Collocations

Introduction

- A COLLOCATION is an expression consisting of two or more words that correspond to some conventional way of saying thing
- Collocations of a given word are statements of the habitual or customary place of that word
- Why we say *a stiff breeze* but not *a stiff wind*

Introduction

- 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

Introduction

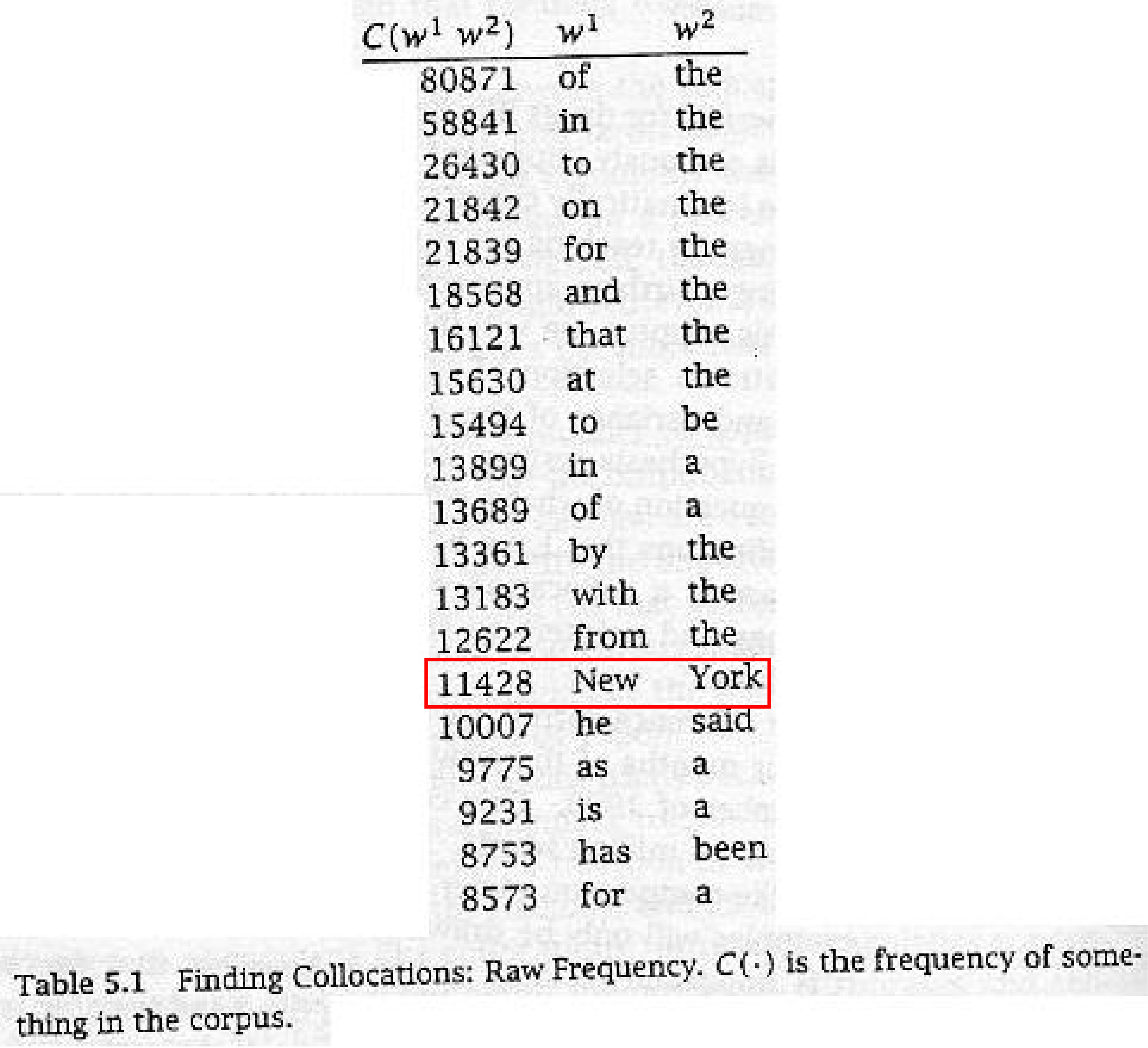
- Idioms are the most extreme examples of non-compositionality
- Idioms like *to kick the bucket* or *to hear it through the grapevine* only have an indirect historical relationship to the meanings of the parts of the expression
- Halliday’s example of strong vs. powerful tea. It is a convention in English to talk about *strong tea*, not *powerful tea*

Introduction

- Finding collocations: frequency, mean and variance, hypothesis testing, and mutual information
- The reference corpus consists of four months of the *New York Times* newswire: 1990/08 ～ 11. 115 Mb of text and 14 million words

Frequency

- The simplest method for finding collocations in a text corpus is counting
- Just selecting the most frequently occurring bigrams is not very interesting as is shown in table 5.1

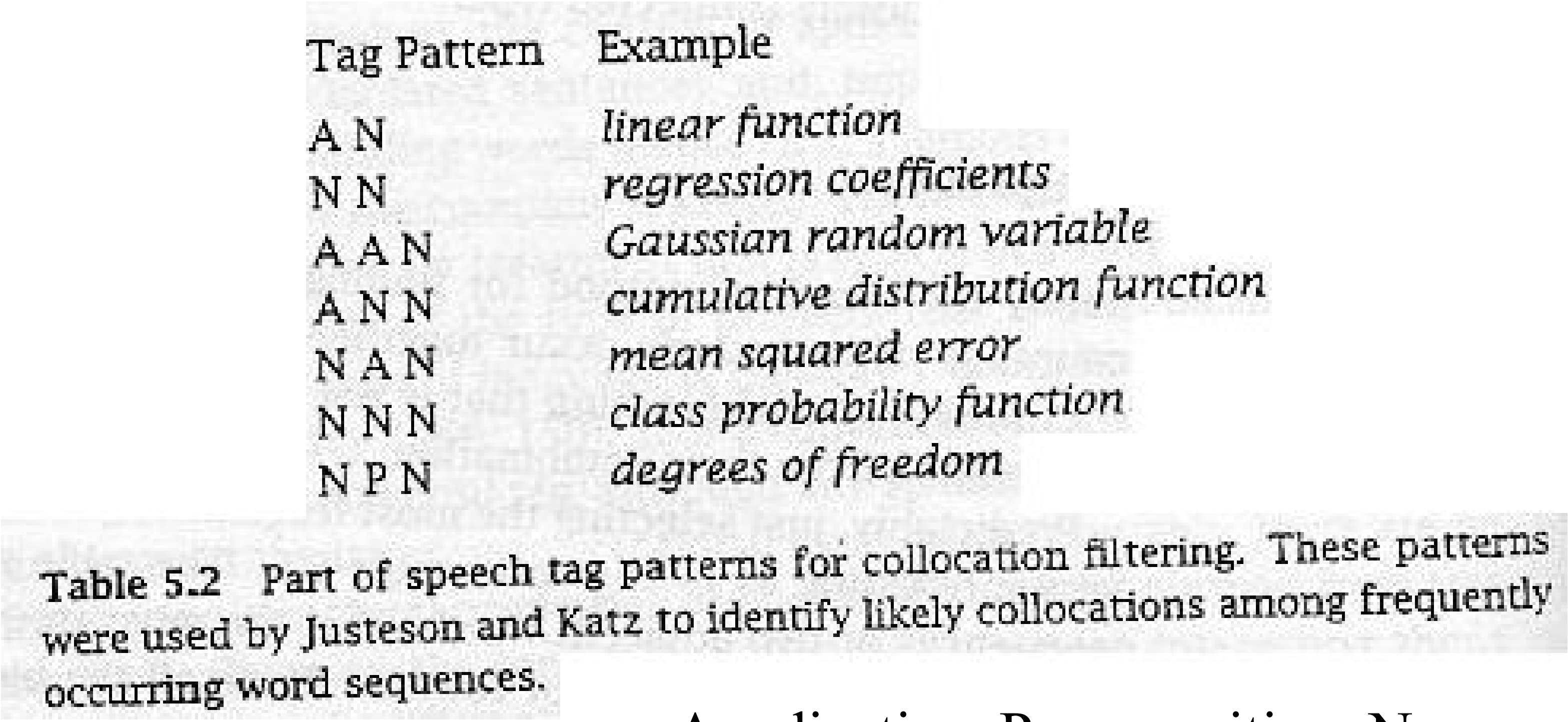


$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

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

Frequency

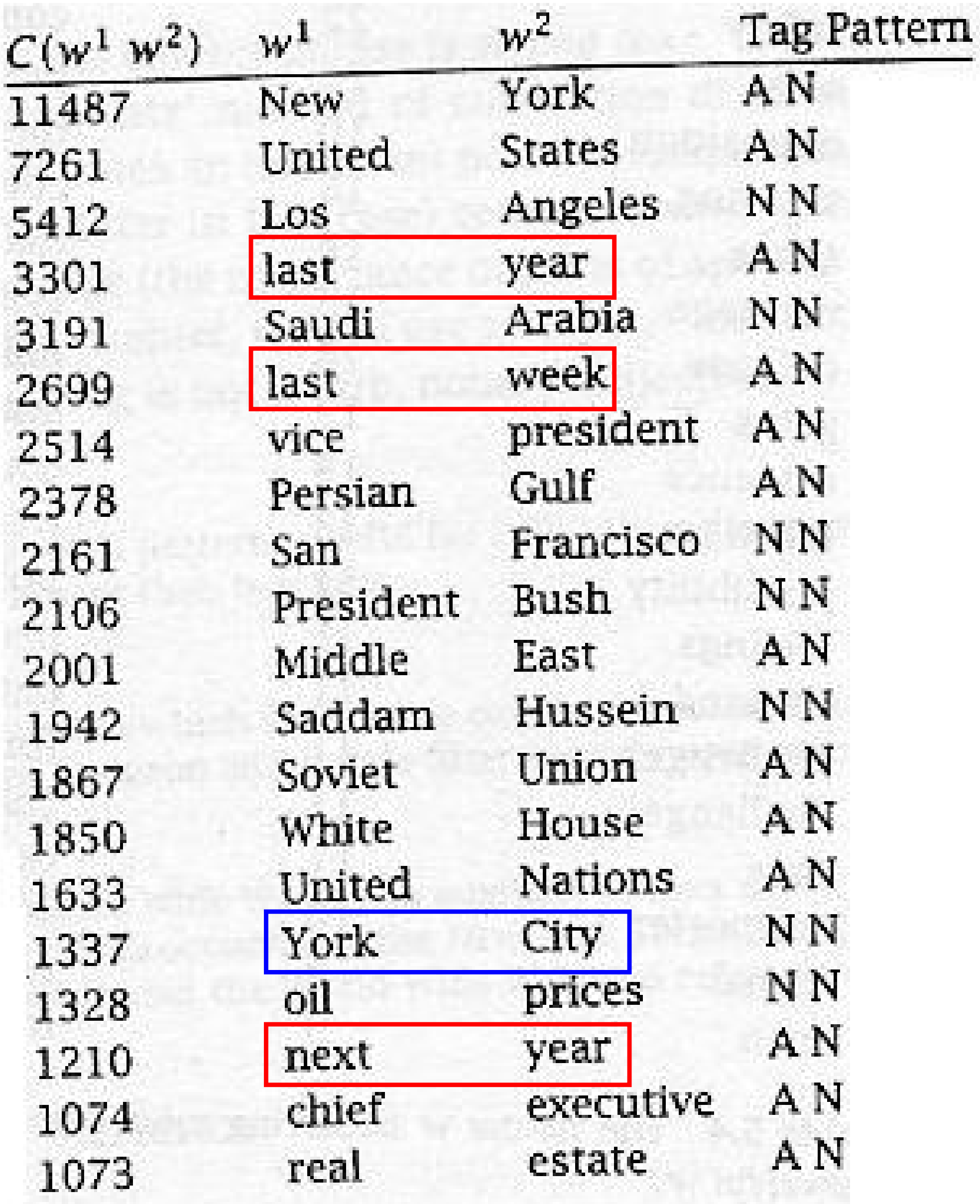
- Pass the candidate phrases through a part-of-speech filter



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.

A: adjective, P: preposition, N: noun



$C(w^1\ w^2)$	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

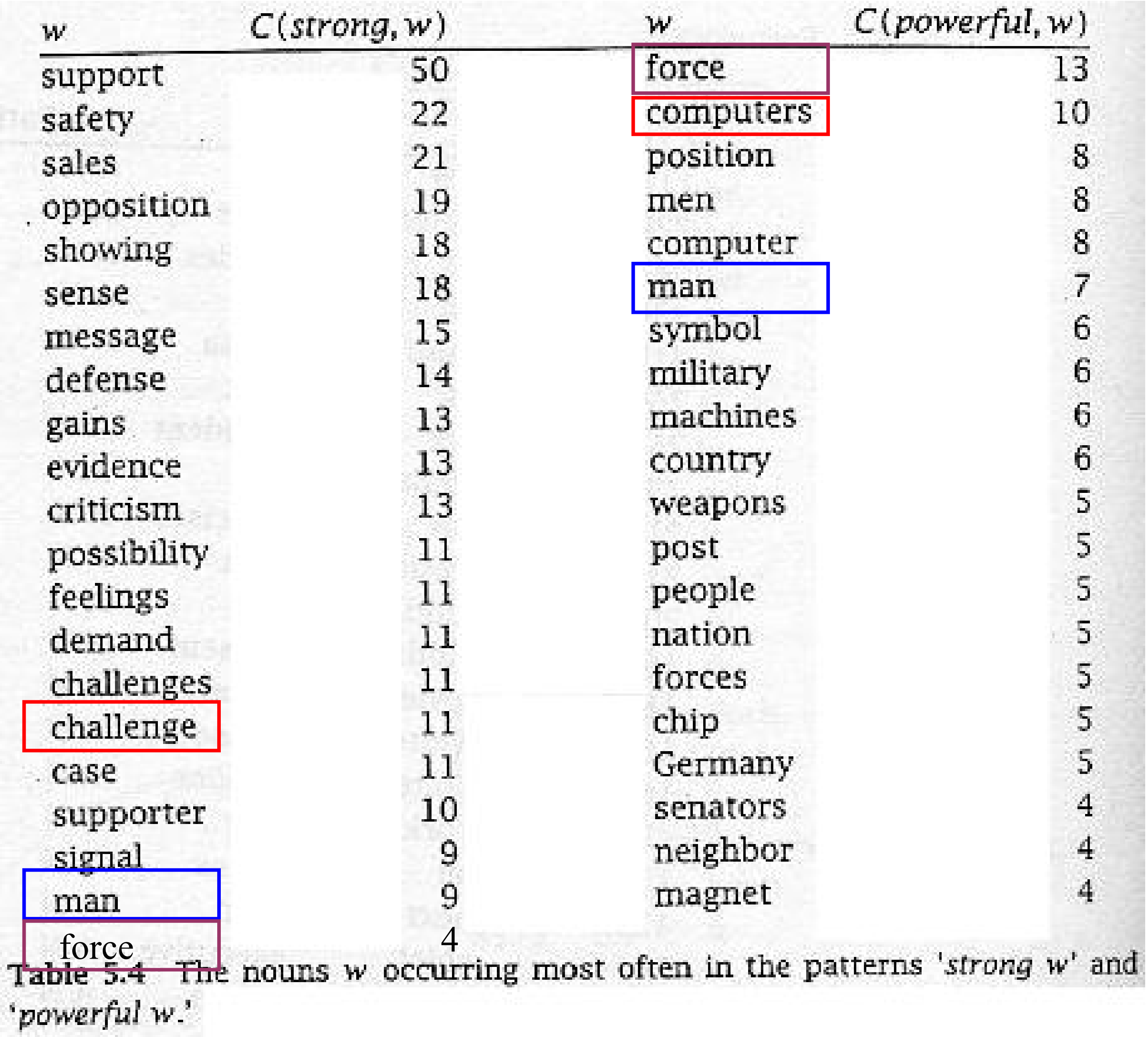
Table 5.3 Finding Collocations: Justeson and Katz' part-of-speech filter.

Frequency

- There are only 3 bigrams that we would not regard as non-compositional phrases: last year, last week, and next year
- *York City* is an artefact of the way we have implemented the 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*

Frequency

- Table 5.4 show the 20 highest ranking phrases containing strong and powerful all have the form AN (where A is either strong or powerful)
- *Strong challenge* and *powerful computers* are correct whereas *powerful challenge* and *strong computers* are not
- Neither *strong tea* nor *powerful tea* occurs in *New York Times* corpus. However, searching the larger corpus of the WWW we find 799 examples of *strong tea* and 17 examples of *powerful tea*



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
force	4		

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

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*

  a. she knocked on his door

  b. they knocked at the door

  c. 100 women knocked on Donaldson's door

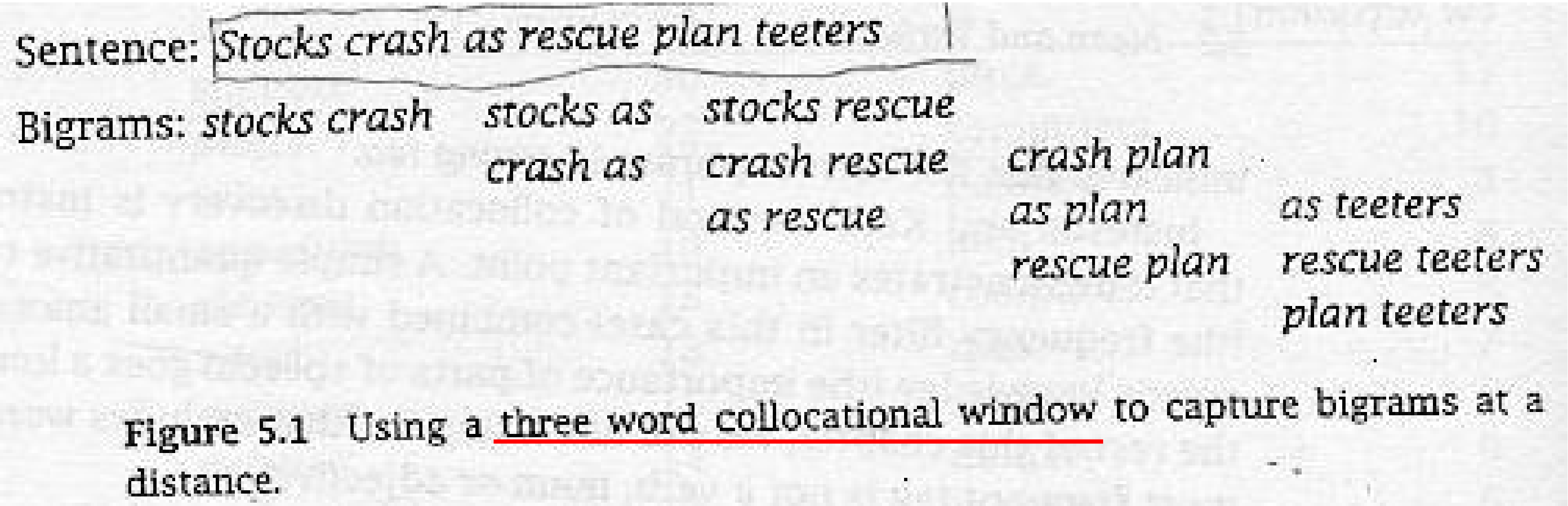
  d. a man knocked on the metal front door

Mean and Variance

- 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
- 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

Mean and Variance

- We use a *collocational window*, and we enter every word pair in there as a collocational bigram



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.

Mean and Variance

- The mean is simply the average offset. We compute the mean offset between *knocked* and *door* as follows:

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

- Variance

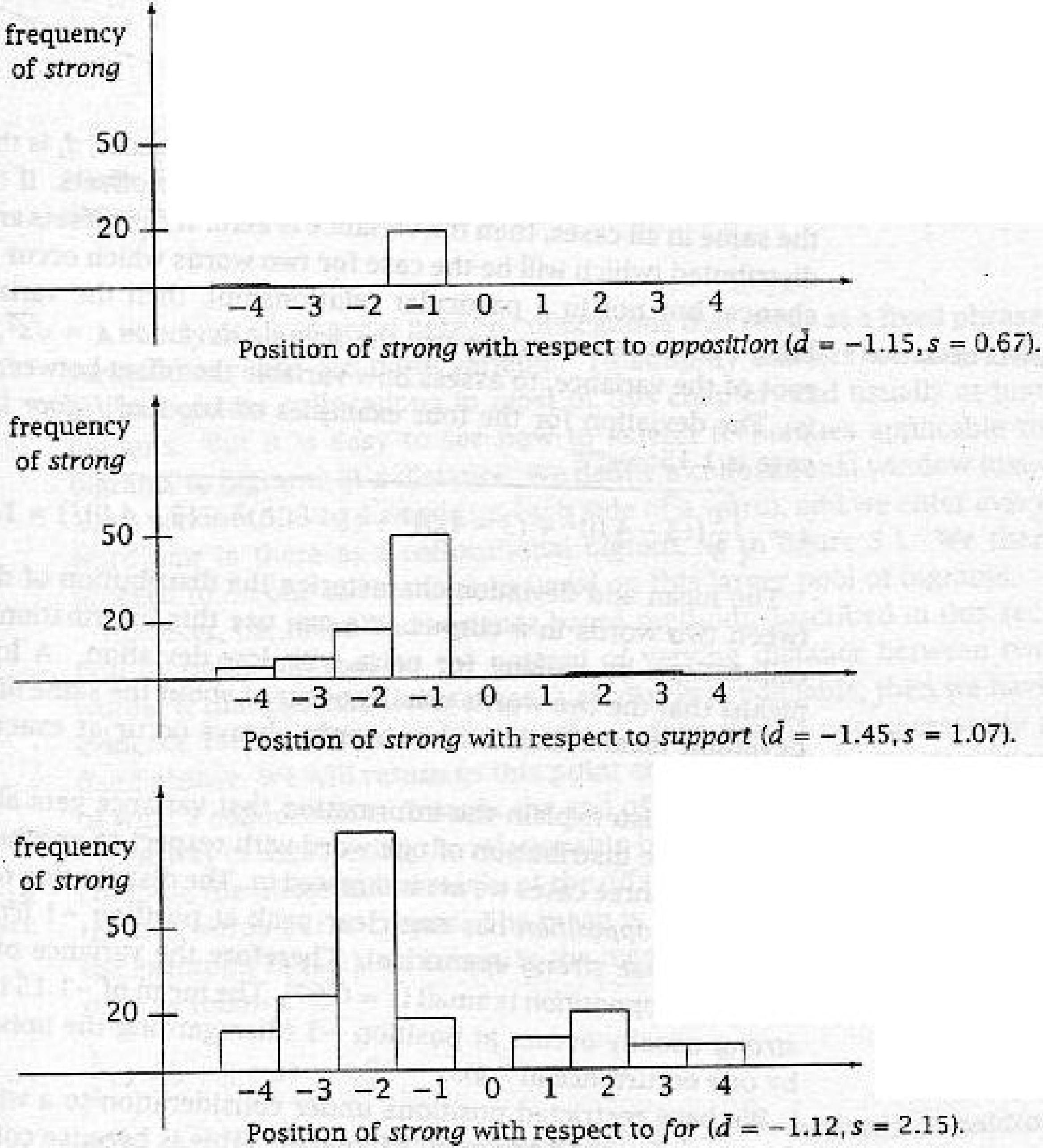
$$s^2=\frac{\sum_{i=1}^{n}(d_i-\overline{d})^2}{n-1}$$

- We use the *sample deviation* to access how variable the offset between two words is. The deviation for the four examples of *knocked* / *door* is

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

Mean and Variance

- We can discover collocations by looking for pairs with low deviation
- A low deviation means that the two words usually occur at about the same distance
- We can also explain the information that variance gets at in terms of peaks

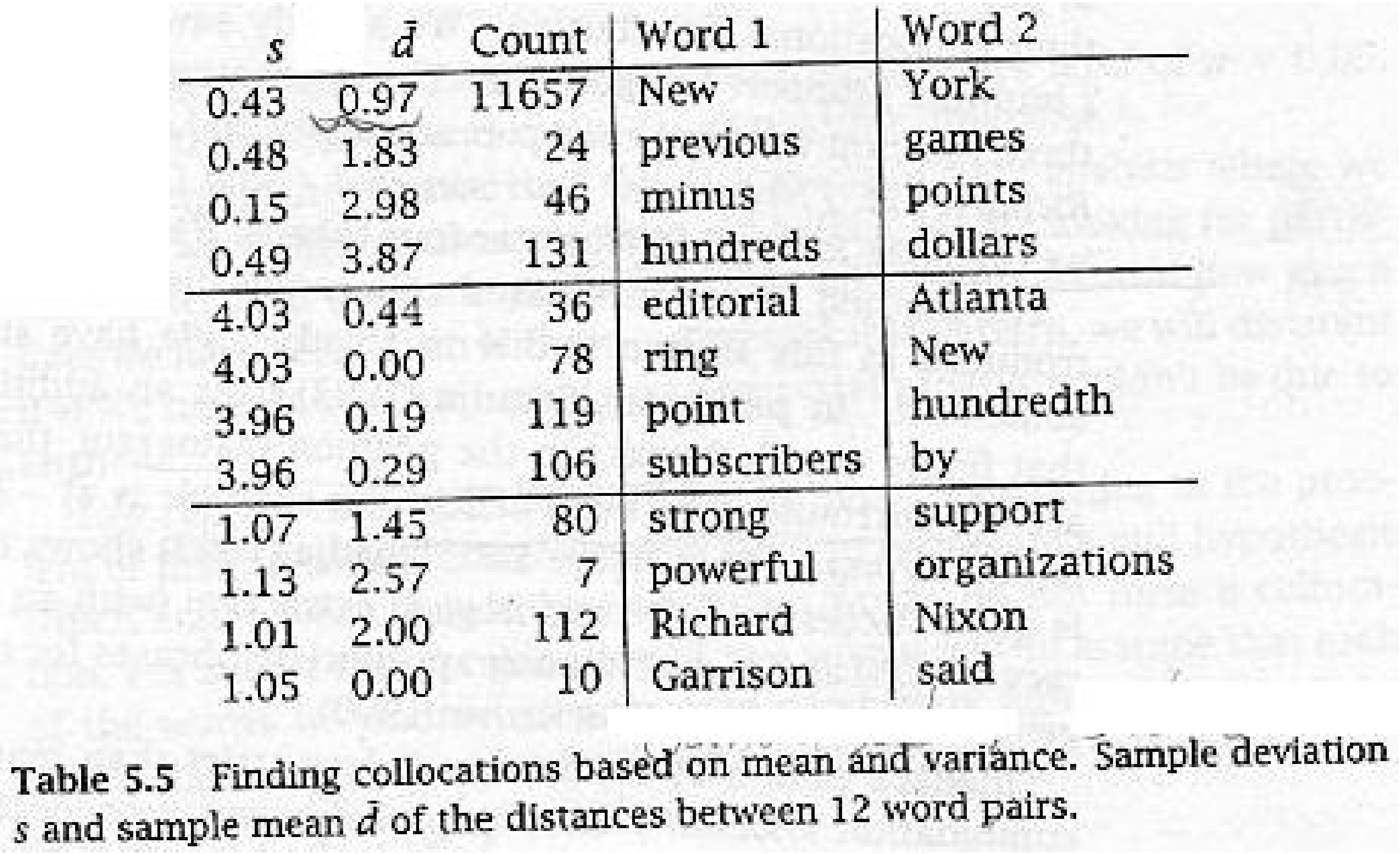


Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).

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	$\bar{d}$	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 $\bar{d}$ of the distances between 12 word pairs.

$\bar{d}$ = 0.00 表示 (word1,word2) 跟 (word2,word1) 出現次數一樣多

Mean and Variance

- If the mean is close to 1.0 and the deviation low, like *New York*, then we have the type of phrase that Justeson and Katz’ frequency-based approach will also discover
- High deviation indicates that the two words of the pair stand in no interesting relationship

Hypothesis Testing

- High frequency and low variance can be accidental
- If the two constituent words of a frequent bigram like *new companies* are frequently occurring words, 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 to know is whether two words occur together more often than chance
- We formulate a *null hypothesis* H_0 that there is no association between the words beyond chance occurrences

Hypothesis Testing

- Free combination: each of the words w^1 and w^2 is generated completely independently, so their chance of coming together is simply given bt $P(w^1w^2) = P(w^1)P(w^2)$

Hypothesis Testing

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

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

$\overline{x}$ is the sample mean, s^2 is the sample variance, N is the sample size, and μ is the mean of the distribution

Hypothesis Testing

The *t* test

- 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.

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

Confidence level of $\alpha = 0.005$, we fine 2.576
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%

Hypothesis Testing
The *t* test

- How to use the *t* test for finding collocations? There is a way of extending the *t* test for use with proportions or counts.

$$P(new) = \frac{15828}{14307668} \quad P(companies) = \frac{4675}{14307668}$$

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

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

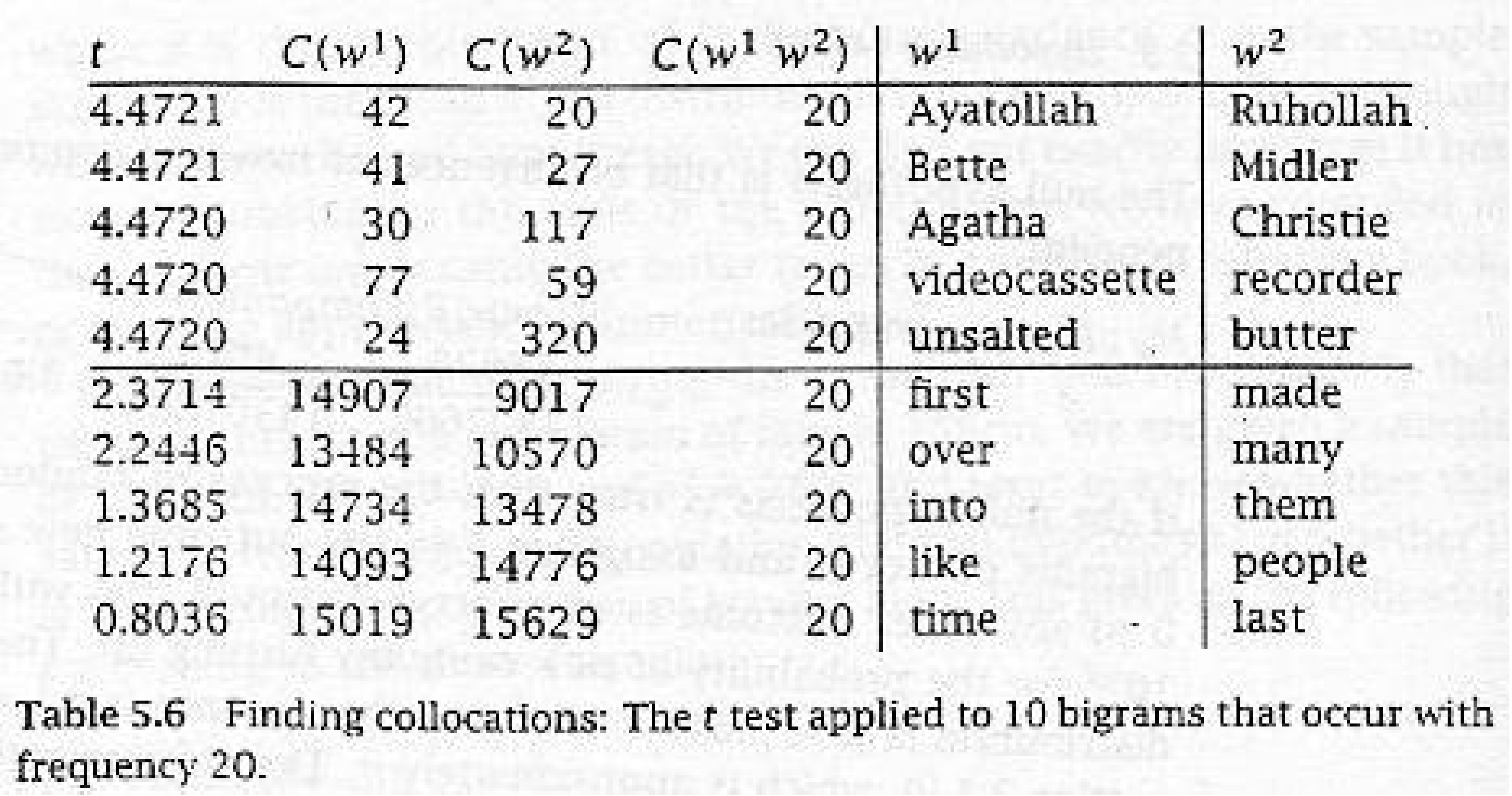
Hypothesis Testing
The *t* test

- $\mu = 3.615*10^{-7}$ and the variance is $\sigma^2 = p(1\text{-}p)$, which is approximately *p* (since for most bigram *p* is small)
- There are actually 8 occurrences of *new companies* among the 14,307,668 bigrams in our corpus, so
$$\bar{x} = \frac{8}{14307668} \approx 5.591 \times 10^{-7}$$
- Now we can compute
$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{5.591 \times 10^{-7} - 3.615 \times 10^{-7}}{\sqrt{\frac{5.591 \times 10^{-7}}{14307668}}} \approx 0.999932$$

Hypothesis Testing

The *t* test

- This *t* value of 0.999932 is not larger than 2.576, so we cannot reject the null hypothesis that *new* and *companies* occur independently and do not form a collocation
- Table 5.6 shows *t* values for ten bigrams that occur exactly 20 times in the corpus



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.

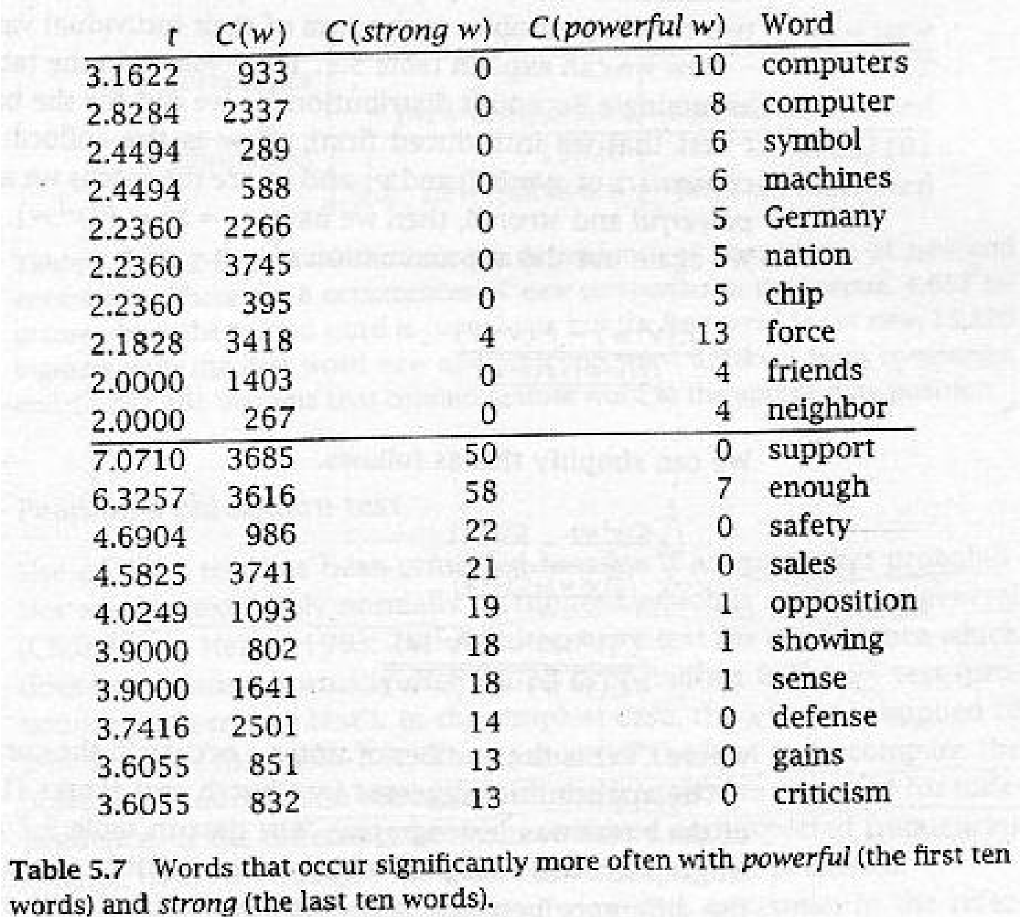
For the top five bigrams, we can reject the null hypothesis. They are good candidates for collocations

Hypothesis Testing

Hypothesis testing of differences

- To find words whose co-occurrence patterns best distinguish between two words

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



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
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).

Hypothesis Testing
Hypothesis testing of differences

- Here the null hypothesis is that the average difference is 0 ($\mu=0$)

$$\bar{x}-\mu=\bar{x}=\frac{1}{N}\sum(x_{1i}-x_{2i})=\bar{x}_1-\bar{x}_2$$

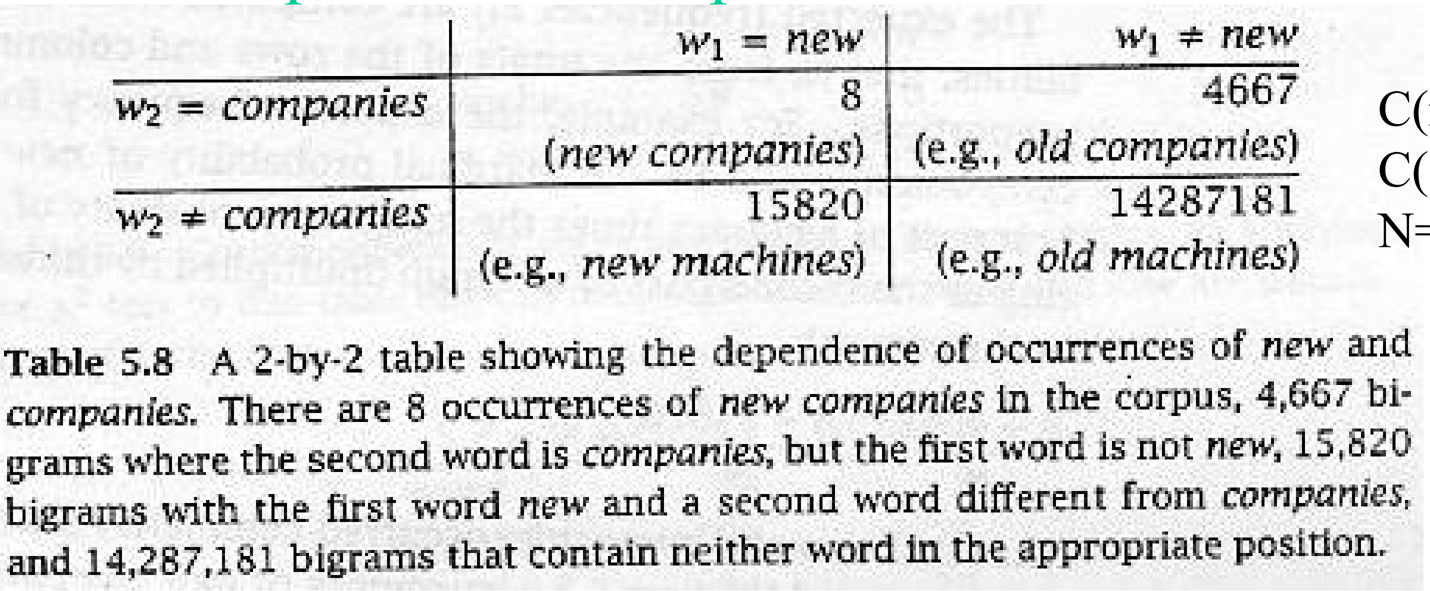
- If *w* is the collocate of interest (e.g., *computers*) and v^1 and v^2 are the words we are comparing (e.g., *powerful* and *strong*), then we have $\bar{x}_1=s_1^2=P(v^1w), \bar{x}_2=s_2^2=P(v^2w)$

$$s^2=p-p^2\approx p$$

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

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
- The essence of χ^2 test is to compare the observed frequencies in the table with the frequencies expected for independence



	$w_1 = new$	$w_1 \neq new$
$w_2 =$ *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.

C(new)=15828
C(companies)=4675
N=14307668

Pearson's chi-square test

- If the difference between observed and expected frequencies is large, then we can reject the null hypothesis of independence

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

- where *i* ranges over rows of the table, *j* ranges over columns, O_{ij} is the oberved value for cell (*i*, *j*) and E_{ij} is the expected value

Pearson’s chi-square test

- The expected frequencies E_{ij} are computed from the marginal probabilities
- Expected frequency for cell (1,1) (*new companies*) would be *new* 發生在第一個位置的機率＊*companies*發生在第二個位置的機率＊corpus中bigram的數目

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

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

Pearson's chi-square test

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

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

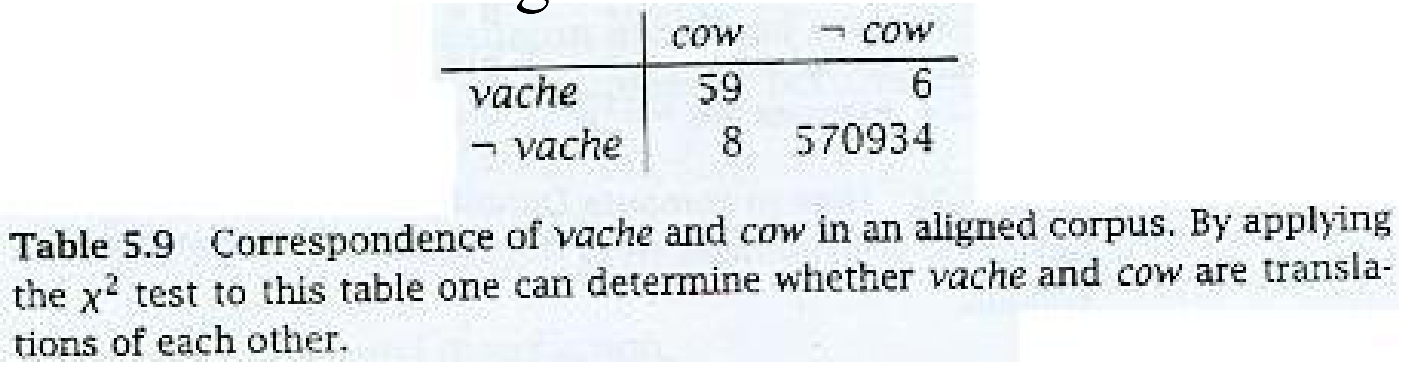
- χ^2 value for table 5.8:

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

- Looking up the χ^2 distribution, we find that at a probability level of α=0.05 the critical value is χ^2=3.841. 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

Pearson’s chi-square test

- One of the early uses of the χ^2 test in Statistical NLP was the identification of translation pairs in aligned corpora
- Table 5.9 strongly suggest that *vahce* is the French translation of English *cow*



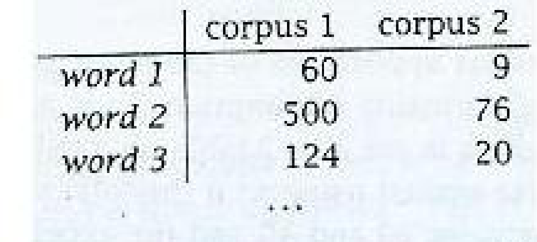
	cow	¬ *cow*
vache	59	6
¬ *vache*	8	570934

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.

χ^2 value is very high, $\chi^2 = 456400$

Pearson’s chi-square test

- An interesting application of χ^2 is as a metric for corpus similarity
- Here we compile an *n*-by-two table for a large *n*, for example *n*=500. The two columns correspond to the two corpora
- In table 5.10, the ratio of the counts are about the same, each word occurs roughly 6 times more often in corpus 1 than in corpus 2. So we cannot reject the null hypothesis that both corpora are drawn from the same underlying source



	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.

Likelihood ratios

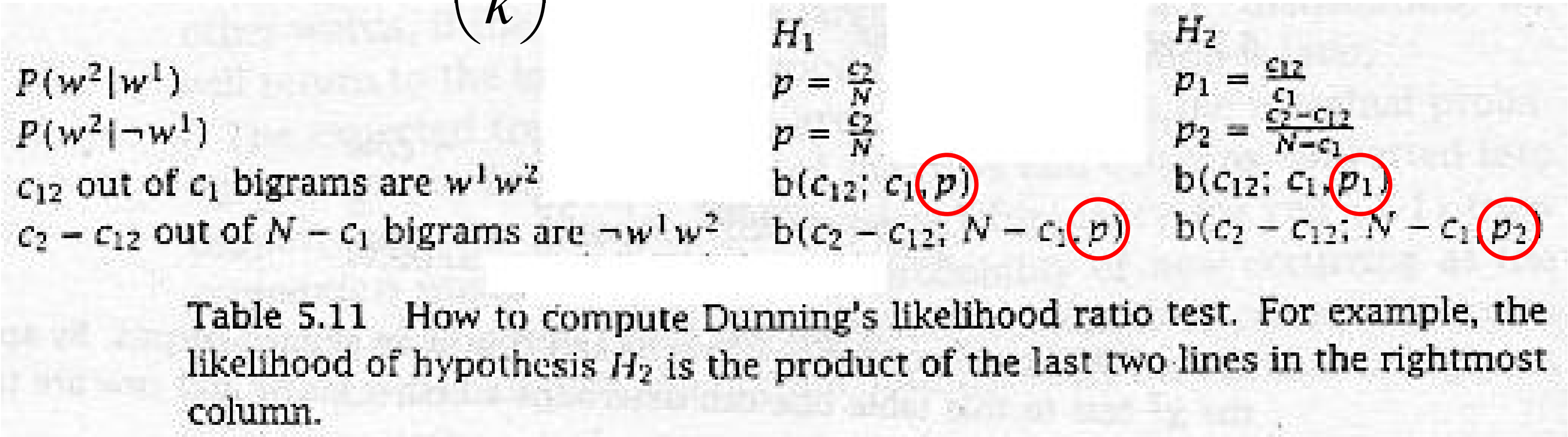
- Hypothesis 1. $P(w^2 \mid w^1) = p = P(w^2 \mid \neg w^1)$
- Hypothesis 2. $P(w^2 \mid w^1) = p_1 \neq p_2 = P(w^2 \mid \neg w^1)$
- Hypothesis 1 is a formalization of independence, hypothesis 2 is a formalization of dependence which is good evidence for an interesting collocation
- We use the usual MLE 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 corpus

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

Likelihood ratios

- Assuming a binomial distribution:

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



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

$$L(H_1) = b(c_{12}; c_1, p)\, b(c_2 - c_{12}; N - c_1, p)$$

$$L(H_2) = b(c_{12}; c_1, p_1)\, b(c_2 - c_{12}; N - c_1, p_2)$$

Likelihood ratios

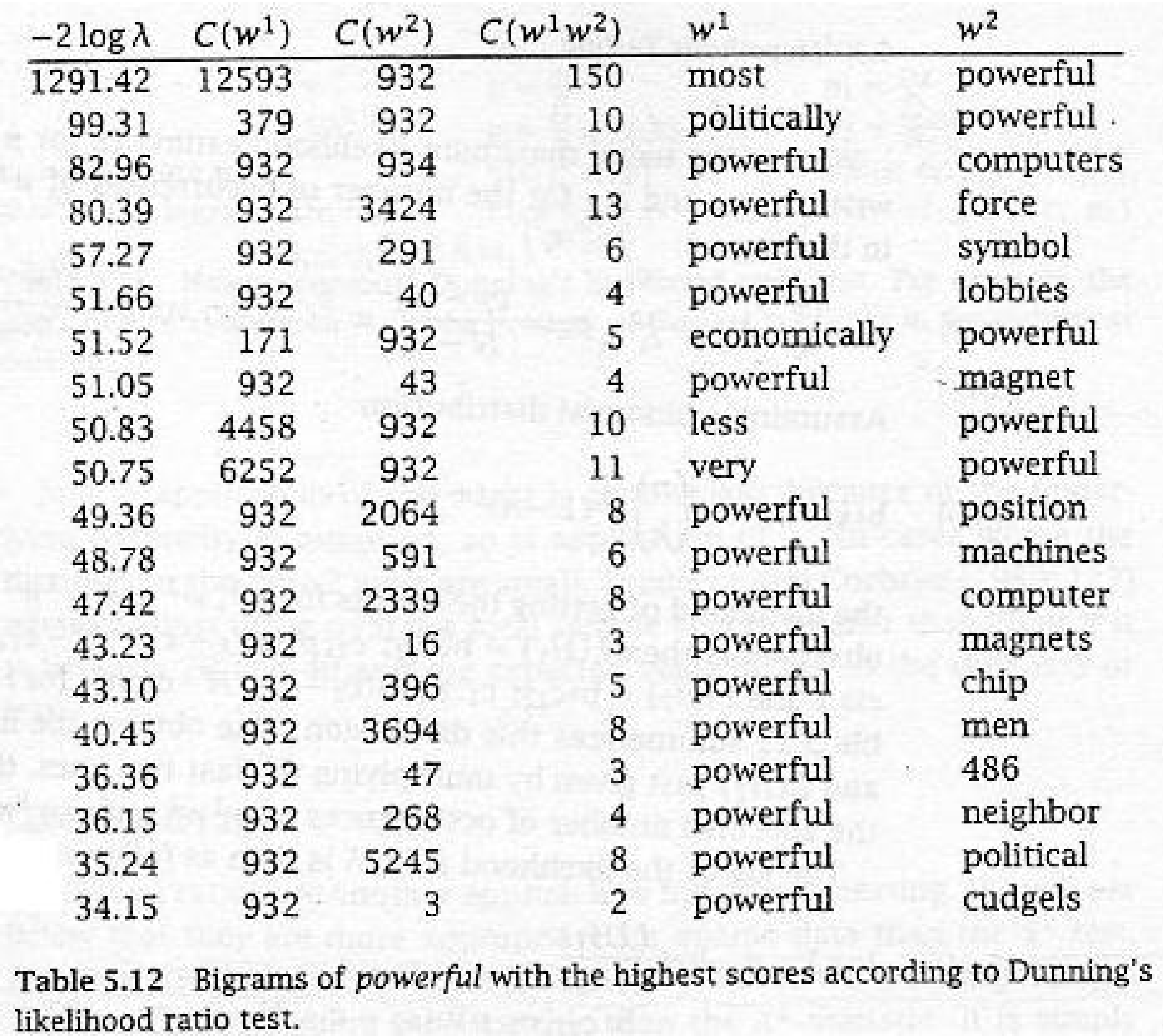
$$\log\lambda = \log\frac{L(H_1)}{L(H_2)}$$

$$= \log\frac{b(c_{12}, c_1, p)b(c_2 - c_{12}, N - c_1, p)}{b(c_{12}, c_1, p_1)b(c_2 - c_{12}, N - c_1, p_2)}$$

$$= \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p)$$

$$- \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2)$$

Where $L(k,n,x) = x^k(1\text{-}x)^{n\text{-}k}$



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

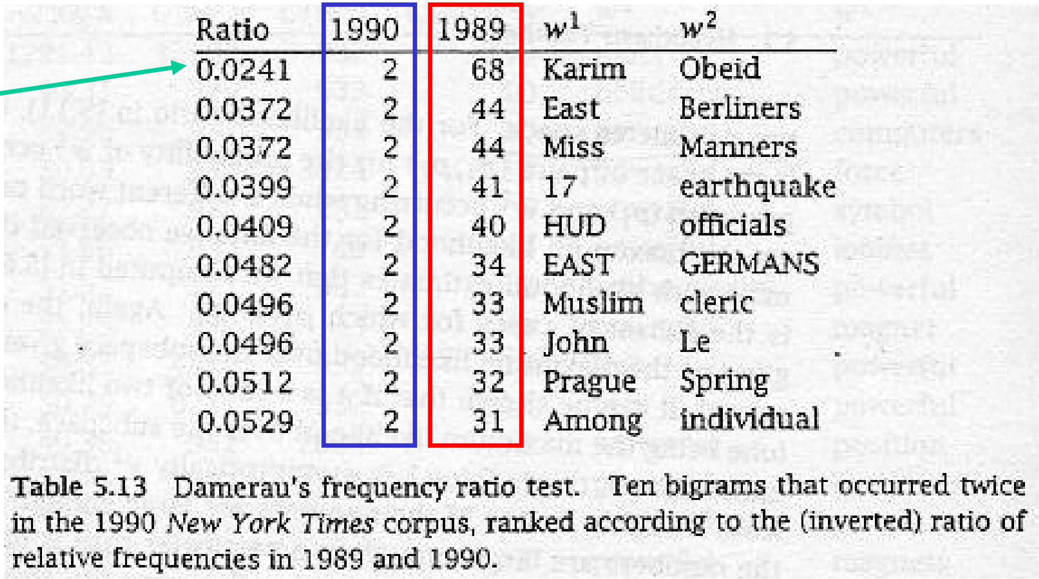
Likelihood ratios

- 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 value in table 5.12 to test the null hypothesis H_1 against the alternative hypothesis H_2
- 34.15 for *powerful cudgels* in the table 5.12 and reject H_1 for this bigram on a confidence level of $\alpha=0.005$ ($\chi^2 = 7.88$, $34.15>7.88$)

Relative frequency ratios

- Table 5.13 shows ten bigrams that occur exactly twice in our reference corpus

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



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.

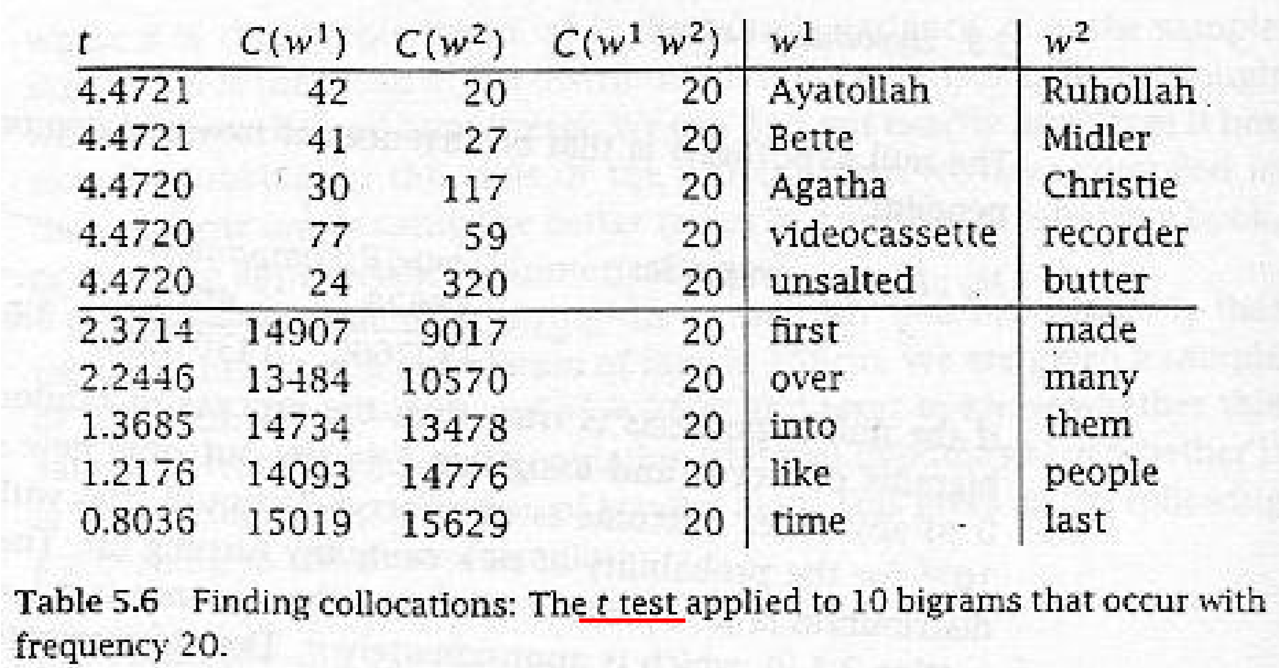
Mutual Information

- Fano (1961:27-28) originally defined mutual information between particular events *x*’ and *y*’, in our case the occurrence of particular words, as follow:

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

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

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



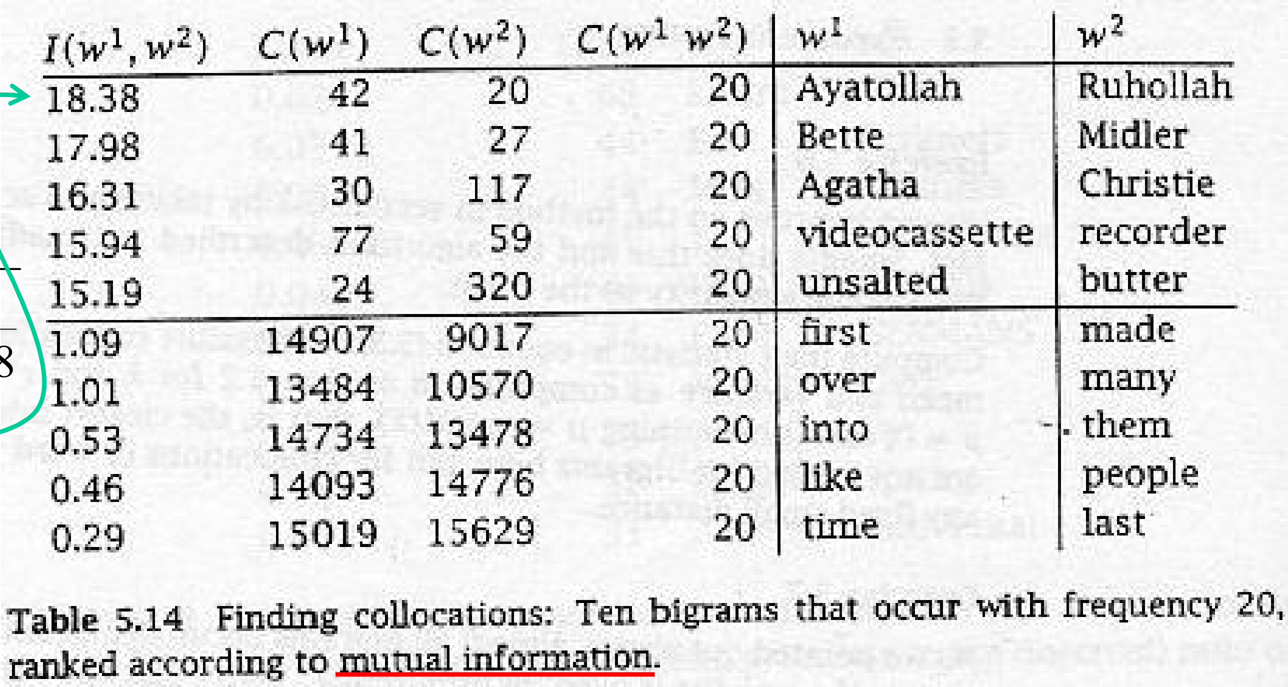
t	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	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.

$$I(Ayatollah, Ruhollah)$$

$$= \log_2 \frac{\frac{20}{14307668}}{\frac{42}{14307668} \times \frac{20}{14307668}}$$

$$\approx 18.38$$



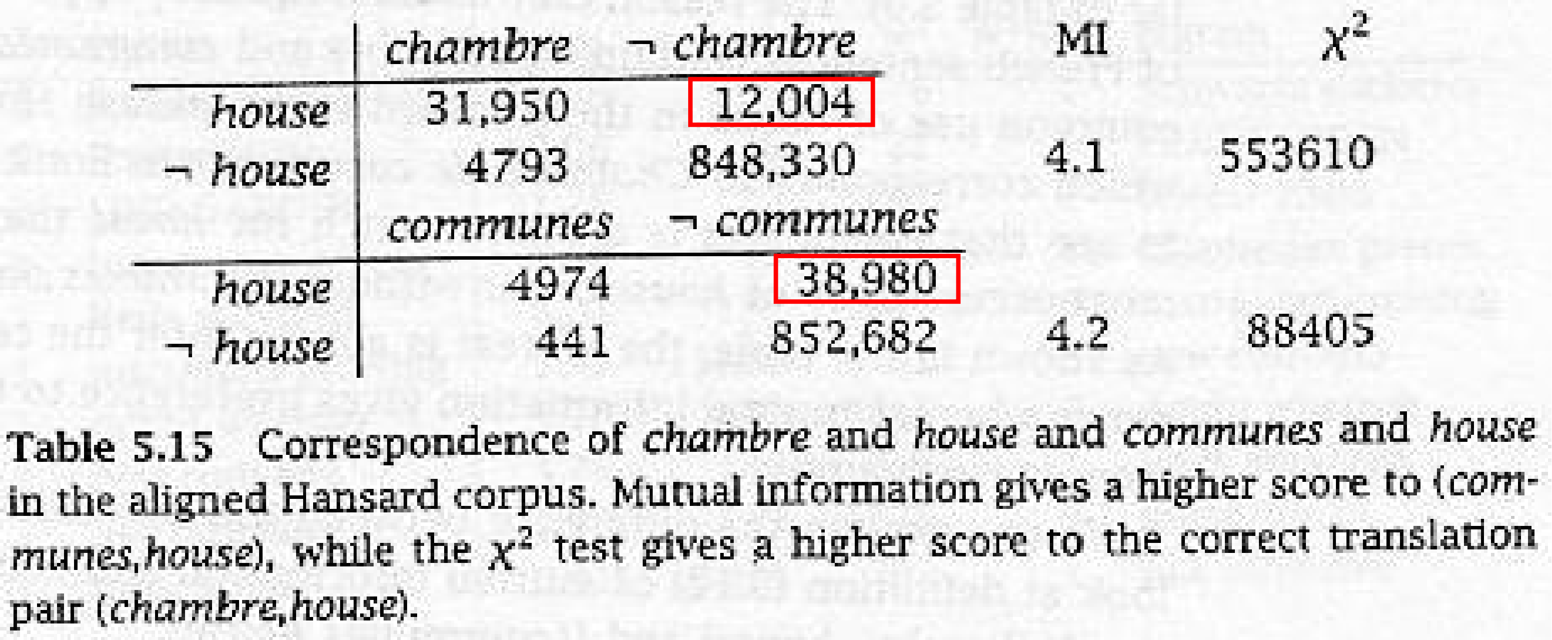
$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13478	20	into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

Mutual Information

- 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)]
- 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

Mutual Information



	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

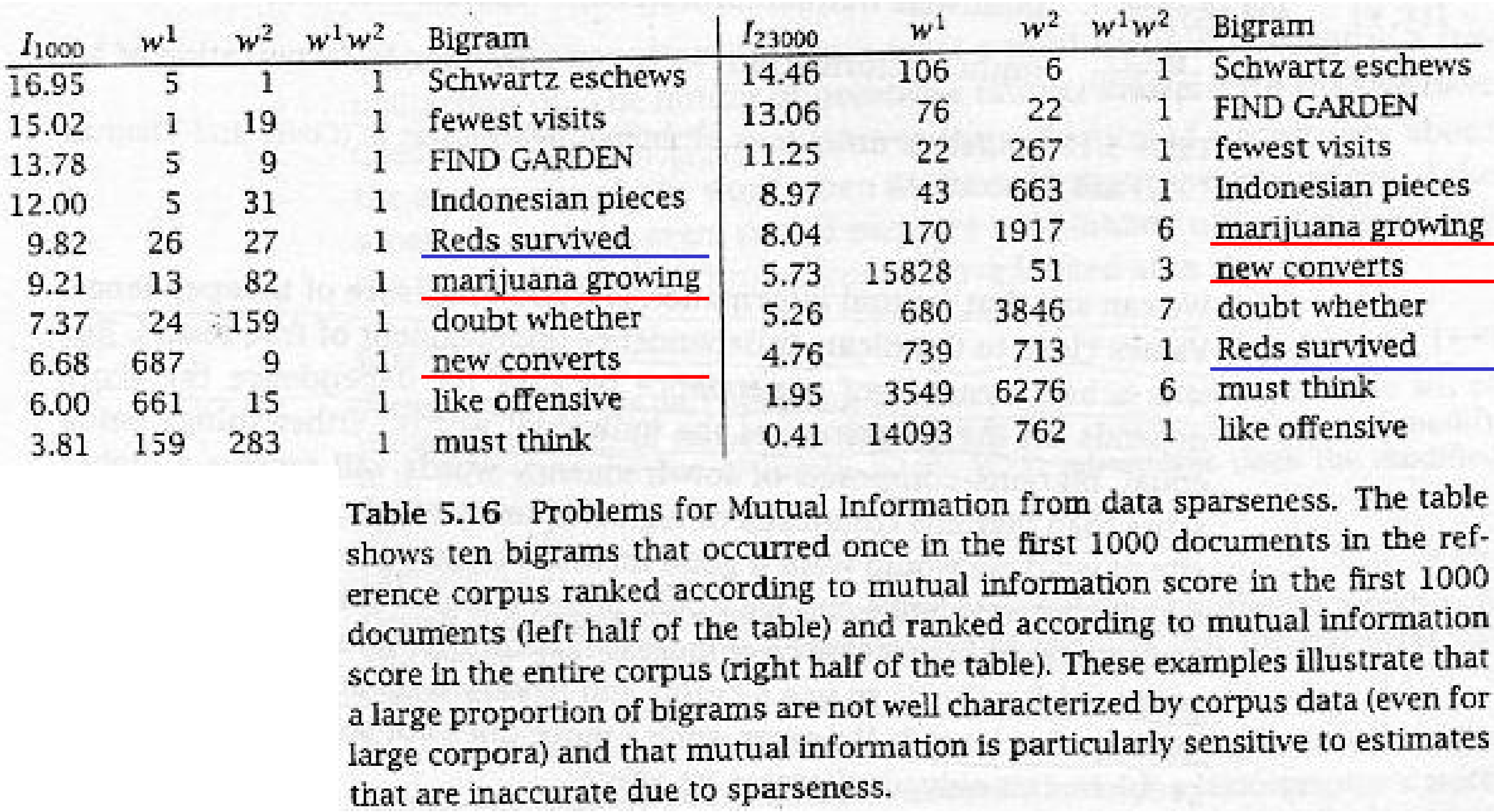
Table 5.15 Correspondence of *chambre* and *house* and *communes* and *house* in the aligned Hansard corpus. Mutual information gives a higher score to (*communes,house*), while the χ^2 test gives a higher score to the correct translation pair (*chambre,house*).

- House of Commons <-> Chambre de communes
- 由紅色框框中可看出 (house, chambre)才是對的，且χ^2 test 結果也是正確的，但mutual information卻是錯誤的。

Mutual Information

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

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



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.

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

Mutual Information

- None of the measures we have seen works very well for low-frequency events
- Perfect dependence

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

as x or y get rarer, their mutual information *increases*

- Perfect independence

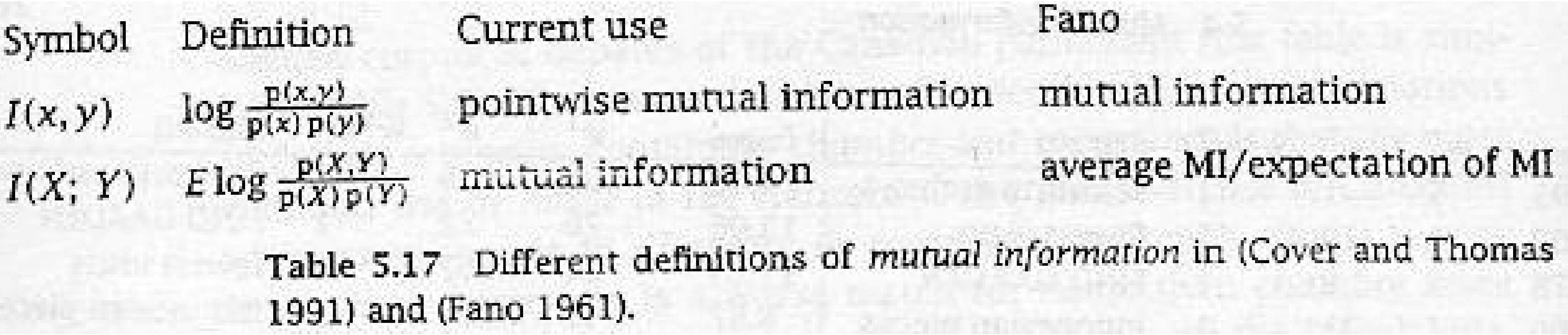
$$I(x,y) = \log \frac{P(xy)}{P(x)P(y)} = \log \frac{P(x)P(y)}{P(x)P(y)} = \log 1 = 0$$

we can say that mutual information is a good measure of independence. Value close to 0 indicate independence

Mutual Information

- But it is a bad measure of dependence because for dependence the score depends on the frequency of the individual word
  → 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
- Mutual information in Information Theory refers to the *expectation* of the quantity

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



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).

The notion of pointwise mutual information that we have used here measures the reduction of uncertainty about the occurrence of one word when we are told about the occurrence of the other

The Notion of Collocation

- 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

The Notion of Collocation

- **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 (idioms like *kick the bucket*) or there is a connotation or added element of meaning that cannot be predicted from the parts (e.g., *white wine*)

The Notion of Collocation

- **Non-substitutability**
  We cannot substitute other words for the components of a collocation even if 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)

The Notion of Collocation

- 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 a ugly frog in one’s throat* although usually nouns like *frog* can be modified by adjectives like *ugly*

The Notion of Collocation

- 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
  make a decision into French one word at a time we get *faire une decision* witch is incorrect (*prendre une decision*)

The Notion of Collocation

- Light verbs, *make, take* and *do*
- Verb particle constructions or phrasal verbs, *fell off ,go down*
- Proper nouns
- Terminological expression