Chapter NLP:II

II. Corpus Linguistics

- Empirical Research
- Text Corpora
- Text Statistics
- Text Statistics in IR
- Data Acquisition

Empirical Research

1. Quantitative research based on numbers and statistics.
2. Studies phenomena and research qestions by analyzing data.
3. Derives knowledge from experience rather than from theory or belief.

Empirical Research

1. Quantitative research based on numbers and statistics.
2. Studies phenomena and research qestions by analyzing data.
3. Derives knowledge from experience rather than from theory or belief.

Quantitative versus qualitative research:

- Quantitative. Characterized by objective measurements.
- Qualitative. Emphasizes the understanding of human experience.

Empirical Research

1. Quantitative research based on numbers and statistics.
2. Studies phenomena and research qestions by analyzing data.
3. Derives knowledge from experience rather than from theory or belief.

Quantitative versus qualitative research:

- Quantitative. Characterized by objective measurements.
- Qualitative. Emphasizes the understanding of human experience.

Descriptive versus inferential statistics:

- Descriptive. Procedures for summarizing and comprehending a sample or distribution of values. Used to describe phenomena.

  1 2 2 2 $\rightarrow$ mean M = 1.75

- Inferential. Procedures that help draw conclusions based on values. Used to generalize inferences beyond a given sample.

  The average number is significantly greater than 1.

Empirical Research

Research Questions

What is a good research question? [Bartos 1992]

- Asks about the relationship between two or more variables.
- Is testable (i.e., it is possible to collect data to answer the question).
- Is stated clearly and in the form of a question.
- Does not pose an ethical or moral problem for implementation.
- Is specific and restricted in scope.
- Identifies exactly what is to be solved.

Empirical Research

Research Questions

What is a good research question? [Bartos 1992]

- Asks about the relationship between two or more variables.
- Is testable (i.e., it is possible to collect data to answer the question).
- Is stated clearly and in the form of a question.
- Does not pose an ethical or moral problem for implementation.
- Is specific and restricted in scope.
- Identifies exactly what is to be solved.

Example of a poorly formulated question:

> *"What is the effectiveness of parent education when given problem children?"*

Empirical Research

Research Questions

What is a good research question? [Bartos 1992]

- Asks about the relationship between two or more variables.
- Is testable (i.e., it is possible to collect data to answer the question).
- Is stated clearly and in the form of a question.
- Does not pose an ethical or moral problem for implementation.
- Is specific and restricted in scope.
- Identifies exactly what is to be solved.

Example of a poorly formulated question:

“What is the effectiveness of parent education when given problem children?”

Example of a well-formulated question:

“What is the effect of the STEP parenting program on the ability of parents to use natural, logical consequences (as opposed to punishment) with their child who has been diagnosed with bipolar disorder?”

Empirical Research

Empirical Research in NLP

- Corpus linguistics.

  NLP is studied in a corpus-linguistics manner; i.e., approaches are developed and evaluated on collections of text.

- Evaluation measures.

  An evaluation of the quality of an approach is important, especially of its effectiveness.

- Experiments.

  The quality of an approach is empirically evaluated on test corpora and compared to alternative approaches.

- Hypothesis testing.

  Methods which verify whether results of an experiment are meaningful/valid by estimating the odds that the results happened by chance.

Empirical Research

Empirical Research in NLP

- Corpus linguistics.

  NLP is studied in a corpus-linguistics manner; i.e., approaches are developed and evaluated on collections of text.

- Evaluation measures.

  An evaluation of the quality of an approach is important, especially of its effectiveness.

- Experiments.

  The quality of an approach is empirically evaluated on test corpora and compared to alternative approaches.

- Hypothesis testing.

  Methods which verify whether results of an experiment are meaningful/valid by estimating the odds that the results happened by chance.

Text Corpora

Corpus Linguistics

- The study of language as expressed in principled collections of natural language texts, called text corpora.
- Aims to derive knowledge and rules from real-world text.
- Covers both manual and automatic analysis of text.

Text Corpora

Corpus Linguistics

- The study of language as expressed in principled collections of natural language texts, called text corpora.
- Aims to derive knowledge and rules from real-world text.
- Covers both manual and automatic analysis of text.

Three main techniques:

1. Analysis. Developing and evaluating methods based on a corpus.
2. Annotation. Coding data with categories to facilitate data-driven research.
3. Abstraction. Mapping of annotated texts to a theory-based model.

→ Need for text corpora: Without a corpus, it's hard to develop a strong approach—and impossible to reliably evaluate it.

"It's often not the one who has the best algorithm that wins. It's who has the most data."

Text Corpora

Definition 1 (Text Corpus [Butler 2004]**)**

A text corpus is (an electronically stored) collection of data designed with according to specific corpus design criteria to be maximally representative of (a particular variety of) language or other semiotic systems.

The basic unit for representing text is typically a word (captures meaning).

Examples:

- 200,000 product reviews for sentiment analysis
- 1,000 news articles for part-of-speech tagging

Corpora in NLP:

- NLP approaches are developed and evaluated on text corpora.
- Usually, the corpora contain annotations of the output information type to be inferred.

Text Corpora

On Representativeness

- *“extent to which a sample includes the full range of variability in a population”* [Biber 1993]

  Here: Sample is our corpus, population is all of the language variety.

- *“A corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety.”* [Leech 1991]

  Question: If we find certain features in the corpus, are we likely to find the same features in further data of that type?

- But—what is representative to the users of language?

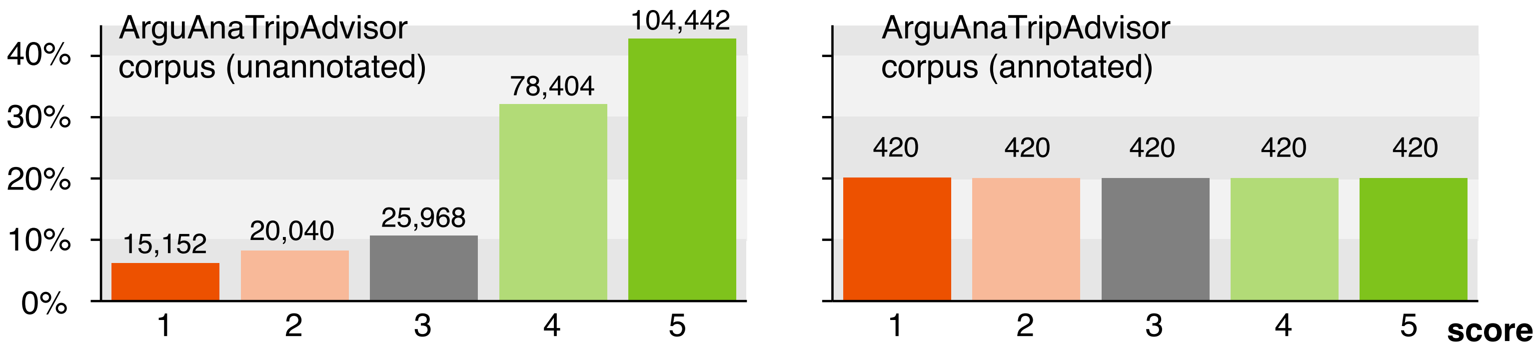
  “According to claims, the most likely document that an ordinary English citizen will cast his or her eyes over is The Sun newspaper” [Sinclair 2005]

  Keyword: reception versus production

- Corpus representativeness is important for generalization, since the corpus governs what can be learned about a given domain.

Text Corpora

Representative Data versus Balanced Data



- A corpus is representative for some output information type C, if it includes the full range of variability of texts with respect to C.
- The distribution of texts over the values of C should be representative for the real distribution.
- Balance with respect to a feature means that no value/level of the feature dominates; equally distributed with respect to a feature (e.g. genre, category of linguistic phenomena).
- A balanced distribution, where all values are evenly represented, may be favorable (particularly for machine learning).

Text Corpora

Character Encoding [detailed in WT:III-163 ff.]

- Character encoding is a mapping between bits and *code points*, where each code point is associated with a glyph.
  - Glyphs are graphical representations of symbols a $\leftarrow$ **a**, A, **A**, *a*, *A*
  - Getting from bits in a file to characters on a screen can be a major source of incompatibility.

- Charset for English: ASCII
  - Encodes 128 letters, numbers, special characters, and control characters in 7 bits, extended with an extra bit for storage in bytes.

- Charset for other European: Latin-1 (ISO-8859-1)

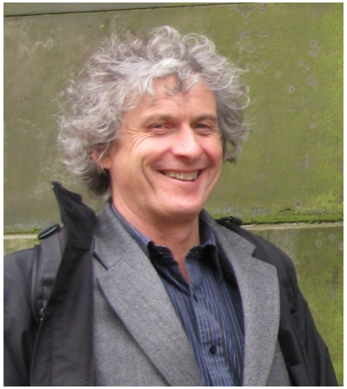
Text Corpora

Character Encoding [detailed in WT:III-163 ff.]

- Character encoding is a mapping between bits and *code points*, where each code point is associated with a glyph.
  - Glyphs are graphical representations of symbols a ← **a**, A, **A**, *a*, *A*
  - Getting from bits in a file to characters on a screen can be a major source of incompatibility.

- Charset for English: ASCII
  - Encodes 128 letters, numbers, special characters, and control characters in 7 bits, extended with an extra bit for storage in bytes.

- Charset for other European: Latin-1 (ISO-8859-1)



"Even when documents say they are in ASCII or ISO 8859-1, you have to assume that they are lying, because it's extremely common for such documents to be actually encoded in Windows-1252."

[David Hawking]

Text Corpora

Character Encoding (Fortsetzung) [detailed in WT:III-163 ff.]

ASCII symbols, hexadecimal notation

Code	...0	...1	...2	...3	...4	...5	...6	...7	...8	...9	...A	...B	...C	...D	...E	...F
0...	*NUL*	*SOH*	*STX*	*ETX*	*EOT*	*ENQ*	*ACK*	*BEL*	*BS*	*HT*	*LF*	*VT*	*FF*	*CR*	*SO*	*SI*
1...	*DLE*	*DC1*	*DC2*	*DC3*	*DC4*	*NAK*	*SYN*	*ETB*	*CAN*	*EM*	*SUB*	*ESC*	*FS*	*GS*	*RS*	*US*
2...	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3...	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4...	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5...	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6...	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7...	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Text Corpora

Character Encoding (Fortsetzung) [detailed in WT:III-163 ff.]

What could go wrong?

Text Corpora

Character Encoding (Fortsetzung) [detailed in WT:III-163 ff.]

- Other languages can have many more glyphs:
  - Chinese has more than 40,000 characters, with over 3,000 in common use.
- Many languages have multiple encoding schemes:
  - the CJK (Chinese-Japanese-Korean) family of East Asian languages, Hindi, Arabic
  - must specify encoding, cannot have multiple languages in one file

➜ Solution: Unicode

Text Corpora

Character Encoding (Fortsetzung) [detailed in WT:III-163 ff.]

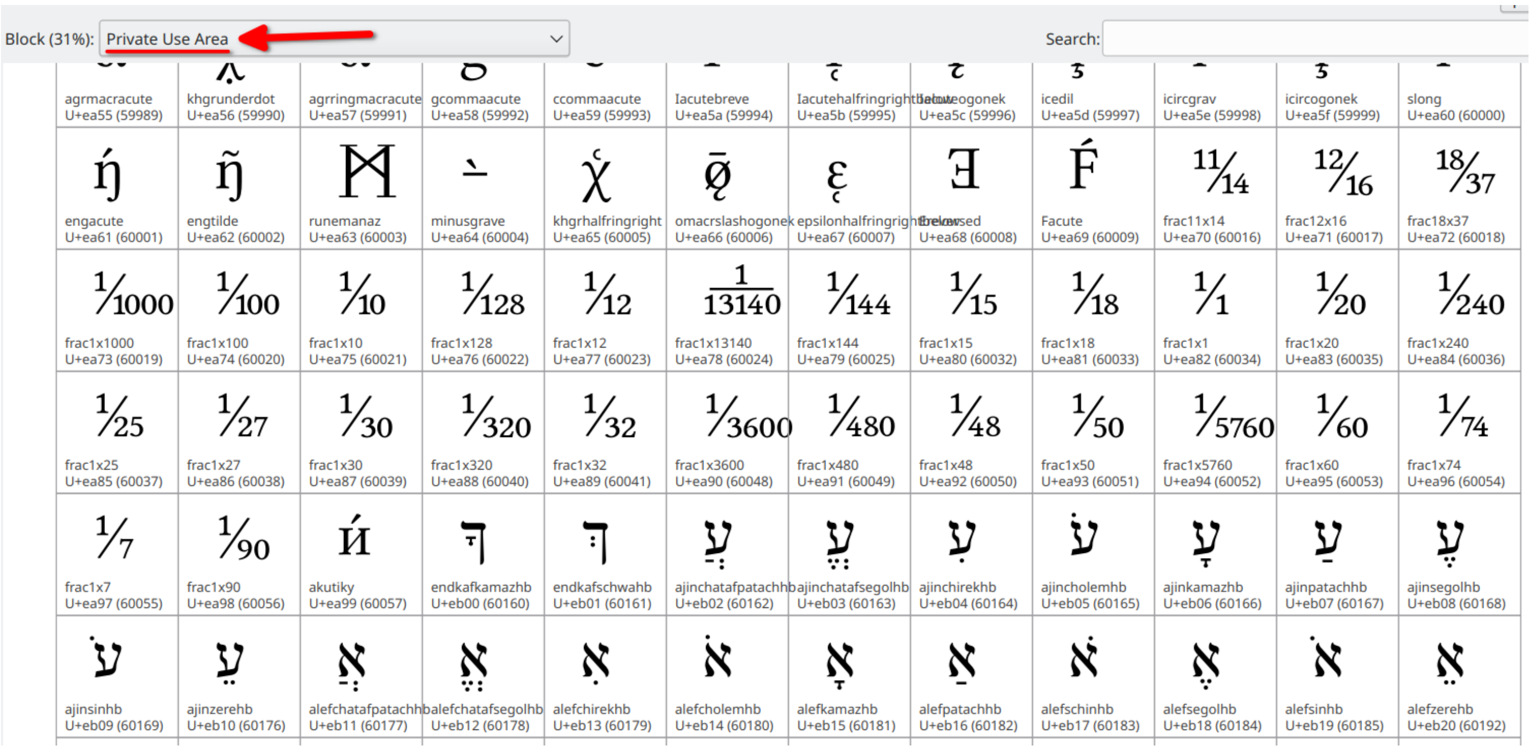
Unicode:

- All-encompassing charset and encoding for most writing systems.
- Allows for multiple languages in one file.
- Tailored encoding schemes to translate code points to a byte representation:
  - UTF-8 uses one byte for English (ASCII), and as many as 4 bytes for some traditional Chinese characters (variable length encoding).
  - UTF-16 uses 2 or 4 bytes for every character.
  - UTF-32 uses 4 bytes for every character.
- Applications may use UTF-32 for internal encoding (fast random lookup) and UTF-8 for disk storage (less space).

Text Corpora

Character Encoding (Fortsetzung)

Unicode – Private Use Areas (here: KompLett)



Text Corpora

Text as data

String: concatenation of alphabet elements

- "Hello world!", "", "0001011110001 0101", "To be or not to be..."
- essential, elementary data type in computer linguistics
- common operations: e.g.
  - concatenation: "Hello" + "World!" + "!" $\rightarrow$ "Hello World!"
  - splitting: split("Hello World!", " ") $\rightarrow$ {"Hello", "World!"}
  - case conversion: uppercase("Hello") $\rightarrow$ "HELLO"
  - substring: substr("Hello", start = 0, length = 4) $\rightarrow$ "Hell"

Document: compound data type

- (collection of) strings (e.g. title, body) [+ Metadata]

Corpus: collection of documents

Text Corpora

Text as data (Fortsetzung)

Type: (cp. class)

- (abstract) string representing a meaningful concept, e.g. words

Token: (cp. object)

- (concrete) string as instance of a meaningful concept

{ disciplines, distinction, concept, ... } ← "In disciplines such as knowledge representation and philosophy, the type–token distinction is a distinction that separates a concept from the objects which are particular instances of the concept."

(Wikipedia → Type–token distinction)

Vocabulary:

- complete set of all types occurring in a [document | collection]

Text Corpora

Text as data (Fortsetzung)

Transformation of text into numerical objects. See Text Models for more detail!

$$\left\{ \begin{array}{l} \text{disciplines} \\ \text{distinction} \\ \text{concept} \\ \ldots \end{array} \right\} \qquad \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \qquad \begin{pmatrix} 1 & 2 & 4 & \ldots \\ 2 & 0 & 0 & \ldots \\ 2 & 5 & 1 & \ldots \end{pmatrix}$$

Transformed objects $\rightarrow$ Data Mining

- process of discovering patterns in large data sets

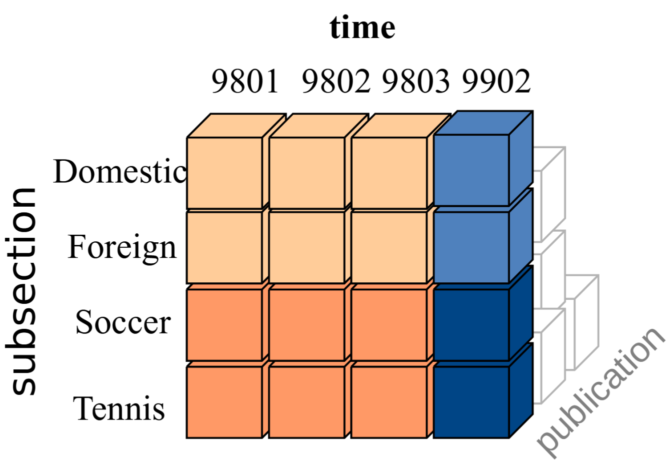
Text Corpora
Metadata

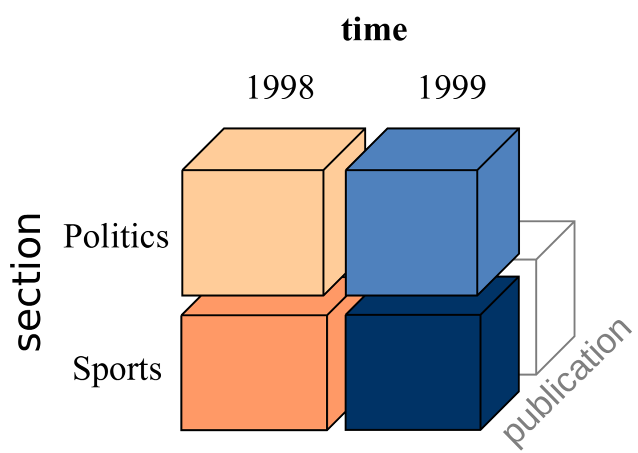
Metadata = text external context / covariate

Metadata = data facet

- Subselections of sources
- Aggregation / differentiation of results

context → contrast → meaning

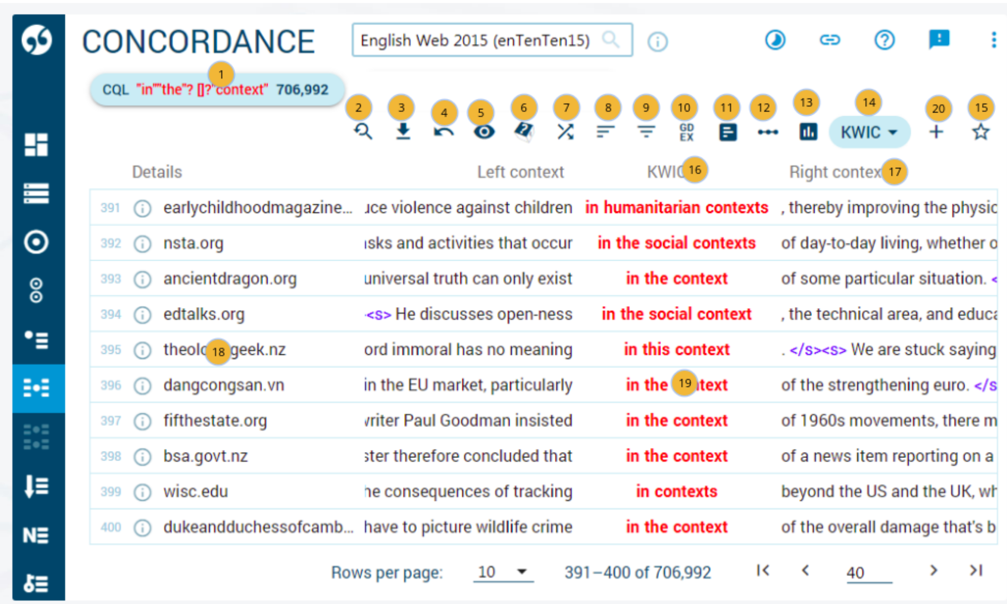




Text Corpora

Research in Language Use

Concordance: (alphabetical) list of principal words (or phrases) used in a book (nowadays: corpus) listing every instance of each with immediate context



[www.sketchengine.eu]

Text Corpora

Research in Language Use (Fortsetzung)

Compare usages of a word, analyse keywords, analyse frequencies, find phrases, idioms, etc.

Find the best synonym

Use the hash sign in front of a word to check which of its synonyms are commonly written.

waiting * #response

waiting for an answer	110,000	35%
waiting for a reply	71,000	22%
waiting for a response	59,000	18%
waiting for reply	15,000	4.6%
waiting for your reply	13,000	4.1%
waiting for the answer	12,000	4.0%
waiting for response	10,000	3.2%
waiting to answer	9,600	3.0%
waiting for your answer	7,500	2.3%
waiting for his answer	7,300	2.3%
waiting for my answer	6,400	2.0%

[netspeak.org]

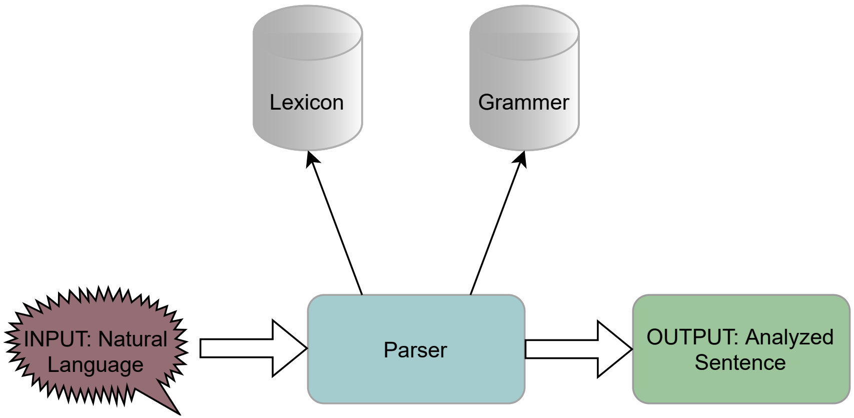
Chapter NLP:II

II. Corpus Linguistics

- ❑ Empirical Research
- ❑ Text Corpora
- ❑ Text Statistics
- ❑ Text Statistics in IR
- ❑ Data Acquisition

Text Statistics

Classic processing model for language:



Statistical aspects of language

- The lexical entries are not used equally often
- The grammatical rules are not used equally often
- The expected value of certain word forms or word form combinations depends on the technical language used (Sub Language)

Text Statistics

Questions:

- How many words are there?
- How do we count?

  bank$^{(1)}$ (the financial institution),
  bank$^{(2)}$ (land along the side of a river or lake),
  banks$^{(1)}$, banks$^{(2)}$, ...

- How often does each word occur?

Text Statistics

Questions:

- How many words are there?
- How do we count?

  bank$^{(1)}$ (the financial institution),
  bank$^{(2)}$ (land along the side of a river or lake),
  banks$^{(1)}$, banks$^{(2)}$, ...

- How often does each word occur?

Experiment:

- Read a text left to right (beginning to end); make a tally of every new word seen.
- n words seen in total, $v(n)$ different words so far.
- How does the vocabulary V (set of distinct words) grow? $\rightarrow$ Plot $v(n)$.

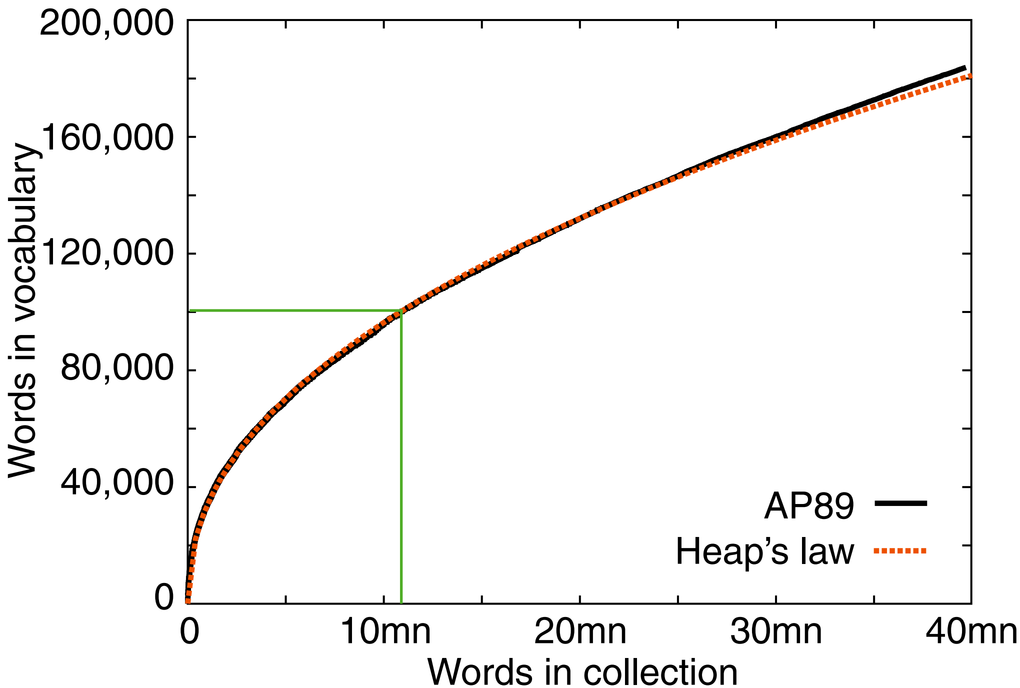
Text Statistics

Vocabulary Growth: Heaps' Law

The vocabulary V of a collection of documents grows with the collection.
Vocabulary growth can be modeled with Heaps' Law:

$$|V| = k \cdot n^{\beta},$$

where n is the number of non-unique words, and k and β are collection parameters.



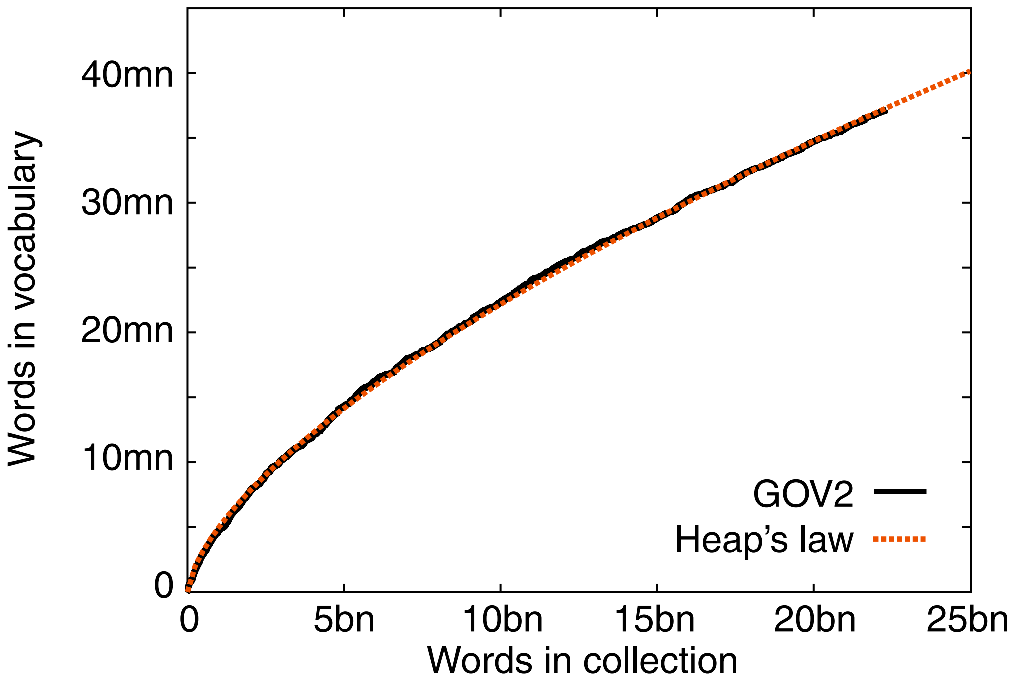
- Corpus: AP89
- $k = 62.95$, $\beta = 0.455$
- At $10,879,522$ words: $100,151$ predicted, $100,024$ actual.
- At $< 1,000$ words: poor predictions

Vocabulary Growth: Heaps' Law

The vocabulary V of a collection of documents grows with the collection. Vocabulary growth can be modeled with Heaps' Law:

$$|V| = k \cdot n^{\beta},$$

where n is the number of non-unique words, and k and β are collection parameters.

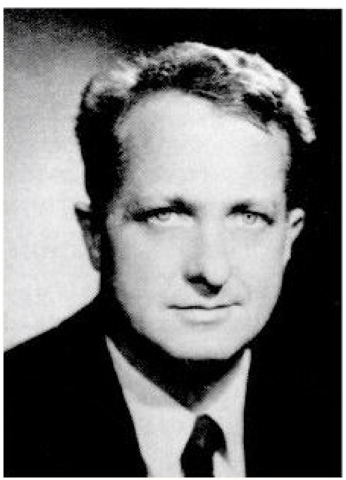


- Corpus: GOV2
- $k = 7.34$, $\beta = 0.648$
- Vocabulary continuously grows in large collections
- New words include spelling errors, invented words, code, other languages, email addresses, etc.

Text Statistics

Term Frequency: Zipf's Law

- The distribution of word frequencies is very *skewed*: Few words occur very frequently, many words hardly ever.
- For example, the two most common English words (the, of) make up about 10% of all word occurrences in text documents. In large text samples, about 50% of the unique words occur only once.



George Kingsley Zipf, an American linguist, was among the first to study the underlying statistical relationship between the frequency of a word and its rank in terms of its frequency, formulating what is known today as Zipf's law.

For natural language, the "Principle of Least Effort" applies.

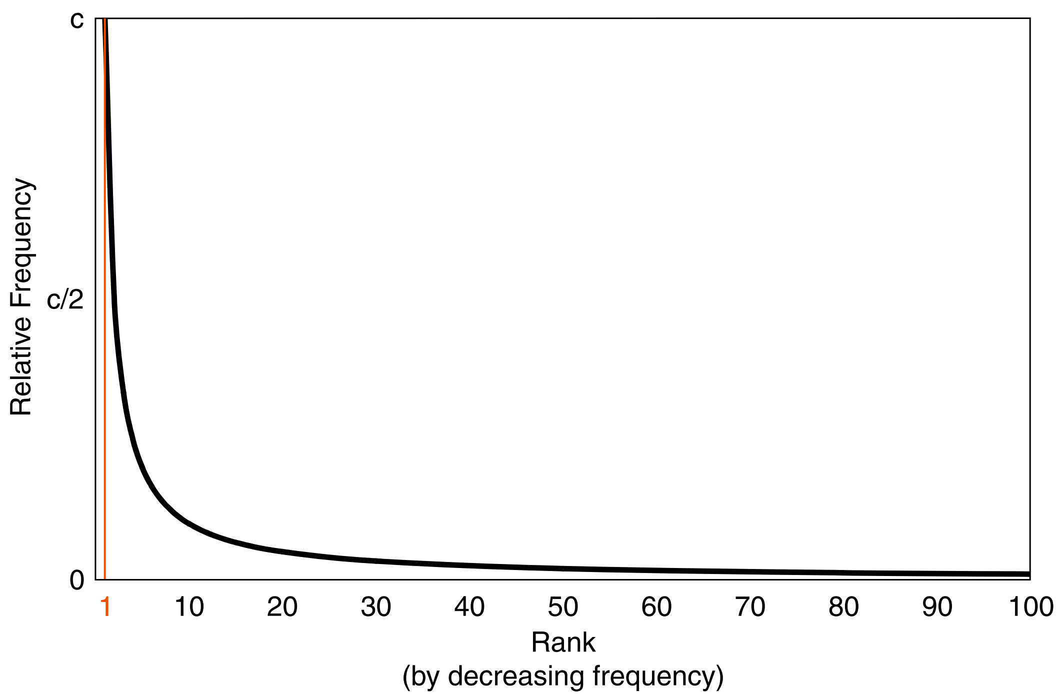
Text Statistics

Term Frequency: Zipf's Law (Fortsetzung)

The relative frequency $P(w)$ of a word w in a sufficiently large text (collection) inversely correlates with its frequency rank $r(w)$ in a power law:

$$P(w) = \frac{c}{(r(w))^a} \qquad \Leftrightarrow \qquad P(w) \cdot r(w)^a = c,$$

where c is a constant and the exponent a is language-dependent; often $a \approx 1$.



Text Statistics

Term Frequency: Zipf's Law (Fortsetzung)

Example: Top 50 most frequent words from AP89. Have a guess at c?

r	w	frequency	$P \cdot 100$	$P \cdot r$
1	the	2,420,778	6.09	0.061
2	of	1,045,733	2.63	0.053
3	to	968,882	2.44	0.073
4	a	892,429	2.25	0.090
5	and	865,644	2.18	0.109
6	in	847,825	2.13	0.128
7	said	504,593	1.27	0.089
8	for	363,865	0.92	0.073
9	that	347,072	0.87	0.079
10	was	293,027	0.74	0.074
11	on	291,947	0.73	0.081
12	he	250,919	0.63	0.076
13	is	245,843	0.62	0.080
14	with	223,846	0.56	0.079
15	at	210,064	0.53	0.079
16	by	209,586	0.53	0.084
17	it	195,621	0.49	0.084
18	from	189,451	0.48	0.086
19	as	181,714	0.46	0.087
20	be	157,300	0.40	0.079
21	were	153,913	0.39	0.081
22	an	152,576	0.38	0.084
23	have	149,749	0.38	0.087
24	his	142,285	0.36	0.086
25	but	140,880	0.35	0.089

r	w	frequency	$P \cdot 100$	$P \cdot r$
26	has	136,007	0.34	0.089
27	are	130,322	0.33	0.089
28	not	127,493	0.32	0.090
29	who	116,364	0.29	0.085
30	they	111,024	0.28	0.084
31	its	111,021	0.28	0.087
32	had	103,943	0.26	0.084
33	will	102,949	0.26	0.085
34	would	99,503	0.25	0.085
35	about	92,983	0.23	0.082
36	i	92,005	0.23	0.083
37	been	88,786	0.22	0.083
38	this	87,286	0.22	0.083
39	their	84,638	0.21	0.083
40	new	83,449	0.21	0.084
41	or	81,796	0.21	0.084
42	which	80,385	0.20	0.085
43	we	80,245	0.20	0.087
44	more	76,388	0.19	0.085
45	after	75,165	0.19	0.085
46	us	72,045	0.18	0.083
47	percent	71,956	0.18	0.085
48	up	71,082	0.18	0.086
49	one	70,266	0.18	0.087
50	people	68,988	0.17	0.087

Example: Top 50 most frequent words from AP89. For English: $c \approx 0.1$.

r	w	frequency	$P \cdot 100$	$P \cdot r$
1	the	2,420,778	6.09	0.061
2	of	1,045,733	2.63	0.053
3	to	968,882	2.44	0.073
4	a	892,429	2.25	0.090
5	and	865,644	2.18	0.109
6	in	847,825	2.13	0.128
7	said	504,593	1.27	0.089
8	for	363,865	0.92	0.073
9	that	347,072	0.87	0.079
10	was	293,027	0.74	0.074
11	on	291,947	0.73	0.081
12	he	250,919	0.63	0.076
13	is	245,843	0.62	0.080
14	with	223,846	0.56	0.079
15	at	210,064	0.53	0.079
16	by	209,586	0.53	0.084
17	it	195,621	0.49	0.084
18	from	189,451	0.48	0.086
19	as	181,714	0.46	0.087
20	be	157,300	0.40	0.079
21	were	153,913	0.39	0.081
22	an	152,576	0.38	0.084
23	have	149,749	0.38	0.087
24	his	142,285	0.36	0.086
25	but	140,880	0.35	0.089

r	w	frequency	$P \cdot 100$	$P \cdot r$
26	has	136,007	0.34	0.089
27	are	130,322	0.33	0.089
28	not	127,493	0.32	0.090
29	who	116,364	0.29	0.085
30	they	111,024	0.28	0.084
31	its	111,021	0.28	0.087
32	had	103,943	0.26	0.084
33	will	102,949	0.26	0.085
34	would	99,503	0.25	0.085
35	about	92,983	0.23	0.082
36	i	92,005	0.23	0.083
37	been	88,786	0.22	0.083
38	this	87,286	0.22	0.083
39	their	84,638	0.21	0.083
40	new	83,449	0.21	0.084
41	or	81,796	0.21	0.084
42	which	80,385	0.20	0.085
43	we	80,245	0.20	0.087
44	more	76,388	0.19	0.085
45	after	75,165	0.19	0.085
46	us	72,045	0.18	0.083
47	percent	71,956	0.18	0.085
48	up	71,082	0.18	0.086
49	one	70,266	0.18	0.087
50	people	68,988	0.17	0.087

Remarks:

- Collection statistics for AP89:

Total documents	84,678
Total word occurrences	39,749,179
Vocabulary size	198,763
Words occurring > 1000 times	4,169
Words occurring once	70,064

Text Statistics

Term Frequency: Zipf's Law (Fortsetzung)

For relative frequencies, c can be estimated as follows:

$$1 = \sum_{i=1}^{n} P(w_i) = \sum_{i=1}^{n} \frac{c}{r(w_i)} = c \sum_{i=1}^{n} \frac{1}{r(w_i)} = c \cdot H_t, \quad \rightsquigarrow \quad c = \frac{1}{H_t} \approx \frac{1}{\ln(t)}$$

where t is the size $|V|$ of the vocabulary V, and H_n is the n-th harmonic number. Constant c is text-independent, but language-dependent.

Thus, the expected average number of occurrences of a word w in a document d of length m is

$$m \cdot P(w),$$

since $P(w)$ can be interpreted as a term occurrence probability.

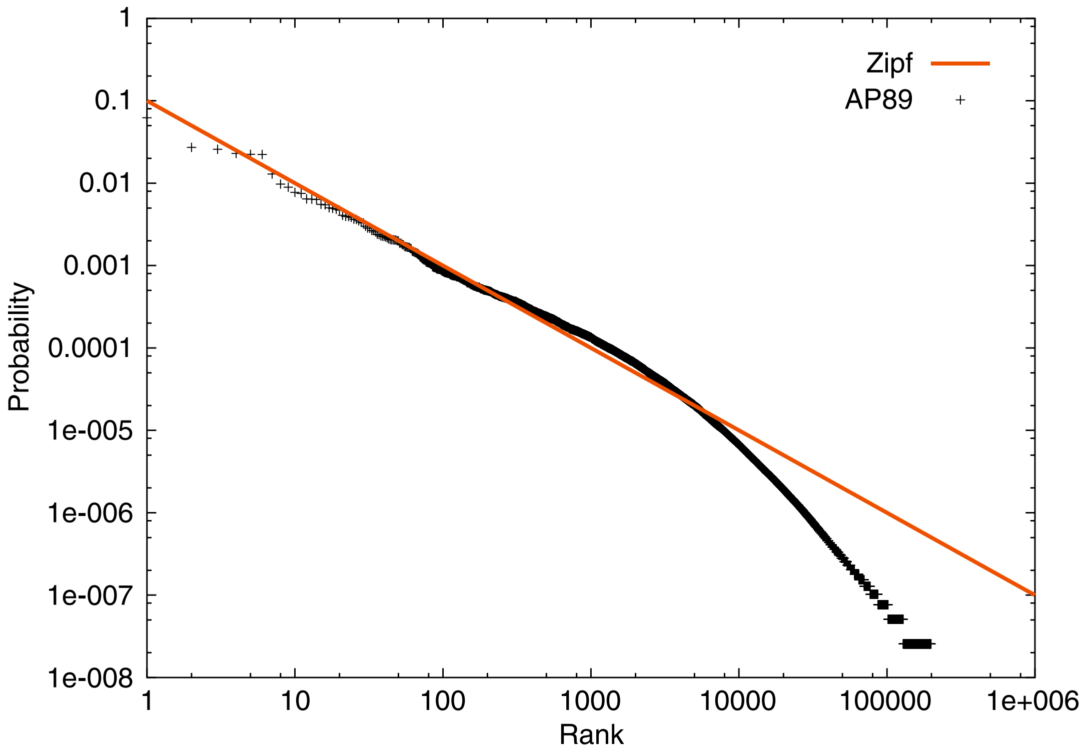
Text Statistics

Term Frequency: Zipf's Law (Fortsetzung)

By logarithmization a linear form is obtained, yielding a straight line in a plot:

$$\log(P(w)) \;=\; \log(c) - a \cdot \log(r(w))$$

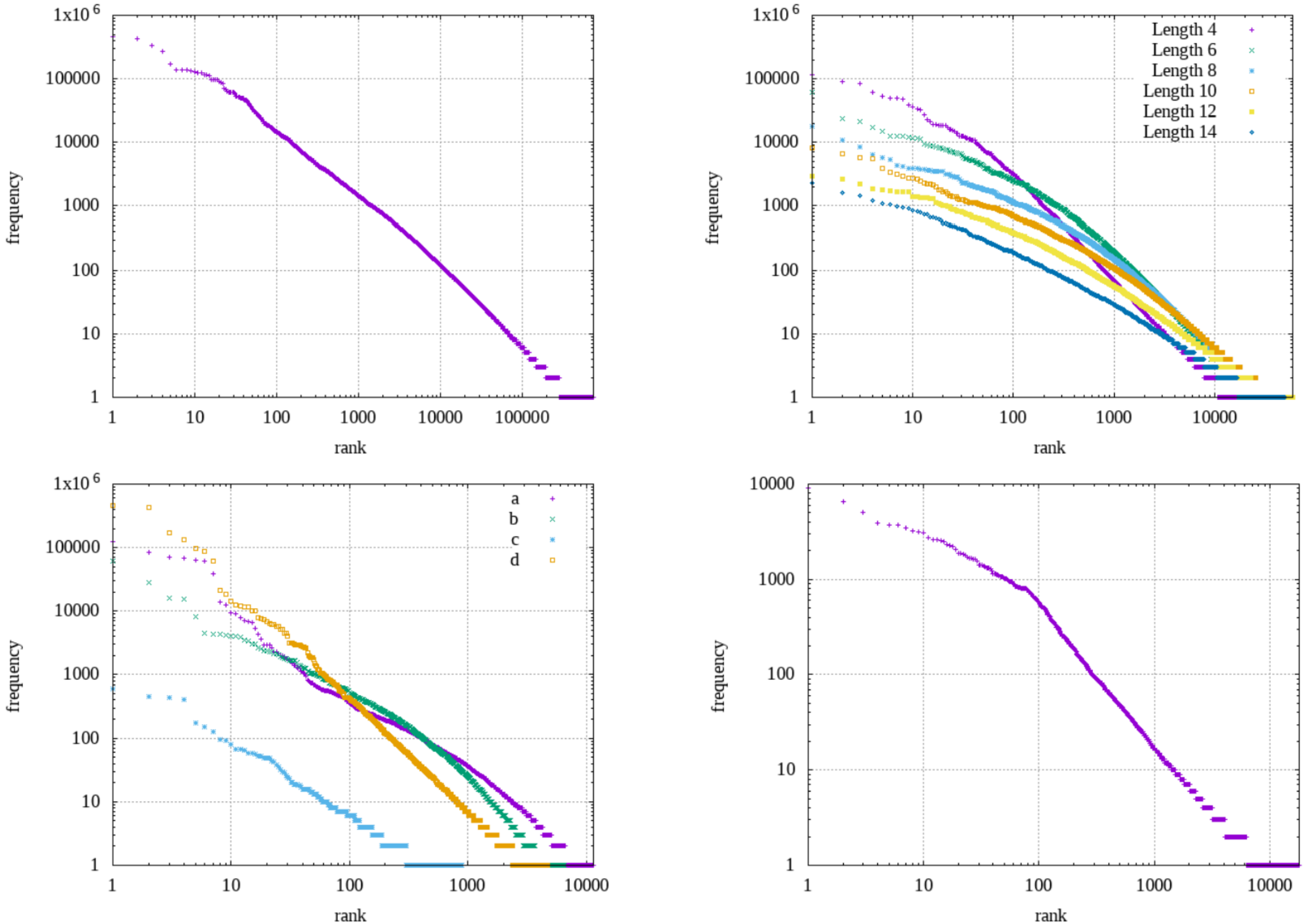
Example for AP89:



Text Statistics

Term Frequency: Zipf's Law (Fortsetzung)

Some variations (based on Wortschatz Leipzig corpus *deu_news_2022_1M*)



Remarks:

- As with all empirical laws, Zipf's law holds only approximately. While mid-range ranks of the frequency distribution fit quite well, this is less so for the lowest ranks and very high ranks (i.e., very infrequent words). The Zipf-Mandelbrot law is an extension of Zipf's law that provides for a better fit.

$$n \approx \frac{1}{(r(w) + c_1)^{1+c_2}}$$

- Interestingly, this relation cannot only be observed for words and letters in human language texts or music score sheets, but for all kinds of natural symbol sequences (e.g., DNA). It is also true for randomly generated character sequences where one character is assigned the role of a blank space. [Li 1992]

- Independently of Zipf's law, a special case is Benford's law, which governs the frequency distribution of first digits in a number.

Text Statistics

Term Frequency: Zipf's Law (Fortsetzung)

For the vocabulary, t (types) is as large as the largest rank of the frequency-sorted list. For words with frequency 1:

$$P(w) = \frac{n_w}{N}, \quad t = r(n_w = 1) = c \times \frac{N}{1} = c \times N$$

Proportion of word forms that occur only n time. For $\mathbf{w}_n$ applies:

$$\mathbf{w}_n = r(n_w) - r(n_w + 1) = c \times \frac{N}{n} - c \times \frac{N}{n+1} = \frac{c \times N}{n(n+1)} = \frac{t}{n(n+1)}$$

For $\mathbf{w}_1$ applies in particular:

$$\mathbf{w}_1 = \frac{t}{2}$$

Half of the vocabulary in a text probably occurs only 1 time.

Text Statistics

Term Frequency: Zipf's Law (Fortsetzung)

The ratio of words with a given absolute frequency n can be estimated by

$$\frac{\mathbf{w}_n}{t} = \frac{\frac{t}{n(n+1)}}{t} = \frac{1}{n(n+1)}$$

Observations:

- Estimations are fairly accurate for small x.
- Roughly half of all words can be expected to be unique.

Applications:

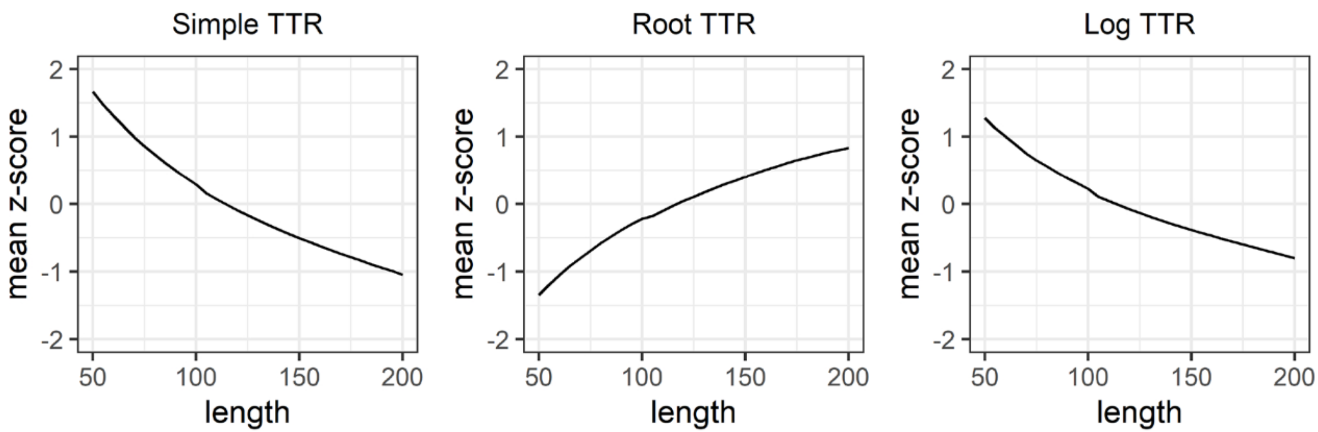
- Estimation of the number of word forms that occur n times in the text.
- Estimation of vocabulary size
- Estimation of vocabulary growth as text volume increases
- Analysis of search queries
- Term extraction (for indexing)
- Difference analysis (comparison of documents)

Text Statistics

Linguistic Diversity (LD)

Type-Token-Ratio is a measure of **LD**. Remember the formula:

$$TTR = \frac{t}{N}$$

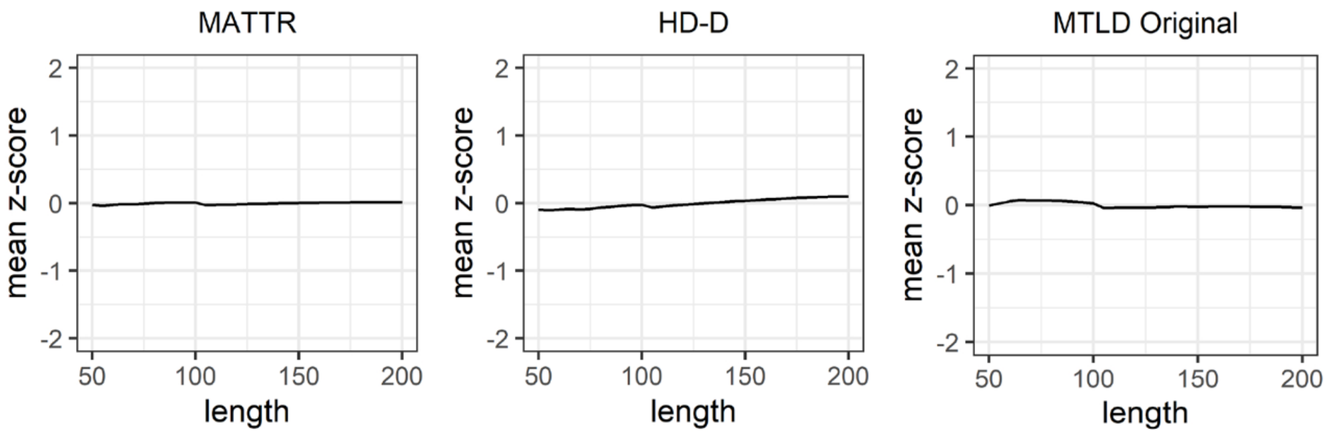


- Observation of [Zenker, Kyle 2021]: TTR is not independent of the corpus size. Thus, different sized corpora **cannot be compared**!
- Note: **z-score** is the number of standard deviations by which the value of a raw score (i.e., an observed value or data point) is above or below the mean value of what is being observed or measured

Text Statistics

Linguistic Diversity (LD) (Fortsetzung)

[Zenker, Kyle 2021] compared different measures for LD and identified 3 length-independent canditates.



- **MATTR:** Moving-average TTR calculates the moving average for all segments of a given length. For a segment length of 50 tokens, TTR is calculated on tokens 1–50, 2–51, 3–52, etc., and the resulting TTR measurements are averaged to produce the final MATTR value.
- **HD-D:** The hypergeometric distribution diversity index calculates for each wordtype, using the hypergeometric distribution, the probability of encountering one of its tokens in a random sample of 42 tokens. These probabilities are then added together to produce the final HD-D value for the text.
- **MTLD:** The measure of textual lexical diversity is based on the average number of tokens it takes to reach a given TTR (TTR is calculated on tokens 1–2, 1–3, . . . ,1–n, 2–3, . . .). The final MTLD value is the average of all the lengths where the TTR $<$ threshold

Chapter NLP:II

II. Corpus Linguistics

- ❑ Empirical Research
- ❑ Text Corpora
- ❑ Text Statistics
- ❑ Text Statistics in IR
- ❑ Data Acquisition

Data acquisition

Data Sources

Digitally available texts

- natively digital / born digital
- retro-digitzed

Metadata: “data about data”

- structural metadata
- descriptive metadata

“Big Data”

- 15,3 Mio .de-Domains (31.12.2012)
- 1.9 Mio articles in F.A.Z. Archive in 1949–2011
- 400 million Twitter tweets per day (2013)

Data acquisition

Data Sources | Newspapers

archive of political public sphere, societal knowledge or public discourse

Properties

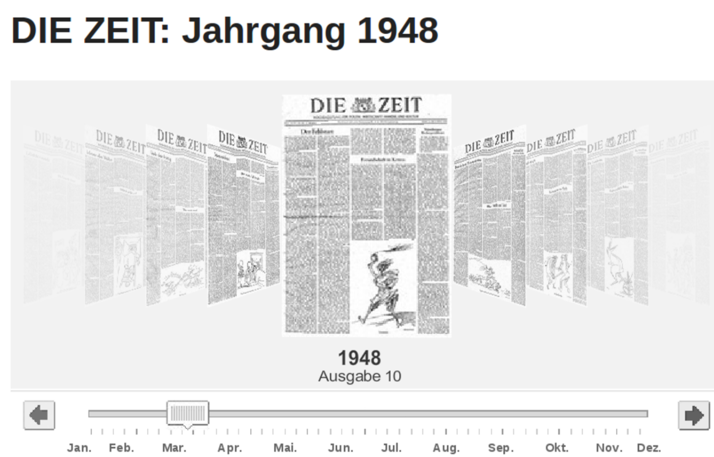
- representativity (?)
- availability improves

Difficulties

- licences
- bad OCR

Example: DIE ZEIT

- `http://www.zeit.de/archiv`
- articles since 1946
- 400.000 articles
- PDF + OCR-ed Text



Date	← 1948-05-12	Date
Author(s)	← {„GH“, „geh“, „Gerda Heller“}	String[]
Page number	← {1, 1-2}	Integer
Section(s)	← {„Sport“, „Leibesübungen“}	String[]
Subsection(s)	← „Handball“	String
News agency	← {true|false; „dpa“}	Boolean

Data acquisition

Data Sources | Blogs & Forums

Extract of (political) public discourse

Properties

- expert generated content
- user generated content (comments)

Properties

- high availability
- lesser license restrictions
- no OCR problems

Difficulties

- identifying relevant content
- representativity of content?
- Crawling + Web scraping

Date ← 2012-11-12 21:40
Author(s) ← {„E. F.“}
Url ← {„http://www.blogactiv.eu/blog/31/123“}
PolicyField ← „Agriculture“
numberOfComments ← 216
numberOfReadings ← 12002

Data acquisition

Data Sources | Social network

controlled public spheres

Properties

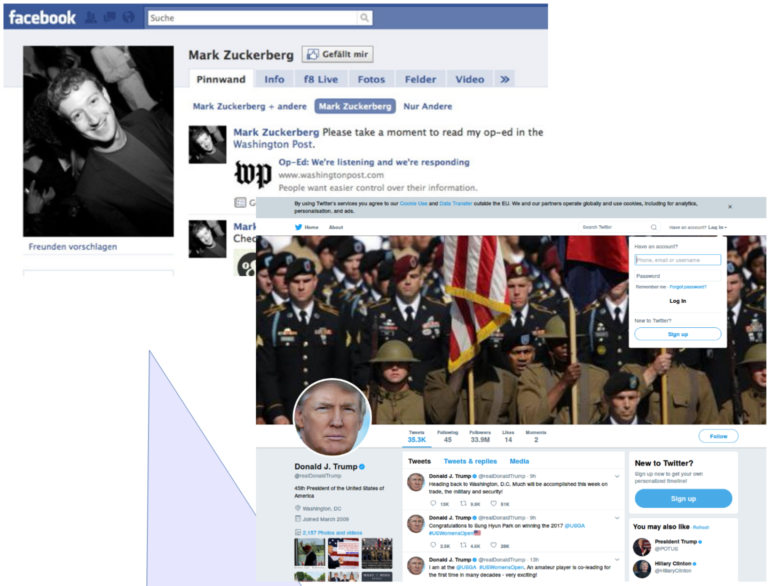
- just in time
- really big data

Difficulties

- very short text snippets
- typos and special language
- representativity?
- Data acquisition may be complicated

Data acquisition via APIs

- Twitter sample API (1%)
- Twitter keyword location search
- Facebook API: retrieve user networks and (public) posts, comments, replies from users



Type ← {post, comment, reply, tweet}
Datetime ← 2014-05-12 12:47
Author ← User_462945
Reactions ← {like:67, angry:472, sad:12}

Data acquisition

Data Sources | Other sources

- Emails
- Parliamentary protocols
- Political documents
  - political speeches
  - party manifestos
  - press releases
- Open questions from (online) surveys
- Literature: distant reading of (world) literature
- Scientific publications: lots of science of science studies

Web crawling and scraping

Crawling data

Crawling = massive automated download of data from the web

Uninformed Approach

- define initial list of URLs (seed URLs)
- download web page from seed URLs
- expand list of URLs by analyzing page source (extracting new URLs)

Informed Approach

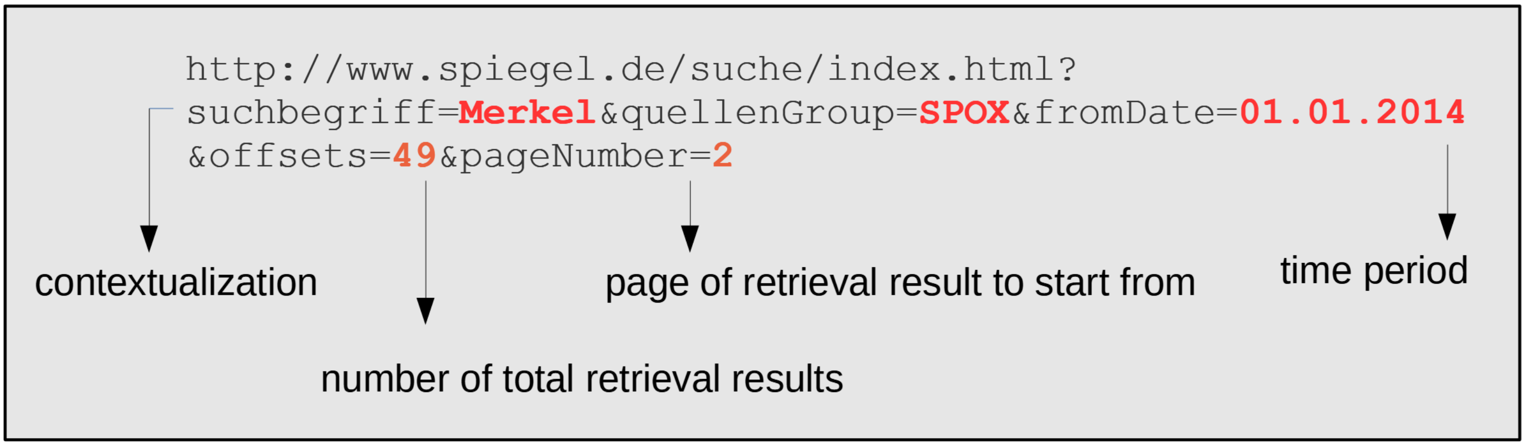
- generate list of URLs by patterns, e.g.

```
http://www.spiegel.de/suche/index.html?
suchbegriff=Merkel&quellenGroup=SPOX&fromDate=01.01.2014
```

Web crawling and scraping

Crawling data

Web crawling from specific sources: Using site internal search e.g. search for “Merkel” on: http://www.spiegel.de/suche



spiegel.de displays 20 results per page → ceil(49 / 20) = 3 result pages

generating a complete list of URLs to retrieve retrieval result links:

```
spiegel.de/suche/index.html?suchbegriff=Merkel&fromDate=01.01.2014&offsets=49&pageNumber=1
spiegel.de/suche/index.html?suchbegriff=Merkel&fromDate=01.01.2014&offsets=49&pageNumber=2
spiegel.de/suche/index.html?suchbegriff=Merkel&fromDate=01.01.2014&offsets=49&pageNumber=3
```

Todo: Download HTML from result page URLs and extract all 49 links to article web pages

Web crawling and scraping

Scraping Data

Scraping

- Extraction of relevant content from downloaded XHTML source
- Page source = XHTML tree where elements have item class identifiers which can be used for content selection

Steps

- (manual) investigation of page source → “inspect element” function in your browser!
- identify relevant elements, e.g.
  - <time>Montag, den 13.04.2014</time>
  - <h2 class="article-title">...</h2>
  - <div class="userCommentBody">...</div>

Web crawling and scraping

File type: XML

XML = Extensible Markup Language

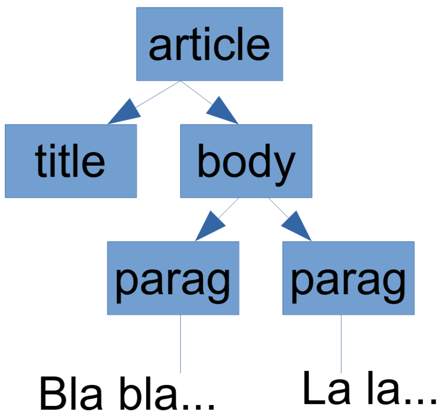
(semi-)structured data in hierarchy

- exactly one root element
- nested child elements in “tree” structure

nested tags:

- elements: opening/closing tags
- attributes: properties of elements
- text: unstructured data entity

```
<article date='2012-03-13'>
  <title>my title</title>
  <body>
    <paragraph>Bla bla ...
    </paragraph>
    <paragraph>La la …
    </paragraph>
  </body>
</article>
```



Web crawling and scraping

File type: XHTML

XHTML

- concrete defined schema on XML to represent web content
- special elements:
  - <html/>, <head/>, <body/>, <p/>, <h1/>, <form/>, <div/> ...
- special attributes:
  - id: e.g. <div **id="main-content"**>...</div>
  - class: e.g. <h2 **class="article-title"**>...</h2>
- Can be processed like any other XML document

Application scenario

- Download all Webpages containing “TTIP” from website www.nytimes.com (→ XHTML files)
- Extract date and headline from XHTML files via XPath
- Write URL, date, headline into CSV file for further analysis

Web crawling and scraping

XPath [detailed in WT:III-311 ff.]

XPath = query language to select items in XML trees by:

- element position
- element name
- attribute value

Examples

- `//`
- `//h2[@class='article-title']`
- `//p`
- `//div[@id='main-content']/p`
- `//div[@class='article-date']`

```
<html>
<head>...</head>
<body>
  <div
class="article-date">2017-04-14</div>
  <div id="main-content">
    <h2 class="article-title">...</h2>
    <p>...</p>
    <p>...</p>
  </div>
  <h2 class="comment-title">...</h2>
  <p>...</p>
  <p class="ad">Buy watches!</p>
</body>
</html>
```

Web crawling – Example

Web Crawl 2021 (for German text)

- Project context: Wortschatz Leipzig
- Used crawler: Heritrix (Internet Archive)

	Raw			After cleaning*	
LCC-DE-2021 subcorpora	**Sources (URLs)**	**Chars**	**Tokens**	**Sources (URLs)**	**Tokens**
deu-de_web_2021	317.400.000	1,4T	182.800.000.000	280.139.875	75.778.591.663
deu-at_web_2021	268.000.000	1,1T	140.400.000.000	225.714.422	45.324.844.491
deu-com_web_2021	1.560.000	7,5G	966.000.000	1.500.000	375.000.000
deu-eu_web_2021	5.300.000	52,2G	7.400.000.000	3.500.000	1.300.000.000
deu-lu_web_2021	3.600.000	12,9G	1.700.000.000	3.000.000	569.000.000
deu-hu_web_2021	400.000	1,1G	151.000.000	328.000	60.200.000
LCC-DE-2021 TOTAL	**596.260.000**	**3,02T**	**333.417.000.000**	**514.000.000**	**123.408.000.000**