Statistical Analysis of Corpus Data with R

You shall know a word by the company it keeps!

Collocation extraction with statistical association measures

— Part 1 —

Designed by Marco Baroni[1] and Stefan Evert[2]

[1]Center for Mind/Brain Sciences (CIMeC)
University of Trento

[2]Institute of Cognitive Science (IKW)
University of Onsabrück

Outline

Collocations & Multiword Expressions (MWE)

- What are collocations?
- Types of cooccurrence

Quantifying the attraction between words

- Contingency tables
- Contingency tables and hypothesis tests in R
- Practice session

What is a collocation?

- Words tend to appear in typical, recurrent combinations:

  day and *night*
  ring and *bell*
  milk and *cow*
  kick and *bucket*
  brush and *teeth*

What is a collocation?

- Words tend to appear in typical, recurrent combinations:

  day and *night*
  ring and *bell*
  milk and *cow*
  kick and *bucket*
  brush and *teeth*

☞ such pairs are called **collocations** (Firth 1957)

- the meaning of a word is in part determined by its characteristic collocations
- *"You shall know a word by the company it keeps!"*

What is a collocation?

- Native speakers have strong & widely shared intuitions about such collocations
- Collocational knowledge is essential for non-native speakers in order to sound natural ➪ “idiomatic English”

An important distinction . . .

. . . which has been the cause of many misunderstandings.

- **collocations** are an empirical linguistic phenomenon
  - can be observed in corpora & quantified
  - provide a window to lexical meaning and word usage
  - applications in language description (Firth 1957) and computational lexicography (Sinclair 1966, 1991)

An important distinction . . .

. . . which has been the cause of many misunderstandings.

- **collocations** are an empirical linguistic phenomenon
  - can be observed in corpora & quantified
  - provide a window to lexical meaning and word usage
  - applications in language description (Firth 1957) and computational lexicography (Sinclair 1966, 1991)

- **multiword expressions** = lexicalised word combinations
  - MWE need to be lexicalised (i.e., stored as units) because of certain idiosyncratic properties
  - non-compositionallity, non-substitutability, non-modifiability (Manning & Schütze 1999)
  - not observable, defined by linguistic tests (e.g. substitution test) and native speaker intuitions

An important distinction . . .

. . . which has been the cause of many misunderstandings.

- **collocations** are an empirical linguistic phenomenon
  - can be observed in corpora & quantified
  - provide a window to lexical meaning and word usage
  - applications in language description (Firth 1957) and computational lexicography (Sinclair 1966, 1991)
- **multiword expressions** = lexicalised word combinations
  - MWE need to be lexicalised (i.e., stored as units) because of certain idiosyncratic properties
  - non-compositionallity, non-substitutability, non-modifiability (Manning & Schütze 1999)
  - not observable, defined by linguistic tests (e.g. substitution test) and native speaker intuitions

☞ the term “collocations” has been used for both concepts

Outline

But what *are* collocations?

- Empirically, collocations are words that show an **attraction** towards each other (or a “mutual expectancy”)
  - in other words, a tendency to occur near each other
  - collocations can also be understood as statistically salient patterns that can be exploited by language learners

But what *are* collocations?

- Empirically, collocations are words that show an **attraction** towards each other (or a "mutual expectancy")
  - in other words, a tendency to occur near each other
  - collocations can also be understood as statistically salient patterns that can be exploited by language learners

- Linguistically, collocations are an **epiphenomenon** ...

But what *are* collocations?

- Empirically, collocations are words that show an **attraction** towards each other (or a “mutual expectancy”)
  - in other words, a tendency to occur near each other
  - collocations can also be understood as statistically salient patterns that can be exploited by language learners

- Linguistically, collocations are an **epiphenomenon** ...

  ... some might also say a hotchpotch ...

But what *are* collocations?

- Empirically, collocations are words that show an **attraction** towards each other (or a "mutual expectancy")
  - in other words, a tendency to occur near each other
  - collocations can also be understood as statistically salient patterns that can be exploited by language learners

- Linguistically, collocations are an **epiphenomenon** ...

  ... some might also say a hotchpotch ...

  ... of many different linguistic causes that lie behind the observed surface attraction.

Collocates of *bucket* (n.)

noun	*f*
water	183
spade	31
plastic	36
slop	14
size	41
mop	16
record	38
bucket	18
ice	22
seat	20
coal	16
density	11
brigade	10
algorithm	9
shovel	7
container	10
oats	7
sand	12
Rhino	7
champagne	10

verb	*f*
throw	36
fill	29
randomize	9
empty	14
tip	10
kick	12
hold	31
carry	26
put	36
chuck	7
weep	7
pour	9
douse	4
fetch	7
store	7
drop	9
pick	11
use	31
tire	3
rinse	3

adjective	*f*
large	37
single-record	5
cold	13
galvanized	4
ten-record	3
full	20
empty	9
steaming	4
full-track	2
multi-record	2
small	21
leaky	3
bottomless	3
galvanised	3
iced	3
clean	7
wooden	6
old	19
ice-cold	2
anti-sweat	1

Collocates of *bucket* (n.)

- opaque **idioms** (*kick the bucket*, but often used literally)
- **proper names** (*Rhino Bucket*, a hard rock band)
- noun **compounds**, lexicalised or productively formed (*bucket shop*, *bucket seat*, *slop bucket*, *champagne bucket*)
- **lexical collocations** = semi-compositional combinations (*weep buckets*, *brush one's teeth*, *give a speech*)
- cultural **stereotypes** (*bucket and spade*)
- **semantic compatibility** (*full, empty, leaky bucket*; *throw, carry, fill, empty, kick, tip, take, fetch a bucket*)
- **semantic fields** (*shovel, mop*; hypernym *container*)
- **facts** of life (*wooden bucket*; *bucket of water, sand, ice, . . .*)
- often sense-specific (*bucket size*, *randomize to a bucket*)

Operationalising collocations

- Firth introduced collocations as an essential component of his methodology, but without any clear definition

  > *Moreover, these and other technical words are given their 'meaning' by the restricted language of the theory, and by applications of the theory in quoted works.* (Firth 1957, 169)

- Empirical concept needs to be formalised and quantified
  - intuition: collocates are "attracted" to each other, i.e. they tend to occur near each other in text
  - definition of "nearness" ➪ **cooccurrence**
  - quantify the strength of attraction between collocates based on their recurrence ➪ cooccurrence **frequency**

☞ We will consider word pairs (w_1, w_2) such as (*brush*, *teeth*)

Outline

Different types of cooccurrence

1. **Surface cooccurrence**
   - criterion: surface distance measured in word tokens
   - words in a *collocational span* around the node word, may be symmetric (L5, R5) or asymmetric (L2, R0)
   - traditional approach in lexicography and corpus linguistics

Different types of cooccurrence

1. **Surface cooccurrence**
   - criterion: surface distance measured in word tokens
   - words in a *collocational span* around the node word, may be symmetric (L5, R5) or asymmetric (L2, R0)
   - traditional approach in lexicography and corpus linguistics

2. **Textual cooccurrence**
   - words cooccur if they are in the same text segment (sentence, paragraph, document, Web page, ...)
   - often used in Web-based research (➪ Web as corpus)

Different types of cooccurrence

1. **Surface cooccurrence**
    - criterion: surface distance measured in word tokens
    - words in a *collocational span* around the node word, may be symmetric (L5, R5) or asymmetric (L2, R0)
    - traditional approach in lexicography and corpus linguistics
2. **Textual cooccurrence**
    - words cooccur if they are in the same text segment (sentence, paragraph, document, Web page, ...)
    - often used in Web-based research (➪ Web as corpus)
3. **Syntactic cooccurrence**
    - words in a specific syntactic relation, e.g.
        - adjective modifying noun
        - subject / object noun of verb
        - N *of* N and similar patterns
    - suitable for extraction of MWE (Krenn & Evert 2001)

Types of cooccurrence: examples

Surface cooccurrence

- **Surface cooccurrences** of w_1 = *hat* with w_2 = *roll*
  - symmetric window of four words (L4, R4)
  - limited by sentence boundaries

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a **hat**. A man must not be precipitate, or he runs over it ; he must not rush into the opposite extreme, or he loses it altogether. [. . .] There was a fine gentle wind, and Mr. Pickwick's **hat** *rolled* sportively before it. The wind puffed, and Mr. Pickwick puffed, and the **hat** *rolled* over and over as merrily as a lively porpoise in a strong tide ; and on it might have *rolled*, far beyond Mr. Pickwick's reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate.

Types of cooccurrence: examples

Surface cooccurrence

- **Surface cooccurrences** of w_1 = *hat* with w_2 = *roll*
  - symmetric window of four words (L4, R4)
  - limited by sentence boundaries

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a **hat**. A man must not be precipitate, or he runs over it ; he must not rush into the opposite extreme, or he loses it altogether. [. . .] There was a fine gentle wind, and Mr. Pickwick's **hat** *rolled* sportively before it. The wind puffed, and Mr. Pickwick puffed, and the **hat** *rolled* over and over as merrily as a lively porpoise in a strong tide ; and on it might have *rolled*, far beyond Mr. Pickwick's reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate.

- **cooccurrence frequency** $f = 2$
- **marginal frequencies** $f_1 = f_2 = 3$

Types of cooccurrence: examples

Textual cooccurrence

- **Textual cooccurrences** of w_1 = *hat* and w_2 = *over*
  - textual units = sentences
  - multiple occurrences within a sentence ignored

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat.	hat	—
A man must not be precipitate, or he runs *over* it ;	—	over
he must not rush into the opposite extreme, or he loses it altogether.	—	—
There was a fine gentle wind, and Mr. Pickwick's hat rolled sportively before it.	hat	—
The wind puffed, and Mr. Pickwick puffed, and the hat rolled *over* and *over* as merrily as a lively porpoise in a strong tide ;	hat	over

Types of cooccurrence: examples

Textual cooccurrence

- **Textual cooccurrences** of w_1 = *hat* and w_2 = *over*
  - textual units = sentences
  - multiple occurrences within a sentence ignored

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat.	hat	—
A man must not be precipitate, or he runs *over* it ;	—	over
he must not rush into the opposite extreme, or he loses it altogether.	—	—
There was a fine gentle wind, and Mr. Pickwick's hat rolled sportively before it.	hat	—
The wind puffed, and Mr. Pickwick puffed, and the hat rolled *over* and *over* as merrily as a lively porpoise in a strong tide ;	hat	over

- coocurrence frequency $f = 1$
- marginal frequencies $f_1 = 3$, $f_2 = 2$

Types of cooccurrence: examples

Syntactic cooccurrence

- **Syntactic cooccurrences** of adjectives and nouns
  - every instance of the syntactic relation of interest is extracted as a **pair token**

In an *open barouche* [...] stood a *stout old gentleman*, in a *blue coat* and *bright buttons*, corduroy breeches and top-boots; two *young ladies* in scarfs and feathers; a *young gentleman* apparently enamoured of one of the *young ladies* in scarfs and feathers; a lady of *doubtful age*, probably the aunt of the aforesaid; and [...]

➜

open	barouche
stout	gentleman
old	gentleman
blue	coat
bright	button
young	lady
young	gentleman
young	lady
doubtful	age

Types of cooccurrence: examples

Syntactic cooccurrence

- **Syntactic cooccurrences** of adjectives and nouns
  - every instance of the syntactic relation of interest is extracted as a **pair token**

In an *open barouche* [. . .] stood a *stout old gentleman*, in a *blue coat* and *bright buttons*, corduroy breeches and top-boots ; two *young ladies* in scarfs and feathers ; a *young gentleman* apparently enamoured of one of the *young ladies* in scarfs and feathers ; a lady of *doubtful age*, probably the aunt of the aforesaid ; and [. . .]

➜

open	barouche
stout	gentleman
old	gentleman
blue	coat
bright	button
young	lady
young	gentleman
young	lady
doubtful	age

Cooccurrency frequency data for *young gentleman*:

- coocurrence frequency $f = 1$
- marginal frequencies $f_1 = f_2 = 3$

Outline

Collocations & Multiword Expressions (MWE)

- What are collocations?
- Types of cooccurrence

Quantifying the attraction between words

- Contingency tables
- Contingency tables and hypothesis tests in R
- Practice session

Quantifying attraction

- Quantitative measure for attraction between words based on their recurrence ➪ **cooccurrence frequency**

Quantifying attraction

- Quantitative measure for attraction between words based on their recurrence ➪ **cooccurrence frequency**
- But cooccurrence frequency is not sufficient
  - bigram *is to* occurs $f = 260$ times in Brown corpus
  - but both components are so frequent ($f_1 \approx 10{,}000$ and $f_2 \approx 26{,}000$) that one would also find the bigram 260 times if words in the text were arranged in completely random order

Quantifying attraction

- Quantitative measure for attraction between words based on their recurrence ➪ **cooccurrence frequency**
- But cooccurrence frequency is not sufficient
  - bigram *is to* occurs $f = 260$ times in Brown corpus
  - but both components are so frequent ($f_1 \approx 10{,}000$ and $f_2 \approx 26{,}000$) that one would also find the bigram 260 times if words in the text were arranged in completely random order

  ☞ take **expected frequency** into account as "baseline"
- Statistical model required to bring in notion of "chance cooccurrence" and to adjust for sampling variation

Quantifying attraction

- Quantitative measure for attraction between words based on their recurrence ➪ **cooccurrence frequency**
- But cooccurrence frequency is not sufficient
  - bigram *is to* occurs $f = 260$ times in Brown corpus
  - but both components are so frequent ($f_1 \approx 10{,}000$ and $f_2 \approx 26{,}000$) that one would also find the bigram 260 times if words in the text were arranged in completely random order

☞ take **expected frequency** into account as "baseline"

- Statistical model required to bring in notion of "chance cooccurrence" and to adjust for sampling variation
  - ☞ NB: bigrams can be understood either as syntactic cooccurrences (adjacency relation) or as surface cooccurrences (L1, R0 or L0, R1)

Attraction as statistical association

- Tendency of events to cooccur = **statistical association**
  - statistical measures of association are available for **contingency tables**, resulting from a **cross-classification** of a set of “items” according to two (binary) factors
  - cross-classifying factors represent the two events

Attraction as statistical association

- Tendency of events to cooccur = **statistical association**
  - statistical measures of association are available for **contingency tables**, resulting from a **cross-classification** of a set of "items" according to two (binary) factors
  - cross-classifying factors represent the two events
- Application to word cooccurrence data
  - most natural for **syntactic cooccurrences**
  - "items" are pair tokens = instances of syntactic relation
  - factor 1: Is first component of pair token an instance of word type w_1?
  - factor 2: Is second component of pair token an instance of word type w_2?

Outline

Contingency table of observed frequencies

For **syntactic** cooccurrences

	$* \mid w_2$	$* \mid \neg w_2$	
$w_1 \mid *$	O_{11}	O_{12}	$= f_1$
$\neg w_1 \mid *$	O_{21}	O_{22}	
	$= f_2$		$= N$

	$* \mid$gent.	$* \mid \neg$gent.	
young $\mid *$	1	2	$= 3$
$\neg$young $\mid *$	2	4	
	$= 3$		$= 9$

In an *open barouche* [. . .] stood a *stout old gentleman*, in a *blue coat* and *bright buttons*, corduroy breeches and top-boots ; two *young ladies* in scarfs and feathers ; a *young gentleman* apparently enamoured of one of the *young ladies* in scarfs and feathers ; a lady of *doubtful age*, probably the aunt of the aforesaid ; and [. . .]

➜

open	barouche
stout	gentleman
old	gentleman
blue	coat
bright	button
young	lady
young	gentleman
young	lady
doubtful	age

Contingency table of observed frequencies

For **textual** cooccurrences (sentence windows)

	$w_2 \in S$	$w_2 \notin S$	
$w_1 \in S$	O_{11}	O_{12}	$= f_1$
$w_1 \notin S$	O_{21}	O_{22}	
	$= f_2$		$= N$

	over $\in S$	over $\notin S$	
hat $\in S$	1	2	$= 3$
hat $\notin S$	1	1	
	$= 2$		$= 5$

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat.	hat	—
A man must not be precipitate, or he runs *over* it ;	—	over
he must not rush into the opposite extreme, or he loses it altogether.	—	—
There was a fine gentle wind, and Mr. Pickwick's hat rolled sportively before it.	hat	—
The wind puffed, and Mr. Pickwick puffed, and the hat rolled *over* and *over* as merrily as a lively porpoise in a strong tide ;	hat	over

Contingency table of observed frequencies

For **surface** cooccurrences (L4, R4)

	w_2	$\neg w_2$	
$near(w_1)$	O_{11}	O_{12}	$\approx k \cdot f_1$
$\neg\, near(w_1)$	O_{21}	O_{22}	
	$= f_2$		$= N - f_1$

	roll	$\neg$roll	
$near$(hat)	2	18	$= 20$
$\neg\, near$(hat)	1	87	
	$= 3$		$= 108$

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a **hat**. A man must not be precipitate, or he runs over it ; he must not rush into the opposite extreme, or he loses it altogether. [. . .] There was a fine gentle wind, and Mr. Pickwick's **hat** *rolled* sportively before it. The wind puffed, and Mr. Pickwick puffed, and the **hat** *rolled* over and over as merrily as a lively porpoise in a strong tide ; and on it might have *rolled*, far beyond Mr. Pickwick's reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate.

More details: Section 5.1 of Evert, S. (2008, in press). Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 57. Mouton de Gruyter, Berlin.

Measuring association in contingency tables

A) Measures of **significance**

- apply statistical hypothesis test with null hypothesis H_0: independence of rows and columns
- H_0 implies there is no association between w_1 and w_2
- **association score** = test statistic or p-value
- one-sided vs. two-sided tests

☞ amount of evidence for association between w_1 and w_2

B) Measures of **effect-size**

- compare observed frequencies O_{ij} to **expected frequencies** E_{ij} under H_0 (➪ later)
- or estimate conditional prob. $\Pr(w_2 \mid w_1)$, $\Pr(w_1 \mid w_2)$, etc.
- maximum-likelihood estimates or confidence intervals

☞ strength of the attraction between w_1 and w_2

Outline

Contingency tables in R

- Contingency table is represented as a **matrix** in R, i.e. a rectangular array of numbers
  - looks like numeric data frame, but different internally
- E.g. for the following observed frequencies: $O_{11} = 9$, $O_{12} = 47$, $O_{21} = 82$, $O_{22} = 956$

Contingency tables in R

- Contingency table is represented as a **matrix** in R, i.e. a rectangular array of numbers
  - looks like numeric data frame, but different internally
- E.g. for the following observed frequencies: $O_{11} = 9$, $O_{12} = 47$, $O_{21} = 82$, $O_{22} = 956$

```
> A <- matrix(c(10,47,82,956),
  nrow=2, ncol=2, byrow=TRUE)
> A
```

```
# construct matrix from row (or column) vectors
> A <- rbind(c(10,47), c(82,956))
```

Independence tests in R

```
# chi-squared test is the standard independence test
> chisq.test(A)

# use test statistic as association score, p-value for interpretation

# Is there significant evidence for a collocation?

# Fisher's exact test works better for small samples and skewed tables
> fisher.test(A)
```

Interpreting hypothesis tests as association scores

- Establishing significance
  - p-value = probability of observed (or more "extreme") contingency table if H_0 is true
  - theory: H_0 can be rejected if p-value is below accepted **significance level** (commonly .05, .01 or .001)
  - practice: nearly all word pairs are highly significant

Interpreting hypothesis tests as association scores

- Establishing significance
  - p-value = probability of observed (or more "extreme") contingency table if H_0 is true
  - theory: H_0 can be rejected if p-value is below accepted **significance level** (commonly .05, .01 or .001)
  - practice: nearly all word pairs are highly significant
- Test statistic = significance association score
  - **convention** for association scores: high scores indicate strong attraction between words
  - satisfied by **test statistic** X^2, but not by p-value
  - Fisher's test: transform p-value, e.g. $-\log_{10} p$

Interpreting hypothesis tests as association scores

- Establishing significance
  - p-value = probability of observed (or more "extreme") contingency table if H_0 is true
  - theory: H_0 can be rejected if p-value is below accepted **significance level** (commonly .05, .01 or .001)
  - practice: nearly all word pairs are highly significant
- Test statistic = significance association score
  - **convention** for association scores: high scores indicate strong attraction between words
  - satisfied by **test statistic** X^2, but not by p-value
  - Fisher's test: transform p-value, e.g. $-\log_{10} p$
- Odds ratio as measure of effect size
  - Fisher's test also provides estimate for **odds ratio** θ, an effect-size measure for association strength
  - log odds ratio $\log \theta$ as effect-size association score (0 for independence, large values indicate strong attraction)
  - conservative estimate = lower bound of confidence interval

Association scores from hypothesis tests

```
# chi-squared statistic X^2 as association score
> chisq.test(A)$statistic

# p-value of Fisher's test and corresponding association score
> fisher.test(A)$p.value
> -log10(fisher.test(A)$p.value)

# NB: chi-squared and Fisher scores are not on same scale

# log odds ratio and conservative estimate
> log(fisher.test(A)$estimate)
> log(fisher.test(A)$conf.int[1])

> str(fisher.test(A))    # or read help page carefully
```

Association scores from hypothesis tests

```
# define two further (invented) contingency tables
> B1 <- rbind(c(16,84), c(84,816))
> B2 <- rbind(c(1,99), c(99,801))

# calculate chi-squared and Fisher scores for the two tables,
# as well as estimates for their log odds ratios

# Do the results look plausible to you? What is wrong?
```

One-sided vs. two-sided association scores

- Chi-squared and Fisher are **two-sided** tests
  - calculate high association scores (= low p-values) both for strong positive association (**attraction**) and for strong negative association (**repulsion**)
  - we are usually interested in attraction only (unless we are looking for "anti-collocations")

One-sided vs. two-sided association scores

- Chi-squared and Fisher are **two-sided** tests
  - calculate high association scores (= low p-values) both for strong positive association (**attraction**) and for strong negative association (**repulsion**)
  - we are usually interested in attraction only (unless we are looking for “anti-collocations”)
- Fisher can be applied as **one-sided** test
  - we are only interested in the **alternative** to H_0 that there is greater than chance cooccurrence, not in the alternative of less than chance cooccurrence

```
> fisher.test(B1, alternative="greater")
# high scores (significance and log odds ratio)
> fisher.test(B2, alternative="greater")
# low scores (significance and log odds ratio)
```

Outline

Practice: bigrams in the Brown corpus

- Data set of bigrams with $f \geq 5$ in the Brown corpus
  - available on course homepage as `brown_bigrams.tbl`
- 24,167 rows (= bigrams) with variables:
  - **id** = numeric ID of bigram
  - **word1** = first word (e.g. *long* for *long time*)
  - **pos1** = part-of-speech code (e.g. `J` for adjective)
  - **word2** = second word (e.g. *time* for *long time*)
  - **pos2** = part-of-speech code (e.g. `N` for noun)
  - **O11** = observed cooccurrence frequency O_{11}
  - **O12** = observed frequency O_{12}
  - **O21** = observed frequency O_{21}
  - **O22** = observed frequency O_{22}

Practice: bigrams in the Brown corpus

```
> Brown <- read.delim("brown_bigrams.tbl")

# Now select a number of bigrams (e.g. low and high cooccurrence
# frequency, or specific part-of-speech combinations), construct
# the corresponding contingency tables in matrix form,
# and calculate the different association scores you know.
# Can you find a bigram with strong negative association?

# NB: You can use the same tests for corpus frequency comparisons.
# Assume that a certain expression occurs 50 times in the 100,000
# tokens of corpus A, and twice in the 1,000 tokens of corpus B.
# What is an appropriate contingency table for these data, and what
# results do you obtain from the chi-squared and Fisher test?
```