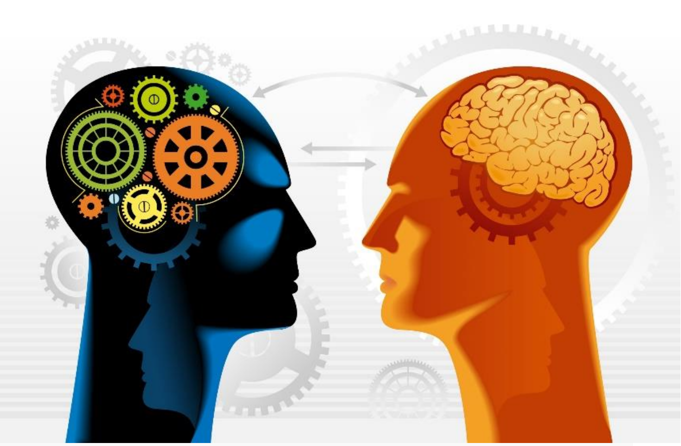
NHH

# CORPUS APPROACHES TO TEXTUAL DATA

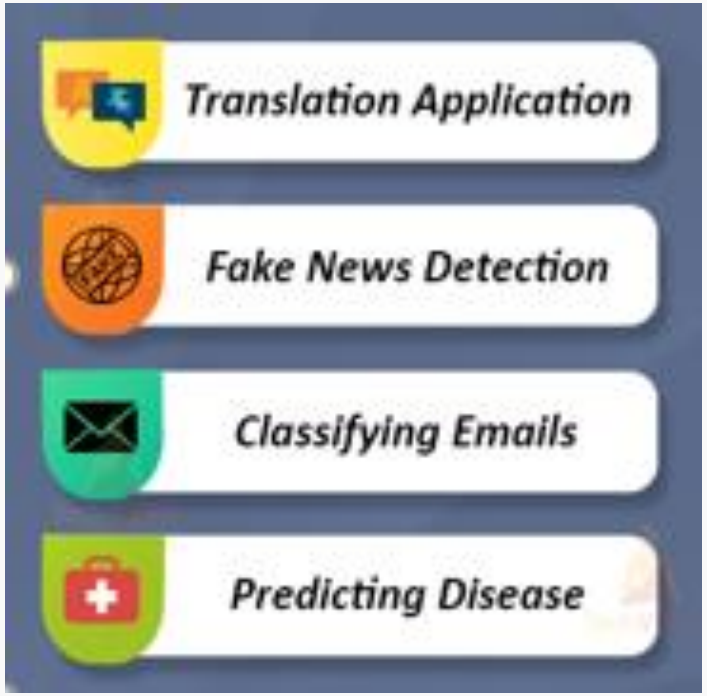
BAN432 APPLIED TEXTUAL DATA FOR BUSINESS AND FINANCE
GUEST LECTURE
GISLE ANDERSEN

# Text data analysis and related fields

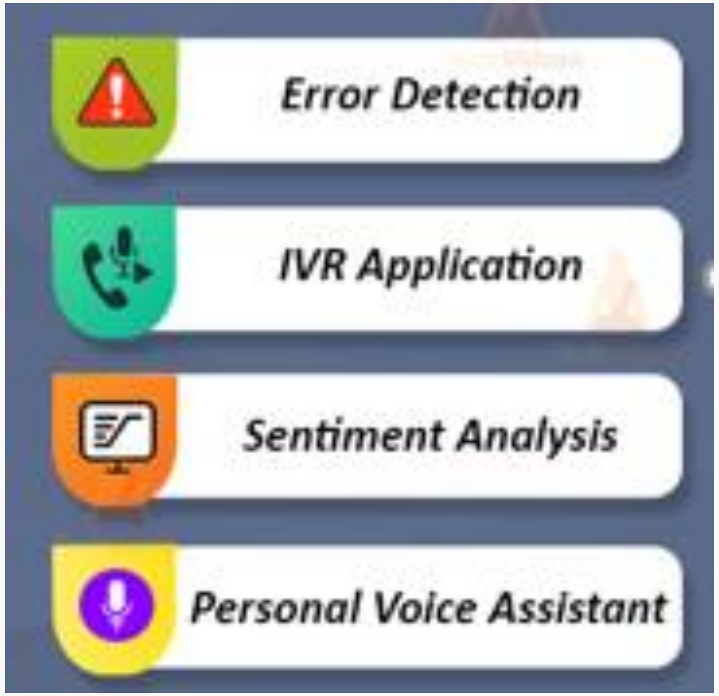


- Corpus linguistics
- Computational linguistics
- Language technology (speech/text)
- Natural language processing
- Artificial intelligence

# Use cases of NLP



- Translation Application
- Fake News Detection
- Classifying Emails
- Predicting Disease



- Error Detection
- IVR Application
- Sentiment Analysis
- Personal Voice Assistant

# Example of relevant research

- Phone discussions in quarterly earnings conference calls involving CEOs/CFOs
- Linguistically-based analysis
- Correlated calls with financial statements
- Labelled call data as “truthful” or “deceptive”

Larcker & Zakolyukina. 2012. Detecting Deceptive Discussions in Conference Calls. *Jrnl of Accounting Research*. 50 (2), 495-540.

- Language of deceptive executives: more references to general knowledge, fewer non-extreme positive emotions and fewer refs to shareholder value.
- Deceptive CEOs use significantly more extreme positive emotion words and fewer anxiety words

*Very good;*
*Highly positive*

*I fear that …;*
*… should be regarded with caution*

# Motivational credo

- Linguistic processing/language technology tools are useful.
- Annotating text according to linguistic categories may be a valuable supplement to strictly statistical methods.
- This can potentially increase accuracy in applications such as sentiment analysis, machine translation, etc.
- But also have obvious appliances in business and finance.
- E.g. finding out whether tweets about a certain brand/product/company are truly positive or negative
- And contributing more generally to our understanding of how texts are used to achieve strategic objectives.

# Outline

- Language (&) technology and the relevance of linguistic processing
  - Brief student activity
- Keywords/keyness – term extraction
- Collocation – the systematic co-occurrence of words
  - Demo and application: The Sketch Engine
  - Resources in R
- Concluding remarks

# Outline

- Language (&) technology and the relevance of linguistic processing
  - Brief student activity
- Keywords/keyness – term extraction
- Collocation – the systematic co-occurrence of words
  - Demo and application: The Sketch Engine
  - Resources in R
- Concluding remarks

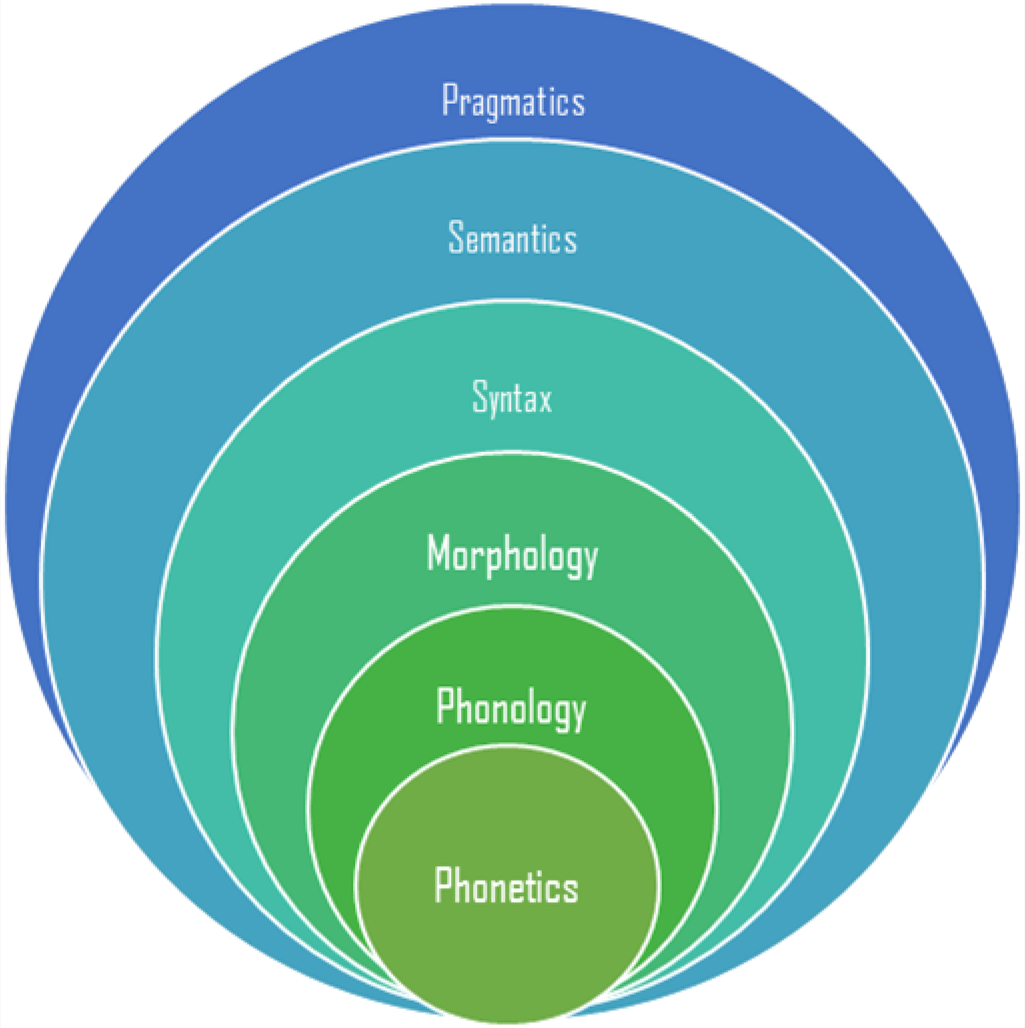
# Main learning objectives

- Understand the relevance of methods in NLP/CL for business studies.
- Learn about basic concepts and methods in linguistics/NLP.
- Familiarise ourselves with the concepts keywords/keyness
  - i.e. the fact that certain words and phrases occur significantly more frequently in one text collection than another.
- Familiarise ourselves with the significance of collocation
  - (NO=*samforekomst*) in text, i.e. the fact that two or more words occur together more often than by chance.
- Briefly hear about some useful resources in R
- Get acquainted with the infrastructure Sketch Engine

CORPUS APPROACHES TO TEXTUAL DATA

# BASIC CONCEPTS IN CORPUS LINGUISTICS AND NATURAL LANGUAGE PROCESSING

# Levels of analysis and modelling



Pragmatics: reference assignment, disambiguation

Semantics: *he* = male person;
*bolt* = “a stout pin for fastening” OR “a sudden spring or start”

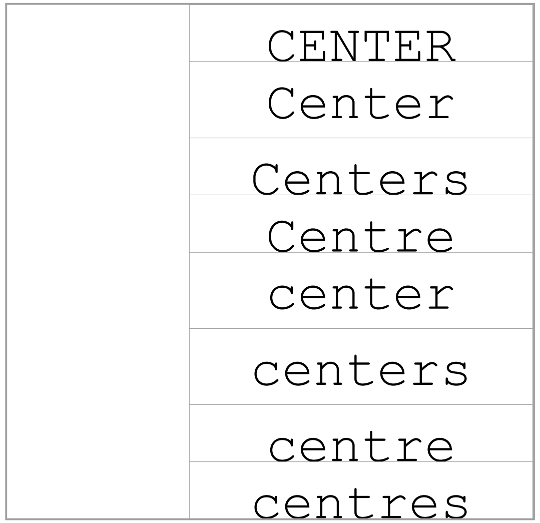
```
[(he, ‘PP’), (made, ‘VVD’),
(a, ‘DT’), (bolt, NN),
(for, ‘IN’), (the, ‘DT’),
(door, ‘NN’)]
```

```
/hI meId @ b@Ult fQ D@ dO:/
```

He made a bolt for the door.

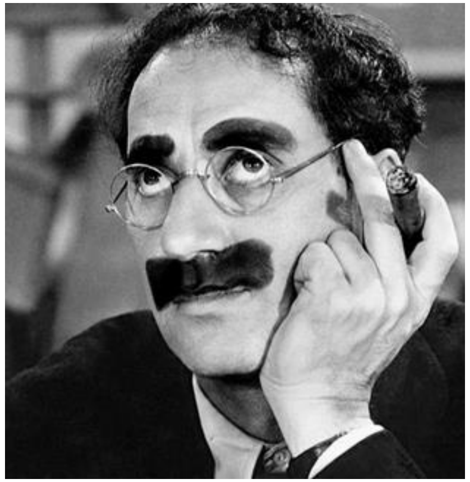
# Aspects of linguistic processing in text data analysis

- Lemmatisation (grouping: *center* – *centre* – *centres …*)
- Word class tagging, disambiguation (*center* noun/adj/verb)
- Semantic annotation (e.g. ‘middle point’ vs. ‘location’)
- Phraseological analysis (e.g. *centre of attention/balance/excellence*)
- Syntactic parsing (subj. obj., etc.)
- Discourse annotation (new/given information, speech acts)



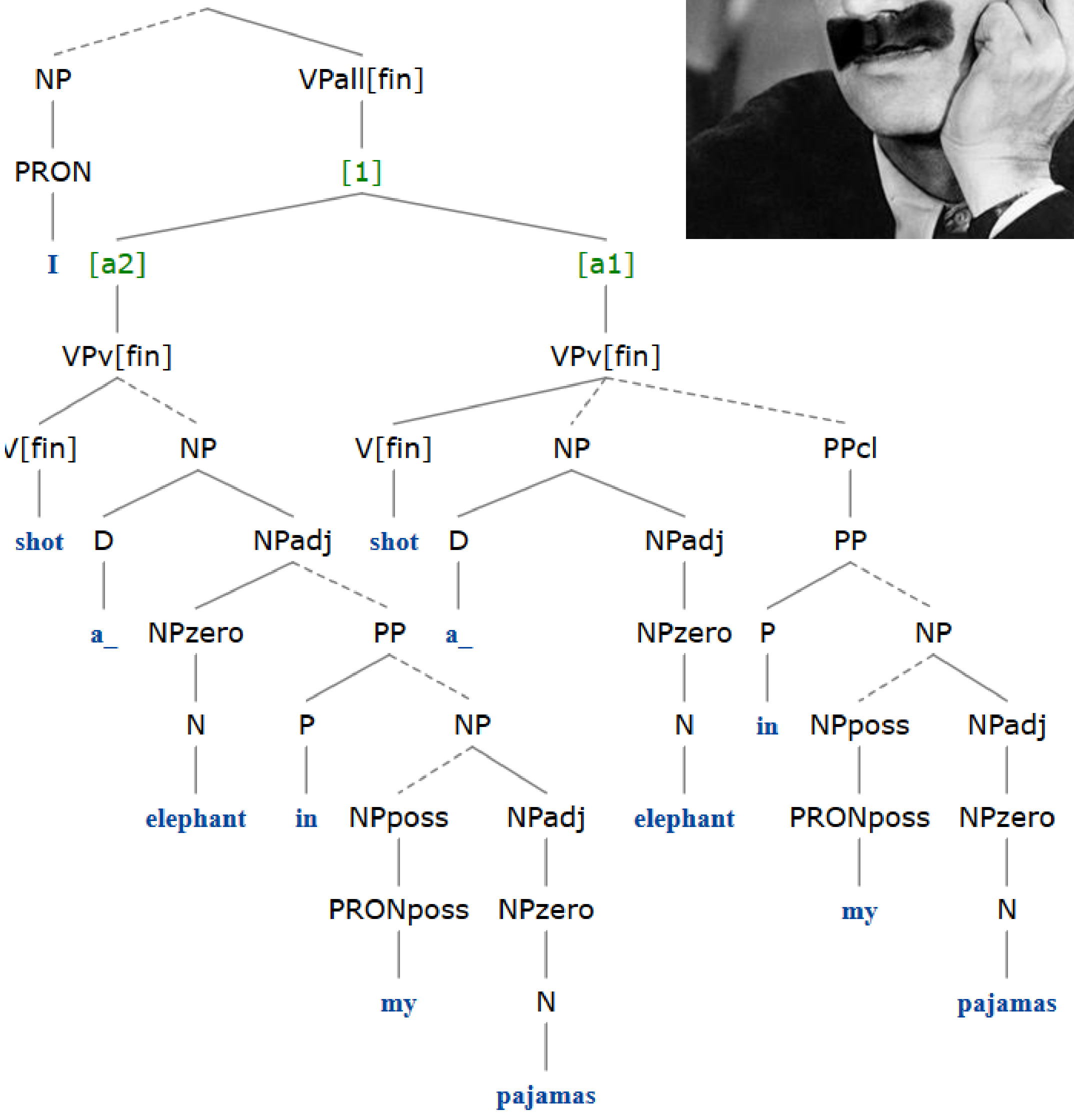
| | |
|---|---|
| | CENTER |
| | Center |
| | Centers |
| | Centre |
| | center |
| | centers |
| | centre |
| | centres |

# Syntactic parsing / ambiguity resolution



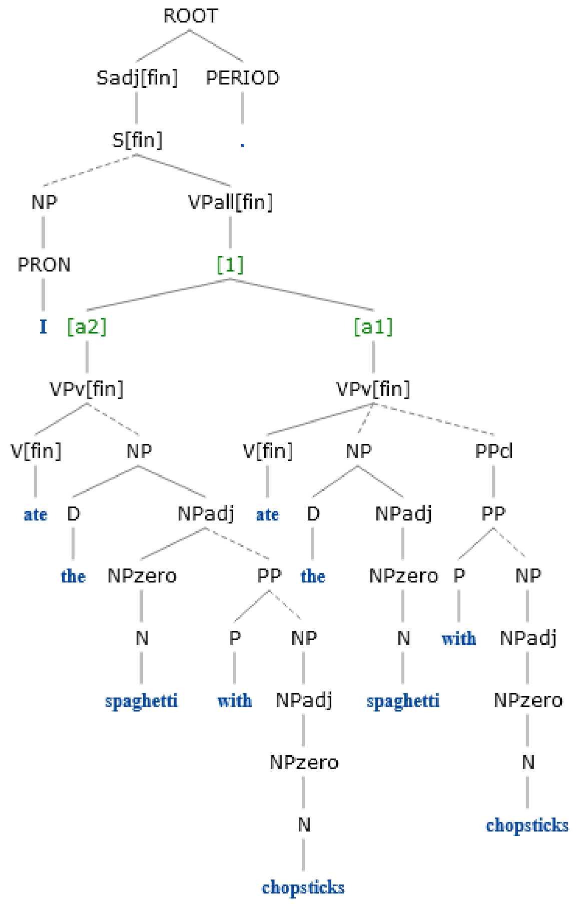
- *One morning I shot an elephant in my pyjamas.*
- *How he got into my pyjamas I'll never know.*

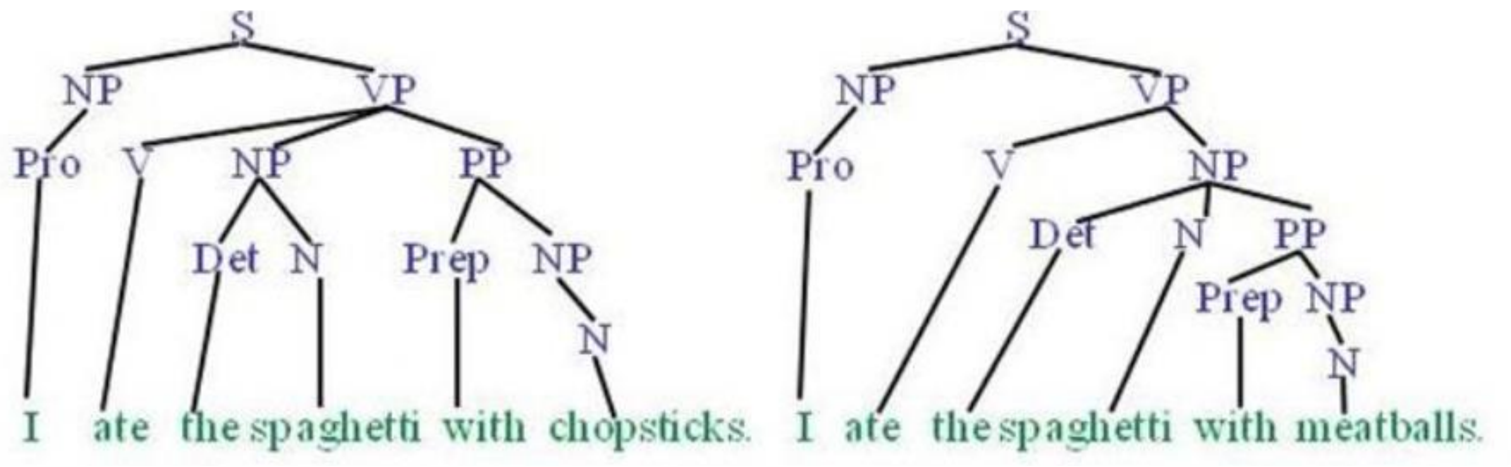
- **Syntactic parsing**: assigning syntactic structure to a sentence
  - phrase structure
  - subject/object
  - active/passive
  - *wh*-cleft; *it*/*there*; etc.
- Expressed (visualised) as parse tree or symbolic output (string)
- Often many outputs of same sentence (structural ambiguity)



# Why parsing?

- To get the <u>meaning</u> right!
- S1: “I ate the spaghetti with chopsticks.”
- [(I, ‘PP’), (ate, ‘VVD’), (the, ‘DT’), (spaghetti, NN), (with, ‘IN’), (chopsticks, ‘NNS’)]
- S2: “I ate the spaghetti with meatballs.”
- [(I, ‘PP’), (ate, ‘VVD’), (the, ‘DT’), (spaghetti, NN), (with, ‘IN’), (meatballs, ‘NNS’)]





# Task: check your own mother tongue

- Discuss in pairs/groups of three.

1. Translate the two “spaghetti” sentences into your mother tongue.
   a) To what extent does your language display the same syntactic ambiguity as in English?
   b) What happens to the preposition “with”?

2. Think of a lexically ambiguous word in your language.
3. Try to construct a syntactically ambiguous sentence in a language of your choice.
4. Can you think of translations problems resulting from these ambiguities?

Corpus approaches to textual data

# KEYWORDS/ KEYNESS AND TERM EXTRACTION

# Keywords

- Keywords: words that occur relatively more often in a te corpus being analysed than in some reference corpus
- Reference corpus usually large, general language (e.g. BNC)
- Frequency difference must be statistically significant
- A way of highlighting lexical saliency
- i.e. identifying words that stand out as signposts, thus identifying the ‘aboutness’
- Relevant for the analysis of trends and developments
- Useful technique for term extraction

# General vs. domain-specific corpus
# Case: economics (Bondi 2010)

- Keywords: *employment, interest, money, rate, investment*
- Key clusters: *the rate of interest, (the) marginal efficiency of capital, the quantity of money*
- Organisational key phrases: *in terms of, is equal to, it follows that, as a whole*
- Grammatical words: *of*, *which*
  - (*income/factors/consumption*)(,) *which*
  - *the* X *of* Y

- Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. New York: Harcourt & Brace.
- Corpus of economics articles
  - cf. Bondi, M. 2010. Perspectives on keywords and keyness. In Bondi, M. & M. Scott. *Keyness in texts*. Amsterdam: John Benjamins, 1-18.

# Male vs. female speech (Rayson et al. 1997)

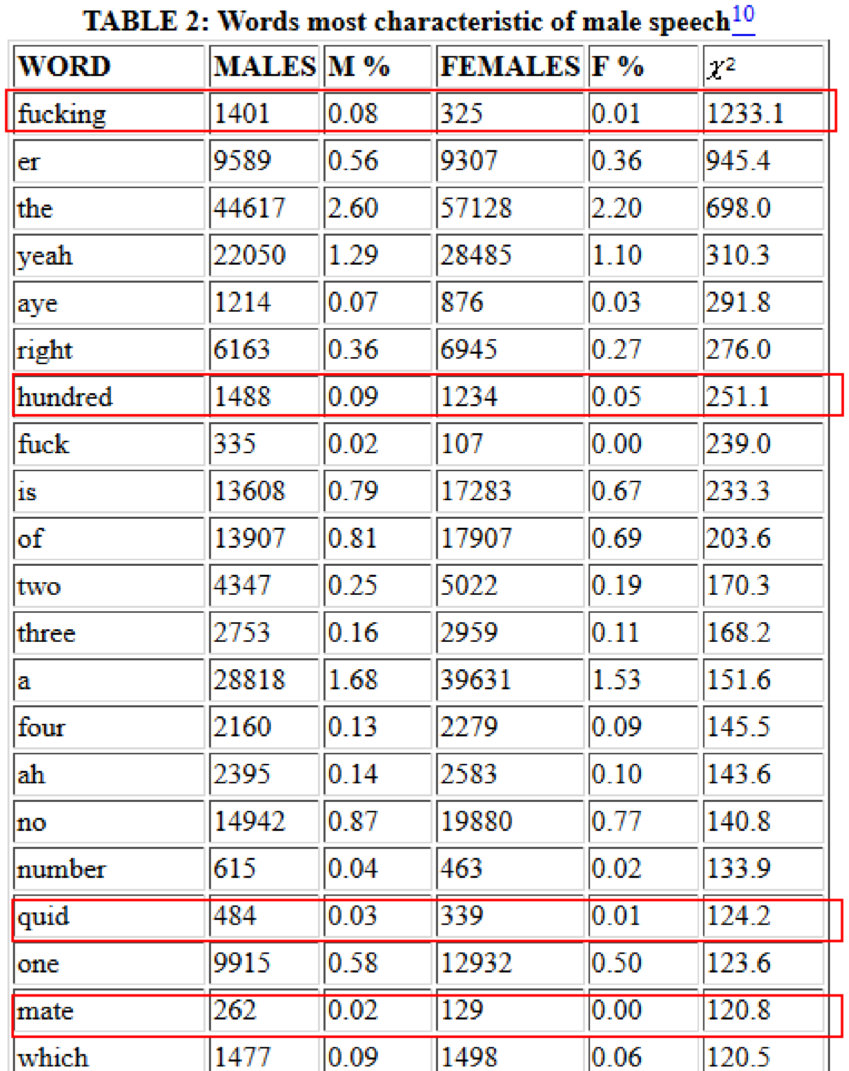


TABLE 2: Words most characteristic of male speech[10]

| WORD | MALES | M % | FEMALES | F % | $\chi^2$ |
|---|---|---|---|---|---|
| fucking | 1401 | 0.08 | 325 | 0.01 | 1233.1 |
| er | 9589 | 0.56 | 9307 | 0.36 | 945.4 |
| the | 44617 | 2.60 | 57128 | 2.20 | 698.0 |
| yeah | 22050 | 1.29 | 28485 | 1.10 | 310.3 |
| aye | 1214 | 0.07 | 876 | 0.03 | 291.8 |
| right | 6163 | 0.36 | 6945 | 0.27 | 276.0 |
| hundred | 1488 | 0.09 | 1234 | 0.05 | 251.1 |
| fuck | 335 | 0.02 | 107 | 0.00 | 239.0 |
| is | 13608 | 0.79 | 17283 | 0.67 | 233.3 |
| of | 13907 | 0.81 | 17907 | 0.69 | 203.6 |
| two | 4347 | 0.25 | 5022 | 0.19 | 170.3 |
| three | 2753 | 0.16 | 2959 | 0.11 | 168.2 |
| a | 28818 | 1.68 | 39631 | 1.53 | 151.6 |
| four | 2160 | 0.13 | 2279 | 0.09 | 145.5 |
| ah | 2395 | 0.14 | 2583 | 0.10 | 143.6 |
| no | 14942 | 0.87 | 19880 | 0.77 | 140.8 |
| number | 615 | 0.04 | 463 | 0.02 | 133.9 |
| quid | 484 | 0.03 | 339 | 0.01 | 124.2 |
| one | 9915 | 0.58 | 12932 | 0.50 | 123.6 |
| mate | 262 | 0.02 | 129 | 0.00 | 120.8 |
| which | 1477 | 0.09 | 1498 | 0.06 | 120.5 |

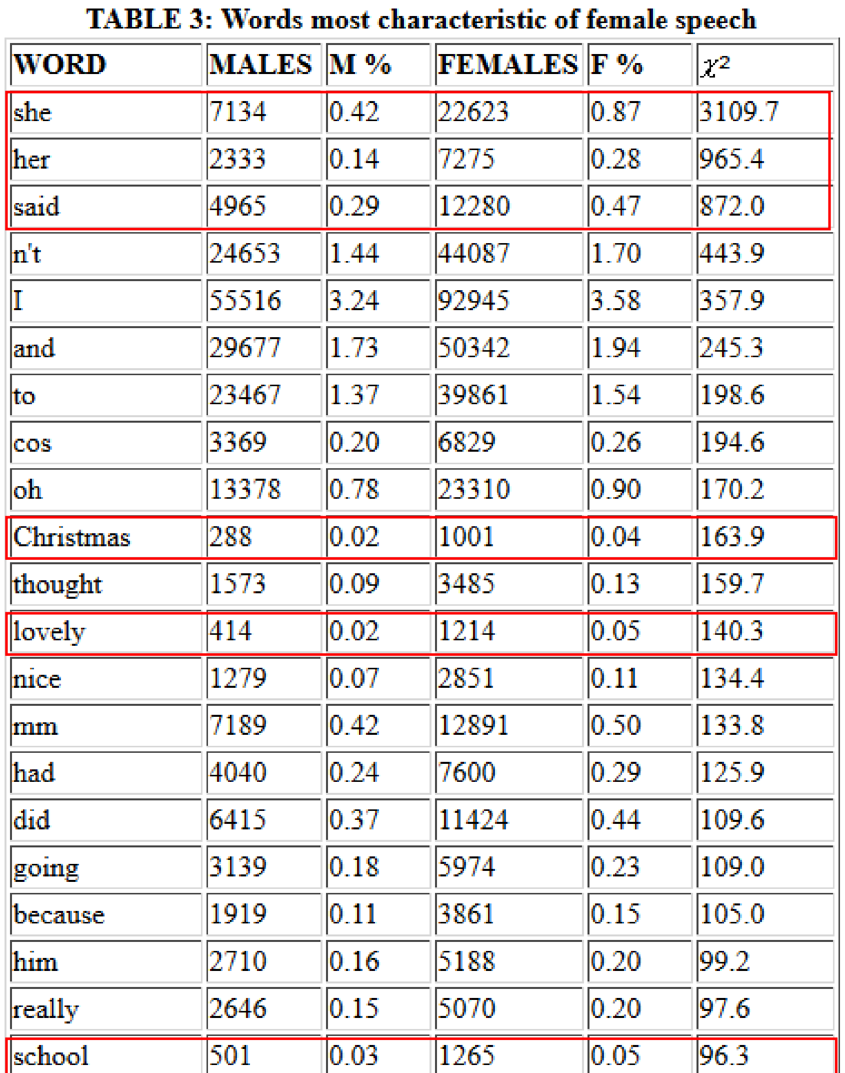


TABLE 3: Words most characteristic of female speech

| WORD | MALES | M % | FEMALES | F % | $\chi^2$ |
|---|---|---|---|---|---|
| she | 7134 | 0.42 | 22623 | 0.87 | 3109.7 |
| her | 2333 | 0.14 | 7275 | 0.28 | 965.4 |
| said | 4965 | 0.29 | 12280 | 0.47 | 872.0 |
| n't | 24653 | 1.44 | 44087 | 1.70 | 443.9 |
| I | 55516 | 3.24 | 92945 | 3.58 | 357.9 |
| and | 29677 | 1.73 | 50342 | 1.94 | 245.3 |
| to | 23467 | 1.37 | 39861 | 1.54 | 198.6 |
| cos | 3369 | 0.20 | 6829 | 0.26 | 194.6 |
| oh | 13378 | 0.78 | 23310 | 0.90 | 170.2 |
| Christmas | 288 | 0.02 | 1001 | 0.04 | 163.9 |
| thought | 1573 | 0.09 | 3485 | 0.13 | 159.7 |
| lovely | 414 | 0.02 | 1214 | 0.05 | 140.3 |
| nice | 1279 | 0.07 | 2851 | 0.11 | 134.4 |
| mm | 7189 | 0.42 | 12891 | 0.50 | 133.8 |
| had | 4040 | 0.24 | 7600 | 0.29 | 125.9 |
| did | 6415 | 0.37 | 11424 | 0.44 | 109.6 |
| going | 3139 | 0.18 | 5974 | 0.23 | 109.0 |
| because | 1919 | 0.11 | 3861 | 0.15 | 105.0 |
| him | 2710 | 0.16 | 5188 | 0.20 | 99.2 |
| really | 2646 | 0.15 | 5070 | 0.20 | 97.6 |
| school | 501 | 0.03 | 1265 | 0.05 | 96.3 |

# How is keyness calculated?

- The ‘classical’ method in CL: Pearson’s chi-squared test

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

- $O_i$ : observed frequency
- $E_i$ : expected frequency
- *lovely*, observed: expected ((R-tot * C-tot) / N):
- Result: $X^2 = 140.3$ (significant at $p<0.01$; d.f. = 1)

| | *lovely* | All other words | Total |
|---|---|---|---|
| **Males** | 414 | 1714029 | 1714443 |
| **Females** | 1214 | 2592238 | 2593452 |
| **Total** | 1628 | 4306267 | 4307895 |

| | *lovely* | All other words | Total |
|---|---|---|---|
| **Males** | 647.91 | 1713795.09 | 1714443 |
| **Females** | 980.09 | 2592471.91 | 2593452 |
| **Total** | 1628 | 4306267 | 4307895 |

# Another common method

- Log-likelihood ratio (Rayson & Garside 2000)

$$-2\ln\lambda = 2\sum_{i} O_i \ln\left(\frac{O_i}{E_i}\right)$$

- Chi-square value unreliable when $E_i < 5$; overestimates with high-frq.ws.
- LL = 2*((a*log(a/E1)) + (b*log(b/E2)))
  - 95th percentile; 5% level; $p < 0.05$; critical value = 3.84
  - 99th percentile; 1% level; $p < 0.01$; critical value = 6.63
  - 99.9th percentile; 0.1% level; $p < 0.001$; critical value = 10.83
  - 99.99th percentile; 0.01% level; $p < 0.0001$; critical value = 15.13
- http://ucrel.lancs.ac.uk/llwizard.html (Log likelihood)
- http://www.thegrammarlab.com/?p=193 (Chi-square)

# Increasingly more common: Effect size/log ratio

cf. Hardie (2014); Liffijit et al. (2016)

$$log\ ratio = log_2 \frac{norm.freq.c1}{norm.freq.c2}$$

- Binary logarithm of the ratio of the normalized frequencies
- Easy to interpret, compared to other effect size measures
  - Log ratio of 3 means that the word is $2^3$ = 8 times more frequent in corpus 1
  - Log ratio of -3 means that the word is 8 times more frequent in corpus 2

- “significance testing can be used to find consequential differences between corpora, but that assuming independence between all words may lead to overestimating the significance of the observed differences, especially for poorly dispersed words” (Liffijit et al. 2016)

# A case in point: the FOMC corpus

Andersen & Langerfeld (2021, 2022)

- What is the FOMC?
  - Federal Open Market Committee
  - Part of the Federal Reserve System
  - Makes key decisions about the interest rates

- About the meetings:
  - Members of the Federal Reserve Board and 5 presidents from the regional Federal Reserve Banks
  - 8 meetings over the course of a year

- The Sketch Engine: [https://www.sketchengine.eu/](https://www.sketchengine.eu/)
- Corpus-based term extraction

UNITED STATES • FEDERAL RESERVE SYSTEM •

NHH

## CORPUS APPROACHES TO TEXTUAL DATA

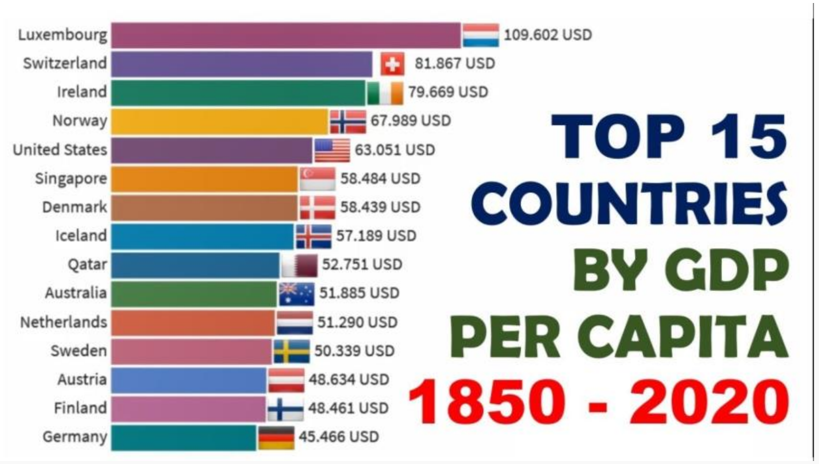
## COLLOCATION – THE SYSTEMATIC CO-OCCURRENCE OF WORDS

# Collocations

- Words do not occur at random but tend to collocate – co-occur for a number of reasons.
- Collocation – the tendency for words to co-occur, i.e. occur together more often than expected by chance.
- Sinclair (1996) – The search for units of meaning

- Sentence/clause elements:
  - *That's good.*
- Compounds:
  - *carbon footprint, central bank*
- Phrasal verbs:
  - *look up, freak out*

NHH



# Collocations and phraseology

- Fixed expressions:
  - *per capita*
  - *Trojan horse*
  - *de facto*
- Technical terms:
  - *central bank*
  - *large hadron collider*
  - *de facto standard*

- Idioms/stock phrases:
  - *kick the bucket*
  - *not exactly rocket science*
  - *to be fair*
  - *as far as X is concerned*
  - *When in Rome, do as the Romans.*

NHH

# Co-occurrence of words

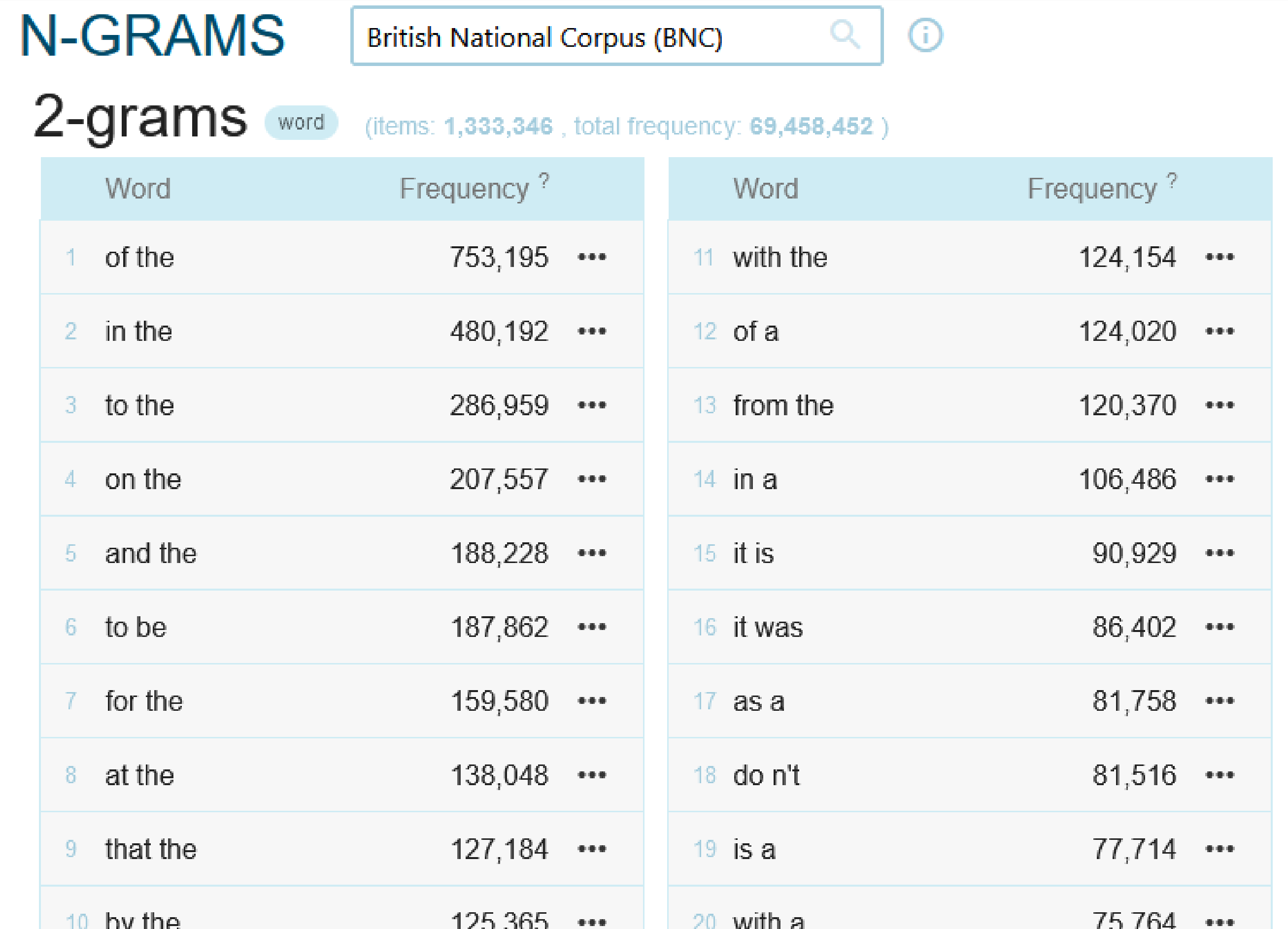
- n-gram: a sequence of n elements (usually words) that occur directly one after another in a corpus

| Length | name | Example 1 | Example 2 |
|---|---|---|---|
| n = 1 | “unigram” | end | large |
| n = 2 | “bigram” | end of | large hadron |
| n = 3 | “trigram” | end of the | large hadron collider |
| n = 4 | “tetragram, 4-gram” | end of the day | Large Hadron Collider (LHC) |

# Identifying collocations based on n-grams

## Example: BNC

- Sketch Engine: https://app.sketchengine.eu/



N-GRAMS British National Corpus (BNC)

2-grams word (items: **1,333,346** , total frequency: **69,458,452** )

| | Word | Frequency ? | | Word | Frequency ? |
|---|---|---|---|---|---|
| 1 | of the | 753,195 | 11 | with the | 124,154 |
| 2 | in the | 480,192 | 12 | of a | 124,020 |
| 3 | to the | 286,959 | 13 | from the | 120,370 |
| 4 | on the | 207,557 | 14 | in a | 106,486 |
| 5 | and the | 188,228 | 15 | it is | 90,929 |
| 6 | to be | 187,862 | 16 | it was | 86,402 |
| 7 | for the | 159,580 | 17 | as a | 81,758 |
| 8 | at the | 138,048 | 18 | do n't | 81,516 |
| 9 | that the | 127,184 | 19 | is a | 77,714 |
| 10 | by the | 125,365 | 20 | with a | 75,764 |

# Identifying strong collocations

- Inspecting n-grams only has little value if we want to systematically identify the building blocks of language, i.e. the terms and vocabulary it contains.
- So we need a way of capturing the most important n-grams in order to sort ‘the wheat from the chaff’.
- i.e. to find out what are statistically significant co-occurrences
- i.e. calculate observed and expected frequencies and use a statistical association measure
  (e.g. collocations.de; Lyse & Andersen 2012)

# Observed frequencies

## Contingency table for observed frequencies, e.g. *per capita*

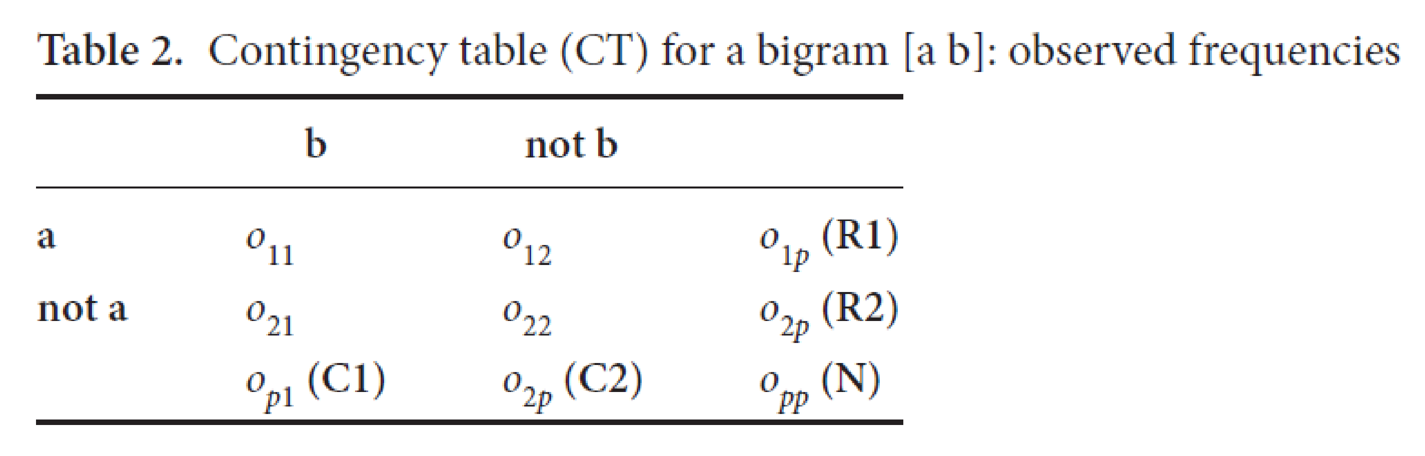


Table 2. Contingency table (CT) for a bigram [a b]: observed frequencies

| | b | not b | |
|---|---|---|---|
| a | $o_{11}$ | $o_{12}$ | $o_{1p}$ (R1) |
| not a | $o_{21}$ | $o_{22}$ | $o_{2p}$ (R2) |
| | $o_{p1}$ (C1) | $o_{2p}$ (C2) | $o_{pp}$ (N) |

- $o_{11}$: bigram *per capita*
- $o_{12}$: sum of all other bigrams like *per cent*, etc.
- $o_{21}$: sum of all other bigrams like *international capita*, etc.
- $o_{22}$: sum of all other bigrams like *international project*, etc.

# Expected frequencies

## Contingency table for expected frequencies, e.g. *per capita*

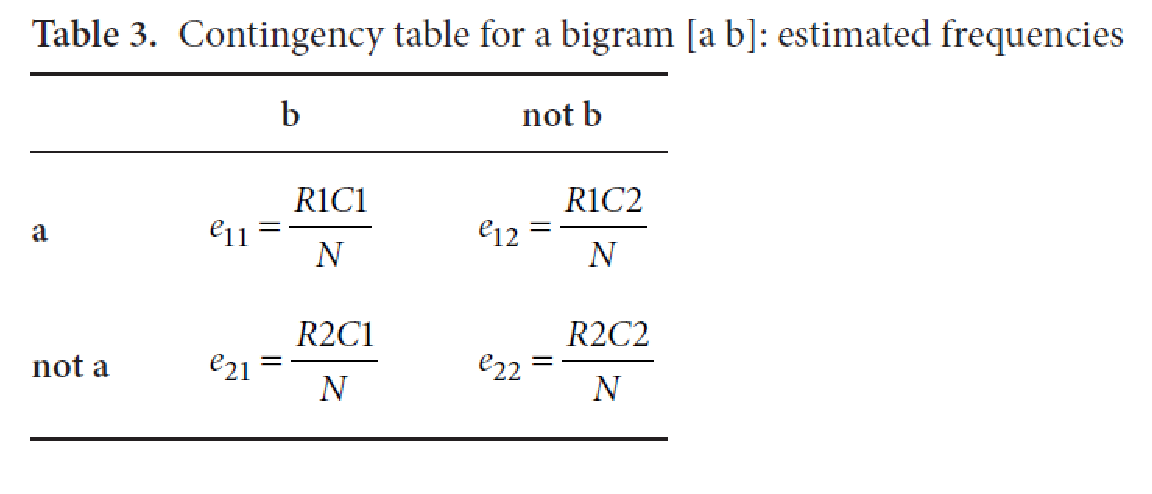


Table 3. Contingency table for a bigram [a b]: estimated frequencies

| | b | not b |
|---|---|---|
| a | $e_{11} = \frac{R1C1}{N}$ | $e_{12} = \frac{R1C2}{N}$ |
| not a | $e_{21} = \frac{R2C1}{N}$ | $e_{22} = \frac{R2C2}{N}$ |

- $e_{11}$: expected *per capita* if no association (null hypothesis)
- $e_{12}$: expected all other bigrams *per cent*, etc.
- $e_{21}$: expected all other bigrams *international capita*, etc.
- $e_{22}$: expected all other bigrams *international project*, etc.

# Association measures

- The strength of an association between two words is calculated by a statistical significance measure.
- This means performing a mathematical operation on the figures $o_{11\text{-}22}$ and $e_{11\text{-}22}$ shown above.
- There are many different such measures, Mutual Information, Chi-square, Log-likelihood being the most common ones.

# Weaknesses/strengths of AMs

- NB! contingency tables used for collocations are (virtually always) highly skewed ($o_{22} > (o_{12}|o_{21}) > o_{11}$)
- Log-likelihood assumed to perform better than chi square, esp. for lexical (as opposed to grammatical) word collocations, which tend to have a low o11.
- Mutual Information: biased towards low-frequency n-grams, i.e. where o11 is low.
- As for Z-score, when the e11 is low, Z-score values can become very large, leading to highly inflated scores for low-frequency pair types.
- Lyse & Andersen 2012; http://www.collocations.de/

# A closer look at co-occurrence relations

- Collocation
- Colligation
- Semantic preference
- Semantic prosody

## Sinclair (1996); Stubbs (2001)

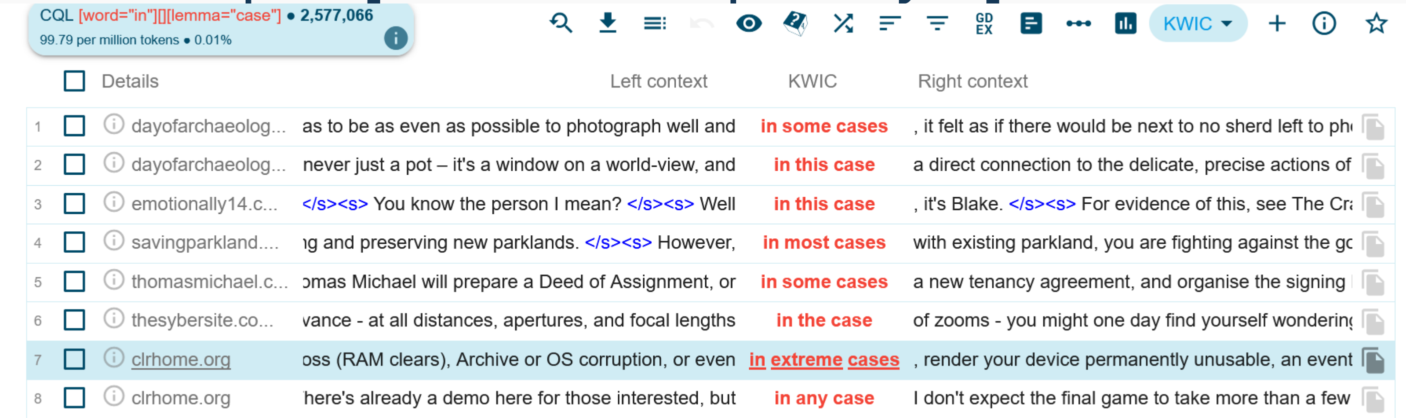
- frequent co-occurrence of word forms
- co-occurrence of word with grammatical category
- lexical set of frequently occurring collocates that share a semantic feature
- tendency for word to be used in contexts with positive or negative associations/ connotations

NHH

# Colligation

## Tendency for word to co-occur with particular grammatical category

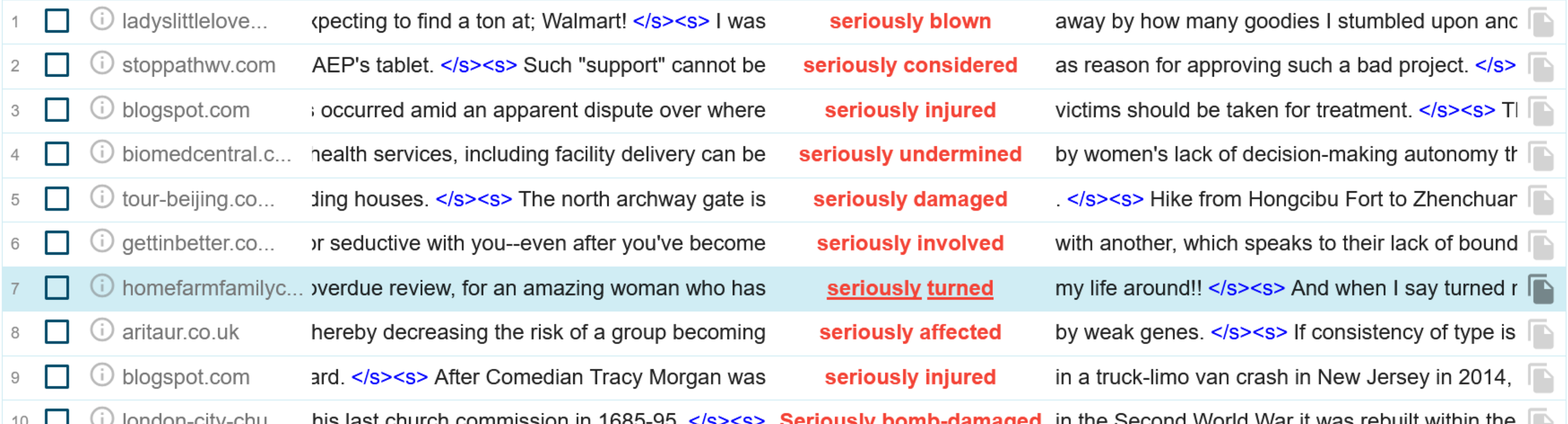
- Example: [determiner/quantifyer] + *case*



CQL [word="in"][][lemma="case"] • 2,577,066
99.79 per million tokens • 0.01%

| | □ | Details | Left context | KWIC | Right context |
|---|---|---|---|---|---|
| 1 | □ | dayofarchaeolog... | as to be as even as possible to photograph well and | in some cases | , it felt as if there would be next to no sherd left to phı |
| 2 | □ | dayofarchaeolog... | never just a pot – it's a window on a world-view, and | in this case | a direct connection to the delicate, precise actions of |
| 3 | □ | emotionally14.c... | </s><s> You know the person I mean? </s><s> Well | in this case | , it's Blake. </s><s> For evidence of this, see The Crı |
| 4 | □ | savingparkland.... | ıg and preserving new parklands. </s><s> However, | in most cases | with existing parkland, you are fighting against the gc |
| 5 | □ | thomasmichael.c... | ɔmas Michael will prepare a Deed of Assignment, or | in some cases | a new tenancy agreement, and organise the signing |
| 6 | □ | thesybersite.co... | vance - at all distances, apertures, and focal lengths | in the case | of zooms - you might one day find yourself wondering |
| 7 | □ | clrhome.org | oss (RAM clears), Archive or OS corruption, or even | in extreme cases | , render your device permanently unusable, an event |
| 8 | □ | clrhome.org | here's already a demo here for those interested, but | in any case | I don't expect the final game to take more than a few |

- Example: *seriously* + [V-ed]



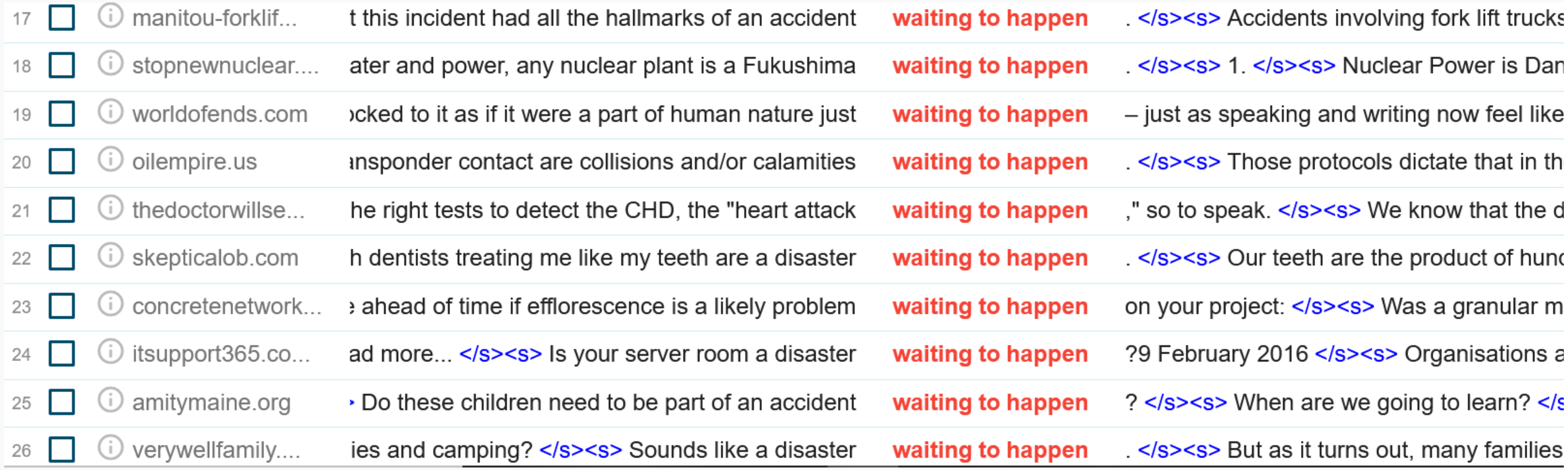
| | □ | | | | |
|---|---|---|---|---|---|
| 1 | □ | ladyslittlelove... | ‹pecting to find a ton at; Walmart! </s><s> I was | seriously blown | away by how many goodies I stumbled upon anc |
| 2 | □ | stoppathwv.com | AEP's tablet. </s><s> Such "support" cannot be | seriously considered | as reason for approving such a bad project. </s> |
| 3 | □ | blogspot.com | ; occurred amid an apparent dispute over where | seriously injured | victims should be taken for treatment. </s><s> Tl |
| 4 | □ | biomedcentral.c... | ıealth services, including facility delivery can be | seriously undermined | by women's lack of decision-making autonomy tl |
| 5 | □ | tour-beijing.co... | ding houses. </s><s> The north archway gate is | seriously damaged | . </s><s> Hike from Hongcibu Fort to Zhenchuar |
| 6 | □ | gettinbetter.co... | ɔr seductive with you--even after you've become | seriously involved | with another, which speaks to their lack of bound |
| 7 | □ | homefarmfamilyc... | ɔverdue review, for an amazing woman who has | seriously turned | my life around!! </s><s> And when I say turned r |
| 8 | □ | aritaur.co.uk | hereby decreasing the risk of a group becoming | seriously affected | by weak genes. </s><s> If consistency of type is |
| 9 | □ | blogspot.com | ard. </s><s> After Comedian Tracy Morgan was | seriously injured | in a truck-limo van crash in New Jersey in 2014, |
| 10 | □ | london-city-chu... | his last church commission in 1685-95. </s><s> | Seriously bomb-damaged | in the Second World War it was rebuilt within the |

# Semantic preference

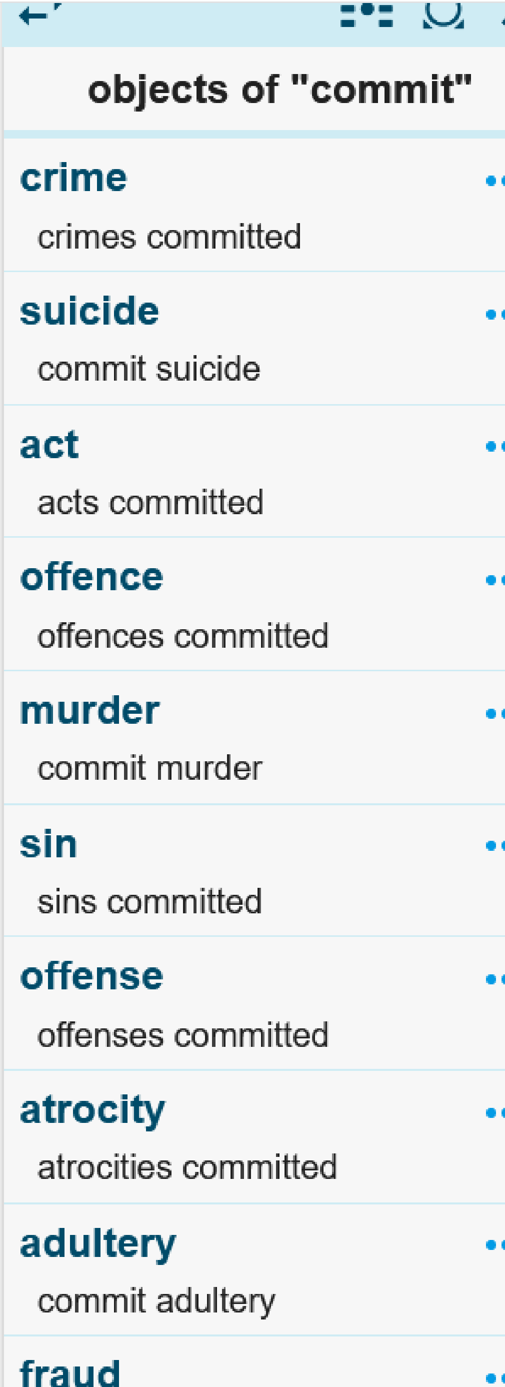
## Tendency for word to occur with member of set of lexically similar words

- Example: objects of *commit*

- Example: collocates of *waiting to happen*



| | | | | | |
|---|---|---|---|---|---|
| 17 | □ | ⓘ manitou-forklif... | t this incident had all the hallmarks of an accident | waiting to happen | . </s><s> Accidents involving fork lift trucks |
| 18 | □ | ⓘ stopnewnuclear.... | ater and power, any nuclear plant is a Fukushima | waiting to happen | . </s><s> 1. </s><s> Nuclear Power is Dan |
| 19 | □ | ⓘ worldofends.com | ›cked to it as if it were a part of human nature just | waiting to happen | – just as speaking and writing now feel like |
| 20 | □ | ⓘ oilempire.us | ınsponder contact are collisions and/or calamities | waiting to happen | . </s><s> Those protocols dictate that in th |
| 21 | □ | ⓘ thedoctorwillse... | he right tests to detect the CHD, the "heart attack | waiting to happen | ," so to speak. </s><s> We know that the d |
| 22 | □ | ⓘ skepticalob.com | h dentists treating me like my teeth are a disaster | waiting to happen | . </s><s> Our teeth are the product of hunc |
| 23 | □ | ⓘ concretenetwork... | ɛ ahead of time if efflorescence is a likely problem | waiting to happen | on your project: </s><s> Was a granular m |
| 24 | □ | ⓘ itsupport365.co... | ad more... </s><s> Is your server room a disaster | waiting to happen | ?9 February 2016 </s><s> Organisations a |
| 25 | □ | ⓘ amitymaine.org | › Do these children need to be part of an accident | waiting to happen | ? </s><s> When are we going to learn? </s |
| 26 | □ | ⓘ verywellfamily.... | ies and camping? </s><s> Sounds like a disaster | waiting to happen | . </s><s> But as it turns out, many families |



objects of "commit"

**crime**
crimes committed

**suicide**
commit suicide

**act**
acts committed

**offence**
offences committed

**murder**
commit murder

**sin**
sins committed

**offense**
offenses committed

**atrocity**
atrocities committed

**adultery**
commit adultery

**fraud**

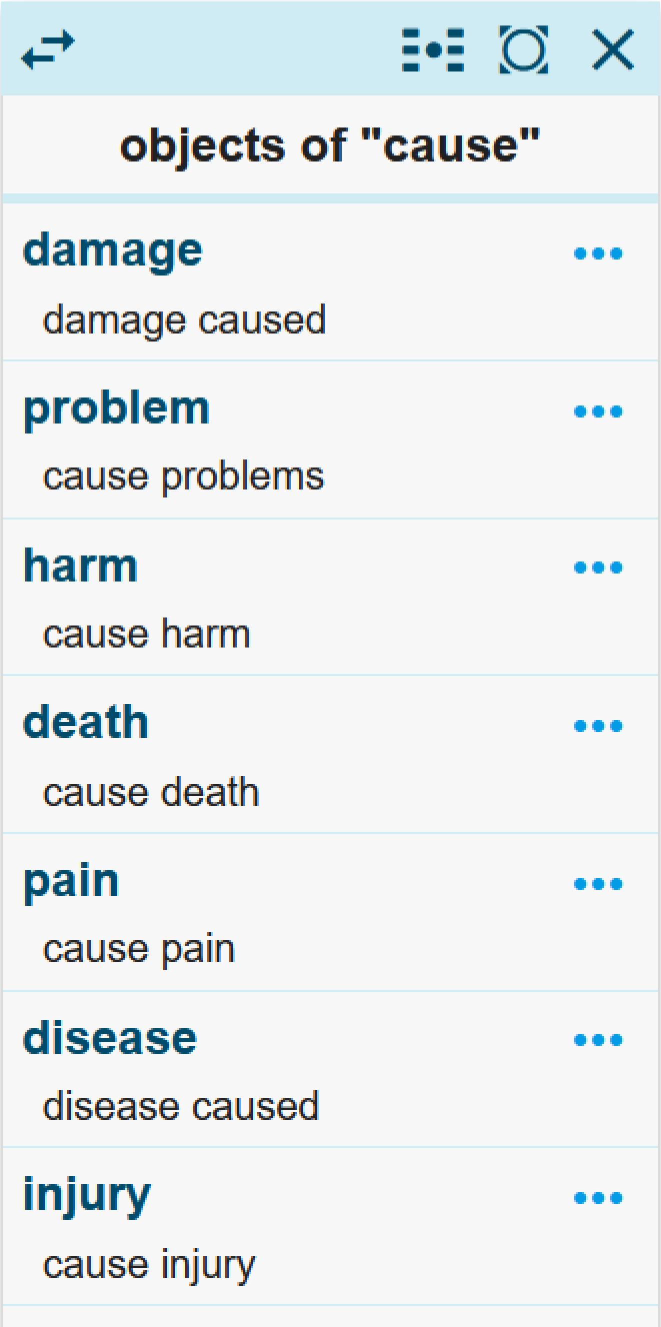
# In other words,

- Very strong association between construction N + *waiting to happen* and negative evaluation
- [*accident/disaster/wreck/tragedy/catastrophe*…] *waiting to happen*
- non-neutral added meaning, negative semantic prosody (pragmatics)

- Incidentally, construction N + *waiting to happen* is an example of a so-called collostruction (Stefanowitsch & Gries 2003)

# Semantic prosody:

- **Semantic prosody** of words is important because it often reveals the **positive or negative associations** of words.
- Has clear implications for **word choice**, e.g. in strategic and political contexts.
- *cause* (v) = *lead to* ?



## objects of "cause"

**damage** •••
damage caused

**problem** •••
cause problems

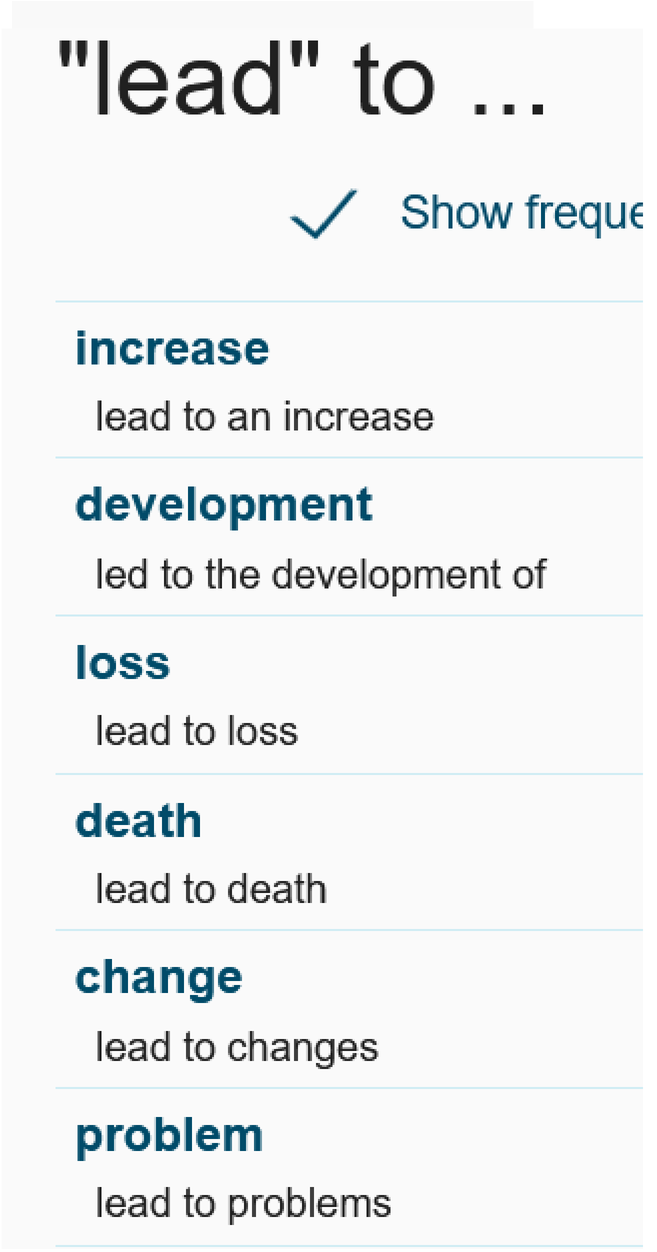
**harm** •••
cause harm

**death** •••
cause death

**pain** •••
cause pain

**disease** •••
disease caused

**injury** •••
cause injury



## "lead" to ...

✓ Show freque

**increase**
lead to an increase

**development**
led to the development of

**loss**
lead to loss

**death**
lead to death

**change**
lead to changes

**problem**
lead to problems

# In conclusion, …

- *cause* (v) and *lead to* are near-synonyms
- similar in meaning, but each with its own collocational profile.
- Choosing the wrong word can lead to unintended associations/**connotations**.
  - ?*cause an increase*

- **Connotation**: the positive/ neutral/negative sentiment (value) that emanates from the totality of concepts that the word tends to be associated with.
- = the “aura” of words

# Sketch Engine

- English Web 2020 (enTenTen20)
- Word sketch
- Word sketch difference

## CORPUS APPROACHES TO TEXTUAL DATA

# CORPUS-LINGUISTIC METHODS IN R

# Corpus linguistic methods in R

Suggested packages

- ngram: https://cran.r-project.org/web/packages/ngram/index.html
- corpora: https://cran.r-project.org/web/packages/corpora/index.html

- ngram: R package for constructing n-grams ("tokenizing"), as well as generating new text based on the n-gram structure of a given text input ("babbling"). The package can be used for serious analysis or for creating "bots" that say amusing things.

# Logical sequence for corpus processing

1. multiread: Read in a collection of text files into a list
2. concatenate: Put multiple files into single string
3. preprocess: A simple text preprocessor for removing punctuation, digits and multiple/trailing spaces
4. ngram: Takes input string and converts it into the internal n-gram representation
5. ngram-print: Various print methods
6. get.phrasetable: Print n-grams into a handy table format

# Some of the functions in package ‘corpora’

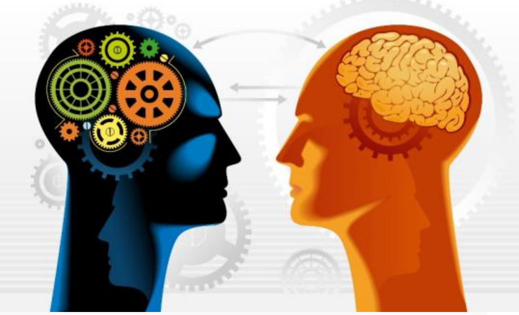
- Data from the British National Corpus (BNC) and other corpora
- Pearson’s chi-square statistic: *chisq / chisq.pval*
  - unlike *chisq.test* accepts vector arguments which allows for multiple comparisons in a single function call
- *cont.table*
- This is a convenience function which constructs 2x2 contingency tables needed for frequency comparisons with *chisq.test*, *fisher.test* and similar functions.
- Random samples from data frames: *sample.df*
- Other statistics: binomial p-value; Fisher’s exact; Z-score …

# Further suggestions

- Combined with what you already know about R programming, you can use these functions to

1. load text files into a corpus (multiread)
2. prepare them for n-gram analysis (concatenate, preprocess)
3. calculate all n-gram frequencies (n=1; n=2; n=3)
4. make contingency tables (cont.table) for observed and expected frequencies, e.g. of all 1-grams in one corpus compared with another
5. run chi-square test (chisq/chisq.pval) on all word/n-gram frequencies to test for significant frequency differences

# Concluding remarks



- The linguistic perspective, complementary to purely statistical analyses
- Concordanses, multiword processing, wordclass tagging, syntactic parsing and semantic/discourse annotation may augment linguistically statistical machine processing
  - and allow for a more sophisticated analysis than what is achievable through mere counting of individual words
- Corpus-based term extraction (keyness) useful for charting the terminology of a specific usage domain
- Collocation analysis needed to explore the “aura” of words
- Connotations important to near in mind when choosing words

# References (1/2)

- Andersen, Gisle & Christian Langerfeld. 2021. Pragmatic markers in deliberations of monetary policy.  IPrA21 conference, Winterthur (digital), 2021-06-29.
- Andersen, Gisle & Christian Langerfeld. 2022. The dynamics of turn-taking in deliberations of monetary policy. 23rd International Conference on Language for Specific Purposes (LSP 2022), Lisbon, Portugal. 2022-09-13.
- Andersen, G. (2012). Semi-automatic approaches to Anglicism detection in Norwegian corpus data. The Anglicization of European Lexis. C. Furiassi, V. Pulcini and F. Rodríguez Gonzáles. Amsterdam, John Benjamins: 111-130.
- Azevedo, Lucas. 2018. Truth or Lie? Automatically Fact Checking News. The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.
- Biber, D., et al. (2004). "If you look at...: Lexical bundles in university teaching and textbooks." Applied linguistics 25(3): 371-405.
- Bondi, M. 2010. Perspectives on keywords and keyness: An introduction. Keyness in Texts. M. Bondi & M. Scott. Amsterdam, John Benjamins: 1-18.
- Conroy, Niall J. Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology 52, 1 (2015), 1–4.
- Gries, S. Th. 2008. Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics 13:4, 403-437. http://www.linguistics.ucsb.edu/faculty/stgries/research/2008_STG_Dispersion_IJCL.pdf
- Hardie, Andrew. 2014. "Log ratio – an informal introduction." Post on the website of the ESRC Centre for Corpus Approaches to Social Science CASS. Retrieved from http://cass.lancs.ac.uk/?p=1133
- Higdon, Nolan. 2020. The Anatomy of Fake News: A Critical News Literacy Education. Oakland, CA: University of California Press.

# References (2/2)

- Jurafsky D. & J.H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd ed. Upper Saddle River: Prentice Hall.
- Larcker & Zakolyukina. 2012. Detecting Deceptive Discussions in Conference Calls. Journal of Accounting Research. 50 (2), 495-540.
- Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki and Heikki Mannila (2016). “Significance testing of word frequencies in corpora”. Digital Scholarship in the Humanities 31(2): 374–397
- Lyse, G. I. and G. Andersen (2012). Collocations and statistical analysis of n-grams. Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian. G. Andersen. Amsterdam, John Benjamins: 79-110.
- Meurer, P. (2012). Corpuscle – a new corpus management platform for annotated corpora. Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian. G. Andersen. Amsterdam, John Benjamins: 31-50.
- Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1-8 October 2000, Hong Kong, pp. 1 – 6. http://eprints.lancs.ac.uk/11882/1/rg_acl2000.pdf
- Rayson, P., Leech, G., and Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. International Journal of Corpus Linguistics. Volume 2, number 1. pp 133 - 152. John Benjamins, Amsterdam/Philadelphia. ISSN 1384-665.
- Sag, I. A., et al. (2002). "Multiword expressions : a pain in the neck for NLP." Lecture notes in computer science 2276: 1-15.