Chapter NLP:II

II. Corpus Linguistics

- ❑ Empirical Research
- ❑ Hypothesis Testing
- ❑ Text Corpora
- ❑ Data Acquisition
- ❑ Data Annotation

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

Text as Data

Bits: A sequence of bits that symbolize text when decoded into glyphs [cf WT:II-166 ff.]
String: concatenation of glyphs (alphabet elements)

- “Hello world!”, “”, “0001011110001010 1”, “To be or not to be...”
- essential, elementary data type in computer linguistics
- common operations: e.g.
  - concatenation: “Hello” + “World!” + “!” → “Hello World!”
  - splitting: split(“Hello World!”, “ ”) → {“Hello”, “World!”}
  - case conversion: uppercase(“Hello”) → “HELLO”
  - substring: substr(“Hello”, start = 0, length = 4) → “Hell”

Document: compound data type

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

Corpus: collection of documents

Text Corpora

Text as Data (continued)

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

Metadata

Metadata = text external context / covariate

Metadata = data facet

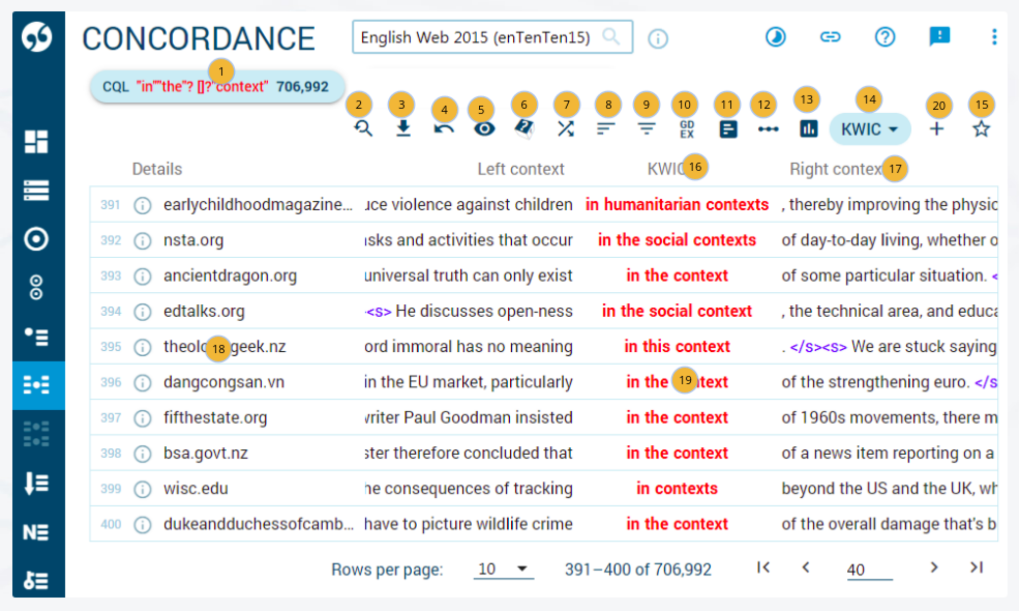
- ❑ Subselections of sources
- ❑ Aggregation / differentiation of results

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



CQL "in""the"? []?"context" 706,992

		Details	Left context	KWIC	Right context
391	ⓘ	earlychildhoodmagazine...	ıce violence against children	in humanitarian contexts	, thereby improving the physic
392	ⓘ	nsta.org	ısks and activities that occur	in the social contexts	of day-to-day living, whether o
393	ⓘ	ancientdragon.org	universal truth can only exist	in the context	of some particular situation. <
394	ⓘ	edtalks.org	<s> He discusses open-ness	in the social context	, the technical area, and educa
395	ⓘ	theolc geek.nz	ord immoral has no meaning	in this context	. </s><s> We are stuck saying
396	ⓘ	dangcongsan.vn	in the EU market, particularly	in the text	of the strengthening euro. </s
397	ⓘ	fifthestate.org	vriter Paul Goodman insisted	in the context	of 1960s movements, there m
398	ⓘ	bsa.govt.nz	ster therefore concluded that	in the context	of a news item reporting on a
399	ⓘ	wisc.edu	he consequences of tracking	in contexts	beyond the US and the UK, wh
400	ⓘ	dukeandduchessofcamb...	have to picture wildlife crime	in the context	of the overall damage that's b

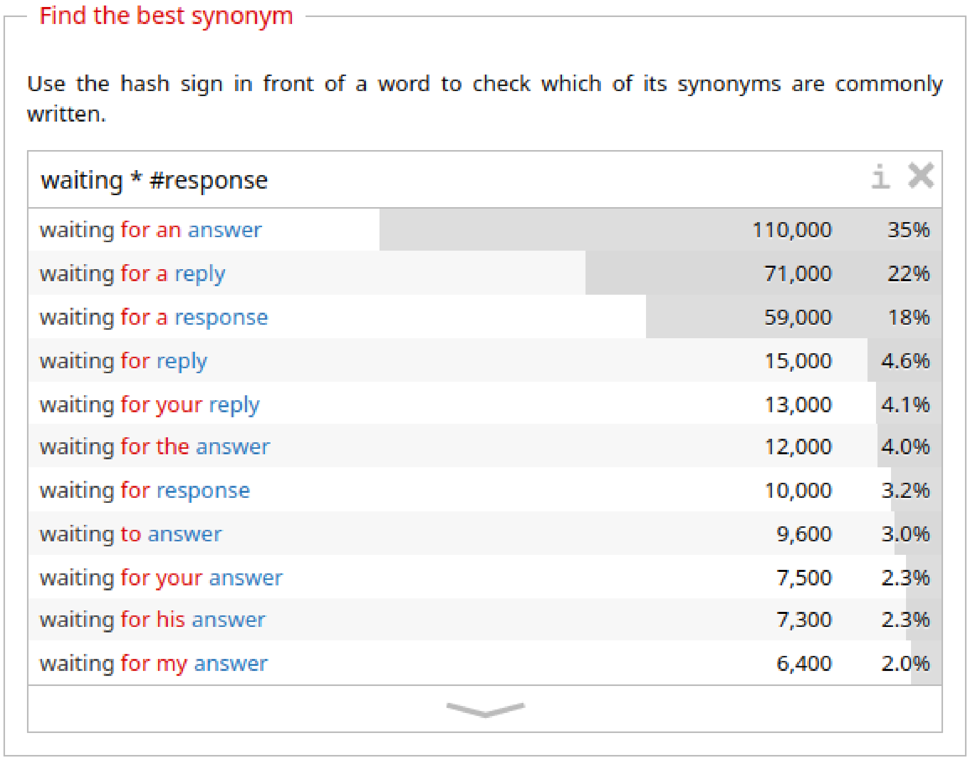
Rows per page: 10 ▾ 391–400 of 706,992 |< < 40 > >|

[www.sketchengine.eu]

Text Corpora

Research in Language Use (continued)

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]

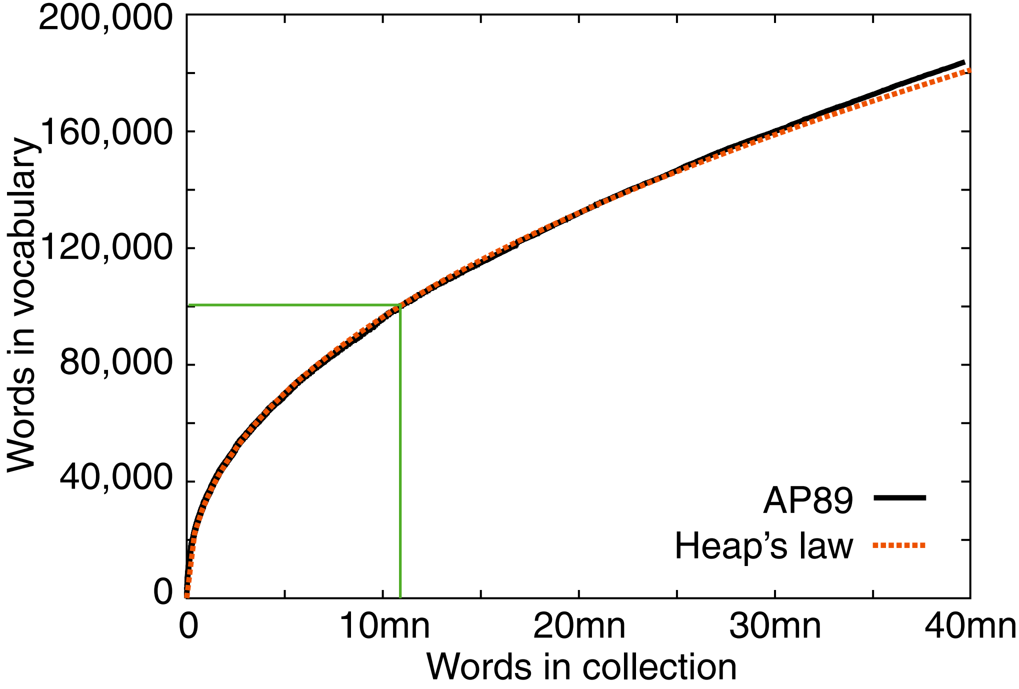
Text Corpora

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

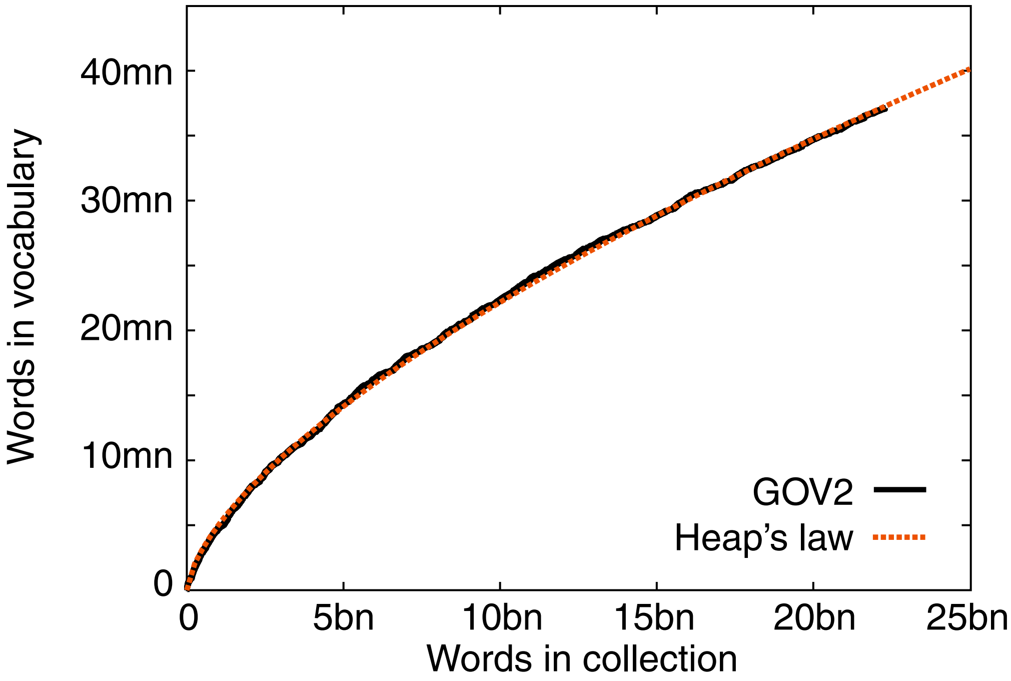
Text Corpora

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 Corpora

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.

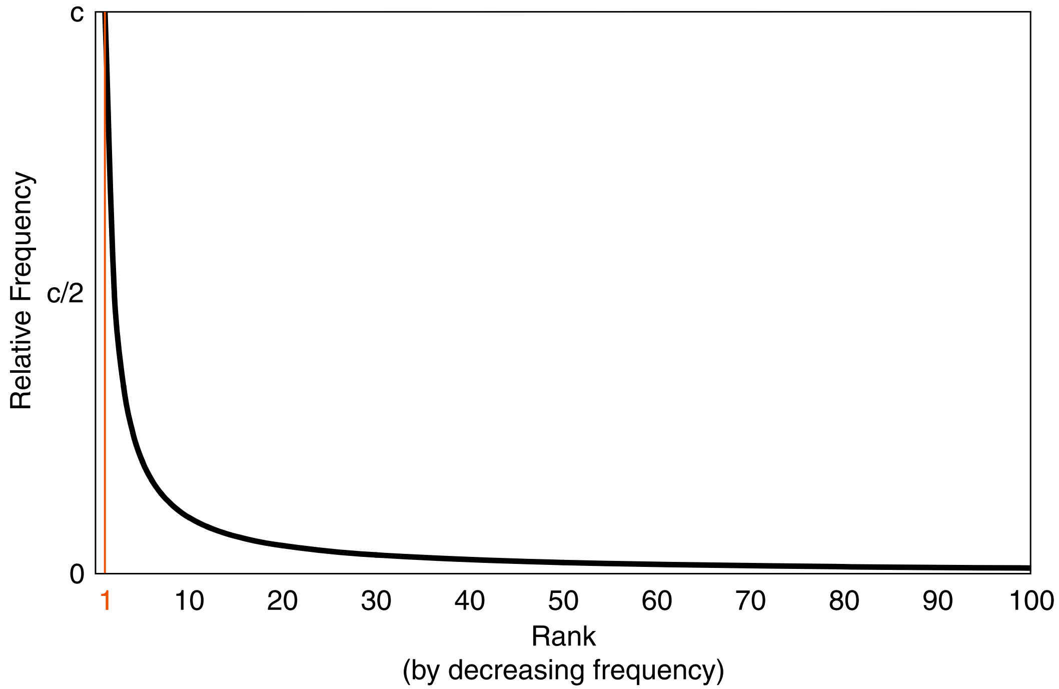
Text Corpora

Term Frequency: Zipf's Law (continued)

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 Corpora

Term Frequency: Zipf's Law (continued)

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
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

Text Corpora

Term Frequency: Zipf's Law (continued)

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 Corpora

Term Frequency: Zipf's Law (continued)

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 language-dependent; e.g., for German $c = 1/ln(7.841.459) \approx 0.063$. [Wortschatz Leipzig]

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.

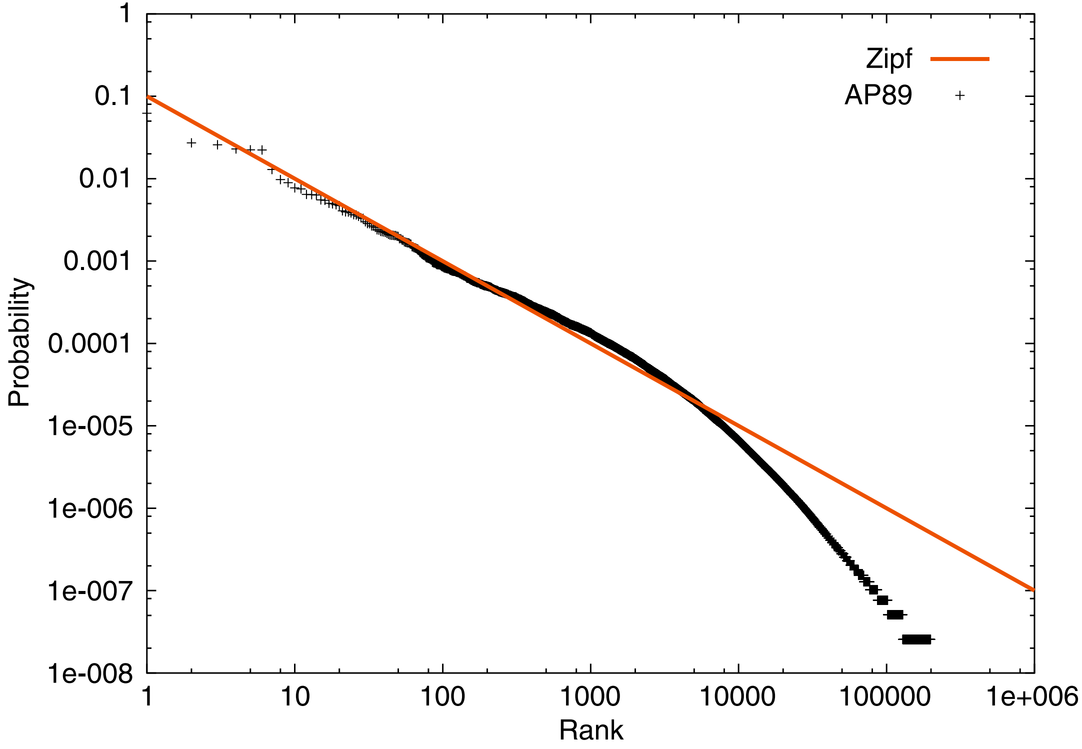
Text Corpora

Term Frequency: Zipf's Law (continued)

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:



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 Corpora

Term Frequency: Zipf's Law (continued)

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

$$P(w) = \frac{n_w}{N}, \quad t = r(n_w = 1) = c \times \frac{N}{1} = c \times N \approx \; e^{1/c}$$

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 Corpora

Term Frequency: Zipf's Law (continued)

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 Corpora

n-grams

Word n-grams represent text as overlapping n-length subsequences.

- For a sequence of $m \geq n$ tokens, the number of n-grams is $(m - n) + 1$.

Example: `the quick brown fox jumps over the lazy dog`

- 1-grams: `the`, `quick`, `brown`, `fox`, . . . , `dog` For English: $c \approx 0.1$.

Text Corpora

n-grams

Word n-grams represent text as overlapping n-length subsequences.

- For a sequence of $m \geq n$ tokens, the number of n-grams is $(m - n) + 1$.

Example: `the quick brown fox jumps over the lazy dog`

- 1-grams: `the`, `quick`, `brown`, `fox`, ..., `dog`
- 2-grams: `The quick`, `quick brown`, `brown fox`, ..., `lazy dog` For English: $c \approx 0.1$.

Text Corpora

n-grams

Word n-grams represent text as overlapping n-length subsequences.

- For a sequence of $m \geq n$ tokens, the number of n-grams is $(m - n) + 1$.

Example: `the quick brown fox jumps over the lazy dog`

- 1-grams: `the`, `quick`, `brown`, `fox`, ..., `dog`
- 2-grams: `The quick`, `quick brown`, `brown fox`, ..., `lazy dog`
- 3-grams: `The quick brown`, `quick brown fox`, ..., `the lazy dog`

Text Corpora

n-gram Corpora

Google: “All Our N-Grams are Belong to You”

- Google Web 1T 5-gram Version 1: [LDC 2006]

Tokens	1,024,908,267,229
Sentences	95,119,665,584
Unigrams	13,588,391
Bigrams	314,843,401
Trigrams	977,069,902
Fourgrams	1,313,818,354
Fivegrams	1,176,470,663

- In general, stop word-only n-grams do not dominate on the web.
- n-grams with less than 40 occurrences are not included (200 for $n = 1$). Web search engines index n-grams for $n > 1$
- Primary use cases for n-gram frequency datasets is language modeling, i.e., training $(n-1)$-order Markov models to predict the next word in a sequence.

Text Corpora

n-gram Corpora

Example n-gram counts from *Google Web 1T*:

1-grams	**2-grams**	**3-grams**	**4-grams**	**5-grams**	**Tokens**	**Sentences**
13.6 million	314.8 million	977.1 million	1.3 billion	1.2 billion	1.0 trilion	95.1 billion

Observations:

- ❑ The most frequent 3-gram on the English web is `all rights reserved`.
- ❑ n-grams for $n \geq 1$ combined fit Zipf's law better than just words. [Williams 2015]
- ❑ Heap's law does not apply to $n > 1$; other models are required. [Silva 2016]
- ❑ A search engine's index size grows linearly with n.
- ❑ For $n > 1$, n-gram frequency reveals phrases in common use.
- ❑ Indexing n-grams speeds up search query processing (esp. for stop words).

Chapter NLP:II

II. Corpus Linguistics

- ❑ Empirical Research
- ❑ Hypothesis Testing
- ❑ Text Corpora
- ❑ Data Acquisition
- ❑ Data Annotation

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
Newspapers

archive of political public sphere, societal knowledge or public discourse

Properties

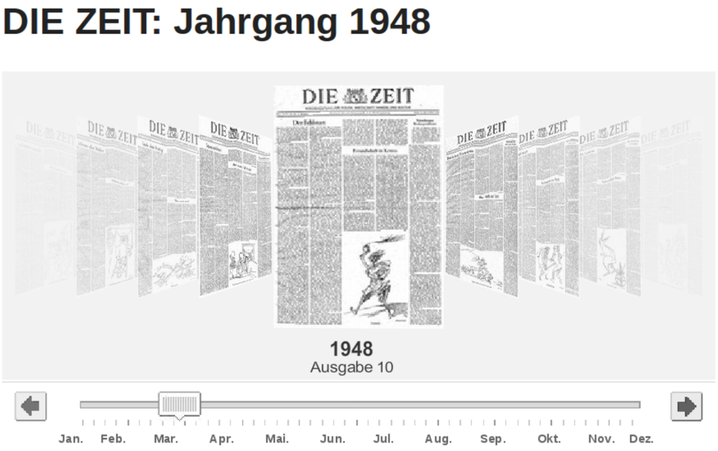
- representativity (?)
- availability improves

Difficulties

- licences
- bad OCR

Example: DIE ZEIT

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



DIE ZEIT: Jahrgang 1948

1948
Ausgabe 10

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

Blogs and Forums

Extract of (political) public discourse

Properties

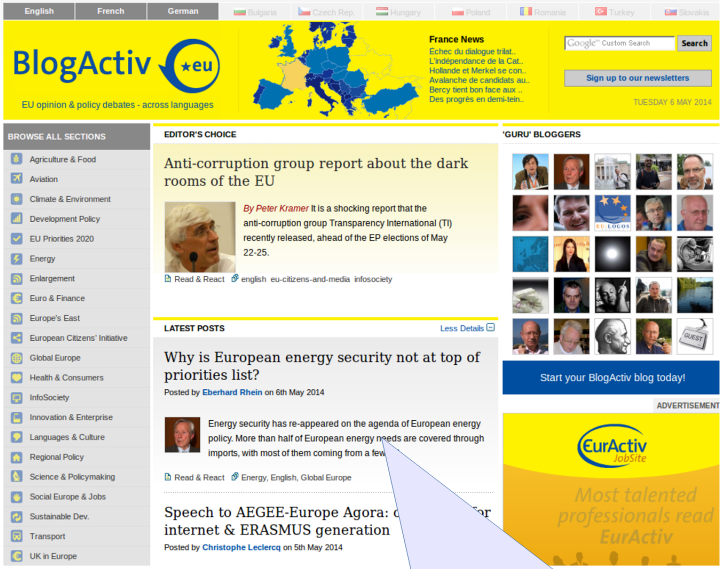
- expert generated content
- user generated content (comments)

Properties

- high availability
- lesser license restrictions
- no OCR problems

Difficulties

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



EU opinion & policy debates - across languages

France News
Échec du dialogue trilat..
L'indépendance de la Cat..
Hollande et Merkel se con..
Avalanche de candidats au..
Bercy tient bon face aux ..
Des progrès en demi-tein..

TUESDAY 6 MAY 2014

EDITOR'S CHOICE

Anti-corruption group report about the dark rooms of the EU

By Peter Kramer It is a shocking report that the anti-corruption group Transparency International (TI) recently released, ahead of the EP elections of May 22-25.

Read & React · english eu-citizens-and-media infosociety

LATEST POSTS · Less Details

Why is European energy security not at top of priorities list?

Posted by Eberhard Rhein on 6th May 2014

Energy security has re-appeared on the agenda of European energy policy. More than half of European energy needs are covered through imports, with most of them coming from a few

Read & React · Energy, English, Global Europe

Speech to AEGEE-Europe Agora: c … for internet & ERASMUS generation

Posted by Christophe Leclercq on 5th May 2014

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

Social network

controlled public spheres

Properties

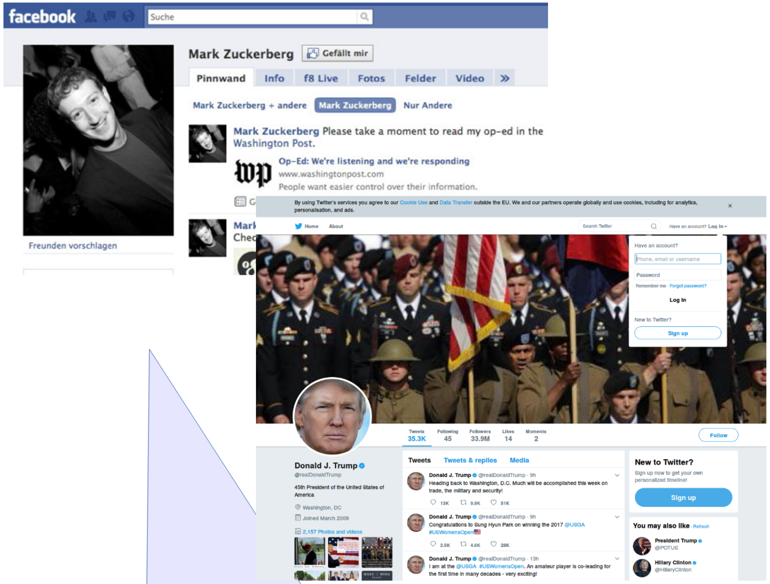
- ❑ just in time
- ❑ really big data

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



Mark Zuckerberg Please take a moment to read my op-ed in the Washington Post.

Op-Ed: We're listening and we're responding
www.washingtonpost.com
People want easier control over their information.

Donald J. Trump
@realDonaldTrump
45th President of the United States of America
Washington, DC
Joined March 2009
2,157 Photos and videos

Donald J. Trump @realDonaldTrump · 9h
Heading back to Washington, D.C. Much will be accomplished this week on trade, the military and security!

Donald J. Trump @realDonaldTrump · 9h
Congratulations to Sung Hyun Park on winning the 2017 @USGA #USWomensOpen🇺🇸

Donald J. Trump @realDonaldTrump · 13h
I am at the @USGA #USWomensOpen. An amateur player is co-leading for the first time in many decades - very exciting!

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

Data Acquisition

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