our calculations as usual on this larger pool of bigrams.

However, the mean and variance based methods described in this section by definition look at the pattern of varying distance between two words. If that pattern of distances is relatively predictable, then we have evidence for a collocation like *knock ... door* that is not necessarily a fixed phrase. We will return to this point and a more in-depth discussion of what a collocation is towards the end of this chapter.

MEAN
VARIANCE

One way of discovering the relationship between *knocked* and *door* is to compute the *mean* and *variance* of the offsets (signed distances) between the two words in the corpus. The mean is simply the average offset. For the examples in (5.1), we compute the mean offset between *knocked* and *door* as follows:

$$\frac{1}{4}(3+3+5+5) = 4.0$$

(This assumes a tokenization of *Donaldson's* as three words *Donaldson*, apostrophe, and *s*, which is what we actually did.) If there was an occurrence of *door* before *knocked*, then it would be entered as a negative number. For example, -3 for *the door that she knocked on*. We restrict our analysis to positions in a window of size 9 around the focal word *knocked*.

The variance measures how much the individual offsets deviate from the mean. We estimate it as follows.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1} \tag{5.2}$$

where n is the number of times the two words co-occur, d_i is the offset for co-occurrence i, and μ is the mean. If the offset is the same in all cases, then the variance is zero. If the offsets are randomly distributed (which will be the case for two words which occur together by chance, but not in a particular relationship), then the variance will be high. As is customary, we use the *standard deviation* $\sigma = \sqrt{\sigma^2}$, the square root of the variance, to assess how variable the offset between two words is. The standard deviation for the four examples of *knocked* / *door* in the above case is 1.15:

STANDARD DEVIATION

$$\sigma = \sqrt{\frac{1}{3}\left((3-4.0)^2 + (3-4.0)^2 + (5-4.0)^2 + (5-4.0)^2\right)} \approx 1.15$$

The mean and standard deviation characterize the distribution of distances between two words in a corpus. We can use this information to discover collocations by looking for pairs with low standard deviation. A low

standard deviation means that the two words usually occur at about the same distance. Zero standard deviation means that the two words always occur at exactly the same distance.

We can also explain the information that variance gets at in terms of peaks in the distribution of one word with respect to another. Figure 5.2 shows the three cases we are interested in. The distribution of *strong* with respect to *opposition* has one clear peak at position -1 (corresponding to the phrase *strong opposition*). Therefore the variance of *strong* with respect to *opposition* is small ($\sigma = 0.67$). The mean of -1.15 indicates that *strong* usually occurs at position -1 (disregarding the noise introduced by one occurrence at -4).

We have restricted positions under consideration to a window of size 9 centered around the word of interest. This is because collocations are essentially a local phenomenon. Note also that we always get a count of 0 at position 0 when we look at the relationship between two different words. This is because, for example, *strong* cannot appear in position 0 in contexts in which that position is already occupied by *opposition*.

Moving on to the second diagram in Figure 5.2, the distribution of *strong* with respect to *support* is drawn out, with several negative positions having large counts. For example, the count of approximately 20 at position -2 is due to uses like *strong leftist support* and *strong business support*. Because of this greater variability we get a higher σ (1.07) and a mean that is between positions -1 and -2 (-1.45).

Finally, the occurrences of *strong* with respect to *for* are more evenly distributed. There is tendency for *strong* to occur before *for* (hence the negative mean of -1.12), but it can pretty much occur anywhere around *for*. The high standard deviation of $\sigma = 2.15$ indicates this randomness. This indicates that *for* and *strong* don't form interesting collocations.

The word pairs in Table 5.5 indicate the types of collocations that can be found by this approach. If the mean is close to 1.0 and the standard deviation low, as is the case for *New York*, then we have the type of phrase that Justeson and Katz' frequency-based approach will also discover. If the mean is much greater than 1.0, then a low standard deviation indicates an interesting phrase. The pair *previous* / *games* (distance 2) corresponds to phrases like *in the previous 10 games* or *in the previous 15 games*; *minus* / *points* corresponds to phrases like *minus 2 percentage points*, *minus 3 percentage points* etc; *hundreds* / *dollars* corresponds to *hundreds of billions of dollars* and *hundreds of millions of dollars*.

High standard deviation indicates that the two words of the pair stand in no interesting relationship as demonstrated by the four high-variance

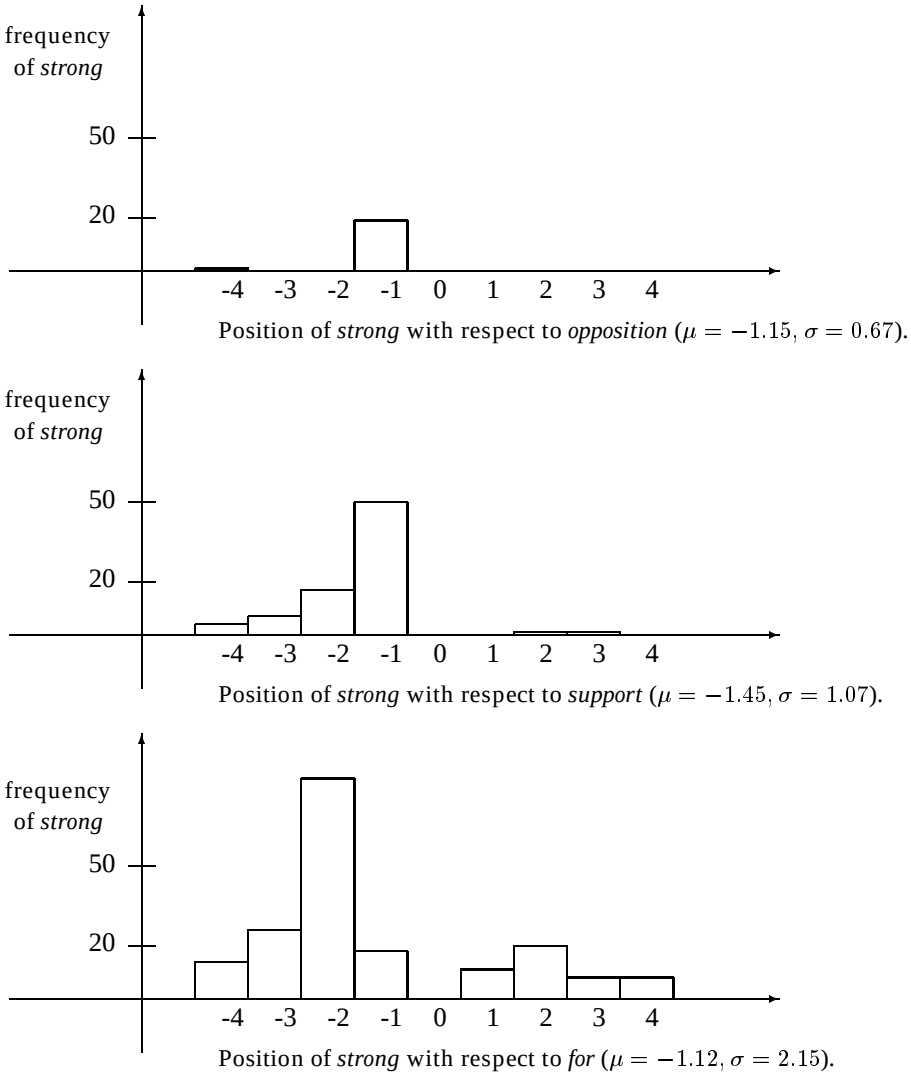

Figure 5.2 Histograms of the position of *strong* relative to three words.

σ	μ	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

Table 5.5 Finding collocations based on mean and variance. Standard Deviation σ and mean μ of the distances between 12 word pairs.

examples in Table 5.5. Note that means tend to be close to zero here as one would expect for a uniform distribution. More interesting are the cases in between, word pairs that have large counts for several distances in their collocational distribution. We already saw the example of *strong { business } support* in Figure 5.2. The alternations captured in the other three medium-variance examples are *powerful { lobbying } organizations*, *Richard { M. } Nixon*, and *Garrison said* / *said Garrison* (remember that we tokenize *Richard M. Nixon* as four tokens: *Richard*, *M*, ., *Nixon*).

The method of variance-based collocation discovery that we have introduced in this section is due to Smadja. We have simplified things somewhat. In particular, Smadja (1993) uses an additional constraint that filters out "flat" peaks in the position histogram, that is, peaks that are not surrounded by deep valleys (an example is at -2 for the combination *strong* / *for* in Figure 5.2). Smadja (1993) shows that the method is quite successful at terminological extraction (with an estimated accuracy of 80%) and at determining appropriate phrases for natural language generation (Smadja and McKeown 1990).

Smadja's notion of collocation is less strict than many others'. The combination *knocked* / *door* is probably not a collocation we want to classify as terminology – although it may be very useful to identify for the purpose of text generation. Variance-based collocation discovery is the appropriate method if we want to find this type of word combination, combinations

of words that are in a looser relationship than fixed phrases and that are variable with respect to intervening material and relative position.

5.3 Hypothesis Testing

One difficulty that we have glossed over so far is that high frequency and low variance can be accidental. If the two constituent words of a frequent bigram like *new companies* are frequently occurring words (as *new* and *companies* are), then we expect the two words to co-occur a lot just by chance, even if they do not form a collocation.

What we really want to know is whether two words occur together more often than chance. Assessing whether or not something is a chance event is one of the classical problems of statistics. It is usually couched in terms of hypothesis testing. We formulate a *null hypothesis* H_0 that there is no association between the words beyond chance occurrences, compute the probability p that the event would occur if H_0 were true, and then reject H_0 if p is too low (typically if beneath a *significance level* of $p < 0.05, 0.01, 0.005$, or 0.001) and retain H_0 as possible otherwise.[3]

NULL HYPOTHESIS

SIGNIFICANCE LEVEL

It is important to note that this is a mode of data analysis where we look at two things at the same time. As before, we are looking for particular patterns in the data. But we are also taking into account how much data we have seen. Even if there is a remarkable pattern, we will discount it if we haven't seen enough data to be certain that it couldn't be due to chance.

How can we apply the methodology of hypothesis testing to the problem of finding collocations? We first need to formulate a null hypothesis which states what should be true if two words do not form a collocation. For such a free combination of two words we will assume that each of the words w^1 and w^2 is generated completely independently of the other, and so their chance of coming together is simply given by:

$$P(w^1 w^2) = P(w^1)P(w^2)$$

The model implies that the probability of co-occurrence is just the product of the probabilities of the individual words. As we discuss at the end of this section, this is a rather simplistic model, and not empirically accurate, but for now we adopt independence as our null hypothesis.

3. Significance at a level of 0.05 is the weakest evidence that is normally accepted in the experimental sciences. The large amounts of data commonly available for Statistical NLP tasks means the we can often expect to achieve greater levels of significance.

5.3.1 The t test

Next we need a statistical test that tells us how probable or improbable it is that a certain constellation will occur. A test that has been widely used for collocation discovery is the t test. The t test looks at the mean and variance of a sample of measurements, where the null hypothesis is that the sample is drawn from a distribution with mean μ. The test looks at the difference between the observed and expected means, scaled by the variance of the data, and tells us how likely one is to get a sample of that mean and variance (or a more extreme mean and variance) assuming that the sample is drawn from a normal distribution with mean μ. To determine the probability of getting our sample (or a more extreme sample), we compute the t statistic:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \tag{5.3}$$

where $\bar{x}$ is the sample mean, s^2 is the sample variance, N is the sample size, and μ is the mean of the distribution. If the t statistic is large enough we can reject the null hypothesis. We can find out exactly how large it has to be by looking up the table of the t distribution we have compiled in the appendix (or by using the better tables in a statistical reference book, or by using appropriate computer software).

Here's an example of applying the t test. Our null hypothesis is that the mean height of a population of men is 158cm. We are given a sample of 200 men with $\bar{x} = 169$ and $s^2 = 2600$ and want to know whether this sample is from the general population (the null hypothesis) or whether it is from a different population of smaller men. This gives us the following t according to the above formula:

$$t = \frac{169 - 158}{\sqrt{\frac{2600}{200}}} \approx 3.05$$

If you look up the value of t that corresponds to a confidence level of $\alpha = 0.005$, you will find 2.576.[4] Since the t we got is larger than 2.576, we can reject the null hypothesis with 99.5% confidence. So we can say that the sample is not drawn from a population with mean 158cm, and our probability of error is less than 0.5%.

To see how to use the t test for finding collocations, let us compute the t value for *new companies*. What is the sample that we are measuring the

4. A sample of 200 means 199 degress of freedom, which corresponds to about the same t as ∞ degrees of freedom. This is the row of the table where we looked up 2.576.

mean and variance of? There is a standard way of extending the t test for use with proportions or counts. We think of the text corpus as a long sequence of N bigrams, and the samples are then indicator random variables that take on the value 1 when the bigram of interest occurs, and are 0 otherwise.

Using maximum likelihood estimates, we can compute the probabilities of *new* and *companies* as follows. In our corpus, *new* occurs 15,828 times, *companies* 4,675 times, and there are 14,307,668 tokens overall.

$$P(\textit{new}) = \frac{15828}{14307668}$$

$$P(\textit{companies}) = \frac{4675}{14307668}$$

The null hypothesis is that occurrences of *new* and *companies* are independent.

$$\begin{aligned} H_0 : P(\textit{new companies}) &= P(\textit{new})P(\textit{companies}) \\ &= \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7} \end{aligned}$$

If the null hypothesis is true, then the process of randomly generating bigrams of words and assigning 1 to the outcome *new companies* and 0 to any other outcome is in effect a Bernoulli trial with $p = 3.615 \times 10^{-7}$ for the probability of *new company* turning up. The mean for this distribution is $\mu = 3.615 \times 10^{-7}$ and the variance is $\sigma^2 = p(1-p)$ (see Section 2.1.9), which is approximately p. The approximation $\sigma^2 = p(1-p) \approx p$ holds since for most bigrams p is small.

It turns out that there are actually 8 occurrences of *new companies* among the 14307668 bigrams in our corpus. So, for the sample, we have that the sample mean is: $\bar{x} = \frac{8}{14307668} \approx 5.591 \times 10^{-7}$. Now we have everything we need to apply the *t* test:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{5.59110^{-7} - 3.61510^{-7}}{\sqrt{\frac{5.59110^{-7}}{14307668}}} \approx 0.999932$$

This t value of 0.999932 is not larger than 2.576, the critical value for $\alpha = 0.005$. So we cannot reject the null hypothesis that *new* and *companies* occur independently and do not form a collocation. That seems the right result here: the phrase *new companies* is completely compositional and there is no element of added meaning here that would justify elevating it to the status of collocation. (The t value is suspiciously close to 1.0, but that is a coincidence. See Exercise 5-5.)

t	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

Table 5.6 Finding collocations: The t test applied to 10 bigrams that occur with frequency 20.

Table 5.6 shows t values for ten bigrams that occur exactly 20 times in the corpus. For the top five bigrams, we can reject the null hypothesis that the component words occur independently for $\alpha = 0.005$, so these are good candidates for collocations. The bottom five bigrams fail the test for significance, so we will not regard them as good candidates for collocations.

Note that a frequency-based method would not be able to rank the ten bigrams since they occur with exactly the same frequency. Looking at the counts in Table 5.6, we can see that the t test takes into account the number of co-occurrences of the bigram ($C(w^1\ w^2)$) relative to the frequencies of the component words. If a high proportion of the occurrences of both words (*Ayatollah Ruhollah*, *videocassette recorder*) or at least a very high proportion of the occurrences of one of the words (*unsalted*) occurs in the bigram, then its t value is high. This criterion makes intuitive sense.

Unlike most of this chapter, the analysis in Table 5.6 includes some stop words – without stop words, it is actually hard to find examples that fail significance. It turns out that most bigrams attested in a corpus occur significantly more often than chance. For 824 out of the 831 bigrams that occurred 20 times in our corpus the null hypothesis of independence can be rejected. But we would only classify a fraction as true collocations. The reason for this surprisingly high proportion of possibly dependent bigrams ($\frac{824}{831} \approx 0.99$) is that language – if compared with a random word generator – is very regular so that few completely unpredictable events happen. Indeed, this is the basis of our ability to perform tasks like word sense disambiguation and probabilistic parsing that we discuss in other chapters.

The t test and other statistical tests are most useful as a method for *ranking* collocations. The level of significance itself is less useful. In fact, in most publications that we cite in this chapter, the level of significance is never looked at. All that is used is the scores and the resulting ranking.

5.3.2 Hypothesis testing of differences

The t test can also be used for a slightly different collocation discovery problem: to find words whose co-occurrence patterns best distinguish between two words. For example, in computational lexicography we may want to find the words that best differentiate the meanings of *strong* and *powerful*. This use of the t test was suggested by Church and Hanks (1989). Table 5.7 shows the ten words that occur most significantly more often with *powerful* than with *strong* (first ten words) and most significantly more often with *strong* than with *powerful* (second set of ten words).

The t scores are computed using the following extension of the t test to the comparison of the means of two normal populations:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} \tag{5.4}$$

Here the null hypothesis is that the average difference is 0 ($\mu = 0$), so we have $\bar{x} - \mu = \bar{x} = \frac{1}{N}\sum(x_{1_i} - x_{2_i}) = \bar{x}_1 - \bar{x}_2$. In the denominator we add the variances of the two populations since the variance of the difference of two random variables is the sum of their individual variances.

Now we can explain Table 5.7. The t values in the table were computed assuming a Bernoulli distribution (as we did for the basic version of the t test that we introduced first). If w is the collocate of interest (e.g., *computers* or *symbol*) and v^1 and v^2 are the words we are comparing (e.g., *powerful* and *strong*), then we have $\bar{x}_1 = s_1^2 = P(v^1 w)$, $\bar{x}_2 = s_2^2 = P(v^2 w)$. We again use the approximation $s^2 = p - p^2 \approx p$:

$$t \approx \frac{P(v^1 w) - P(v^2 w)}{\sqrt{\frac{P(v^1 w) + P(v^2 w)}{N}}}$$

We can simplify this as follows.

$$\begin{aligned} t &\approx \frac{\frac{C(v^1 w)}{N} - \frac{C(v^2 w)}{N}}{\sqrt{\frac{C(v^1 w) + C(v^2 w)}{N^2}}} \\ &= \frac{C(v^1 w) - C(v^2 w)}{\sqrt{C(v^1 w) + C(v^2 w)}} \end{aligned} \tag{5.5}$$

t	$C(w)$	$C(\textit{strong}\ w)$	$C(\textit{powerful}\ w)$	word
3.1622	933	0	10	computers
2.8284	2337	0	8	computer
2.4494	289	0	6	symbol
2.4494	588	0	6	machines
2.2360	2266	0	5	Germany
2.2360	3745	0	5	nation
2.2360	395	0	5	chip
2.1828	3418	4	13	force
2.0000	1403	0	4	friends
2.0000	267	0	4	neighbor
7.0710	3685	50	0	support
6.3257	3616	58	7	enough
4.6904	986	22	0	safety
4.5825	3741	21	0	sales
4.0249	1093	19	1	opposition
3.9000	802	18	1	showing
3.9000	1641	18	1	sense
3.7416	2501	14	0	defense
3.6055	851	13	0	gains
3.6055	832	13	0	criticism

Table 5.7 Words that occur significantly more often with *powerful* (the first ten words) and *strong* (the last ten words).

where $C(x)$ is the number of times x occurs in the corpus.

The application suggested by Church and Hanks (1989) for this form of the t test was lexicography. The data in Table 5.7 are useful to a lexicographer who wants to write precise dictionary entries that bring out the difference between *strong* and *powerful*. Based on significant collocates, Church and Hanks analyze the difference as a matter of intrinsic vs. extrinsic quality. For example, *strong* support from a demographic group means that the group is very committed to the cause in question, but the group may not have any power. So *strong* describes an intrinsic quality. Conversely, a *powerful* supporter is somebody who actually has the power to move things. Many of the collocates we found in our corpus support Church and Hanks' analysis. But there is more complexity to the difference in meaning between the two words since what is extrinsic and intrinsic can depend on subtle matters like cultural attitudes. For example, we talk about *strong tea*

	$w_1 =$ *new*	$w_1 \neq$ *new*
$w_2 =$ *companies*	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq$ *companies*	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

Table 5.8 A 2-by-2 table showing the dependence of occurrences of *new* and *companies*. There are 8 occurrences of *new companies* in the corpus, 4667 bigrams where the second word is *companies*, but the first word is not *new*, 15,820 bigrams with the first word *new* and a second word different from *companies*, and 14,287,181 bigrams that contain neither word in the appropriate position.

on the one hand and *powerful drugs* on the other, a difference that tells us more about our attitude towards tea and drugs than about the semantics of the two adjectives (Church et al. 1991: 133).

5.3.3 Pearson's chi-square test

Use of the *t* test has been criticized because it assumes that probabilities are approximately normally distributed, which is not true in general (Church and Mercer 1993: 20). An alternative test for dependence which does not assume normally distributed probabilities is the χ^2 test (pronounced "chi-square test"). In the simplest case, the χ^2 test is applied to 2-by-2 tables like Table 5.8. The essence of the test is to compare the observed frequencies in the table with the frequencies expected for independence. If the difference between observed and expected frequencies is large, then we can reject the null hypothesis of independence.

Table 5.8 shows the distribution of *new* and *companies* in the reference corpus that we introduced earlier. Recall that $C(\textit{new}) = 15,828$, $C(\textit{companies}) = 4,675$, $C(\textit{new companies}) = 8$, and that there are 14,307,668 tokens in the corpus. That means that the number of bigrams $w_i w_{i+1}$ with the first token being *new* and the second token not being *companies* is $4667 = 4675 - 8$. The two cells in the bottom row are computed in a similar way.

The χ^2 statistic sums the differences between observed and expected values in all squares of the table, scaled by the magnitude of the expected values, as follows:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{5.6}$$

where i ranges over rows of the table, j ranges over columns, O_{ij} is the

observed value for cell (i, j) and E_{ij} is the expected value.

One can show that the quantity X^2 is asymptotically χ^2 distributed. In other words, if the numbers are large, then X^2 has a χ^2 distribution. We will return to the issue of how good this approximation is later.

The expected frequencies E_{ij} are computed from the marginal probabilities, that is from the totals of the rows and columns converted into proportions. For example, the expected frequency for cell $(1, 1)$ (*new companies*) would be the marginal probability of *new* occurring as the first part of a bigram times the marginal probability of *companies* occurring as the second part of a bigram (multiplied by the number of bigrams in the corpus):

$$\frac{8 + 4667}{N} \times \frac{8 + 15820}{N} \times N \approx 5.2$$

That is, if *new* and *companies* occurred completely independently of each other we would expect 5.2 occurrences of *new companies* on average for a text of the size of our corpus.

The χ^2 test can be applied to tables of any size, but it has a simpler form for 2-by-2 tables: (see Exercise 5-9)

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})} \tag{5.7}$$

This formula gives the following χ^2 value for Table 5.8:

$$\frac{14307668(8 \times 14287181 - 4667 \times 15820)^2}{(8 + 4667)(8 + 15820)(4667 + 14287181)(15820 + 14287181)} \approx 1.55$$

Looking up the χ^2 distribution in the appendix, we find that at a probability level of $\alpha = 0.05$ the critical value is $\chi^2 = 3.841$. (the statistic has one degree of freedom for a 2-by-2 table). So we cannot reject the null hypothesis that *new* and *companies* occur independently of each other. Thus *new companies* is not a good candidate for a collocation.

This result is the same as we got with the t statistic. In general, for the problem of finding collocations, the differences between the t statistic and the χ^2 statistic do not seem to be large. For example, the 20 bigrams with the highest t scores in our corpus are also the 20 bigrams with the highest χ^2 scores.

However, the χ^2 test is also appropriate for large probabilities, for which the normality assumption of the t test fails. This is perhaps the reason that the χ^2 test has been applied to a wider range of problems in collocation discovery.

	cow	¬ *cow*
vache	59	6
¬ *vache*	8	570934

Table 5.9 Correspondence of *vache* and *cow* in an aligned corpus. By applying the χ^2 test to this table one can determine whether *vache* and *cow* are translations of each other.

	corpus 1	corpus 2
word 1	60	9
word 2	500	76
word 3	124	20
	...	

Table 5.10 Testing for the independence of words in different corpora using χ^2. This test can be used as a metric for corpus similarity.

One of the early uses of the χ^2 test in Statistical NLP was the identification of translation pairs in aligned corpora (Church and Gale 1991b).[5] The data in Table 5.9 (from a hypothetical aligned corpus) strongly suggest that *vache* is the French translation of English *cow*. Here, 59 is the number of aligned sentence pairs which have *cow* in the English sentence and *vache* in the French sentence etc. The χ^2 value is very high here: $\chi^2 = 456400$. So we can reject the null hypothesis that *cow* and *vache* occur independently of each other with high confidence. This pair is a good candidate for a translation pair.

An interesting application of χ^2 is as a metric for corpus similarity (Kilgarriff and Rose 1998). Here we compile an n-by-two table for a large n, for example $n = 500$. The two columns correspond to the two corpora. Each row corresponds to a particular word. This is schematically shown in Table 5.10. If the ratio of the counts are about the same (as is the case in Table 5.10, each word occurs roughly 6 times more often in corpus 1 than in corpus 2), then we cannot reject the null hypothesis that both corpora are drawn from the same underlying source. We can interpret this as a high degree of similarity. On the other hand, if the ratios vary wildly, then the X^2 score will be high and we have evidence for a high degree of dissimilarity.

5. They actually use a measure they call ϕ^2, which is X^2 multiplied by N. They do this since they are only interested in ranking translation pairs, so that assessment of significance is not important.

	H_1	H_2
$P(w^2\|w^1)$	$p = \frac{c_2}{N}$	$p_1 = \frac{c_{12}}{c_1}$
$P(w^2\|\neg w^1)$	$p = \frac{c_2}{N}$	$p_2 = \frac{c_2 - c_{12}}{N - c_1}$
c_{12} out of c_1 bigrams are $w^1 w^2$	$\mathrm{b}(c_{12};\ c_1, p)$	$\mathrm{b}(c_{12};\ c_1, p_1)$
$c_2 - c_{12}$ out of $N - c_1$ bigrams are $\neg w^1 w^2$	$\mathrm{b}(c_2 - c_{12};\ N - c_1, p)$	$\mathrm{b}(c_2 - c_{12};\ N - c_1, p_2)$

Table 5.11 How to compute Dunning's likelihood ratio test. For example, the likelihood of hypothesis H_2 is the product of the last two lines in the rightmost column.

Just as application of the t test is problematic because of the underlying normality assumption, so is application of χ^2 in cases where the numbers in the 2-by-2 table are small. Snedecor and Cochran (1989: 127) advise against using χ^2 if the total sample size is smaller than 20 or if it is between 20 and 40 and the expected value in any of the cells is 5 or less. In general, the test as described here can be inaccurate if expected cell values are small (Read and Cressie 1988), a problem we will return to below.

5.3.4 Likelihood Ratios

LIKELIHOOD RATIO

Likelihood ratios are another approach to hypothesis testing. We will see below that they are more appropriate for sparse data than the χ^2 test. But they also have the advantage that the statistic we are computing, a *likelihood ratio*, is more interpretable than the X^2 statistic. It is simply a number that tells us how much more likely one hypothesis is than the other.

In applying the likelihood ratio test to collocation discovery, we examine the following two alternative explanations for the occurrence frequency of a bigram $w^1 w^2$ (Dunning 1993):

- **Hypothesis 1.** $P(w^2|w^1) = p = P(w^2|\neg w^1)$
- **Hypothesis 2.** $P(w^2|w^1) = p_1 \neq p_2 = P(w^2|\neg w^1)$

Hypothesis 1 is a formalization of independence (the occurrence of w^2 is independent of the previous occurrence of w^1), Hypothesis 2 is a formalization of dependence which is good evidence for an interesting collocation.[6]

We use the usual maximum likelihood estimates for p, p_1 and p_2 and write c_1, c_2, and c_{12} for the number of occurrences of w^1, w^2 and $w^1 w^2$ in

6. We assume that $p_1 \gg p_2$ if Hypothesis 2 is true. The case $p_1 \ll p_2$ is rare and we will ignore it here.

the corpus:

$$p = \frac{c_2}{N} \quad p_1 = \frac{c_{12}}{c_1} \quad p_2 = \frac{c_2 - c_{12}}{N - c_1} \tag{5.8}$$

Assuming a binomial distribution:

$$\mathrm{b}(k;\ n, x) = \binom{n}{k} x^k (1-x)^{(n-k)} \tag{5.9}$$

the likelihood of getting the counts for w^1, w^2 and $w^1 w^2$ that we actually observed is then $L(H_1) = \mathrm{b}(c_{12};\ c_1, p)\mathrm{b}(c_2 - c_{12};\ N - c_1, p)$ for Hypothesis 1 and $L(H_2) = \mathrm{b}(c_{12};\ c_1, p_1)\mathrm{b}(c_2 - c_{12};\ N - c_1, p_2)$ for Hypothesis 2. Table 5.11 summarizes this discussion. One obtains the likelihoods $L(H_1)$ and $L(H_2)$ just given by multiplying the last two lines, the likelihoods of the specified number of occurrences of $w^1 w^2$ and $\neg w^1 w^2$, respectively.

The log of the likelihood ratio λ is then as follows:

$$\begin{aligned} \log \lambda &= \log \frac{L(H_1)}{L(H_2)} \\ &= \log \frac{b(c_{12}, c_1, p) b(c_2 - c_{12}, N - c_1, p)}{b(c_{12}, c_1, p_1) b(c_2 - c_{12}, N - c_1, p_2)} \\ &= \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) \\ &\quad - \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2) \end{aligned} \tag{5.10}$$

where $L(k, n, x) = x^k (1-x)^{n-k}$.

Table 5.12 shows the twenty bigrams of *powerful* which are highest ranked according to the likelihood ratio when the test is applied to the New York Times corpus. We will explain below why we show the quantity $-2 \log \lambda$ instead of λ. We consider all occurring bigrams here, including rare ones that occur less than six times, since this test works well for rare bigrams. For example, *powerful cudgels*, which occurs 2 times, is identified as a possible collocation.

One advantage of likelihood ratios is that they have a clear intuitive interpretation. For example, the bigram *powerful computers* is $e^{0.5 \times 82.96} \approx 1.3 \times 10^{18}$ times more likely under the hypothesis that *computers* is more likely to follow *powerful* than its base rate of occurrence would suggest. This number is easier to interpret than the scores of the *t* test or the χ^2 test which we have to look up in a table.

But the likelihood ratio test also has the advantage that it can be more appropriate for sparse data than the χ^2 test. How do we use the likelihood ratio for hypothesis testing? If λ is a likelihood ratio of a particular form, then the quantity $-2 \log \lambda$ is asymptotically χ^2 distributed (Mood

$-2 \log \lambda$	$C(w^1)$	$C(w^2)$	$C(w^1 w^2)$	w^1	w^2
1291.42	12593	932	150	most	powerful
99.31	379	932	10	politically	powerful
82.96	932	934	10	powerful	computers
80.39	932	3424	13	powerful	force
57.27	932	291	6	powerful	symbol
51.66	932	40	4	powerful	lobbies
51.52	171	932	5	economically	powerful
51.05	932	43	4	powerful	magnet
50.83	4458	932	10	less	powerful
50.75	6252	932	11	very	powerful
49.36	932	2064	8	powerful	position
48.78	932	591	6	powerful	machines
47.42	932	2339	8	powerful	computer
43.23	932	16	3	powerful	magnets
43.10	932	396	5	powerful	chip
40.45	932	3694	8	powerful	men
36.36	932	47	3	powerful	486
36.15	932	268	4	powerful	neighbor
35.24	932	5245	8	powerful	political
34.15	932	3	2	powerful	cudgels

Table 5.12 Bigrams of *powerful* with the highest scores according to Dunning's likelihood ratio test.

et al. 1974: 440). So we can use the values in Table 5.12 to test the null hypothesis H_1 against the alternative hypothesis H_2. For example, we can look up the value of 34.15 for *powerful cudgels* in the table and reject H_1 for this bigram on a confidence level of $\alpha = 0.005$. (The critical value (for one degree of freedom) is 7.88. See the table of the χ^2 distribution in the appendix.)

The particular form of the likelihood ratio that is required here is that of a ratio between the maximum likelihood estimate over a subpart of the parameter space and the maximum likelihood estimate over the entire parameter space. For the likelihood ratio in (5.11), this space is the space of pairs (p_1, p_2) for the probability of w^2 occurring when w^1 preceded (p_1) and w^2 occurring when a different word preceded (p_2). We get the maximum likelihood for the data we observed if we assume the maximum likelihood estimates that we computed in (5.8). The subspace is the subset of cases for which $p_1 = p_2$. Again, the estimate in (5.8) gives us the maximum

ratio	1990	1989	w^1	w^2
0.0241	2	68	Karim	Obeid
0.0372	2	44	East	Berliners
0.0372	2	44	Miss	Manners
0.0399	2	41	17	earthquake
0.0409	2	40	HUD	officials
0.0482	2	34	EAST	GERMANS
0.0496	2	33	Muslim	cleric
0.0496	2	33	John	Le
0.0512	2	32	Prague	Spring
0.0529	2	31	Among	individual

Table 5.13 Damerau's frequency ratio test. Ten bigrams that occurred twice in the 1990 New York Times corpus, ranked according to the (inverted) ratio of relative frequencies in 1989 and 1990.

likelihood over the subspace given the data we observed. It can be shown that if λ is a ratio of two likelihoods of this type (one being the maximum likelihood over the subspace, the other over the entire space), then $-2 \log \lambda$ is asymptotically χ^2 distributed. "Asymptotically" roughly means "if the numbers are large enough". Whether or not the numbers are large enough in a particular case is hard to determine, but Dunning has shown that for small counts the approximation to χ^2 is better for the likelihood ratio in (5.11) than, for example, for the X^2 statistic in (5.6). Therefore, the likelihood ratio test is in general more appropriate than Pearson's χ^2 test for collocation discovery.[7]

RELATIVE FREQUENCIES

Relative frequency ratios So far we have looked at evidence for collocations within one corpus. Ratios of *relative frequencies* between two or more different corpora can be used to discover collocations that are characteristic of a corpus when compared to other corpora (Damerau 1993). Although ratios of relative frequencies do not fit well into the hypothesis testing paradigm, we treat them here since they can be interpreted as likelihood ratios.

Table 5.13 shows ten bigrams that occur exactly twice in our reference corpus (the 1990 New York Times corpus). The bigrams are ranked according to the ratio of their relative frequencies in our 1990 reference corpus

7. However, even $-2 \log \lambda$ is not approximated well by χ^2 if the expected values in the 2-by-2 contingency table are less than 1.0 (Read and Cressie 1988; Pedersen 1996).

versus their frequencies in a 1989 corpus (again drawn from the months August through November). For example, *Karim Obeid* occurs 68 times in the 1989 corpus. So the relative frequency ratio r is:

$$r = \frac{\frac{2}{14307668}}{\frac{68}{11731564}} \approx 0.024116$$

The bigrams in the Table are mostly associated with news items that were more prevalent in 1989 than in 1990: The Muslim cleric Sheik Abdul Karim Obeid (who was abducted in 1989), the disintegration of communist Eastern Europe (*East Berliners*, *EAST GERMANS*, *Prague Spring*), the novel *The Russia House* by *John Le Carre*, a scandal in the Department of Housing and Urban Development (HUD), and the October 17 earthquake in the San Francisco Bay Area. But we also find artefacts like *Miss Manners* (whose column the New York Times News Wire stopped carrying in 1990) and *Among individual*. The reporter Phillip H. Wiggins liked to use the latter phrase for his stock market reports (*Among individual Big Board issues* ...), but he stopped writing for the *Times* in 1990.

The examples show that frequency ratios are mainly useful to find *subject-specific* collocations. The application proposed by Damerau is to compare a general text with a subject-specific text. Those words and phrases that on a relative basis occur most often in the subject-specific text are likely to be part of the vocabulary that is specific to the domain.

Exercise 5-4
Identify the most significantly non-independent bigrams according to the t test in a corpus of your choice.

Exercise 5-5
It is a coincidence that the t value for *new companies* is close to 1.0. Show this by computing the t value of *new companies* for a corpus with the following counts. $C(\textit{new}) = 30,000$, $C(\textit{companies}) = 9,000$, $C(\textit{new companies}) = 20$, and corpus size $N = 15,000,000$.

Exercise 5-6
We can also improve on the method in the previous section (Section 5.2) by taking into account variance. In fact, Smadja does this and the algorithm described in (Smadja 1993) therefore bears some similarity to the t test.

Compute the t statistic in equation (5.3) for possible collocations by substituting mean and variance as computed in Section 5.2 for $\bar{x}$ and s^2 and a) assuming $\mu = 0$, and b) assuming $\mu = \text{round}(\bar{x})$, that is, the closest integer. Note that we are not testing for bigrams here, but for collocations of word pairs that occur at any fixed small distance.

Exercise 5-7
As we pointed out above, almost all bigrams occur significantly more often than

chance if a stop list is used for prefiltering. Verify that there is a large proportion of bigrams that occur less often than chance if we do not filter out function words.

Exercise 5-8
Apply the t test of differences to a corpus of your choice. Work with the following word pairs or with word pairs that are appropriate for your corpus: *man* / *woman*, *blue* / *green*, *lawyer* / *doctor*.

Exercise 5-9
Derive (5.7) from (5.6).

Exercise 5-10
Find terms that distinguish best between first and second part of a corpus of your choice.

Exercise 5-11
Repeat the above exercise with random selection. Now you should find that fewer terms are significant. But some still are. Why? Shouldn't there be no differences between corpora drawn from the same source? Do this exercise for different significance levels.

Exercise 5-12
Compute a measure of corpus similarity between two corpora of your choice.

Exercise 5-13
Kilgarriff and Rose's corpus similarity measure can also be used for assessing corpus homogeneity. This is done by constructing a series of random divisions of the corpus into a pair of subcorpora. The test is then applied to each pair. If most of the tests indicated similarity, then it is a homogeneous corpus. Apply this test to a corpus of your choice.

5.4 Mutual Information

POINTWISE MUTUAL INFORMATION

An information-theoretically motivated measure for discovering interesting collocations is *pointwise mutual information* (Church et al. 1991; Church and Hanks 1989; Hindle 1990). Fano (1961: 27–28) originally defined mutual information between particular events x' and y', in our case the occurrence of particular words, as follows:

$$I(x', y') = \log_2 \frac{P(x'y')}{P(x')P(y')} \tag{5.11}$$

$$= \log_2 \frac{P(x'|y')}{P(x')} \tag{5.12}$$

$$= \log_2 \frac{P(y'|x')}{P(y')} \tag{5.13}$$

$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13478	20	into	them
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

Table 5.14 Finding collocations: Ten bigrams that occur with frequency 20, ranked according to mutual information.

This type of mutual information, which we introduced in Section 2.2.3, is roughly a measure of how much one word tells us about the other, a notion that we will make more precise shortly.

In information theory, mutual information is more often defined as holding between *random variables*, not *values of random variables* as we have defined it here (see the standard definition in Section 2.2.3). We will see below that these two types of mutual information are quite different creatures.

When we apply this definition to the 10 collocations from Table 5.6, we get the same ranking as with the t test (see Table 5.14). As usual, we use maximum likelihood estimates to compute the probabilities, for example:

$$I(\textit{Ayatollah}, \textit{Ruhollah}) = \log_2 \frac{\frac{20}{14307668}}{\frac{42}{14307668} \times \frac{20}{14307668}} \approx 18.38$$

So what exactly is (pointwise) mutual information, $I(x', y')$, a measure of? Fano writes about definition (5.12):

> The amount of information provided by the occurrence of the event represented by [y'] about the occurrence of the event represented by [x'] is defined as [(5.12)].

For example, the mutual information measure tells us that the amount of information we have about the occurrence of *Ayatollah* at position i in the corpus increases by 18.38 bits if we are told that *Ruhollah* occurs at position $i + 1$. Or, since (5.12) and (5.13) are equivalent, it also tells us that the amount of information we have about the occurrence of *Ruhollah* at position $i + 1$ in the corpus increases by 18.38 bits if we are told that *Ayatollah*

	chambre	¬ *chambre*	MI	χ^2
house	31,950	12,004		
¬ *house*	4,793	848,330	4.1	553610
	communes	¬ *communes*		
house	4,974	38,980		
¬ *house*	441	852,682	4.2	88405

Table 5.15 Correspondence of *chambre* and *house* and *communes* and *house* in the aligned Hansard corpus. Mutual information gives a higher score to (*communes*,*house*), while the χ^2 test gives a higher score to the correct translation pair (*chambre*,*house*).

occurs at position i. We could also say that our uncertainty is reduced by 18.38 bits. In other words, we can be much more certain that *Ruhollah* will occur next if we are told that *Ayatollah* is the current word.

Unfortunately, this measure of "increased information" is in many cases not a good measure of what an interesting correspondence between two events is, as has been pointed out by many authors. (We base our discussion here mainly on (Church and Gale 1991b) and (Maxwell III 1992).) Consider the two examples in Table 5.15 of counts of word correspondences between French and English sentences in the Hansard corpus, an aligned corpus of debates of the Canadian parliament (the table is similar to Table 5.9). The reason that *house* frequently appears in translations of French sentences containing *chambre* and *communes* is that the most common use of *house* in the Hansard is the phrase *House of Commons* which corresponds to *Chambre de communes* in French. But it is easy to see that *communes* is a worse match for *house* than *chambre* since most occurrences of *house* occur without *communes* on the French side. As shown in the table, the χ^2 test is able to infer the correct correspondence whereas mutual information gives preference to the incorrect pair (*communes*,*house*).

We can explain the difference between the two measures easily if we look at definition (5.12) of mutual information and compare $I(\textit{chambre}, \textit{house})$ and $I(\textit{communes}, \textit{house})$:

$$\begin{aligned} & \log \frac{P(\textit{house}|\textit{chambre})}{P(\textit{house})} = \log \frac{\frac{31950}{31950+4793}}{P(\textit{house})} \approx \log \frac{0.87}{P(\textit{house})} \\ < \quad & \log \frac{0.92}{P(\textit{house})} \approx \log \frac{\frac{4974}{4974+441}}{P(\textit{house})} = \log \frac{P(\textit{house}|\textit{communes})}{P(\textit{house})} \end{aligned}$$

The word *communes* in the French makes it more likely that *house* occurred in the English than *chambre* does. The higher mutual information value for

I_{1000}	w^1	w^2	w^1w^2	bigram	I_{23000}	w^1	w^2	w^1w^2	bigram
16.95	5	1	1	Schwartz eschews	14.46	106	6	1	Schwartz eschews
15.02	1	19	1	fewest visits	13.06	76	22	1	FIND GARDEN
13.78	5	9	1	FIND GARDEN	11.25	22	267	1	fewest visits
12.00	5	31	1	Indonesian pieces	8.97	43	663	1	Indonesian pieces
9.82	26	27	1	Reds survived	8.04	170	1917	6	marijuana growing
9.21	13	82	1	marijuana growing	5.73	15828	51	3	new converts
7.37	24	159	1	doubt whether	5.26	680	3846	7	doubt whether
6.68	687	9	1	new converts	4.76	739	713	1	Reds survived
6.00	661	15	1	like offensive	1.95	3549	6276	6	must think
3.81	159	283	1	must think	0.41	14093	762	1	like offensive

Table 5.16 Problems for Mutual Information from data sparseness. The table shows ten bigrams that occurred once in the first 1000 documents in the reference corpus ranked according to mutual information score in the first 1000 documents (left half of the table) and ranked according to mutual information score in the entire corpus (right half of the table). These examples illustrate that a large proportion of bigrams are not well characterized by corpus data (even for large corpora) and that mutual information is particularly sensitive to estimates that are inaccurate due to sparseness.

communes reflects the fact that *communes* causes a larger decrease in uncertainty here. But as the example shows decrease in uncertainty does not correspond well to what we want to measure. In contrast, the χ^2 is a direct test of probabilistic dependence, which in this context we can interpret as the degree of association between two words and hence as a measure of their quality as translation pairs and collocations.

Table 5.16 shows a second problem with using mutual information for finding collocations. We show ten bigrams that occur exactly once in the first 1000 documents of the reference corpus and their mutual information score based on the 1000 documents. The right half of the table shows the mutual information score based on the entire reference corpus (about 23,000 documents).

The larger corpus of 23,000 documents makes some better estimates possible, which in turn leads to a slightly better ranking. The bigrams *marijuana growing* and *new converts* (arguably collocations) have moved up and *Reds survived* (definitely not a collocation) has moved down. However, what is striking is that even after going to a 10 times larger corpus 6 of the bigrams still only occur once and, as a consequence, have inaccurate maximum likelihood estimates and artificially inflated mutual information scores. All 6 are not collocations and we would prefer a measure which

ranks them accordingly.

None of the measures we have seen works very well for low-frequency events. But there is evidence that sparseness is a particularly difficult problem for mutual information. To see why, notice that mutual information is a log likelihood ratio of the probability of the bigram $P(w^1 w^2)$ and the product of the probabilities of the individual words $P(w^1)P(w^2)$. Consider two extreme cases: perfect dependence of the occurrences of the two words (they only occur together) and perfect independence (the occurrence of one does not give us any information about the occurrence of the other). For perfect dependence we have:

$$I(x,y) = \log \frac{P(xy)}{P(x)P(y)} = \log \frac{P(x)}{P(x)P(y)} = \log \frac{1}{P(y)}$$

That is, among perfectly dependent bigrams, as they get rarer, their mutual information *increases*.

For perfect independence we have:

$$I(x,y) = \log \frac{P(xy)}{P(x)P(y)} = \log \frac{P(x)P(y)}{P(x)P(y)} = \log 1 = 0$$

We can say that mutual information is a good measure of independence. Values close to 0 indicate independence (independent of frequency). But it is a bad measure of dependence because for dependence the score depends on the frequency of the individual words. Other things being equal, bigrams composed of low-frequency words will receive a higher score than bigrams composed of high-frequency words. That is the opposite of what we would want a good measure to do since higher frequency means more evidence and we would prefer a higher rank for bigrams for whose interestingness we have more evidence. One solution that has been proposed for this is to use a cutoff and to only look at words with a frequency of at least 3. However, such a move does not solve the underlying problem, but only ameliorates its effects.

Since pointwise mutual information does not capture the intuitive notion of an interesting collocation very well, it is often not used when it is made available in practical applications (Fontenelle et al. 1994: 81) or it is redefined as $C(w^1 w^2)I(w^1, w^2)$ to compensate for the bias of the original defintion in favor of low-frequency events (Fontenelle et al. 1994: 72, Hodges et al. 1996).

As we mentioned earlier, the definition of mutual information used here is common in corpus linguistic studies, but is less common in Information Theory. Mutual information in Information Theory refers to the *expectation*

EXPECTATION

symbol	definition	terminology: current use	terminology: Fano
$I(x, y)$	$\log \frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\,\mathrm{p}(y)}$	pointwise mutual information	mutual information
$I(X;\ Y)$	$E \log \frac{\mathrm{p}(X,Y)}{\mathrm{p}(X)\,\mathrm{p}(Y)}$	mutual information	average MI expectation of MI

Table 5.17 Different definitions of *mutual information* in (Cover and Thomas 1991) and (Fano 1961).

of the quantity that we have used in this section:

$$I(X;\ Y) = E_{\mathrm{p}(x,y)} \log \frac{\mathrm{p}(X,Y)}{\mathrm{p}(X)\,\mathrm{p}(Y)}$$

The definition we have used in this chapter is an older one, termed pointwise mutual information (see Section 2.2.3, Fano 1961: 28, and Gallager 1968). Table 5.17 summarizes the older and newer naming conventions. One quantity is the expectation of the other, so the two types of mutual information are quite different.

The example of mutual information demonstrates what should be self-evident: it is important to check what a mathematical concept is a formalization of. The notion of pointwise mutual information that we have used here $\left(\log \frac{\mathrm{p}(w^1 w^2)}{\mathrm{p}(w^1)\,\mathrm{p}(w^2)}\right)$ measures the reduction of uncertainty about the occurrence of one word when we are told about the occurrence of the other. As we have seen, such a measure is of limited utility for acquiring the types of linguistic properties we have looked at in this section.

Exercise 5-14
Justeson and Katz's part-of-speech filter in Section 5.1 can be applied to any of the other methods of collocation discovery in this chapter. Pick one and modify it to incorporate a part-of-speech filter. What advantages does the modified method have?

Exercise 5-15
Design and implement a collocation discovery tool for a translator's workbench. Pick either one method or a combination of methods that the translator can choose from.

Exercise 5-16
Design and implement a collocation discovery tool for a lexicographer's workbench. Pick either one method or a combination of methods that the lexicographer can choose from.

Exercise 5-17

Many news services tag references to companies in their news stories. For example, all references to the *General Electric Company* would be tagged with the same tag regardless of which variant of the name is used (e.g., *GE*, *General Electric*, or *General Electric Company*). Design and implement a collocation discovery tool for finding company names. How could one partially automate the process of identifying variants?

5.5 The Notion of Collocation

The notion of collocation may be confusing to readers without a background in linguistics. We will devote this section to discussing in more detail what a collocation is.

There are actually different definitions of the notion of collocation. Some authors in the computational and statistical literature define a collocation as two or more *consecutive* words with a special behavior, for example Choueka (1988):

> [A collocation is defined as] a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components.

Most of the examples we have presented in this chapter also assumed adjacency of words. But in most linguistically oriented research, a phrase can be a collocation even if it is not consecutive (as in the example *knock ... door*). The following criteria are typical of linguistic treatments of collocations (see for example Benson (1989) and Brundage et al. (1992)), non-compositionality being the main one we have relied on here.

- **Non-compositionality.** The meaning of a collocation is not a straightforward composition of the meanings of its parts. Either the meaning is completely different from the free combination (as in the case of idioms like *kick the bucket*) or there is a connotation or added element of meaning that cannot be predicted from the parts. For example, *white wine*, *white hair* and *white woman* all refer to slightly different colors, so we can regard them as collocations.

- **Non-substitutability.** We cannot substitute near-synonyms for the components of a collocation. For example, we can't say *yellow wine* instead of *white wine* even though *yellow* is as good a description of the color of white wine as *white* is (it is kind of a yellowish white).

strength	power
to build up ∼	to assume ∼
to find ∼	emergency ∼
to save ∼	discretionary ∼
to sap somebody's ∼	∼ over [several provinces]
brute ∼	supernatural ∼
tensile ∼	to turn off the ∼
the ∼ to [do X]	the ∼ to [do X]
[our staff was] at full ∼	the balance of ∼
on the ∼ of [your recommendation]	fire ∼

Table 5.18 Collocations in the BBI Combinatory Dictionary of English.

- **Non-modifiability.** Many collocations cannot be freely modified with additional lexical material or through grammatical transformations. This is especially true for frozen expressions like idioms. For example, we can't modify *frog* in *to get a frog in one's throat* into *to get an ugly frog in one's throat* although usually nouns like *frog* can be modified by adjectives like *ugly*. Similarly, going from singular to plural can make an idiom ill-formed, for example in *people as poor as church mice*.

A nice way to test whether a combination is a collocation is to translate it into another language. If we cannot translate the combination word by word, then that is evidence that we are dealing with a collocation. For example, translating *make a decision* into French one word at a time we get *faire une décision* which is incorrect. In French we have to say *prendre une décision*. So that is evidence that *make a decision* is a collocation in English.

Some authors have generalized the notion of collocation even further and included cases of words that are strongly associated with each other, but do not necessarily occur in a common grammatical unit and with a particular order, cases like *doctor – nurse* or *plane – airport*. It is probably best to restrict collocations to the narrower sense of grammatically bound elements that occur in a particular order and use the terms *association* and *co-occurrence* for the more general phenomenon of words that are likely to be used in the same context.

ASSOCIATION
CO-OCCURRENCE

It is instructive to look at the types of collocations that a purely linguistic analysis of text will discover if plenty of time and person power is available so that the limitations of statistical analysis and computer technology need be of no concern. An example of such a purely linguistic analysis is the BBI Combinatory Dictionary of English (Benson et al. 1993). In Table 5.18, we

show some of the collocations (or combinations as the dictionary prefers to call them) of *strength* and *power* that the dictionary lists.[8] We can see immediately that a wider variety of grammatical patterns is considered here (in particular patterns involving prepositions and particles). Naturally, the quality of the collocations is also higher than computer-generated lists – as we would expect from a manually produced compilation.

We conclude our discussion of the concept of collocation by going through some subclasses of collocations that deserve special mention.

LIGHT VERBS

Verbs with little semantic content like *make*, *take* and *do* are called *light verbs* in collocations like *make a decision* or *do a favor*. There is hardly anything about the meaning of *make*, *take* or *do* that would explain why we have to say *make a decision* instead of *take a decision* and *do a favor* instead of *make a favor*, but for many computational purposes the correct light verb for combination with a particular noun must be determined and thus acquired from corpora if this information is not available in machine-readable dictionaries. Dras and Johnson (1996) examine one approach to this problem.

VERB PARTICLE CONSTUCTIONS
PHRASAL VERBS

Verb particle constuctions or *phrasal verbs* are an especially important part of the lexicon of English. Many verbs in English like *to tell off* and *to go down* consist of a combination of a main verb and a particle. These verbs often correspond to a single lexeme in other languages (*réprimander*, *descendre* in French). This type of construction is a good example of a collocation with often non-adjacent words.

PROPER NOUNS
PROPER NAMES

Proper nouns (also called *proper names*) are usually included in the category of collocations in computational work although they are quite different from lexical collocations. They are most amenable to approaches that look for fixed phrases that reappear in exactly the same form throughout a text.

TERMINOLOGICAL EXPRESSIONS

Terminological expressions refer to concepts and objects in technical domains. Although they are often fairly compositional (e.g., *hydraulic oil filter*), it is still important to identify them to make sure that they are treated consistently throughout a technical text. For example, when translating a manual, we have to make sure that all instances of *hydraulic oil filter* are translated by the same term. If two different translations are used (even if they have the same meaning in some sense), the reader of the translated manual would get confused and think that two different entities are being described.

As a final example of the wide range of phenomena that the term colloca-

8. We cannot show collocations of *strong* and *powerful* because these adjectives are not listed as entries in the dictionary.

tion is applied to, let us point to the many different degrees of invariability that a collocation can show. At one extreme of the spectrum we have usage notes in dictionaries that describe subtle differences in usage between near-synonyms like *answer* and *reply* (*diplomatic answer* vs. *stinging reply*). This type of collocation is important for generating text that sounds natural, but getting a collocation wrong here is less likely to lead to a fatal error. The other extreme are completely frozen expressions like proper names and idioms. Here there is just one way of saying things and any deviation will completely change the meaning of what is said. Luckily, the less compositional and the more important a collocation, the easier it often is to acquire it automatically.

5.6 Further Reading

See (Stubbs 1996) for an in-depth discussion of the British tradition of "empiricist" linguistics.

The t test is covered in most general statistics books. Standard references are (Snedecor and Cochran 1989: 53) and (Moore and McCabe 1989: 541). Weinberg and Goldberg (1990: 306) and Ramsey and Schafer (1997) are more accessible for students with less mathematical background. These books also cover the χ^2 test, but not some of the other more specialized tests that we discuss here.

One of the first publications on the discovery of collocations was (Church and Hanks 1989), later expanded to (Church et al. 1991). The authors drew attention to an emerging type of corpus-based dictionary (Sinclair 1995) and developed a program of computational lexicography that combines corpus evidence, computational methods and human judgement to build more comprehensive dictionaries that better reflect actual language use.

There are a number of ways lexicographers can benefit from automated processing of corpus data. A lexicographer writes a dictionary entry after looking at a potentially large number of examples of a word. If the examples are automatically presorted according to collocations and other criteria (for example, the topic of the text), then this process can be made much more efficient. For example, phrasal verbs are sometimes neglected in dictionaries because they are not separate words. A corpus-based approach will make their importance evident to the lexicographer. In addition, a balanced corpus will reveal which of the uses are most frequent and hence most important for the likely user of a dictionary. Difference tests like the t test are useful for writing usage notes and for writing ac-

curate definitions that reflect differences in usage between words. Some of these techniques are being used for the next generation of dictionaries (Fontenelle et al. 1994).

Eventually, a new form of dictionary could emerge from this work, a kind of dictionary-cum-corpus in which dictionary entry and corpus evidence support each other and are organized in a coherent whole. The COBUILD dictionary already has some of these characteristics (Sinclair 1995). Since space is less of an issue with electronic dictionaries plenty of corpus examples can be integrated into a dictionary entry for the interested user.

What we have said about the value of statistical corpus analysis for monolingual dictionaries applies equally to bilingual dictionaries, at least if an aligned corpus is available (Smadja et al. 1996).

Another important application of collocations is Information Retrieval (IR). Accuracy of retrieval can be improved if the similarity between a user query and a document is determined based on common collocations (or phrases) instead of common words (Fagan 1989; Evans et al. 1991; Strzalkowski 1995; Mitra et al. 1997). See Lewis and Jones (1996) and Krovetz (1991) for further discussion of the question of using collocation discovery and NLP in Information Retrieval and Nevill-Manning et al. (1997) for an alternative non-statistical approach to using phrases in IR. Steier and Belew (1993) present an interesting study of how the treatment of phrases (for example, for phrase weighting) should change as we move from a subdomain to a general domain. For example, *invasive procedure* is completely compositional and a less interesting collocation in the subdomain of medical articles, but becomes interesting and non-compositional when "exported" to a general collection that is a mixture of many specialized domains.

Two other important applications of collocations, which we will just mention, are natural language generation (Smadja 1993) and cross-language information retrieval (Hull and Grefenstette 1998).

An important area that we haven't been able to cover is the discovery of proper nouns, which can be regarded as a kind of collocation. Proper nouns cannot be exhaustively covered in dictionaries since new people, places, and other entities come into existence and are named all the time. Proper nouns also present their own set of challenges: co-reference (How can we tell that IBM and International Bureau Machines refer to the same entity?), disambiguation (When does AMEX refer to the American Exchange, when to American Express?), and classification (Is this new entity that the text refers to the name of a person, a location or a company?). One of the earliest studies on this topic is (Coates-Stephens 1993). McDonald (1995)

focuses on lexicosemantic patterns that can be used as cues for proper noun detection and classification. Mani and MacMillan (1995) and Paik et al. (1995) propose ways of classifying proper nouns according to type.

z SCORE

One frequently used measure for interestingness of collocations that we did not cover is the *z score*, a close relative of the t test. It is used in several software packages and workbenches for text analysis (Fontenelle et al. 1994; Hawthorne 1994). The z score should only be applied when the variance is known, which arguably is not the case in most Statistical NLP applications.

Fisher's exact test is another statistical test that can be used for judging how unexpected a set of observations is. In contrast to the t test and the χ^2 test, it is appropriate even for very small counts. However, it is hard to compute, and it is not clear whether the results obtained in practice are much different from, for example, the χ^2 test (Pedersen 1996).

Yet another approach to discovering collocations is to search for points in the word stream with either low or high uncertainty as to what the next (or previous) word will be. Points with high uncertainty are likely to be phrase boundaries, which in turn are candidates for points where a collocation may start or end, whereas points with low uncertainty are likely to be located within a collocation. See (Evans and Zhai 1996) and (Shimohata et al. 1997) for two approaches that use this type of information for finding phrases and collocations.