UNIVERSITÄT
ZU KÖLN

Corpora and Basic Word Counting

VL Sprachliche Informationsverarbeitung

Nils Reiter

nils.reiter@uni-koeln.de

October 31, 2024

Winter term 2024/25

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Junichi Tsujii (2021). “Natural Language Processing and Computational Linguistics”. In: *Computational Linguistics* 47.4, pp. 707–727. DOI: 10.1162/coli_a_00420. URL: https://aclanthology.org/2021.cl-4.24

Questions

Section 1

Corpora

Corpora

- (Large) collections of linguistic expressions
- Speech corpora: Spoken language
  - File formats: wav, mp3, ...
- Text corpora: Written language
  - File formats: txt, xml, json, ...

Corpora

- (Large) collections of linguistic expressions
- Speech corpora: Spoken language
  - File formats: wav, mp3, ...
- Text corpora: Written language
  - File formats: txt, xml, json, ...
- Why do we look at corpora?

Corpora

- (Large) collections of linguistic expressions
- Speech corpora: Spoken language
  - File formats: wav, mp3, ...
- Text corpora: Written language
  - File formats: txt, xml, json, ...
- Why do we look at corpora?
  - Making statements about language needs to take into account many language expressions
  - We under-estimate creativity, flexibility and productivity of language use
  - $\rightarrow$ Empiricism

Meta data and annotations

Meta data: Data about the data

- Information about the corpus
- Language, date of creation, author(s), publication source, ...
- Machine-readable: XML, JSON, CSV, ...

Meta data and annotations

Meta data: Data about the data

- Information about the corpus
- Language, date of creation, author(s), publication source, ...
- Machine-readable: XML, JSON, CSV, ...

Annotations: Data about parts of the corpus

- Examples
  - Linguistic annotation: Parts of speech, named entities, syntactic relations, ...
  - Non-linguistic annotation: Sentiment expressions, rhetoric devices, arguments, ...

Meta data and annotations

Meta data: Data about the data

- Information about the corpus
- Language, date of creation, author(s), publication source, ...
- Machine-readable: XML, JSON, CSV, ...

Annotations: Data about parts of the corpus

- Examples
  - Linguistic annotation: Parts of speech, named entities, syntactic relations, ...
  - Non-linguistic annotation: Sentiment expressions, rhetoric devices, arguments, ...
- Explicit location in the corpus: Document/word/character numbers in text, milliseconds in speech

Preparations (for text corpora)

- OCR: Optical Character Recognition Manning/Schütze (1999, 123)
  - Convert images (e.g., from a scan) into text
  - Huge improvements in last five years

Preparations (for text corpora)

- OCR: Optical Character Recognition Manning/Schütze (1999, 123)
  - Convert images (e.g., from a scan) into text
  - Huge improvements in last five years
- Encoding: How to specify characters in a computer
  - Simple: ASCII (7 bit per character, $2^7 = 128$ different characters)
  - Outdated: Latin-1 / ISO-8859 (8 bit, $\Rightarrow 256$ diff. characters)
  - Modern: Unicode (e.g., UTF-8)
    - 1 B/char to 4 B/char
    - $1\,112\,064$ characters can be represented

Tools and Techniques

- Plain text editors
  - We often want to inspect the corpus as it is on disk (i.e., without an editor interfering too much)
  - Mac: Textmate/emacs/vi; Windows: Notepad++/emacs/vi

Tools and Techniques

- Plain text editors
  - We often want to inspect the corpus as it is on disk (i.e., without an editor interfering too much)
  - Mac: Textmate/emacs/vi; Windows: Notepad++/emacs/vi
- Regular expressions
  - The most important tool for corpus analysis
    - Cleanup (e.g., after scraping a corpus from the web)
    - Analysis (e.g., to find all variants of a word or deal with slang)
  - Usable in *all** programming languages and find tools

Tools and Techniques

- Plain text editors
  - We often want to inspect the corpus as it is on disk (i.e., without an editor interfering too much)
  - Mac: Textmate/emacs/vi; Windows: Notepad++/emacs/vi
- Regular expressions
  - The most important tool for corpus analysis
    - Cleanup (e.g., after scraping a corpus from the web)
    - Analysis (e.g., to find all variants of a word or deal with slang)
  - Usable in *all** programming languages and find tools
- Command line
  - Large corpora often cannot be displayed with GUI tools
  - Command line tools faster and more memory efficient

Tokenization

- Segmenting a corpus into individual units
- Tokens: Words, punctuation, numbers, symbols, ...

Tokenization

- Segmenting a corpus into individual units
- Tokens: Words, punctuation, numbers, symbols, …
- Naive: Splitting at white space (space, newline, …)
  - Why naive?

Tokenization

- Segmenting a corpus into individual units
- Tokens: Words, punctuation, numbers, symbols, …
- Naive: Splitting at white space (space, newline, …)
  - Why naive?
- Solved, but complex
  - E.g., syntactic points vs. morphological points
- Sometimes, shortcuts are ok – depends on the use case

Word Counts

Count	Word
585	die
584	und
407	er
404	der
348	zu
311	sich
259	nicht
250	sie
243	in
243	den
233	war
218	Gregor
189	mit
178	das
176	auf
171	es
162	dem
155	hatte
137	ein
136	aber
133	daß
123	als
110	auch
107	Schwester
	...

Word Counts

Count	Word
585	die
584	und
407	er
404	der
348	zu
311	sich
259	nicht
250	sie
243	in
243	den
233	war
218	Gregor
189	mit
178	das
176	auf
171	es
162	dem
155	hatte
137	ein
136	aber
133	daß
123	als
110	auch
107	Schwester
	...

- Number of words in a text
- Most frequent words (MFW) are function words
- ‘Content words’ that appear often indicate text content

Zipf's Law

Manning/Schütze, 1999, 23 ff.

- George Kingsley Zipf (1902-1950): American Linguist
- Basic property of human language
  - Frequency distribution of words (in a corpus) is stable
  - Word frequency is inversely proportional to its position in the ranking

$$f \propto \frac{1}{r}$$

(there is a constant k, such that $f \times r = k$)

Zipf's Law

Manning/Schütze, 1999, 23 ff.

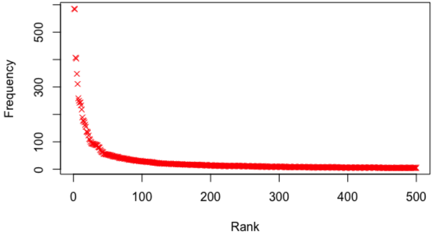

Figure: Words sorted after their frequency (red). Text: Kafka's "Die Verwandlung".

Zipf's Law

Manning/Schütze, 1999, 23 ff.

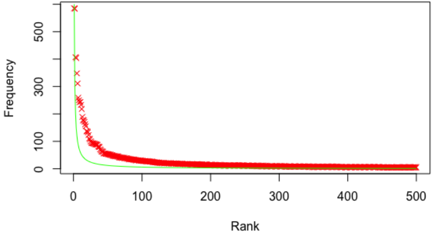

Figure: Words sorted after their frequency (red). Zipf distribution: $y = 600\frac{1}{x}$ (green). Text: Kafka's "Die Verwandlung".

Zipf's Law

Manning/Schütze, 1999, 23 ff.

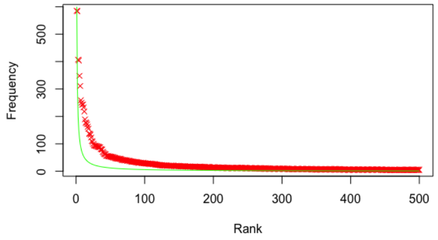

Figure: Words sorted after their frequency (red). Zipf distribution: $y = 600\frac{1}{x}$ (green). Text: Kafka's "Die Verwandlung".

Consequences

- Very few words appear with very high frequency
- The vast majority of words appear only once
  - It's difficult to learn something about these words!

Counting Words

- Absolute numbers are not that interesting
- Insights are only generated through comparison

Abs. number	Word form
20	women
67	woman
31	men
79	family
82	sister
83	friend
99	bath
117	father
133	man
144	sir

Table: Jane Austens's *Persuasion* (nouns)

Abs. number	Word form
0	friend
2	bath
11	women
23	men
30	father
68	woman
83	family
113	sir
121	man
282	sister

Table: Jane Austens's *Sense and Sensibility* (nouns)

Absolute Numbers

Word	Persuasion	Sense
woman	67	68
women	20	11
man	133	121
men	31	23
sister	82	282

...does it make sense to compare absolute numbers? No.

Absolute Numbers

Word	Persuasion	Sense
woman	67	68
women	20	11
man	133	121
men	31	23
sister	82	282

…does it make sense to compare absolute numbers? No.

- The texts/corpora do not have the same size
- Scaling using their length: Division by the total number of words

Absolute Numbers

Word	Persuasion		Sense	
woman	67	$0.000\,79\,\%$	68	$0.000\,55\,\%$
women	20	$0.000\,24\,\%$	11	$0.000\,09\,\%$
man	133	$0.001\,58\,\%$	121	$0.001\,00\,\%$
men	31	$0.000\,37\,\%$	23	$0.000\,19\,\%$
sister	82	$0.000\,97\,\%$	282	$0.002\,33\,\%$

...does it make sense to compare absolute numbers? No.

- The texts/corpora do not have the same size
- Scaling using their length: Division by the total number of words
- Visible changes: Proportion of “sister”: $3.4 \rightarrow 2.4$

Scaling

- Number of words: Result of a measurement
- If measuring in different scenarios, it's important to scale the results
  - "In a text that is much shorter, there are much less chances for a certain word to be used."

Scaling

- Number of words: Result of a measurement
- If measuring in different scenarios, it's important to scale the results
  - "In a text that is much shorter, there are much less chances for a certain word to be used."

Recipe

- Divide the result of the measurement by the theoretical maximum
- How many chances are there for "sister" to be used?
  - As many as there are words in the text
- Thus, we divide by the total number of words

Scaling

- Number of words: Result of a measurement
- If measuring in different scenarios, it's important to scale the results
  - "In a text that is much shorter, there are much less chances for a certain word to be used."

Recipe

- Divide the result of the measurement by the theoretical maximum
- How many chances are there for "sister" to be used?
  - As many as there are words in the text
- Thus, we divide by the total number of words

- It's not always obvious how to scaled
- When reading research: Was it scaled, and how?

Types and Tokens

Manning/Schütze, 1999, 21 f.

- If a text has been tokenized, we can access individual units: Tokens
- Not all tokens are words: Punctuation, detached prefixes, ...

Types and Tokens

Manning/Schütze, 1999, 21 f.

- If a text has been tokenized, we can access individual units: Tokens
- Not all tokens are words: Punctuation, detached prefixes, …
- We are often also interested in different tokens: Types

Types and Tokens

Manning/Schütze, 1999, 21 f.

- If a text has been tokenized, we can access individual units: Tokens
- Not all tokens are words: Punctuation, detached prefixes, ...
- We are often also interested in different tokens: Types

Example

the cat chases the mouse

Types and Tokens

Manning/Schütze, 1999, 21 f.

- If a text has been tokenized, we can access individual units: Tokens
- Not all tokens are words: Punctuation, detached prefixes, ...
- We are often also interested in different tokens: Types

Example

the cat chases the mouse

- Tokens: the, cat, chases, the, mouse
- Types: the, cat, chases, mouse

Type-Token-Ratio (TTR)

- What is the relation between number of tokens and number of types?

Type-Token-Ratio (TTR)

- What is the relation between number of tokens and number of types?
- Construct a sentence with 5 tokens and 5 types!

Type-Token-Ratio (TTR)

- What is the relation between number of tokens and number of types?
- Construct a sentence with 5 tokens and 5 types!
  - “the dog barks loudly .”

Type-Token-Ratio (TTR)

- What is the relation between number of tokens and number of types?
- Construct a sentence with 5 tokens and 5 types!
  - “the dog barks loudly .”
- Construct a sentence with 5 tokens and 4 types!

Type-Token-Ratio (TTR)

- What is the relation between number of tokens and number of types?
- Construct a sentence with 5 tokens and 5 types!
  - "the dog barks loudly ."
- Construct a sentence with 5 tokens and 4 types!
  - "the cat loves the mouse"

Type-Token-Ratio (TTR)

- What is the relation between number of tokens and number of types?
- Construct a sentence with 5 tokens and 5 types!
  - “the dog barks loudly .”
- Construct a sentence with 5 tokens and 4 types!
  - “the cat loves the mouse”
- Construct a sentence with 5 tokens and 1 type!

Type-Token-Ratio (TTR)

- What is the relation between number of tokens and number of types?
- Construct a sentence with 5 tokens and 5 types!
  - "the dog barks loudly ."
- Construct a sentence with 5 tokens and 4 types!
  - "the cat loves the mouse"
- Construct a sentence with 5 tokens and 1 type!
  - "dog dog dog dog dog" (not really a sentence ...)
  - It's not possible to create a 'proper' sentence with 1 type

Type-Token-Ratio (TTR)

- Measure for ‘lexical variability’

$$TTR = \frac{\text{number of types}}{\text{number of tokens}}$$

- Max value: 1

Type-Token-Ratio (TTR)

- Measure for ‘lexical variability’

$$TTR = \frac{\text{number of types}}{\text{number of tokens}}$$

- Max value: 1 (there cannot be more types than tokens)
- Min value: $\epsilon = \frac{1}{\text{very large number}}$

Type-Token-Ratio (TTR)

- Measure for 'lexical variability'

$$TTR = \frac{\text{number of types}}{\text{number of tokens}}$$

- Max value: 1 (there cannot be more types than tokens)
- Min value: $\epsilon = \frac{1}{\text{very large number}}$
- Real (German) texts
  - $10\,000$ words (Wikipedia): $\frac{4021}{10\,000} = 0.4021$

TTR and Text Length

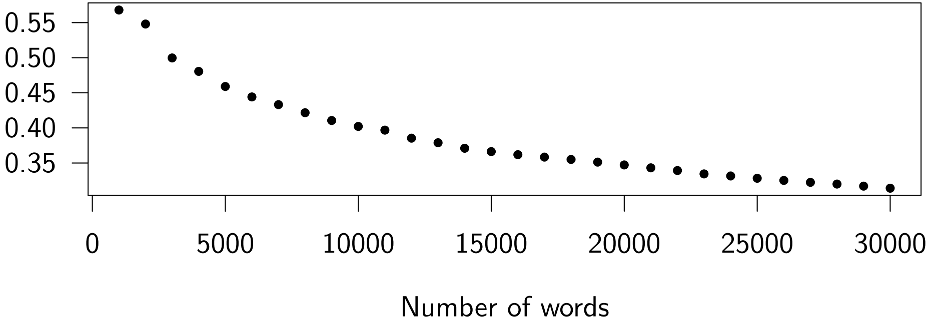

Figure: Type-Token-Ratio for increasing text lengths

TTR and Text Length

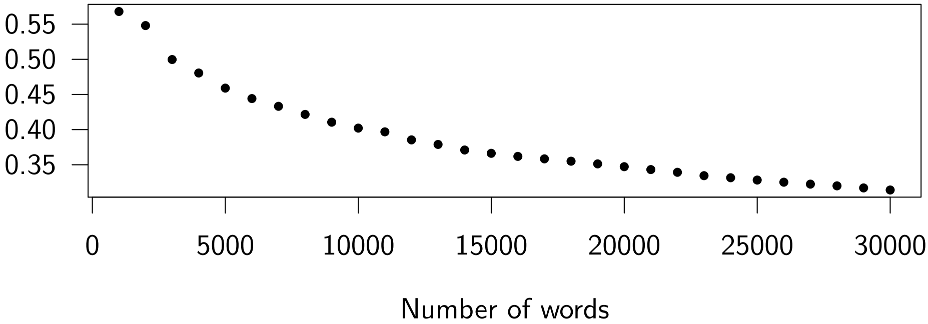

Figure: Type-Token-Ratio for increasing text lengths

- Increasing length $\rightarrow$ lower TTR!
- Why?

TTR and Text Length

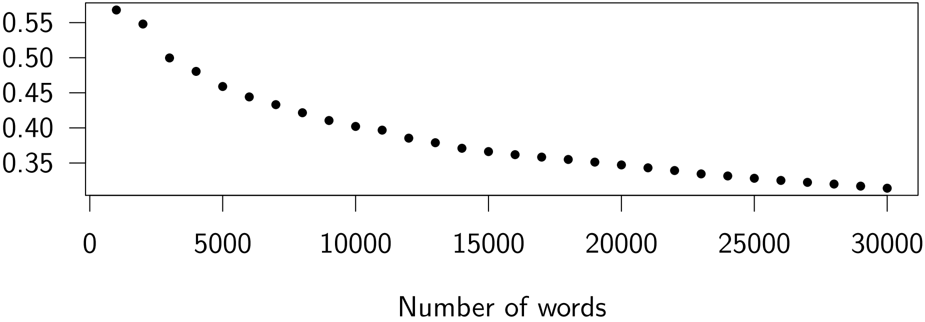

Figure: Type-Token-Ratio for increasing text lengths

- Increasing length → lower TTR!
- Why?– Zipf!

Standardized TTR (STTR)

- Calculate TTR over windows of fixed size (e.g., 1000 words)
- Calculate arithmetic mean over TTR values

Standardized TTR (STTR)

- Calculate TTR over windows of fixed size (e.g., 1000 words)
- Calculate arithmetic mean over TTR values

$$TTR_n = \frac{\text{number of types in } n\text{th window}}{\text{number of tokens in } n\text{th window}}$$

Standardized TTR (STTR)

- Calculate TTR over windows of fixed size (e.g., 1000 words)
- Calculate arithmetic mean over TTR values

$$
\begin{aligned}
TTR_n &= \frac{\text{number of types in } n\text{th window}}{\text{number of tokens in } n\text{th window}} \\
STTR &= \frac{1}{w}\sum_{i=0}^{w} TTR_i
\end{aligned}
$$

n-grams

- So far: Individual tokens
- But: Context is important for linguistic expressions

n-grams

- So far: Individual tokens
- But: Context is important for linguistic expressions
- n-gram: A list of n directly adjacent tokens
  - Popular choices for n: 2 to 4

n-grams

- So far: Individual tokens
- But: Context is important for linguistic expressions
- n-gram: A list of n directly adjacent tokens
  - Popular choices for n: 2 to 4

Example

The dog barks.

- 1-grams: "the", "dog", "barks", "."
- 2-grams (bigrams): "the dog", "dog barks", "barks ."
- 3-grams (trigrams): "the dog barks", "dog barks ."

Counting Bigrams

- Simple idea: We count bigrams (i.e., pairs of subsequent tokens)

Counting Bigrams

- Simple idea: We count bigrams (i.e., pairs of subsequent tokens)
- Corpus: Wikipedia pages (first $10\,000$ sentences)

Bigram	Frequency
wurde er	630
in der	623
wurde die	501
an der	386
mit dem	363
in die	362
in den	329
mit der	312
wurde das	291
wurde der	291
für die	248
er in	193
war er	181
von der	174
wo er	169
bei den	168
bei der	166
und wurde	165
an die	161
und die	150
er die	143
er als	142
er mit	142
wurden die	142
auf dem	135
für den	133
wurde sie	127
er zum	123
auf der	122

Counting Bigrams

- Simple idea: We count bigrams (i.e., pairs of subsequent tokens)
- Corpus: Wikipedia pages (first $10\,000$ sentences)
- Again, there are a lot of function words. Why?

Bigram	Frequency
wurde er	630
in der	623
wurde die	501
an der	386
mit dem	363
in die	362
in den	329
mit der	312
wurde das	291
wurde der	291
für die	248
er in	193
war er	181
von der	174
wo er	169
bei den	168
bei der	166
und wurde	165
an die	161
und die	150
er die	143
er als	142
er mit	142
wurden die	142
auf dem	135
für den	133
wurde sie	127
er zum	123
auf der	122

Counting Bigrams

- Simple idea: We count bigrams (i.e., pairs of subsequent tokens)
- Corpus: Wikipedia pages (first $10\,000$ sentences)
- Again, there are a lot of function words. Why?
- Zipf's law: Two words that are highly frequent have much higher chance to co-occur with high frequency

Bigram	Frequency
wurde er	630
in der	623
wurde die	501
an der	386
mit dem	363
in die	362
in den	329
mit der	312
wurde das	291
wurde der	291
für die	248
er in	193
war er	181
von der	174
wo er	169
bei den	168
bei der	166
und wurde	165
an die	161
und die	150
er die	143
er als	142
er mit	142
wurden die	142
auf dem	135
für den	133
wurde sie	127
er zum	123
auf der	122

Section 2

Summary

Summary

- Language data: Corpora
- Most frequent words are not the most informative words
- Zipf distribution
- Type-token ratio as a measure of lexical diversity
- n-grams: Look at multiple tokens at once

Section 3

Exercise

Übung 1

Besorgen Sie sich auf `https://opendiscourse.de/` Reden von zwei verschiedenen Politiker:innen aus unterschiedlichen Parteien, so dass sie insgesamt pro Person mehr als 10000 Wörter haben. Schreiben Sie dann in einer Programmiersprache Ihrer Wahl ein Programm, das die type-token-ratio für beide berechnet. Abgabe in Ilias bis zum 08.11.