COMP/LING 345

From Natural Language to Data Science

Keywords, Collocations, and Association Metrics

Siva Reddy

McGill UNIVERSITY

Recap

Regular Expressions: Negation in Disjunction

- Negations `[^Ss]`
  - Carat means negation only when first in []

Pattern	Matches	
`[^A-Z]`	Not an upper case letter	`Oyfn pripetchik`
`[^Ss]`	Neither ‘S’ nor ‘s’	`I have no exquisite reason”`
`[^e^]`	Neither e nor ^	`Look here`
`a^b`	The pattern a carat b	`Look up a^b now`

correction from last slides

Corpus Query Language

- Corpus Query Language (CQL) allows us to use regular expressions with linguistic knowledge [1]

- Find all word forms of run:
  [word="[Rr](uns?|an|unning)"]

[1] CQL: https://www.sketchengine.eu/documentation/corpus-querying/

Corpus Query Language

- Corpus Query Language (CQL) allows us to use regular expressions with linguistic knowledge [1]

- Find all word forms of run:
  [word="[Rr](uns?|an|unning)"]

- Find all word forms of run:
  [lemma="run"]

[1] CQL: https://www.sketchengine.eu/documentation/corpus-querying/

Corpus Query Language

- Corpus Query Language (CQL) allows us to use regular expressions with linguistic knowledge [1]

- Find all word forms of run:
  [word="[Rr](uns?|an|unning)"]

- Find all word forms of run:
  [lemma="run"]

Word	↓ Frequency
run	239,819
running	145,217
runs	79,247
ran	48,365
Running	4,204
Run	3,670
Ran	179
Runs	166
RUN	124

[1] CQL: https://www.sketchengine.eu/documentation/corpus-querying/

Linguistic knowledge: Syntactic category

Linguistic knowledge: Syntactic category

- Find all the instances of run as noun.

Linguistic knowledge: Syntactic category

- Find all the instances of run as noun.
- [lemma="run" & tag="N.*"]

Linguistic knowledge: Syntactic category

- Find all the instances of run as noun.
- [lemma="run" & tag="N.*"]

r high-fives or walk-off home	run	hugs. </s><s> Give us reasc
n scored a whopping 13,288	runs	in Test matches and 10,889
in Test matches and 10,889	runs	in One-Day Internationals. <
all reached mid-March on a	run	of three defeats in their last
at counts almost as a strong	run	. </s><s> In 13 matches con
/Ps, it's time for that 75-year	run	to end. </s><s> "Looking ba
ing stretched their unbeaten	run	in all competitions to 15 gam
th minute from a jinking solo	run	capped by a crisp low shot i
ed just 53 more passes than	runs	</s><s> Running backs Be

https://app.sketchengine.eu/#open

Exercise

- Find all the noun phrases containing virus

noun + modifier

the corona virus
a deadly virus
the novel virus
the viruses
a new virus
the new virus
this deadly virus

Exercise

- Find all the noun phrases containing virus

the corona virus
a deadly virus
the novel virus
the viruses
a new virus
the new virus
this deadly virus

Exercise

- Find all the noun phrases containing virus
- [tag="D.*"] [tag="N.*|J.*"]{0,3} [lemma="virus"]

the corona virus
a deadly virus
the novel virus
the viruses
a new virus
the new virus
this deadly virus

Exercise

- What are some things coronavirus has impacted?

Exercise

- What are some things coronavirus has impacted?
- [lemma="coronavirus"] []{0,3} [lemma="impact"] []{0,3} [tag="N.*"]

Exercise

- What are some things coronavirus has impacted?
- [lemma="coronavirus"] []{0,3} [lemma="impact"] []{0,3} [tag="N.*"]

Word	↓ Frequency
economy	135
industry	57
business	53
people	49
businesses	41
demand	31
communities	31
market	30
health	29
finances	28
world	27

Today's outline

- Regular expressions on complex structures
- Keywords
- N-grams
- Collocations

Linguistic knowledge: beyond sequential regular expressions

- Corpus query language works sequentially.
- Even simple queries like finding subjects of a verb require complex patterns (often with many false positives).

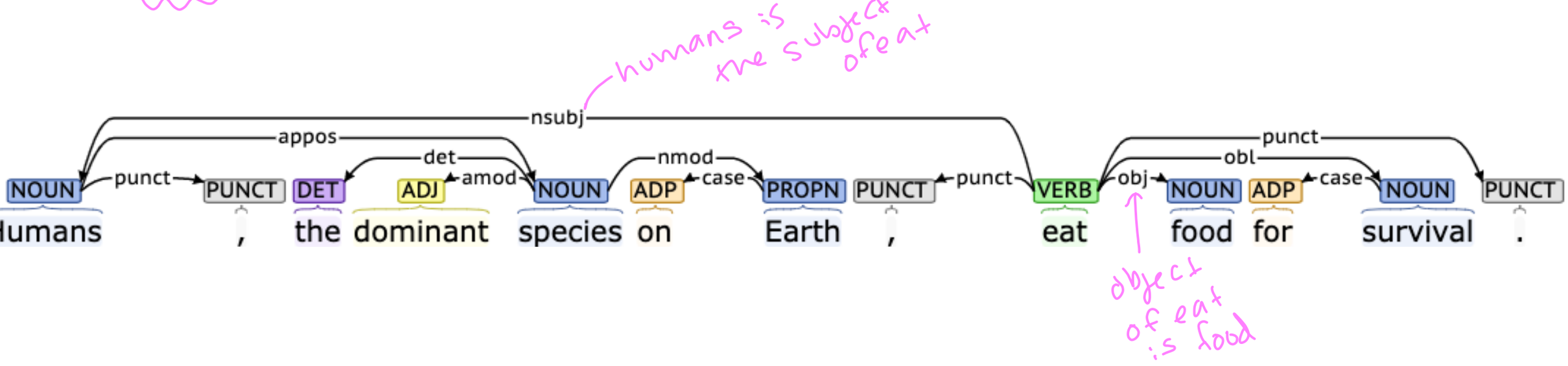
```
Humans eat food for survival.

Humans, the dominant species on Earth, eat food for
survival

The food eaten by humans …
```

Dependency trees

- Dependency trees indicate the syntactic relationship between words



http://corenlp.run/

http://stanza.run/

Dependency trees

- Dependency trees indicate the syntactic relationship between words

kotini aratipandu tinindi

monkey banana eat

Guess the meaning of above Telugu sentence

Dependency trees

- Dependency trees indicate the syntactic relationship between words
- It can help understand the meaning even if you are not familiar with the grammar

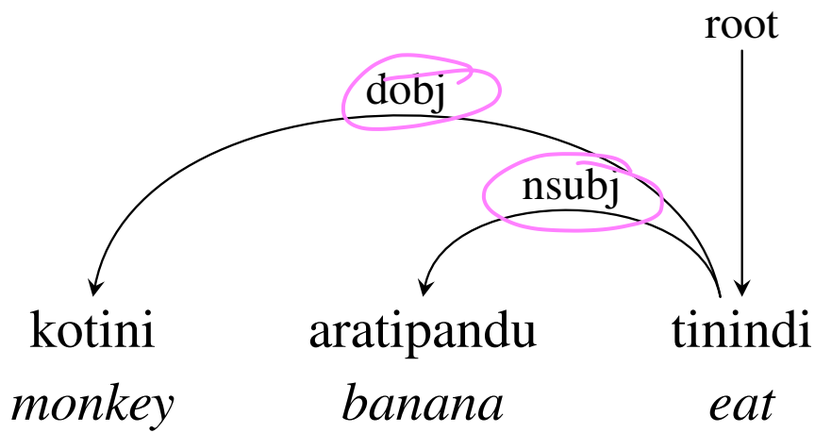
kotini	aratipandu	tinindi
monkey	*banana*	*eat*

it's not monkey eats banana

Your guess is probably wrong :)

Dependency trees

- Dependency trees indicate the syntactic relationship between words
- It can help understand the meaning even if you are not familiar with the grammar

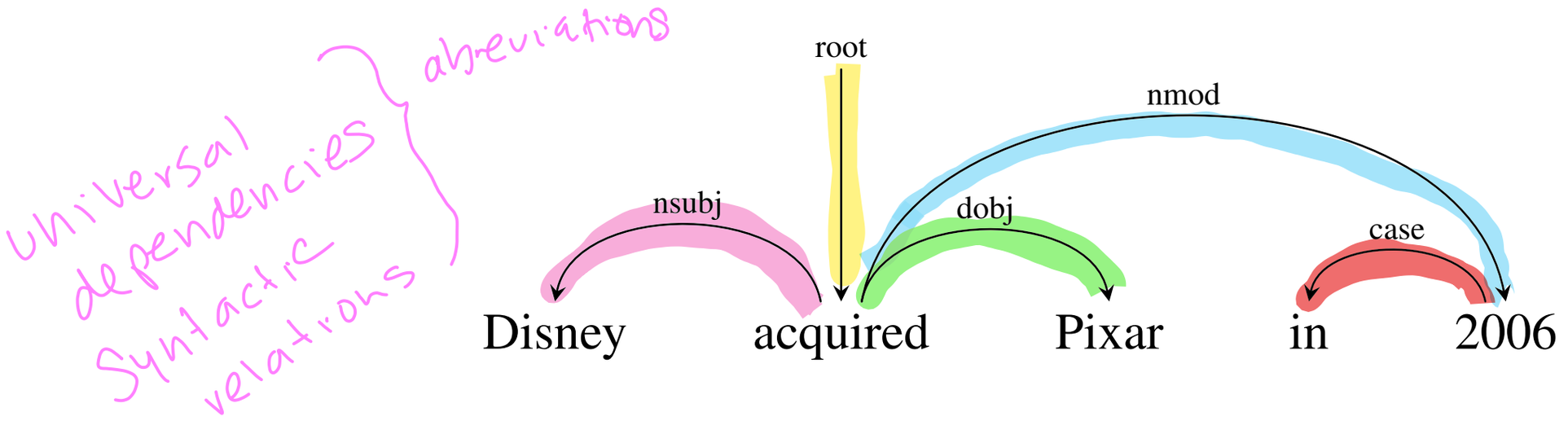


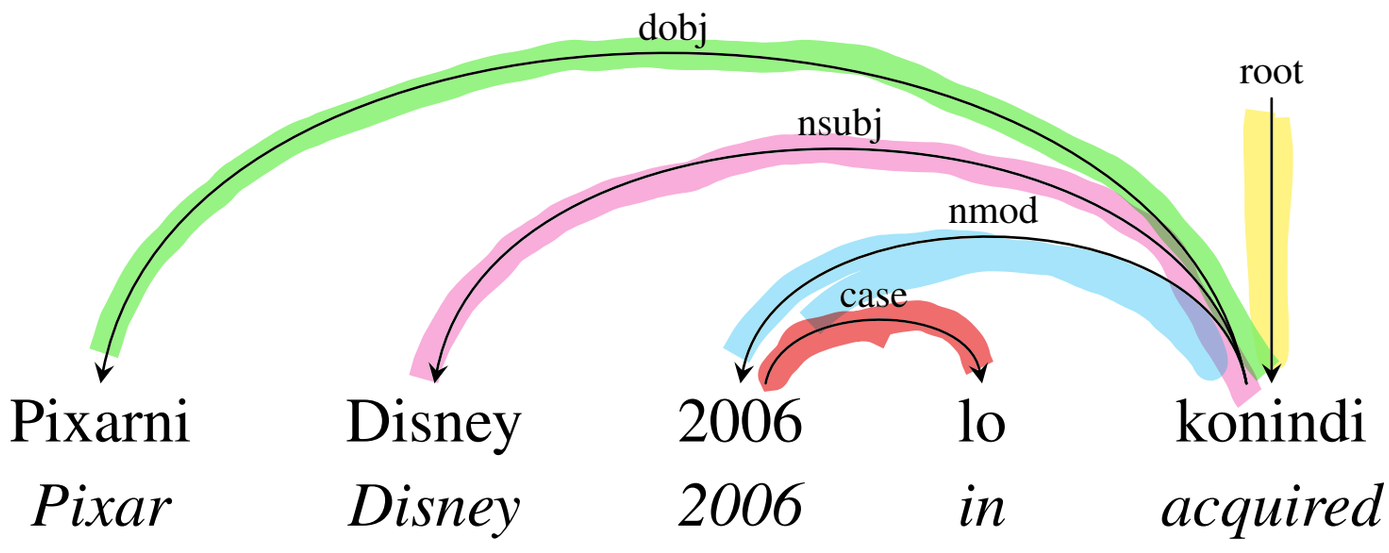
banana is the subject of eat
monkey is the object of eat

banana eats monkey

Dependency trees

- Similar meaning sentences **may** have similar dependency structure





Tree regular expressions

- Tree Regular expressions allow us to run regular expressions on tree structures.

Symbol	Description	Symbol	Description
A > B	A is the parent of B	A >> B	A is an ancestor of B
A $ B	A and B are sisters	A $+ B	B is next sister of A
A $>_i$ B	B is i^{th} child of A	A >: B	B is only child of A
A >># B	A on head path of B	A >>- B	B is rightmost descendent
A .. B	A precedes B in depth-first traversal of tree		
A >+ (C) B	A dominates B via unbroken chain of Cs		

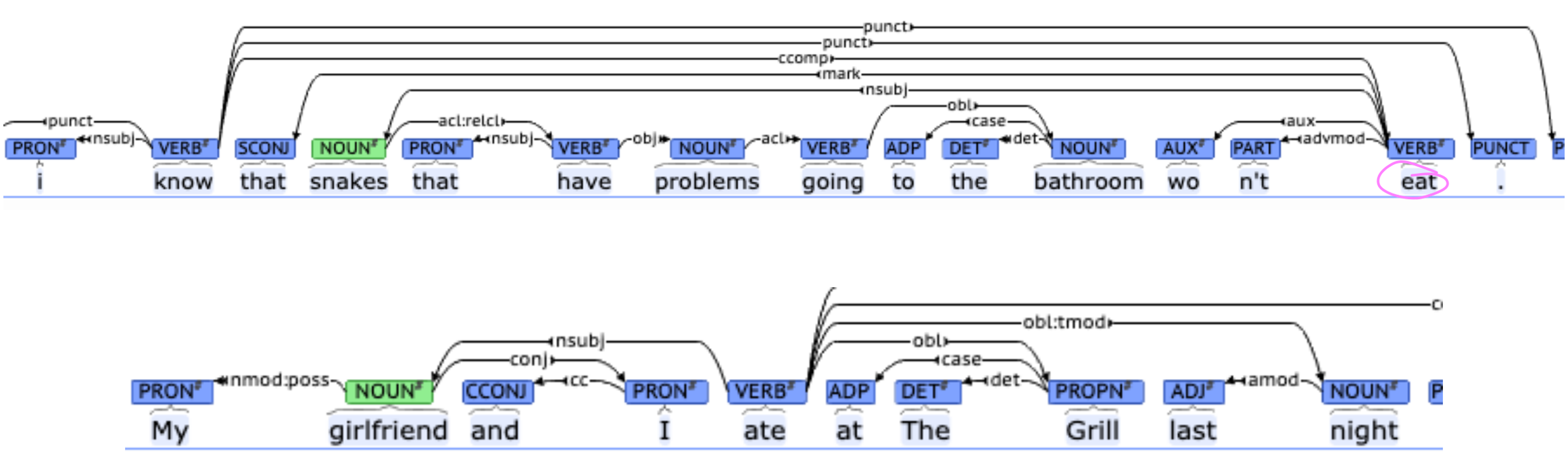
https://nlp.stanford.edu/software/Semgrex.ppt ← documentation CoreNLP.run

Exercise

regex on trees

semgrex

- Find all the subjects of “eat”

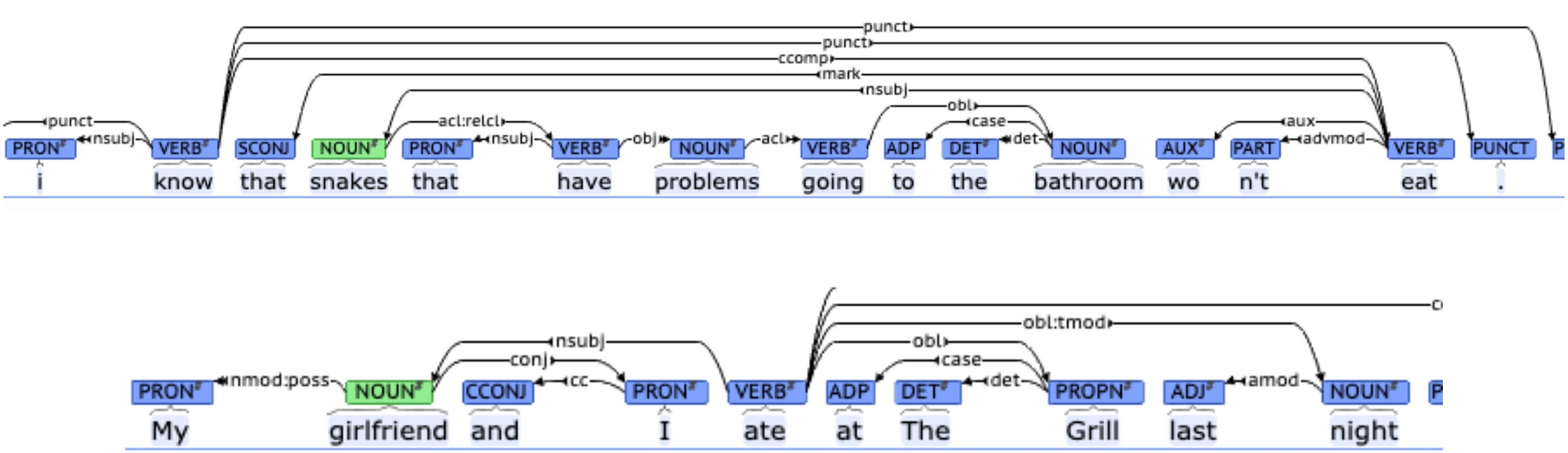


Use https://corenlp.run

Exercise

- Find all the subjects of “eat”
- {pos:/N.*/} <nsubj {word:/eat/}

noun that is a subject of eat



Use https://corenlp.run

How can we study large collections of texts (corpus)?

Word frequency

Gives you non interesting words

Word	↓ Absolute Frequency ?
the	69,971
of	36,412
and	28,870
to	26,158
a	23,225
in	21,343
that	10,787
is	10,206
was	9,969
he	9,801

Brown Corpus (1M words)

Word	↓ Absolute Frequency ?
the	6,054,939
of	3,049,448
and	2,624,147
to	2,599,451
a	2,175,967
in	1,945,533
that	1,120,750
it	1,054,366
is	991,771
was	883,547

British National Corpus (97M words)

A very long tail

- Many words just occur only once
- Out of 96,638 word types, 36,231 occur only once in Europarl corpus

cornflakes, mathematicians, fuzziness, jumbling

pseudo-rapporteur, lobby-ridden, perfunctorily,

Lycketoft, UNCITRAL, H-0695

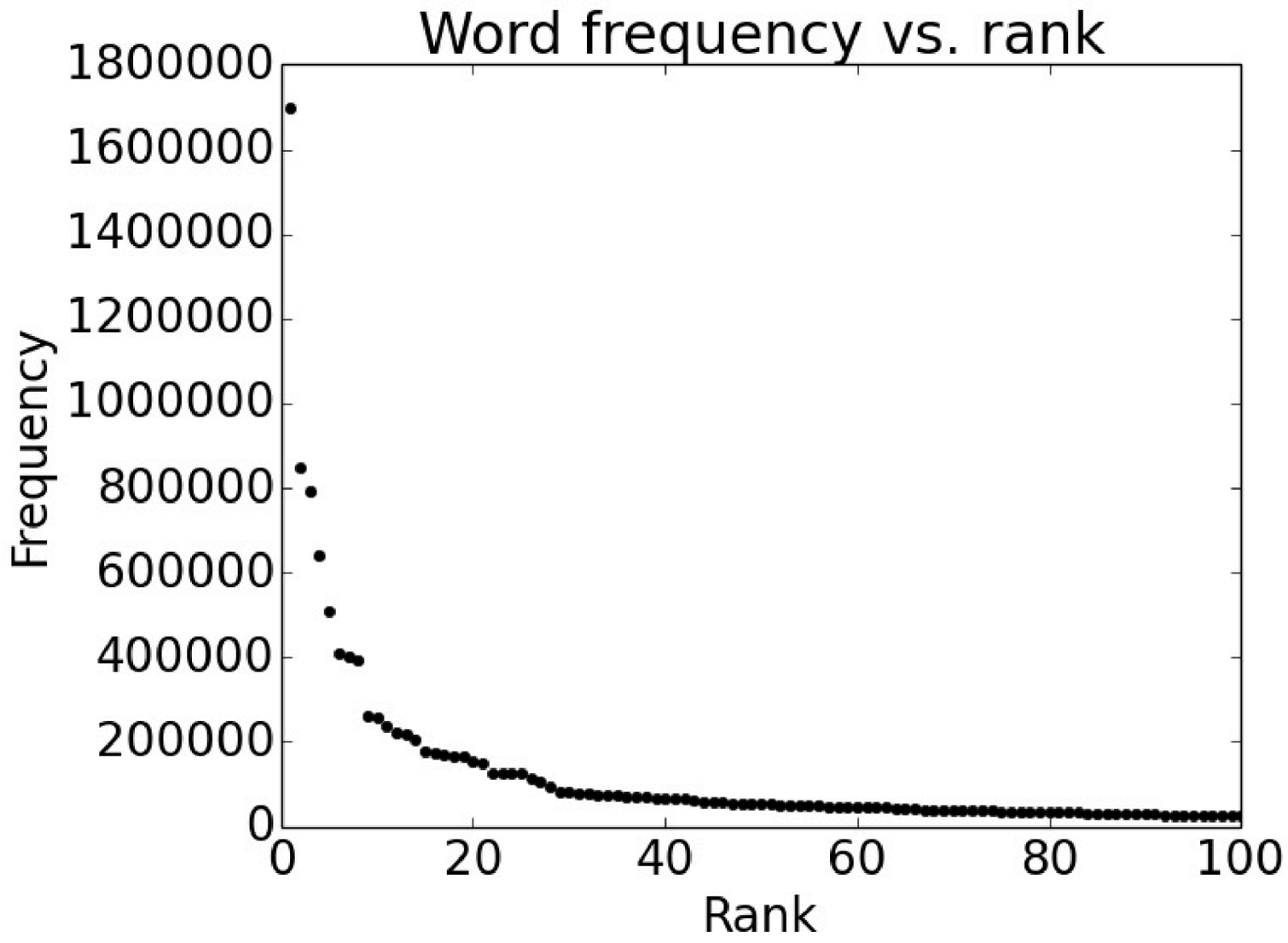
policyfor, Commissioneris, 145.95, 27a

any word

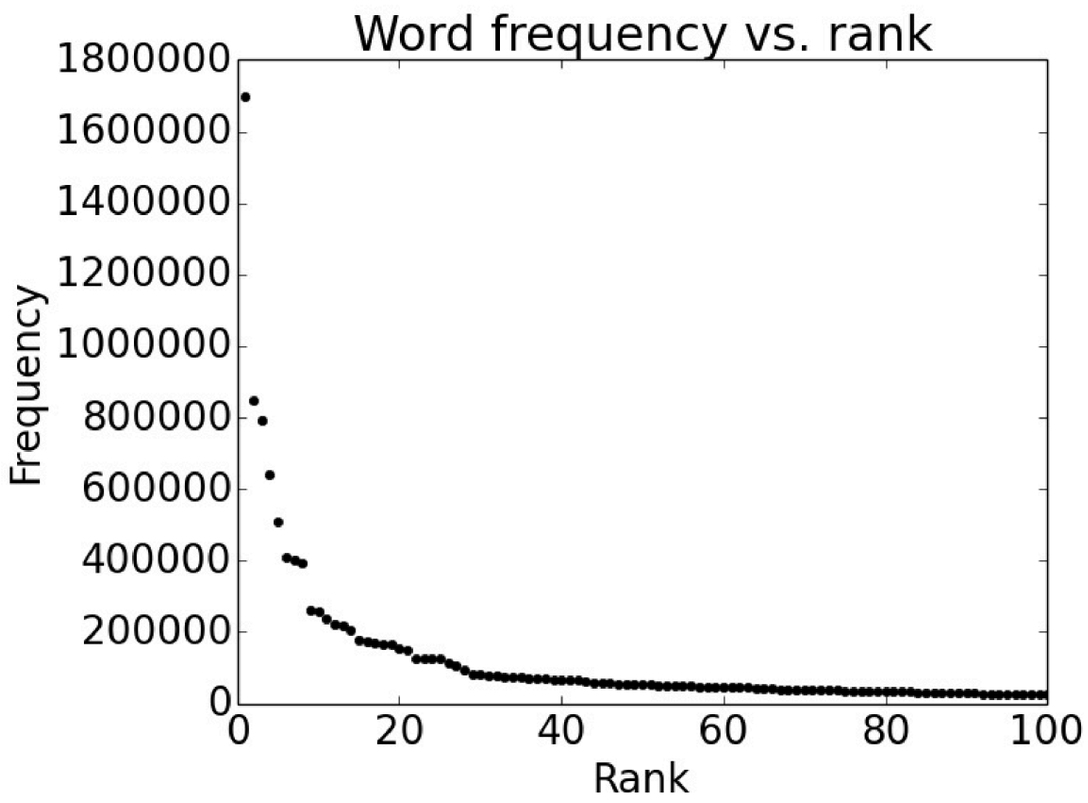
Frequency	Token
1,698,599	the
849,256	of
793,731	to
640,257	and
508,560	in
407,638	that
400,467	is
394,778	a
263,040	I

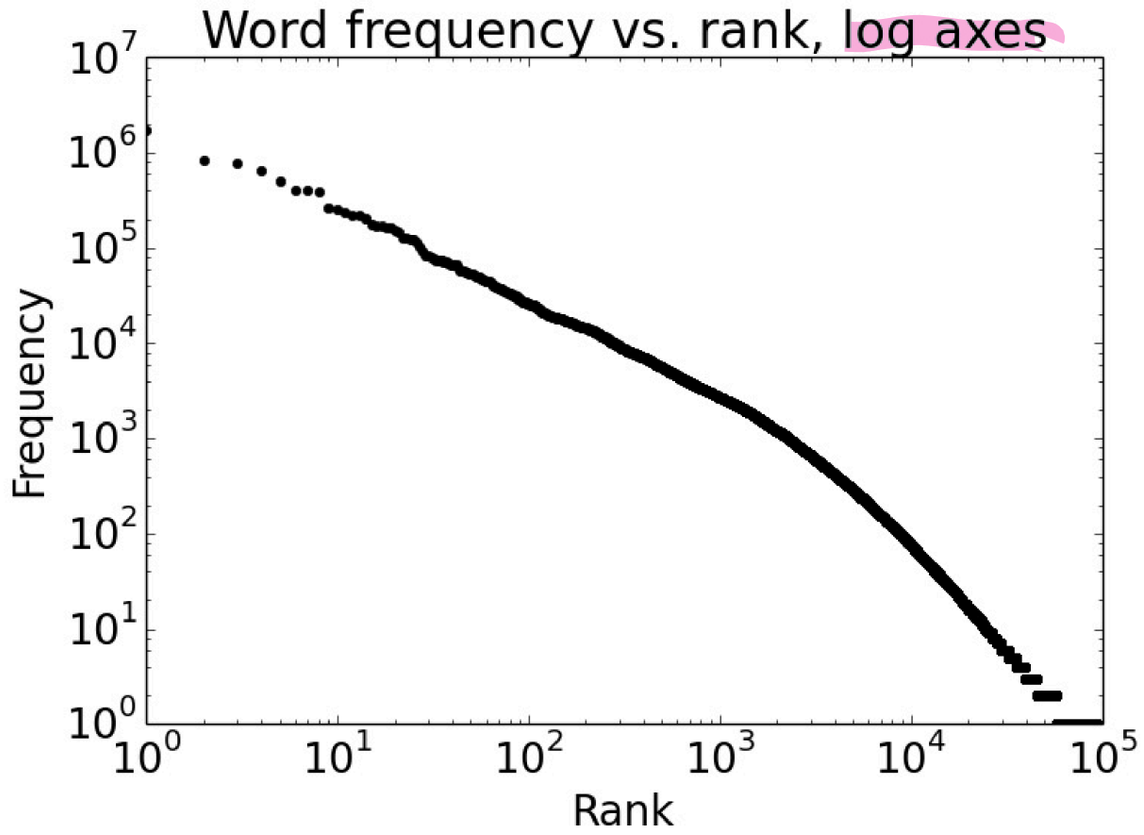
Europarl Corpus (24M words)

Plotting word frequencies



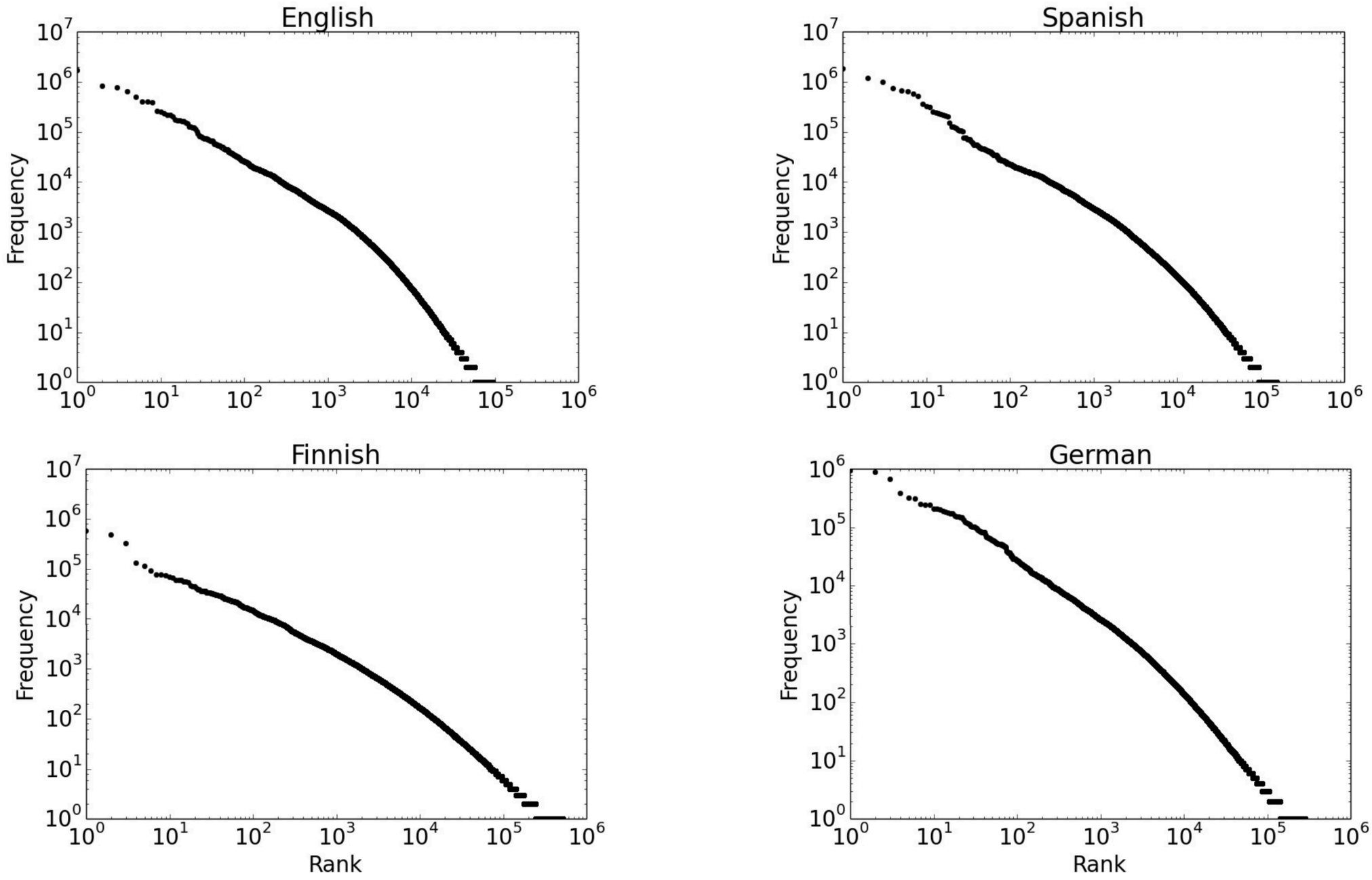
Word distributions follow a surprising pattern





Word distributions follow a surprising pattern

most languages, even rare ones, follow this



Zipf's law

George Zipf (1902 — 1950)

- The frequency of a word is inversely proportional to its rank

$$f \times r \approx k$$

f = frequency of a word

r = rank of a word (if sorted by frequency)

k = a constant

	Word	↓ Absolute Frequency ?
1	the	69,971
2	of	36,412
3	and	28,870

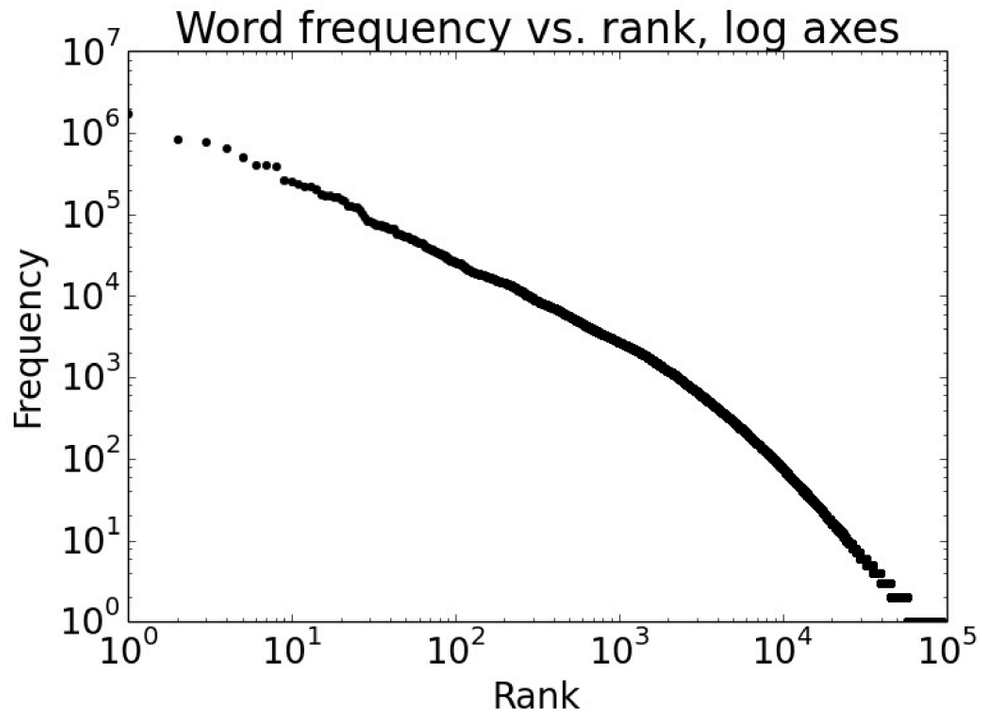
doesn't follow natural distribution

Also related: Benford's law for detecting fraud

Zipf's law

- In log-scale, the pattern emerges as a straight line

$$f \times r \approx k$$

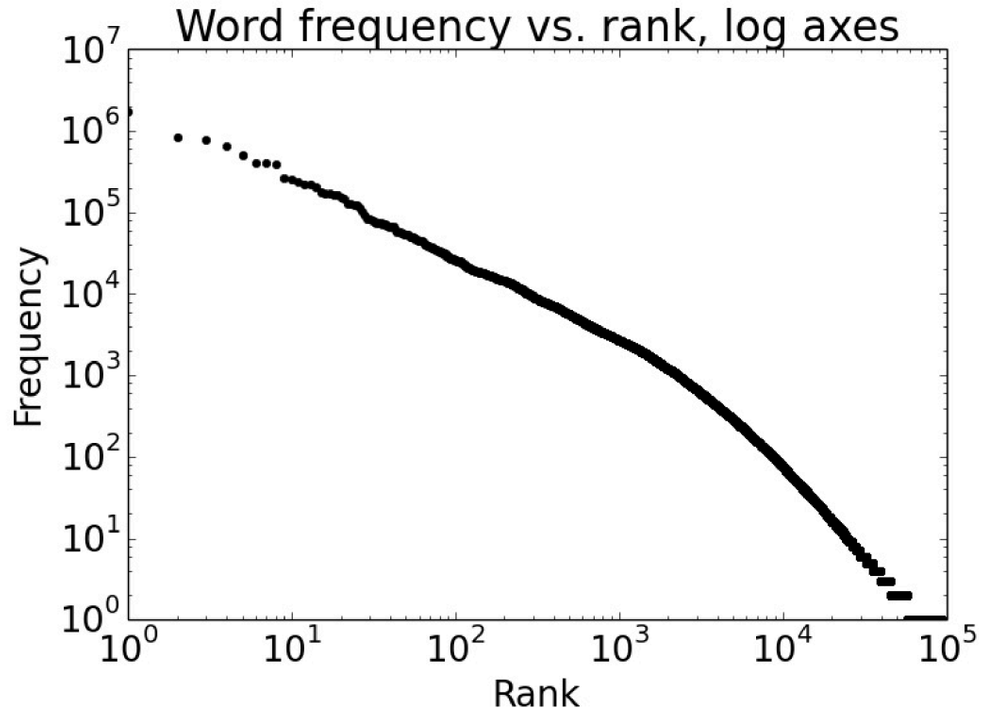


$$fr = k \quad \Rightarrow \quad f = \frac{k}{r} \quad \Rightarrow \quad \log f = \log k - \log r$$

☆ ⇒ natural patterns in human languages ☆

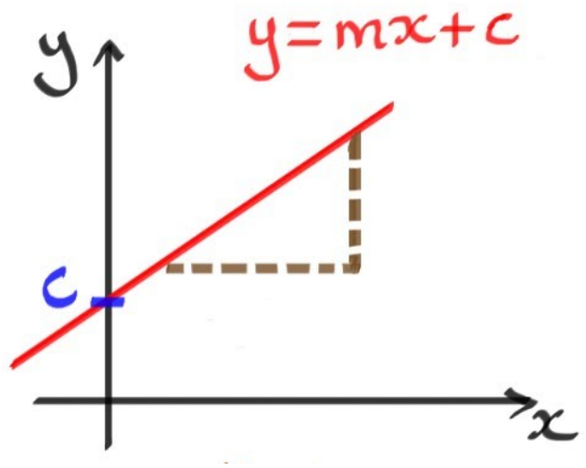
Zipf's law

- In log-scale, the pattern emerges as a straight line



$$f \times r \approx k$$

$$fr = k \quad \Rightarrow \quad f = \frac{k}{r} \quad \Rightarrow \quad \log f = \log k - \log r$$



y is rank
x is frequency
c is k

Is frequency an indicator of importance?

Word	↓ Absolute Frequency [?]
the	6,054,939
of	3,049,448
and	2,624,147
to	2,599,451
a	2,175,967
in	1,945,533
that	1,120,750
it	1,054,366
is	991,771
was	883,547

it can tell us if a corpus is natural or not

Keywords

How do you define important words in a corpus (collection of text documents)?

Keywords

- Words that have higher relative frequency than **expected** are keywords.

Keywords

- Words that have higher relative frequency than **expected** are keywords.

Keywords

- Words that have higher relative frequency than **expected** are keywords.

- How do we get the **expected frequency** of a word?

└ corpus from the internet / large collection of documents

Relative frequency

$$\text{relative frequency} = \frac{\text{absolute frequency}}{\text{number of tokens in corpus}} \text{ x basis for normalization}$$

$$\text{relative frequency of "the"} = \frac{6{,}054{,}950}{97{,}414{,}887} \times 1{,}000{,}000 = 62\text{K per million words}$$

Keywords

- Words that have higher relative frequency than in a **reference corpus** are keywords.

Keywords

- Words that have higher relative frequency than in a **reference corpus** are keywords.

2020

	Relative Frequency in Tweets corpus
people	3,763
case	2,000
day	1,617
lockdown	1,141
corona	1098

Keywords

- Words that have higher relative frequency than in a **reference corpus** are keywords.

	Relative Frequency in Tweets corpus	Relative Frequency in 2015 Web corpus
people	3,763	1,196
case	2,000	447
day	1,617	940
lockdown	1,141	1
corona	1098	2

Keywords

- Words that have higher relative frequency than in a **reference corpus** are keywords.

	Relative Frequency in Tweets corpus	Relative Frequency in 2015 Web corpus	Relative Frequency Ratio
people	3,763	1,196	3.15
case	2,000	447	4.47
day	1,617	940	1.72
lockdown	1,141	1	**1141**
corona	1098	2	**549**

Simple math parameter (SPM) keywords

Simple math parameter (SPM) keywords

	Relative Frequency in Tweets corpus
Covid19	500
mackaysuzie	40
people	3,763

Simple math parameter (SPM) keywords

	Relative Frequency in Tweets corpus	Relative Frequency in 2015 Web corpus
Covid19	500	0
mackaysuzie	40	1
people	3,763	1,196

Simple math parameter (SPM) keywords

	Relative Frequency in Tweets corpus	Relative Frequency in 2015 Web corpus	Relative Frequency Ratio
Covid19	500	0	ERROR
mackaysuzie	40	1	40
people	3,763	1,196	3.15

Simple math parameter (SPM) keywords

	Relative Frequency in Tweets corpus	Relative Frequency in 2015 Web corpus	Relative Frequency Ratio
Covid19	500	0	ERROR
mackaysuzie	40	1	40
people	3,763	1,196	3.15

play with k to see changes

$$\text{relative frequency ratio} = \frac{\text{relative frequency} + \mathbf{k}}{\text{reference relative frequency} + \mathbf{k}}$$

add constant K to normalize, saying every word has some importance

⇒ no division by zero

Keywords from Aug 9-10 tweets

Word	
covid	•••
pandemic	•••
coronavirus	•••
trump	•••
mask	•••
lockdown	•••
distancing	•••
corona	•••
covid19	•••
virus	•••

Keywords from Aug 9-10 tweets

Word	
covid	•••
pandemic	•••
coronavirus	•••
trump	•••
mask	•••
lockdown	•••
distancing	•••
corona	•••
covid19	•••
virus	•••

Word	
quarantine	•••
gonna	•••
idiot	•••
outbreak	•••
pelosi	•••
hoax	•••
biden	•••
damn	•••
hydroxychloroquine	•••
stupid	•••

Exercise

- Let's get keywords on a topic — pick a favorite topic

app.sketchengine.eu

create corpus

Outline

- N-grams
- **Keywords**
- **Collocations**

N-grams

we've only seen keywords limited to 1 token

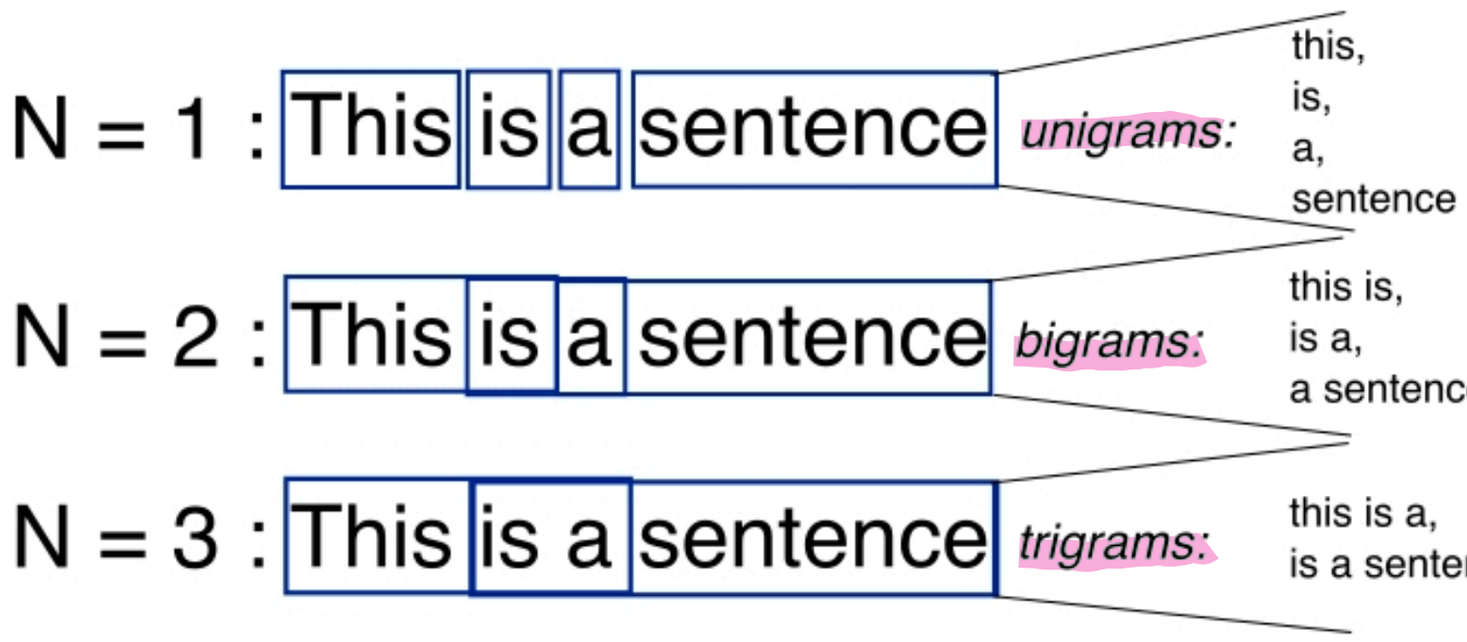
what about key phrases of 1+ tokens?

↓

n-grams

N-grams

N-Grams



N-grams

- Frequency alone is not an indicator of how interesting an n-gram is.

	Word	↓ Count ?	
1	of the	5,350,464	•••
2	in the	4,415,452	•••
3	to the	2,506,002	•••
4	on the	1,825,280	•••
5	for the	1,724,730	•••
6	to be	1,391,050	•••
7	and the	1,347,791	•••
8	with the	1,225,210	•••
9	at the	1,187,112	•••
10	in a	1,019,018	•••

Collocations

- Important n-grams are called collocations.
- Collocation is a series of words or terms that co-occur more often than would be expected by chance.
- "strong tea" v.s. "powerful tea"
- "strong computer" vs. "powerful computer"

How to identify collocations?

this is what we did for keywords

- Keywords metric, i.e., comparing relative frequency of an n-gram with relative frequency in a reference corpus works to some extent.

results in data sparcity
⇒ larger n-grams will occur less andless frequently

Word	
social distancing	•••
college football	•••
face mask	•••
global pandemic	•••
corona virus	•••
hair loss	•••
social distance	•••
coronavirus relief	•••
masked marauder	•••

How to identify collocations?

- Keywords metric, i.e., comparing relative frequency of an n-gram with relative frequency in a reference corpus works to some extent.

- But data sparsity is a problem as the phrase length increases

Word	
social distancing	•••
college football	•••
face mask	•••
global pandemic	•••
corona virus	•••
hair loss	•••
social distance	•••
coronavirus relief	•••
masked marauder	•••

Association metrics for collocations

- Collocation is a series of words or terms that co-occur more often than would be **expected by chance.**

$$p(x, y) > p(x)\ p(y)$$

probability
x and y
occuring together

Pointwise Mutual information

- A measure of the mutual dependence between two words — how much information can you tell about the another word given a word.

$$MI(x, y) = log \frac{p(x, y)}{p(x)\ p(y)}$$

mutual information

if you know about one word what do you know about the other

Pointwise Mutual information

- A measure of the mutual dependence between two words — how much information can you tell about the another word given a word.

$$MI(x,y) = log\frac{p(x,y)}{p(x)\,p(y)} \approx log\frac{p(x|y)}{p(x)}$$

probability of x given y

more probable than seeing x alone

$$p(x,y) = \frac{freq(x,y)}{\sum_x \sum_y freq(x,y)}$$ ← total number of bigrams $\approx N$

(8) a b c d (e) ← 6 words in corpus

} bigrams (5)

total num words in corpus

$$P(x) = \frac{freq(x)}{N}$$

$$P(y) = \frac{freq(y)}{N}$$

Pointwise Mutual information

- A measure of the mutual dependence between two words — how much information can you tell about the another word given a word.

$$MI(x,y) = log\frac{p(x,y)}{p(x)\,p(y)} \quad \approx log\frac{p(x|y)}{p(x)}$$

$$= log\frac{freq(x,y)\,N}{freq(x)\,freq(y)}$$

Pointwise Mutual information

a type of association metric

- A measure of the mutual dependence between two words — how much information can you tell about the another word given a word.

$$MI(x,y) = log\frac{p(x,y)}{p(x)\,p(y)} \approx log\frac{p(x|y)}{p(x)}$$

$$= log\frac{freq(x,y)\,N}{freq(x)\,freq(y)}$$

$$\approx log\frac{obs\_freq(x,y)}{exp\_freq(x,y)}$$

observed frequency

expected frequency

Collocations of pandemic

Lemma	↓ MI
global	6.94
damned	6.26
mid	5.80
worldwide	5.68
ongoing	5.54

Association metrics for collocations

- Collocation is a series of words or terms that co-occur more often than would be **expected by chance.**

observed frequency of x y = obs_freq(x,y) = freq(x,y)

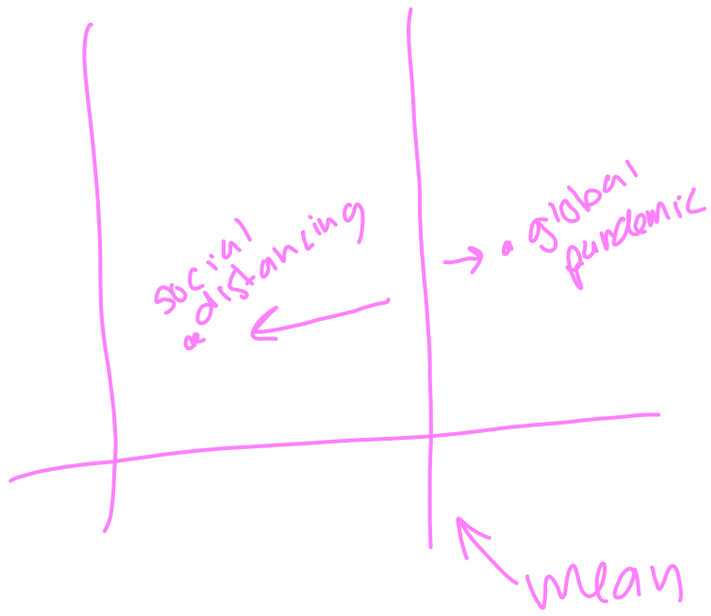
Association metrics for collocations

- Collocation is a series of words or terms that co-occur more often than would be **expected by chance.**

observed frequency of x y = obs_freq(x,y) = freq(x,y)

$$\text{expected frequency of x y} = \text{exp\_freq(x,y)} = \frac{\text{freq(x) freq(y)}}{\text{total number of words}}$$

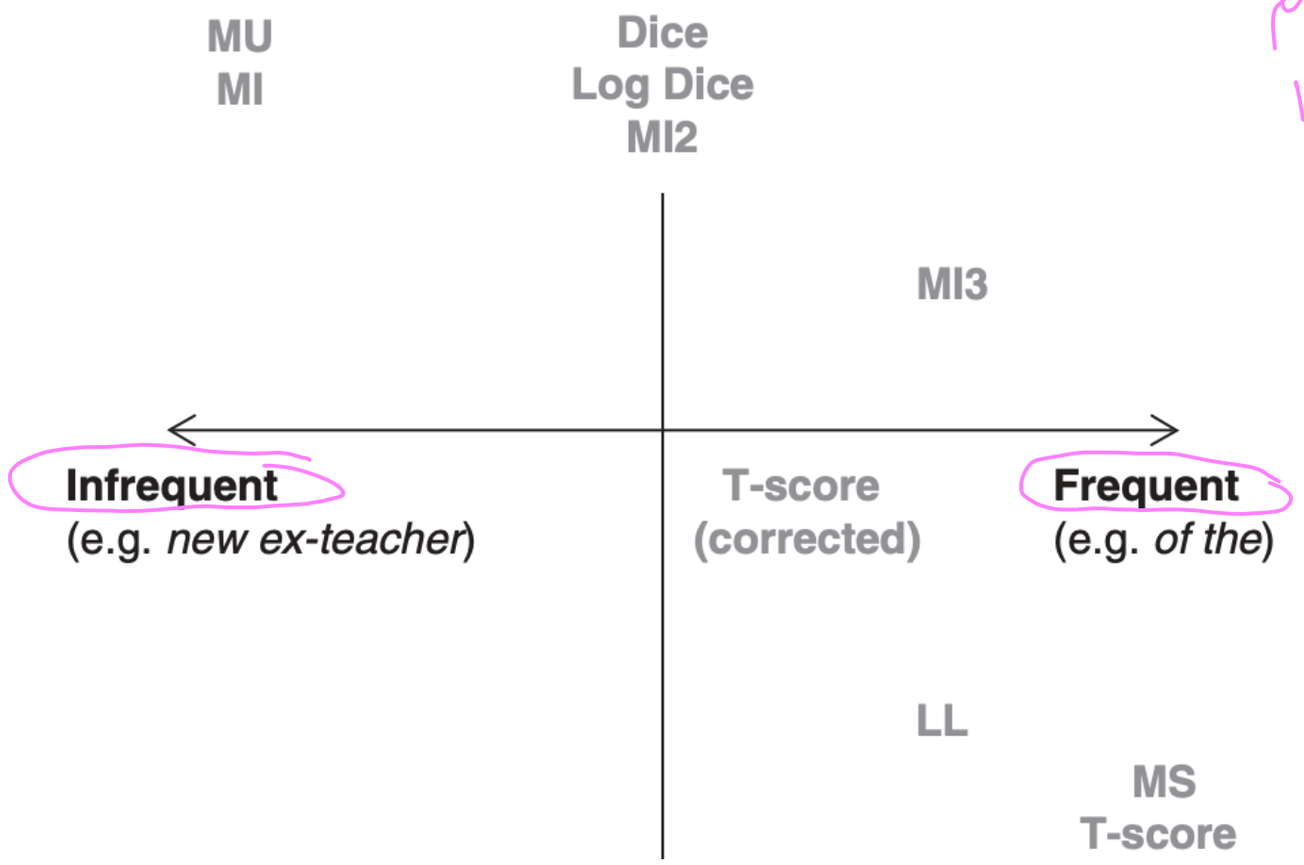
T-score



$$\text{T-score} = \frac{\text{obs\_freq} - \text{exp\_freq}}{\sqrt{\text{obs\_freq}}}$$

Association metrics

MI = mutual information



do we want more frequent or less frequent ones

Collocations of pandemic

mutual information
↳ strings that are
less frequent

Lemma	Cooccurrences ?	Candidates ?	T-score	↓ MI	LogDice	
global	112	187	10.50	6.94	10.31	•••
damned	3	8	1.71	6.26	5.18	•••
mid	3	11	1.70	5.80	5.18	•••
worldwide	9	36	2.94	5.68	6.75	•••
ongoing	5	22	2.19	5.54	5.91	•••
literal	3	14	1.69	5.45	5.18	•••
rage	6	29	2.39	5.40	6.17	•••
Global	3	15	1.69	5.35	5.18	•••
amid	12	61	3.38	5.33	7.16	•••
deadly	11	86	3.19	4.71	7.02	•••

Exercise

- Let’s check out collocations of a word

Collocations of different forms of run

verb

Word	Cooccurrences ?
mate	4,363
out	22,370
away	7,191
through	11,656
back	11,618
into	12,434
backs	2,794

run as noun

Word	Cooccurrences ?
scored	4,517
long	8,844
hits	2,396
unbeaten	1,453
home	7,901
earned	1,540

Exercise

- How would you describe your friend/model?