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 stiffbreeze* but not *??a stiffwind* (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
TERMINOLOGY EXTRACTION

There is considerable overlap between the concept of *collocation* and notions like *term, technical term,* and *terminological phrase.* As these names suggest, the latter three are commonly used when collocations are extracted from technical domains (in a process called *terminology extraction).* The reader should 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.

CONTEXTUAL THEORY OF MEANING

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 a number of approaches to finding collocations: 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

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

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

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. 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

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

w	$C(strong, w)$	*w*	$C(powerful, w)$
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 ***'powerfulw.'***

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. 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.

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.***

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.

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.

$$s^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} \tag{5.2}$$

where n is the number of times the two words co-occur, d_i is the offset for co-occurrence i, and d is the sample mean of the offsets. 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 *sample deviation* $s = \sqrt{s^2}$, the square root of the variance, to assess how variable the offset between two words is. The deviation for the four examples of *knocked / door* in the above case is 1.15:

SAMPLE DEVIATION

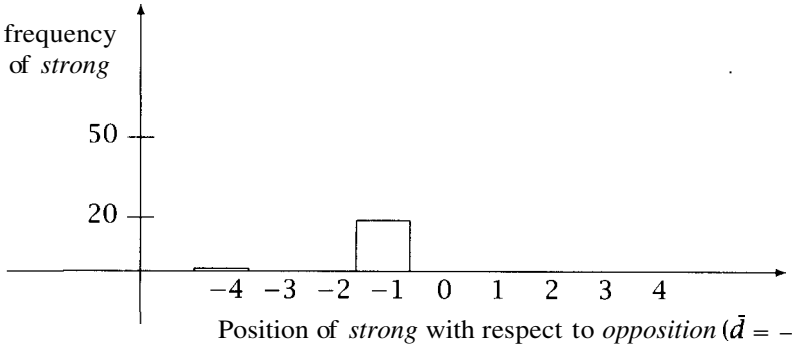
$$s = \sqrt{\frac{1}{3}((3-4.0)^2 + (3-4.0)^2 + (5-4.0)^2 + (5-4.0)^2)} \approx 1.15$$

The mean and 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 deviation. A low deviation means that the two words usually occur at about the same distance. Zero deviation means that the two words always occur at exactly the same distance.

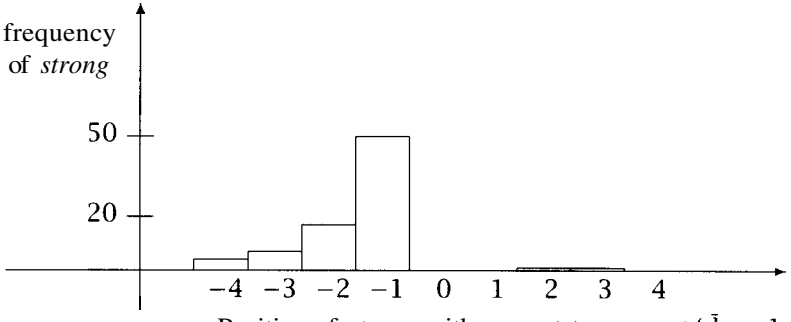
We can also explain the information that variance gets at in terms of peaks in the distribution of one word with respect to another. Figure 5.2 shows the three cases we are interested in. The distribution of *strong* with respect to *opposition* has one clear peak at position -1 (corresponding to the phrase *strong opposition).* Therefore the variance of *strong* with respect to *opposition* is small *(s = 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.*

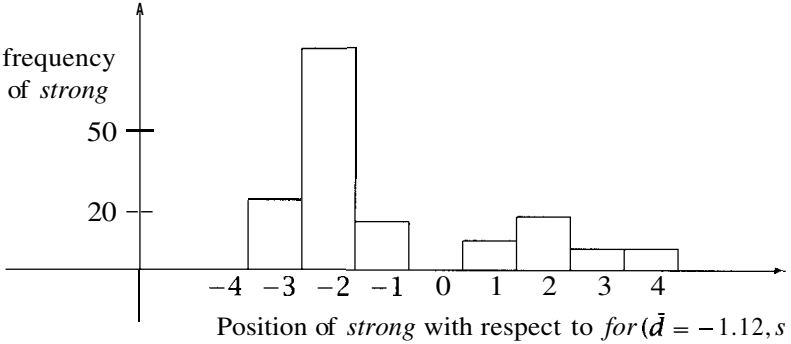
Moving on to the second diagram in figure 5.2, the distribution of *strong* with respect to *support* is drawn out, with several negative positions having large counts. For example, the count of approximately *20*



Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).



Position of *strong* with respect to *support* ($\bar{d} = -1.45, s = 1.07$).



Position of *strong* with respect to *for* ($\bar{d} = -1.12, s = 2.15$).

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

s	d	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. Sample deviation s and sample mean d of the distances between 12 word pairs.

at position −2 is due to uses like ***strong leftist support*** and ***strong business support.*** Because of this greater variability we get a higher s (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 deviation of $s = 2.15$ indicates this variability. 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 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 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 deviation indicates that the two words of the pair stand in no interesting relationship as demonstrated by the four high-variance 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

NULL HYPOTHESIS

SIGNIFICANCE LEVEL

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]

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.

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:

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

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 that we can often expect to achieve greater levels of significance.

where R 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 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.

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

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

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.

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

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

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 14,307,668 bigrams in our corpus. So, for the sample, we have that the sample mean is: $\bar{x} = \frac{8}{14307668} = 5.591 \times 10^{-7}$. 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.)

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

t	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w1	w2
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.

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

t	$C(w)$	$C(strong\ w)$	$C(powerful\ w)$	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
		13	0	criticism

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

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.

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

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 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).

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	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, 4,667 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.

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 $C(new) = 15{,}828$, $C(companies) = 4{,}675$, $C(new\ companies) = 8$, 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 not being *new* and the second token 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.

One of the early uses of the χ^2 test in Statistical NLP was the identifi-