Natural Language Processing

COLLOCATIONS

Updated 04/09

What is a Collocation?

- A COLLOCATION is an expression consisting of two or more words that correspond to some conventional way of saying things.
- The words together can mean more than their sum of parts (*The Times of India, disk drive*)

Examples of Collocations

- 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*.
- *a stiff breeze* but not *a stiff wind* (while either *a strong breeze* or *a strong wind* is okay).
- *broad daylight* (but not *bright daylight* or *narrow darkness*).

Criteria for Collocations

- Typical criteria for collocations: non-compositionality, non-substitutability, non-modifiability.
- Collocations cannot be translated into other languages word by word.
- A phrase can be a collocation even if it is not consecutive (as in the example *knock . . . door*).

Compositionality

- A phrase is compositional if the meaning can 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. Eg. *strong tea.*
- Idioms are the most extreme examples of non-compositionality. Eg. *to hear it through the grapevine.*

Non-Substitutability

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

Non-modifiability

- Many collocations cannot be freely modified with additional lexical material or through grammatical transformations.
- Especially true for idioms, e.g. *frog* in *'to get a frog in ones throat'* cannot be modified into `*green frog*

Linguistic Subclasses of Collocations

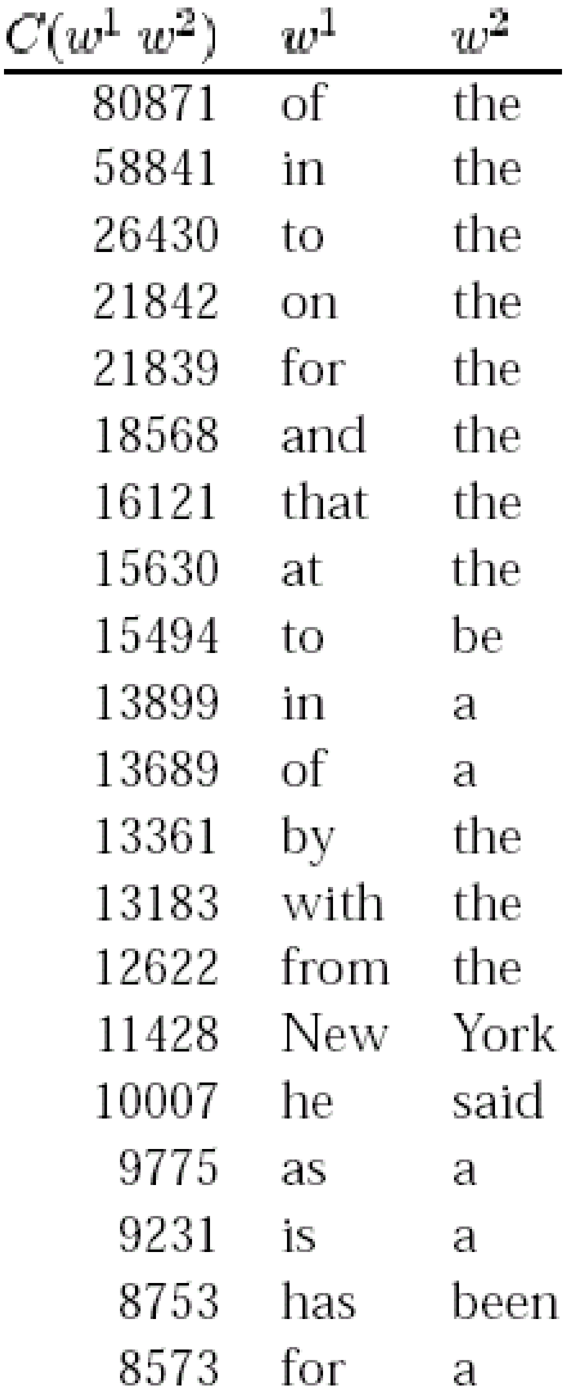
- Light verbs: Verbs with little semantic content like *make*, *take* and *do.*
- Verb particle constructions (*to go down)*
- Proper nouns (*Prashant Aggarwal*)
- Terminological expressions refer to concepts and objects in technical domains. (*Hydraulic oil filter*)

Principal 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**
- **Mutual information**

Frequency

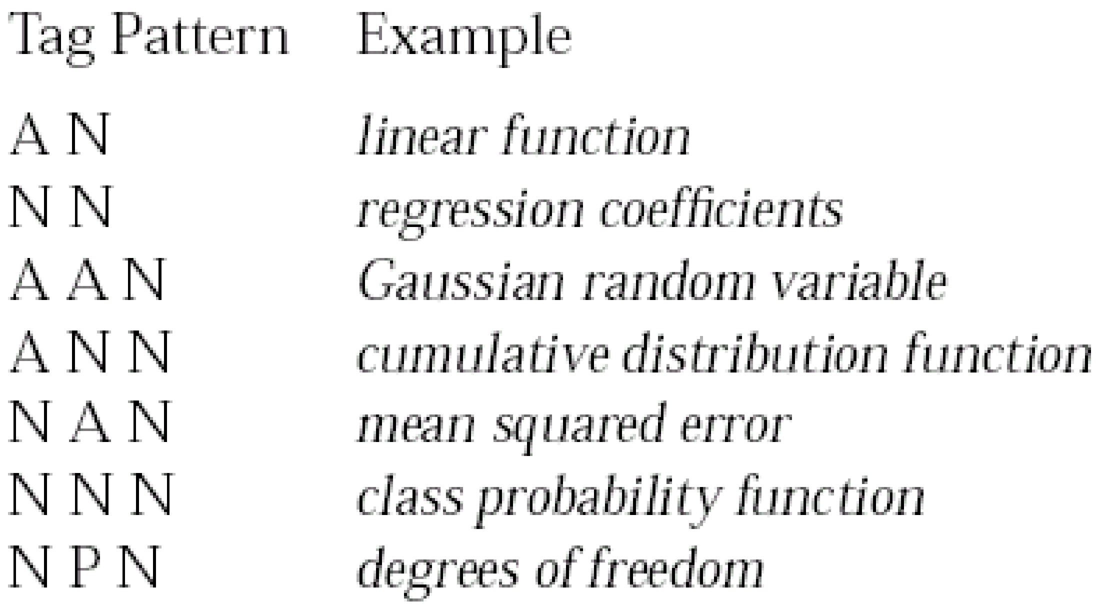
- Finding collocations by counting the number of occurrences.
- Usually results in a lot of function word pairs that need to be filtered out.
- Pass the candidate phrases through a part of-speech filter which only lets through those patterns that are likely to be “phrases”. (Justesen and Katz, 1995)



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

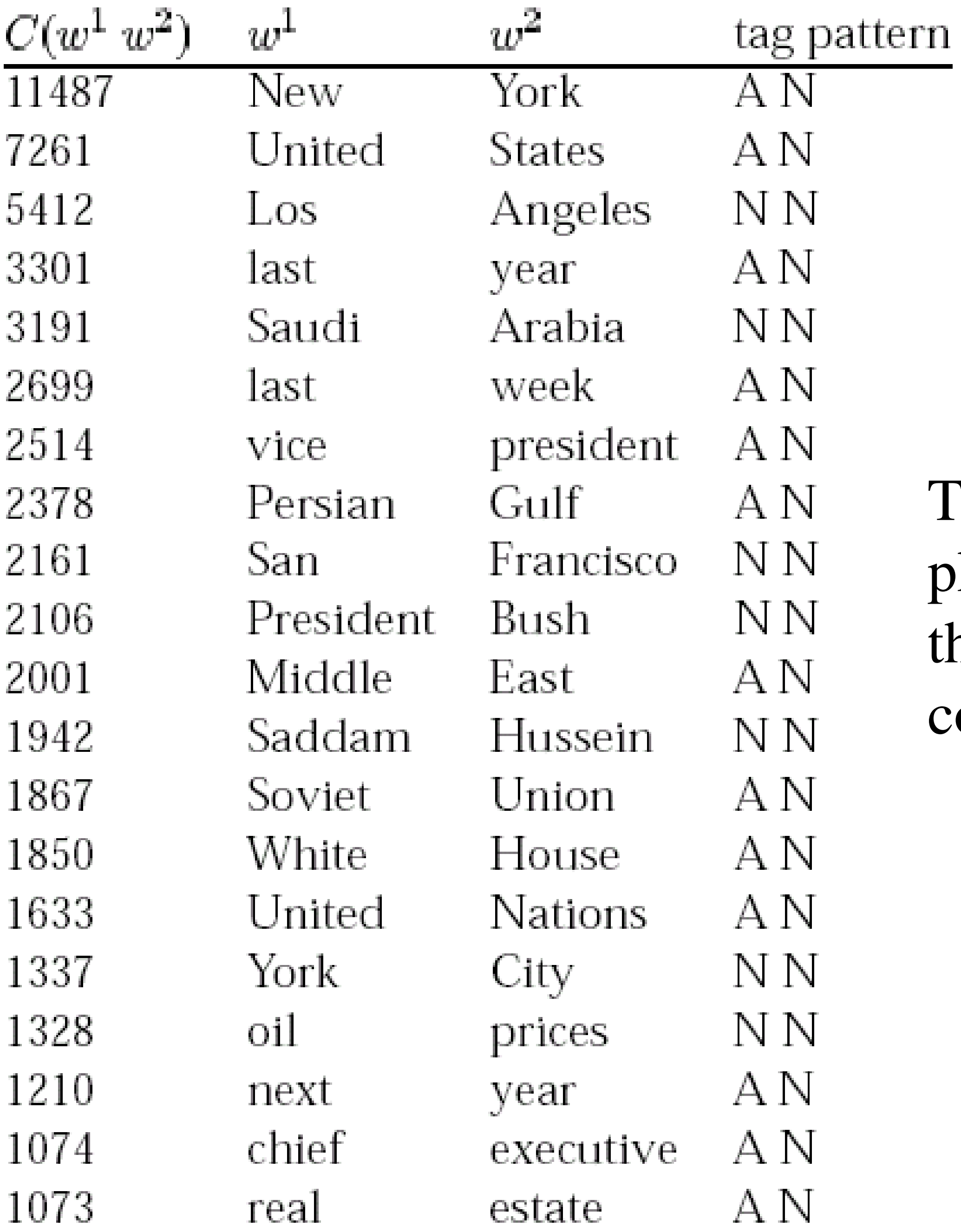
Most frequent bigrams in an Example Corpus

Except for *New York*, all the bigrams are pairs of function words.



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*

Part of speech tag patterns for collocation filtering.



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

The most highly ranked phrases after applying the filter on the same corpus as before.

Collocational Window

- Many collocations occur at variable distances. A collocational window needs to be defined to locate these. Freq based approach can’t be used.
  - she knocked on his door
  - they knocked at the door
  - 100 women knocked on Donaldson’s door
  - a man knocked on the metal front door

Mean and Variance

- The mean μ is the average offset between two words in the corpus.
- The variance σ2

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

where *n* is the number of times the two words co-occur, d_i is the offset for co-occurrence *i*, and μ is the mean.

Mean and Variance: Interpretation

- The mean and variance 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 variance.
- A low variance means that the two words usually occur at about the same distance.

Mean and Variance: An Example

- For the *knock, door* example sentences the mean is:

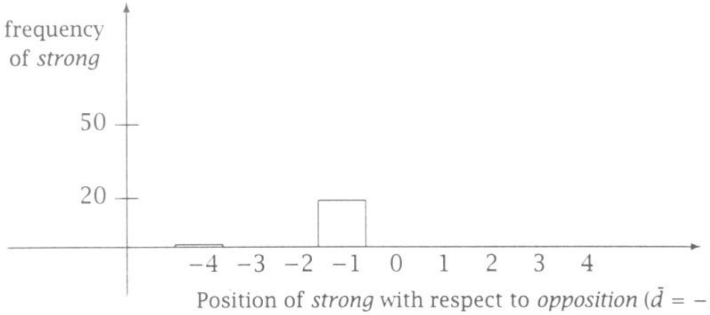
$$\frac{1}{4}(3+3+5+5) = 4.0$$

- And the sample deviation:

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

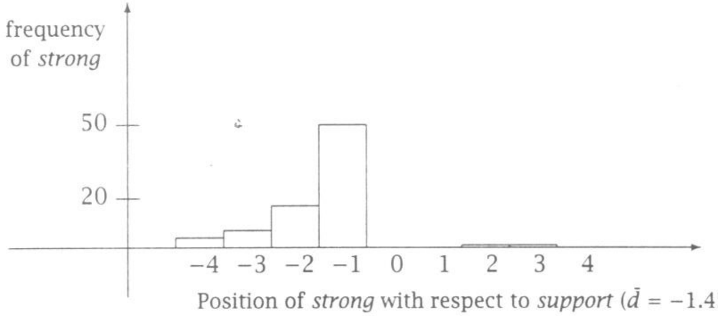
Looking at distribution of distances

strong & opposition



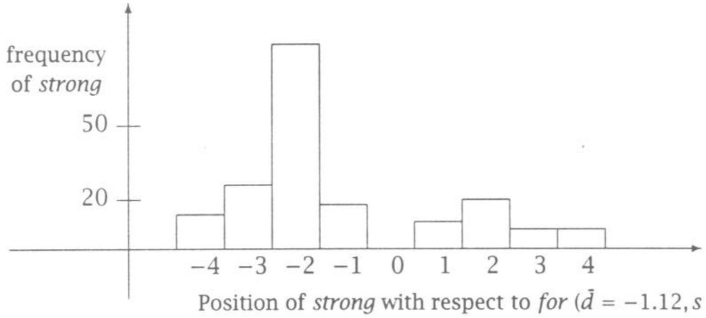
Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).

strong & support



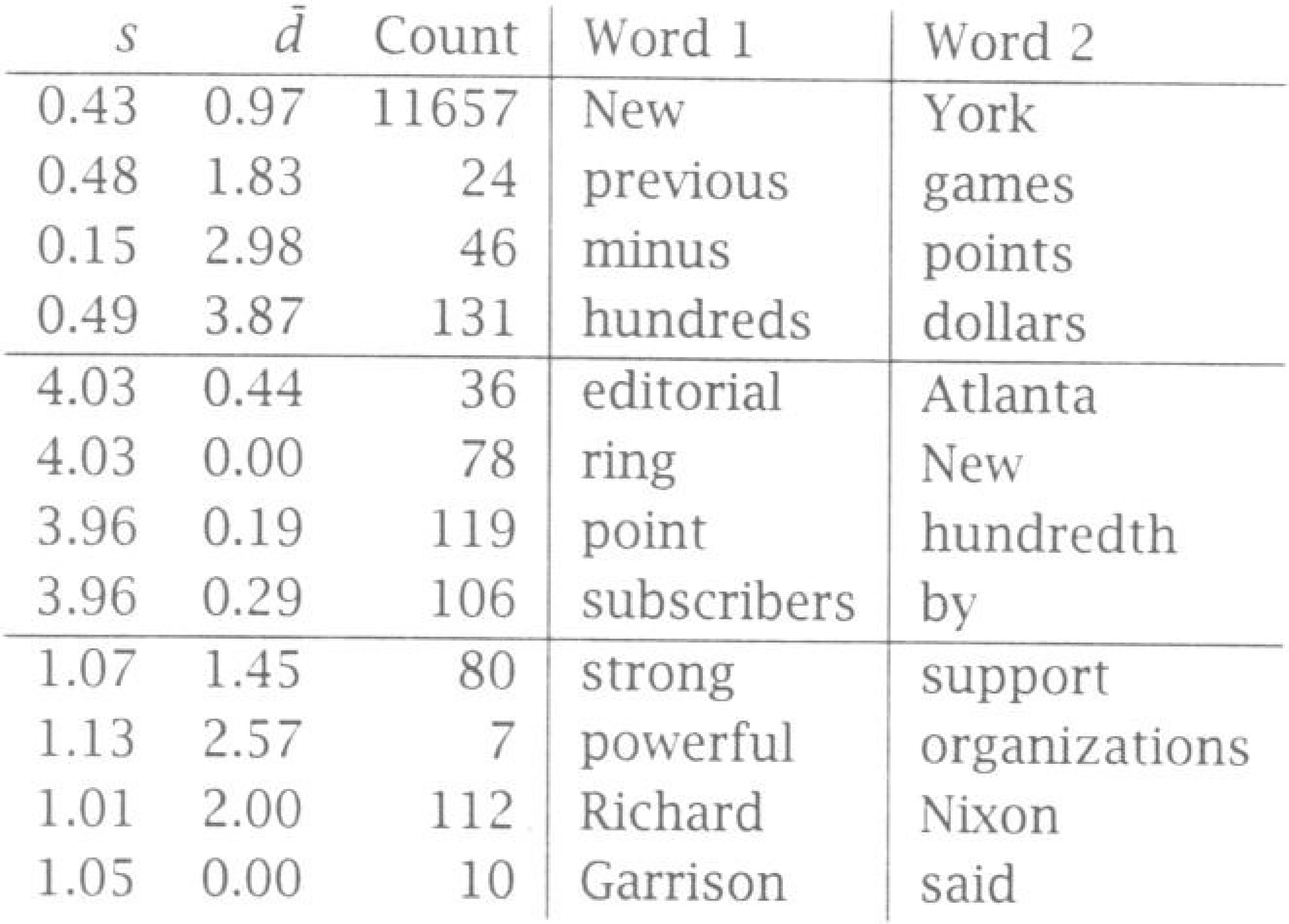
Position of *strong* with respect to *support* ($\bar{d} = -1.45, s = 1.07$).

strong & for



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

Finding collocations based on mean and variance



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

Ruling out Chance

- Two words can co-occur by chance.
- When an independent variable has an effect (two words co-occuring), **Hypothesis Testing** measures the confidence that this was really due to the variable and not just due to chance.

The Null Hypothesis

- We formulate a *null hypothesis* H_0 that there is no association between the words beyond chance occurrences.
- The null hypothesis states what should be true if two words do not form a collocation.

Hypothesis Testing

- Compute the probability p that the event would occur if H_0 were true, and then reject H_0 if p is too low (typically if beneath a ***significance level*** of $p < 0.05, 0.01, 0.005,$ or 0.001) and retain H_0 as possible otherwise.
- In addition to patterns in the data we are also taking into account how much data we have seen.

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.

The *t*-Statistic

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

Where *x* is the sample mean, s^2 is the sample variance, N is the sample size, and λ is the mean of the distribution.

t-Test: Interpretation

- The t-test gives the estimate that the difference between the two means is caused by chance.

t-Test for finding Collocations

- 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.
- The *t*-test and other statistical tests are most useful as a method for ***ranking*** collocations. The level of ***significance*** itself is less useful as language is not completely random.

t-Test: Example

- In our corpus, *new* occurs 15,828 times, *companies* 4,675 times, and there are 14,307,668 tokens overall.
- *new companies* occurs 8 times among the 14,307,668 bigrams
- H0 : P(*new companies*)

  =P(*new*)P(*companies*)

$$= \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7}$$

t-Test: Example (Cont.)

- 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 this distribution $\mu = 3.615 \times 10^{-7}$ and $\sigma^2 = p(1-p)$

t-Test: Example (Cont.)

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

- This t value of 0.999932 is not larger than 2.576, the critical value for α=0.005. So we cannot reject the null hypothesis that *new* and *companies* occur independently and do not form a collocation.

Hypothesis Testing of Differences (Church and Hanks, 1989)

- To find words whose co-occurrence patterns best distinguish between two words.
- For example, in computational lexicography we may want to find the words that best differentiate the meanings of *strong* and *powerful*.
- The *t*-test is extended to the comparison of the means of two normal populations.

Hypothesis Testing of Differences (Cont.)

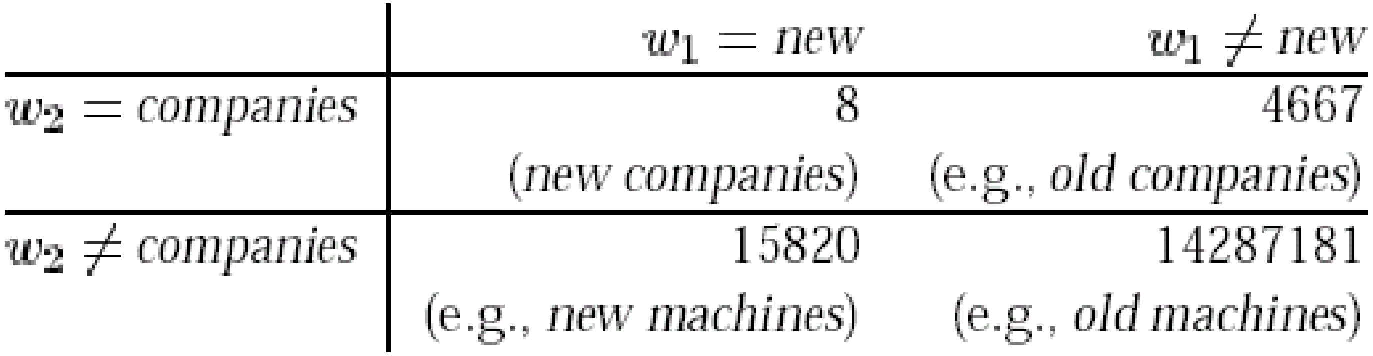
- Here the null hypothesis is that the average difference is 0 (λ=0).
- In the denominator we add the variances of the two populations since the variance of the difference of two random variables is the sum of their individual variances.

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

Pearson’s chi-square test

- The *t*-test assumes that probabilities are approximately normally distributed, which is not true in general. The χ^2 test doesn’t make this assumption.
- The essence of the χ^2 test is to compare the observed frequencies 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.

χ^2 Test: Example – ‘new companies’



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

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:

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

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.

X² - Calculation

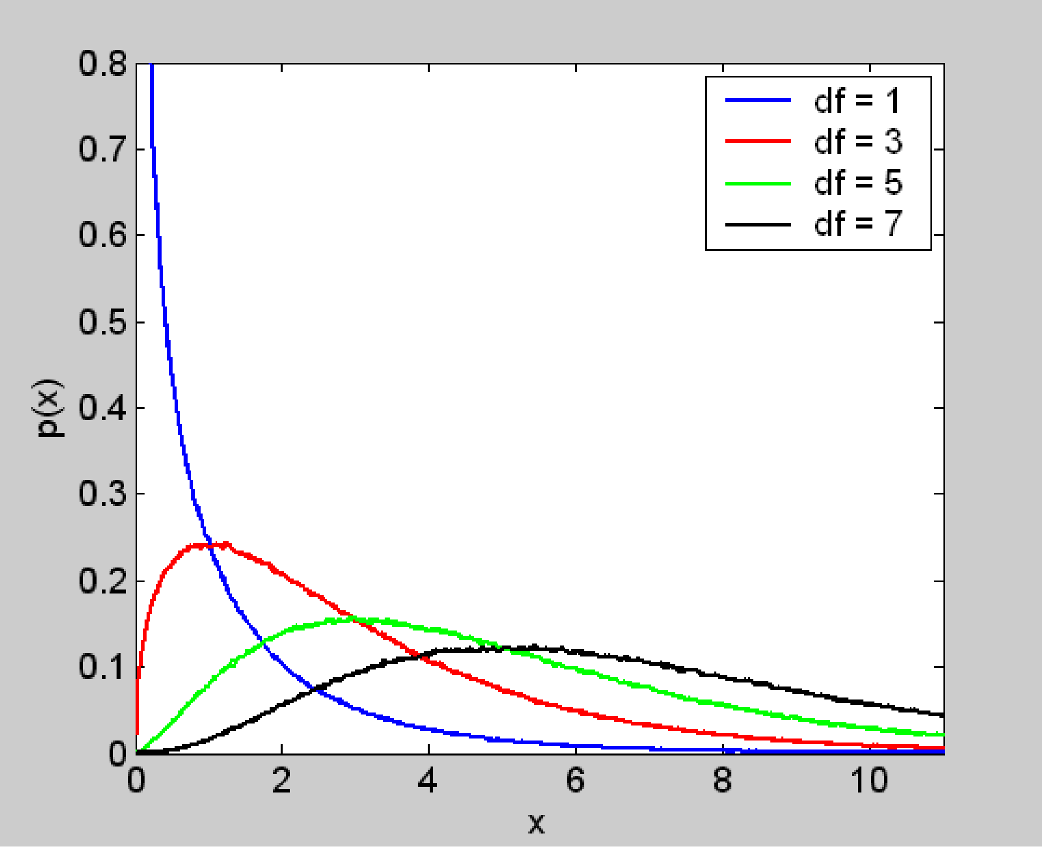
- For a 2*2 table closed form formula

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

- Giving $X^2 = 1.55$

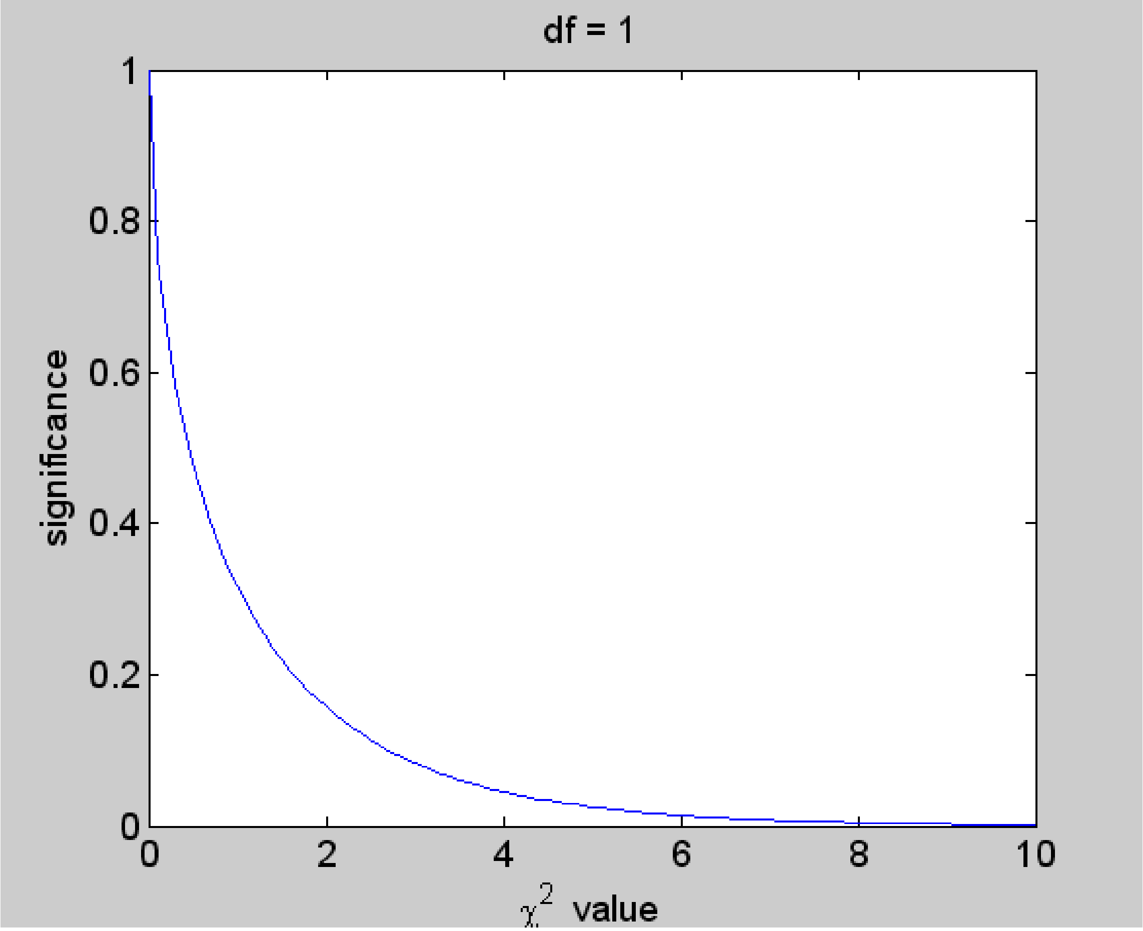
χ^2 distribution

- The χ^2 distribution depends on the parameter *df* = # of degrees of freedom. For a 2*2 table use df =1.



χ^2 Test – significance testing

- $X^2 = 1.55$
- PV = 0.21
- Discard hypothesis



χ^2 Test: Applications

- Identification of translation pairs in aligned corpora (Church and Gale, 1991).
- Corpus similarity (Kilgarriff and Rose, 1998).

Likelihood Ratios

- It is simply a number that tells us how much more likely one hypothesis is than the other.
- More appropriate for sparse data than the χ^2 test.
- A *likelihood ratio*, is more interpretable than the χ^2 or *t* statistic.

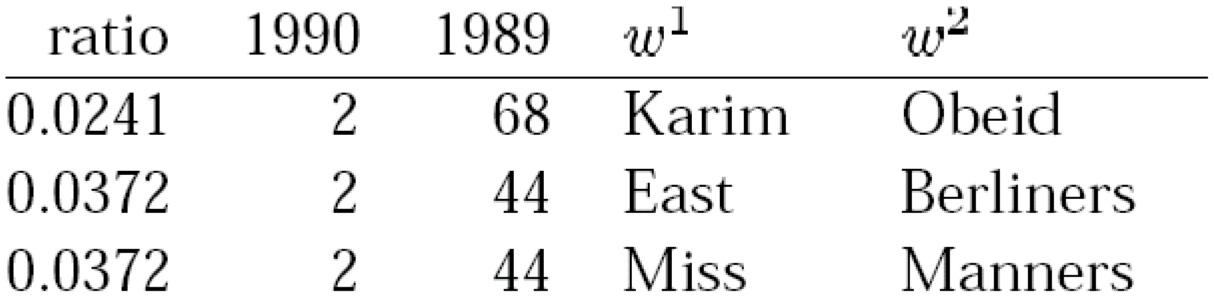
Likelihood Ratios: Within a Single Corpus (Dunning, 1993)

- 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:
- **Hypothesis 1:** The occurrence of w^2 is independent of the previous occurrence of w^1.
- **Hypothesis 2:** The occurrence of w^2 is dependent on the previous occurrence of w^1.
- The log likelihood ratio is then:

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

Relative Frequency Ratios (Damerau, 1993)

- 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.



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

Relative Frequency Ratios: Application

- This approach is most useful for the discovery of 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.

Pointwise Mutual Information

- An information-theoretically motivated measure for discovering interesting collocations is *pointwise mutual information* (Church et al. 1989, 1991; Hindle 1990).
- It is roughly a measure of how much one word tells us about the other.

Pointwise Mutual Information (Cont.)

- Pointwise mutual information between particular events x′ and y′, in our case the occurrence of particular words, is defined as follows:

$$\begin{aligned} I(x', y') &= \log_2 \frac{P(x'y')}{P(x')P(y')} \\ &= \log_2 \frac{P(x'|y')}{P(x')} \\ &= \log_2 \frac{P(y'|x')}{P(y')} \end{aligned}$$

Problems with using Mutual Information

- Decrease in uncertainty is not always a good measure of an interesting correspondence between two events.
- It is a bad measure of dependence.
- Particularly bad with sparse data.