	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.

cation 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^1w^2	$b(c_{12}; c_1, p)$	$b(c_{12}; c_1, p_1)$
$c_2 - c_{12}$ out of $N - c_1$ bigrams are $\neg w^1 w^2$	$b(c_2 - c_{12}; N - c_1, p)$	$b(c_2 - c_{12}; N - c_1, p_2)$

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.

5.3.4 Likelihood ratios

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.

LIKELIHOOD RATIO

In applying the likelihood ratio test to collocation discovery, we examine the following two alternative explanations for the occurrence frequency of a bigram w^1w^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 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.

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^1w^2 in 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:

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

the likelihood of getting the counts for w^1, w^2 and w^1w^2 that we actually observed is then $L(H_1) = b(c_{12}; c_1, p)b(c_2 - c_{12}; N - c_1, p)$ for Hypothesis 1 and $L(H_2) = b(c_{12}; c_1, p_1)b(c_2 - c_{12}; N - c_1, p_2)$ 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^1w^2 and $\neg w^1 w^2$, respectively.

The log of the likelihood ratio A 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 (for which N = 14,307,668). We will explain below why we show the quantity $-2\log\lambda$ instead of A. 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 $e^{0.5\times 82.96} = 1.3 \times 10^{18}$ 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.

$-2 \log \lambda$	$C(w^1)$	$C(w^2)$	$C(w^1 w^2)$	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.

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 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 en-

tire parameter space. For the likelihood ratio in (5.11), the entire 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 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 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 table 5.13 are mostly associated with news items that were more prevalent in 1989 than in 1990: The Muslim cleric Sheik Abdul

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).

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.

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* newswire 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. $C(new) = 30{,}000$, $C(companies) = 9{,}000$, $C(new\ companies) = 20$, and corpus size $N = 15{,}000{,}000$.

Exercise 5.6 [*]

We can improve on the method in 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 $\mu = \text{round}(\%)$ 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 equation (5.7) from equation (5.6).

Exercise 5.10 [**]

Find terms that distinguish best between the 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.

$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	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.

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:

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

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

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

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

	chambre	¬ chambre	MI	χ^2
house	31,950	12,004		
¬ house	4793	848,330	4.1	553610
	communes	¬ communes		
house	4974	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*).

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:

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

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 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 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 the quantities I*(chambre,house)* and *I(communes,house)*:

$$\log \frac{P(house|chambre)}{P(house)} = \log \frac{\frac{31950}{31950+4793}}{P(house)} \approx \log \frac{0.87}{P(house)}$$

$$< \log \frac{0.92}{P(house)} \approx \log \frac{\frac{4974}{4974+441}}{P(house)} = \log \frac{P(house|communes)}{P(house)}$$

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 *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

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.

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

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

Symbol	Definition	Current use	Fano
$I(x,y)$	$\log \frac{p(x,y)}{p(x)\,p(y)}$	pointwise mutual information	mutual information
$I(X;Y)$	$E \log \frac{p(X,Y)}{p(X)\,p(Y)}$	mutual information	average MI/expectation of MI

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

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 $C(w^1w^2)I(w^1,w^2)$ to compensate for the bias of the original definition 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 of the quantity that we have used in this section:

EXPECTATION

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

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 ($\log \frac{p(w^1 w^2)}{p(w^1)\,p(w^2)}$) 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 other words for for the components of a collocation even if, in context, they have the same meaning. 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).

- **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 decision*** which is incorrect. In French we have to say ***prendre une decision***. So that is evidence that ***make a decision*** is a collocation in English.

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 for the words ***strength*** and ***power.***

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.

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

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 CONSTRUCTIONS
PHRASAL VERBS

Verb particle constructions 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 *(reprimander, descendre* in French). This type of construction is a good example of a collocation with often non-adjacent words.

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 or phrases 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 could get confused and think that two different entities are being described.

As a final example of the wide range of phenomena that the term collocation 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 (diplomaticanswer* 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 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 accurate 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 Business 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.

One frequently used measure for interestingness of collocations that

z SCORE

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.