## Collocations

## Prep Course Statistics

Alexandra, Jelke, Jesús, Nikos

## Contents

- Intro (what is a collocation)
- linguistically
- computationally
- Frequency-based methods
- Frequency
- Mean \& Variance
- Association measures
- Chi-square
- Likelihood ratios
- Fisher's exact test (Grammatical collocations)
- Minimum Sensitivity (Collostructions)
- Exercises

What is shown in this picture?


What is shown in this picture?
white wine


## What is shown in this picture?

white wine

yellow
white

How to translate this phrase?

## make a decision

Two approaches:

1. Find the word for make and for decision, then combine them according to the rules of the language:
*faire une décision
2. Find the word for decision, then find the word that performs a similar function to make when combined with it:

## A linguistically-motivated definition

Kahane, Polguere 2001: "a linguistic expression made up of at least two components:

1. the base of the collocation: a full lexical unit (e.g. smoker) which is "freely" chosen by the speaker;
2. the collocate: a lexical unit (e.g. heavy) or a multilexical expression which is chosen in a (partially) arbitrary way to express a given meaning and/or a grammatical structure contingent upon the choice of the base."

A linguistically-motivated definition (cont'd)

Collocations are also recursive:

- adopt a radical attitude towards sth
- play a central role
- conduct a thorough investigation
- an increasingly important concern
- following her strong recommendation
- I find it highly unlikely


## Collocation: a more relaxed definition

Manning \& Schutze, 1999: "an expression consisting of two or more words that correspond to some conventional way of saying things."
Three criteria are mentioned:

1. Non-compositionality (includes idioms):
to sell off
go all the way
throw in the towel
2. Non-substitutability:
white wine vs ??yellow wine
do me a favor vs ??make me a favor make the bed vs ??do the bed
3. Non-modifiability (mainly for idioms):
??throw in the white towel (works in Greek)

## Collocation: a more relaxed definition (cont'd)

The above definition includes:

- Phrasal verbs
- tell off, go down, get up
- Idioms
- Bite the bullet, throw in the towel
- Proper nouns
- New York, Eiffel Tower, The Doors
- Terminological expressions
- Computational Linguistics, power failure, citric acid
- Proper collocations (base-collocate combinations)
- strong coffee, sneaky attack, take a shower


## Collocation is not co-occurrence

- Some authors have generalized collocation to mean all frequently co-occurring words, e.g.:
- teacher-school
- beer-alcohol
- the ... of
- This is not the approach we follow in this presentation
- Instead, collocations are limited to grammatically bound elements that occur in a particular order
- Frequency remains the key for automatically identifying these expressions


## Frequency

- quantitative method
- bigrams in a text corpus
- very simple method

| $C\left(w^{1} w^{2}\right)$ | $w^{1}$ | $w^{2}$ |
| ---: | :--- | :--- |
| 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 |

## Frequency

- improved method by Justeso and Katz (1995)
- they added part-ofspeech patterns
- much better results
- works well for fixed phrase:

| $C\left(w^{1} w^{2}\right)$ | $w^{1}$ | $w^{2}$ | 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 | NN |
| 2106 | President | Bush | $\mathrm{N} N$ |
| 2001 | Middle | East | AN |
| 1942 | Saddam | Hussein | NN |
| 1867 | Soviet | Union | AN |
| 1850 | White | House | AN |
| 1633 | United | Nations | AN |
| 1337 | York | City | $\mathrm{N} N$ |
| 1328 | oil | prices | $\mathrm{N} N$ |
| 1210 | next | year | AN |
| 1074 | chief | executive | AN |
| 1073 | real | estate | AN |

## Mean \& Variance (Smadja 1993)

- works well for words in a more flexible relationship
- determines the distance between two words
- Smadja uses a less strict definition of collocation
- succesful at terminological extraction
(estimated 80\% accuracy)


## Mean \& Variance (Smadja 1993)

- e.g: - they knocked on the door - he knocked on his door knocked on Donaldson's door
- a man women knocked on the metal front door
- how to compute the mean offset:1/4(3+3+5+5)=4.0
- how to compute the variance:

$$
s^{2}=\frac{\sum_{i=1}^{n}\left(d_{i}-\bar{d}\right)^{2}}{n-1}
$$

( $\mathrm{n}=$ =number of times the words co-occur; di= the offset for co-occurence i ; $d=$ the sample mean of the offset)

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

## Mean \& Variance (Smadja 1993)

- low deviation: words usually occur at about the same distance
- zero derivation: words always occur at the same distance
- high derivation: words stand in no particular relationship to one another
- we can also determine the peaks of words:



## Pearson's $\chi^{2}$ Test

- Normally applied to 2-by-2 tables:

|  | $W_{1}=$ new | $W_{1}!=$ new |
| :--- | :--- | :--- |
| $W_{2}=$ companies | 8 | 4667 |
|  | (new companies) | (e.g., old companies) |
| $W_{2}!=$ companies | 15820 | 14287173 |
|  | (e.g., new machines) | (e.g., old machines) |

- "...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."


## Pearson's $\chi^{2}$ Test

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

i -> rows of the table
j -> columns of the table
$\mathrm{O}_{\mathrm{ij}}$ : observed value for cell (i,j)
$\mathrm{E}_{\mathrm{ij}}$ : expected value for cell ( $\mathrm{i}, \mathrm{j}$ )
for 2-by-2 tables:

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

## Pearson's $\chi^{2}$ Test

$\mathrm{E}_{\mathrm{ij}}$ are computed from the marginal probabilities.
In this case:

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

That is, if new and companies occurred completely independently of each other, we would expect 5.2 occurrences of new companies on average.

But since it is a 2-by-2 table we can calculate $X^{2}$ :

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

## Pearson's $\chi^{2}$ Test

Looking up the $\chi^{2}$ distribution,
at a probability level of $\alpha=0.05$

$$
x^{2}=3.841
$$

(the statistic has one degree of freedom for a 2-by-2 table)
$1.55<3.841$---> We cannot deny $\mathrm{H}_{0}$

## Pearson's $\chi^{2}$ Test

Appropiate for:

- Large Probabilities


## Do NOT apply when:

- The numbers in the 2-by-2 table are small.
- Total sample size < 20
- 20 < sample size < 40 and the expected value in any of the cells is 5 or less.


## 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 $\chi^{2}$ test.
- more interpretable.


## Likelihood Ratios

We examine the following two alternative explanations for the occurrence frequency of a bigram $w^{1} w^{2}$ (Dunning 1993):

- Hypothesis 1. $P\left(w^{2} \mid w^{1}\right)=p=P\left(w^{2} \mid \neg w^{1}\right)$
- Hypothesis 2. $P\left(w^{2} \mid w^{1}\right)=p_{1} \neq p_{2}=P\left(w^{2} \mid \neg w^{1}\right)$

The first one is a formalization of independence, the second one a formalization of dependence (collocation).

## Likelihood Ratios

## Assuming a binomial distribution:

$$
\begin{array}{lll}
P\left(w^{2} \mid w^{1}\right) & H_{1} & H_{2} \\
P\left(w^{2} \mid \neg w^{1}\right) & p=\frac{c_{2}}{N} & p_{1}=\frac{c_{12}}{c_{1}} \\
c_{12} \text { out of } c_{1} \text { bigrams are } w^{1} w^{2} & p=\frac{c_{2}}{N} & p_{2}=\frac{c_{2}-c_{12}}{N-c_{1}} \\
c_{2}-c_{12} \text { out of } N-c_{1} \text { bigrams are } \neg w^{1} w^{2} & \mathrm{~b}\left(c_{12} ; c_{1}, p\right) & \mathrm{b}\left(c_{2}-c_{12} ; N-c_{1}, p\right) \\
& \begin{array}{l}
\mathrm{b}\left(c_{12} ; c_{1}, p_{1}\right) \\
\log \lambda= \\
=\log \frac{L\left(c_{2}\right)}{L\left(H_{2}\right)} \\
=\log \frac{\mathrm{b}\left(c_{12}, c_{12} ; N-c_{1}, p\right) \mathrm{b}\left(c_{2}-c_{12}, N-p_{2}\right)}{\mathrm{b}\left(c_{12}, c_{1}, p_{1}\right) \mathrm{b}\left(c_{2}-c_{12}, N-c_{1}, p\right)} \\
=\log L\left(c_{12}, c_{1}, p\right)+\log L\left(c_{2}-c_{12}, N-c_{1}, p\right) \\
\\
\\
\\
\quad-\log L\left(c_{12}, c_{1}, p_{1}\right)-\log L\left(c_{2}-c_{12}, N-c_{1}, p_{2}\right)
\end{array}
\end{array}
$$

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

## Likelihood Ratios

| $-2 \log \lambda$ | $C\left(w^{1}\right)$ | $C\left(w^{2}\right)$ | $C\left(w^{1} w^{2}\right)$ | $w^{1}$ | $w^{2}$ |
| ---: | ---: | ---: | ---: | :--- | :--- |
| 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 |

Bigrams of powertul with the highest scores according to Dunning's likelihood ratio test.

## Likelihood Ratios

If $\lambda$ is a likelihood ratio of a particular form,
then $-2 \log \lambda$ is asymptotically $\chi^{2}$ distributed (Mood et al. 1974:440).

So we can use the values to test the null Hypothesis.
Asymptotically means "if the numbers are large enough"

In general the likelihood ratio test is more appropriate than Pearson's $\chi^{2}$ test for collocation discovery.

## Fisher's Exact Test

- Used to test for associations between two variables
- Identifying dependent bigrams
- Computes a p-value
- Calculates significance exactly (unlike $\chi^{2}$ test)
- Based on hypergeometric distribution
- Drawing from a finite population without replacement


## Using Fisher's Exact Test

- Natural language data is skewed
- Fisher's test does not require a normal distribution of data
- Sparse data problem
- Fisher's test can be used with small sample sizes

However, Fisher's Exact Test is more computationally intensive.

## Example: Determining animacy of

## nouns

- Human: doctor, player, photographer, Englishman
- Inanimate: banana, Netherlands, feeling, crime
- Automatically determine this based on cooccurrence with verbs
- The doctor thought John was right
- The banana thought John was right


## Fisher's Exact Test on animacy data

- Hypothesis: Animate nouns are associated with different verbs than inanimate nouns
- Variables:

1. Verb is "ontstaan" (to start, to arise)
2. Subject is "gevoel" (feeling)

- Binary variables
- 4 classifications


## Contingency table

- The Fisher's exact test is calculated using 2x2 tables
- Totals are fixed

The noun "gevoel" (feeling) as a subject of the verb "ontstaan" (to start, to arise)

|  | gevoel | -gevoel | Row totals |
| :--- | :--- | :--- | :--- |
| ontstaan | $\mathbf{2 9 8}$ | 5927 | $\mathbf{6 2 2 5}$ |
| -ontstaan | 405 | 111952 | 112357 |
| Column totals | $\mathbf{7 0 3}$ | 117879 | $\mathbf{1 1 8 5 8 2}$ |
| p $<\mathbf{0 . 0 0 0 0 1}$ |  |  |  |

## Collocation and its opposite

- The p-value can go both ways: Association strength

The noun "gevoel" (feeling) as a subject of the verb "schrijven" (to write)

|  | gevoel | -gevoel | Row totals |
| :--- | :--- | :--- | :--- |
| schrijven | $\mathbf{1}$ | 299 | $\mathbf{3 0 0}$ |
| -schrijven | 702 | 117578 | 118282 |
| Column totals | $\mathbf{7 0 3}$ | 117879 | $\mathbf{1 1 8 5 8 2}$ |
| p $>0.99999$ |  |  |  |

## Hypothesis

- This p-value can be used as a measure of association strength
- A low value indicates a strong association, a high value indicates none
- HO: The noun $x$ and the verb $y$ are independent in subject relations
- H1: The noun x occurs as a subject of the verb y more often than would be expected by chance


## Calculating the value

- The p-value expresses the total probability of the observed distribution (table) and all the more extreme ones

|  | gevoel | -gevoel |
| :--- | :--- | :--- |
| ontstaan | 298 | 5927 |
| -ontstaan | 405 | 111952 |
|  | gevoel | -gevoel |
|  | 300 | 5925 |
| ontstaan | 403 | 111950 |
| -ontstaan | 4 |  |


|  | gevoel | -gevoel |
| :--- | :--- | :--- |
| ontstaan | 299 | 5926 |
| -ontstaan | 404 | 111951 |
|  | gevoel | -gevoel |
|  | 301 | 5924 |
| ontstaan | 302 | 111949 |
| -ontstaan | 402 |  |

## Calculating the value

|  | gevoel | -gevoel | totals |
| :--- | :--- | :--- | :--- |
| ontstaan | 298 | 5927 | 6225 |
| -ontstaan | 405 | 111952 | 112357 |
| totals | 703 | 117879 | 118582 |

- $P(n)=\frac{6225!* 112357!* 703!* 117879!}{298!* 5927!* 405!* 111952!* 118582!}$
- $P(n+1)=\frac{6225!* 112357!* 703!* 117879!}{299!* 5926!* 404!* 111951!* 118582!}$
- $\mathrm{p}=P(n)+P(n+1)+P(n+2)+\ldots$
- A and B are associated more strongly than would be expected by chance ( $\alpha=0.001$ )


## Association strength

"gevoel" subject relations (inanimate)

| 0.000000000000000 | ontsta | arise |
| :--- | :--- | :--- |
| 0.000000000000830 | heb | have |
| 0.000000000002380 | speel | play |
| 0.000000000501125 | ben | be |
| 0.000000003404273 | zeg | say |
| 0.731409478841741 | krijg | get |
| 0.823487761949459 | spreek | speak |
| 0.853510038160385 | neem | take |
| 0.902189553992116 | ken | know |
| 1.000000000002866 | schrijf | write |

## Association strength

"hippie" subject relations (human)

| 0.001468162077883 | ga | go |
| :--- | :--- | :--- |
| 0.019216198962412 | kom | come |
| 0.048523337414639 | noem | call, name |
| 0.053750193619017 | zeg | say |
| 0.101731760645688 | vind | think, find |
| 0.847872307894773 | heb |  |
| 1.000000000000009 | maak | have |
|  | make |  |

## Application

- Hypothesis: Animate nouns are associated with different verbs than inanimate nouns
- Classification of nouns
- Distinguishing feature: The verbs that they occur with
- Use machine learning to classify nouns based on these features


## Fisher's Exact Test for association strength

- Fisher's Exact Test is a very robust measure
- It is computationally intensive
- Cannot compare data from samples of different sizes
- Does not show effect size


## Minimum Sensitivity

- Handles different sample sizes
- Less computationally demanding
- Measures effect size


## Collostructions

- ,Collostructions: Like collocations, but with constructions and words rather than words and words
- ) [sich V]

Johann und Peter [verteidigen] [sich]. Johann and Peter [defend] [themselves / each other].

- German sich: reflexive and reciprocal construction


## Calculating Minimum Sensitivity

$P$ (verb|construction) and $P$ (construction|verb)

|  | Sich | -Sich | totals |
| :--- | :--- | :--- | :--- |
| Fühlen | 4,603 | 12,550 | 17,153 |
| -Fühlen | 91,272 | $9,647,422$ | $9,738,694$ |
| totals | 95,875 | $9,659,972$ | $9,755,847$ |

$$
\begin{gathered}
S_{w 1}=\frac{4,603}{95,875}=\mathrm{P}(\mathrm{v} \mid \mathrm{c}) \quad S_{w 2}=\frac{4,603}{17,153}=\mathrm{P}(\mathrm{c} \mid \mathrm{v}) \\
M S=\min \left\{S_{w 1} ; S_{w 2}\right\}
\end{gathered}
$$

## Association strength

- Fisher's Exact Test p-value becomes too small with this much data
- These Minimum Sensitivity scores still work, and show effect size

| 1 --- sich<>zeigen | - 0.0236173981499705 |
| :---: | :---: |
| 2 --- sich<>handeln | -0.0196811651249754 |
| 3 --- sich<>machen | - 0.0186971068687266 |
| 4 --- sich<>stellen | 0.0185002952174769 |
| $5---$ sich<>befinden | 0.0159417437512301 |
| 6 --- sich<>füh1en | - 0.0149576854949813 |
| 7 --- sich<>halten | - 0.0147608738437315 |
| 8 --- sich<>setzen | - 0.0131863806337335 |
| 9 --- sich<>einigen | 0.0120055107262350 |
| 10 --- sich<>wenden | - 0.0116118874237355 |

## Exercise

- Task for mean and variance: Compute the mean and variance of these example sentences:
- He drives me mad
- She drives everyone around her mad
- He drives Tom's sister mad
- The disobedient pupil drives his teachers mad

$$
s^{2}=\frac{\sum_{i=1}^{n}\left(d_{i}-\bar{d}\right)^{2}}{n-1}
$$

( $\mathrm{n}=$ =number of times the words co-occur; di= the offset for cooccurence $i$; $d=$ the sample mean of the offset)

## Exercise

- Task for frequency: Add the missing tag patterns:

Tag Pattern Example<br>linear function<br>regression coefficients<br>Gaussian random variable<br>cumulative distribution function<br>mean squared error<br>class probability function<br>degrees of freedom

## Exercise

Using Minimum Sensitivity, calculate which of these verbs is more strongly associated with the past tense.

$$
\mathrm{MS}=\min \{\mathrm{p}(\mathrm{v} \mid \mathrm{c}) ; \mathrm{p}(\mathrm{c} \mid \mathrm{v})\}
$$

|  | Past tense | $\neg$ Past tense | totals |
| :--- | :--- | :--- | :--- |
| Remember | 10 | 7 | 17 |
| -Remember | 90 | 9,893 | 9,983 |
| totals | 100 | 9,900 | 10,000 |
|  | Past tense | $\neg$ Past tense | totals |
|  | 9 | 31 | 40 |
| Plan | 141 | 9,819 | 9,960 |
| -Plan | 150 | 9,850 | 10,000 |
| totals |  |  |  |

