Faculty of Computer Science and Information Technology
West Pomeranian University of Technology, Szczecin

Natural Language processing

PhD. Eng. Joanna Kolodziejczyk
jkolodziejczyk@zut.edu.pl

October 13, 2021

Content

What is Natural Language Processing?

Textbook
e-book

Authors

Dan Jurafsky and James H. Martin

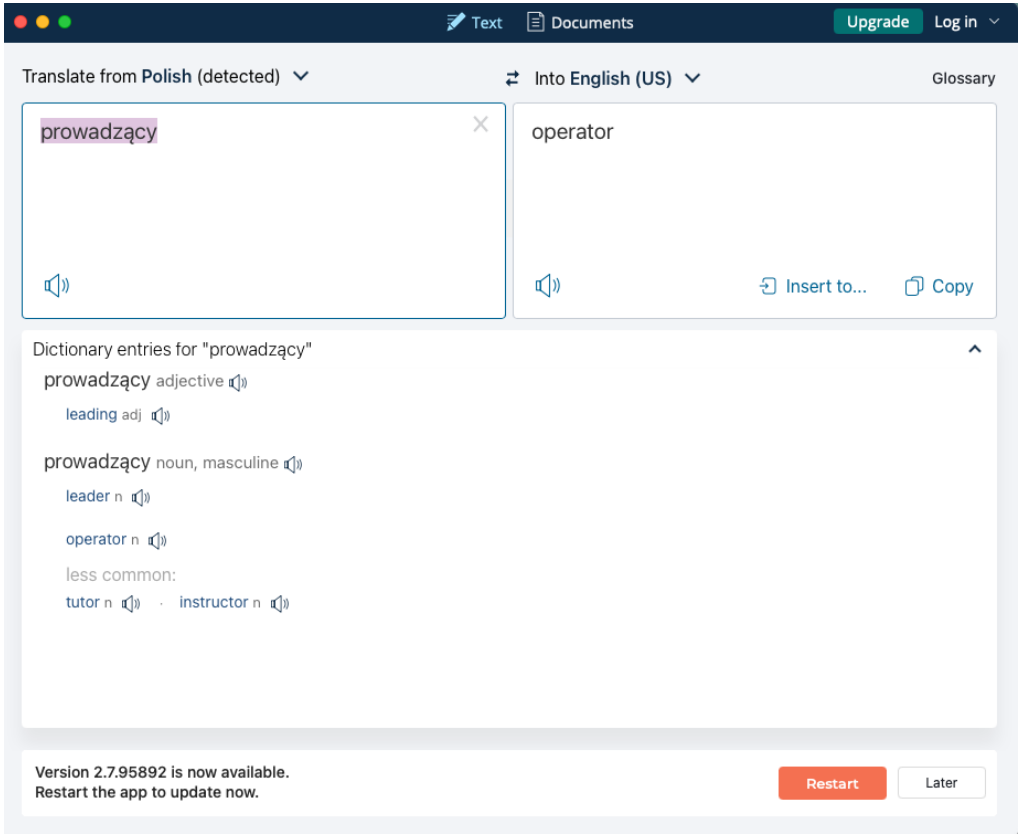
Title

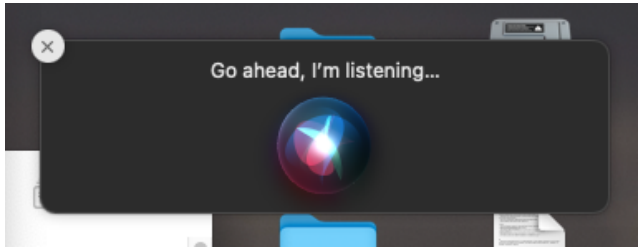
Speech and Language Processing 3rd Edition

Web sources

Newest Version: `https://web.stanford.edu/~jurafsky/slp3/ed3book_sep212021.pdf`

Examples
Translator, speech recognition, understanding

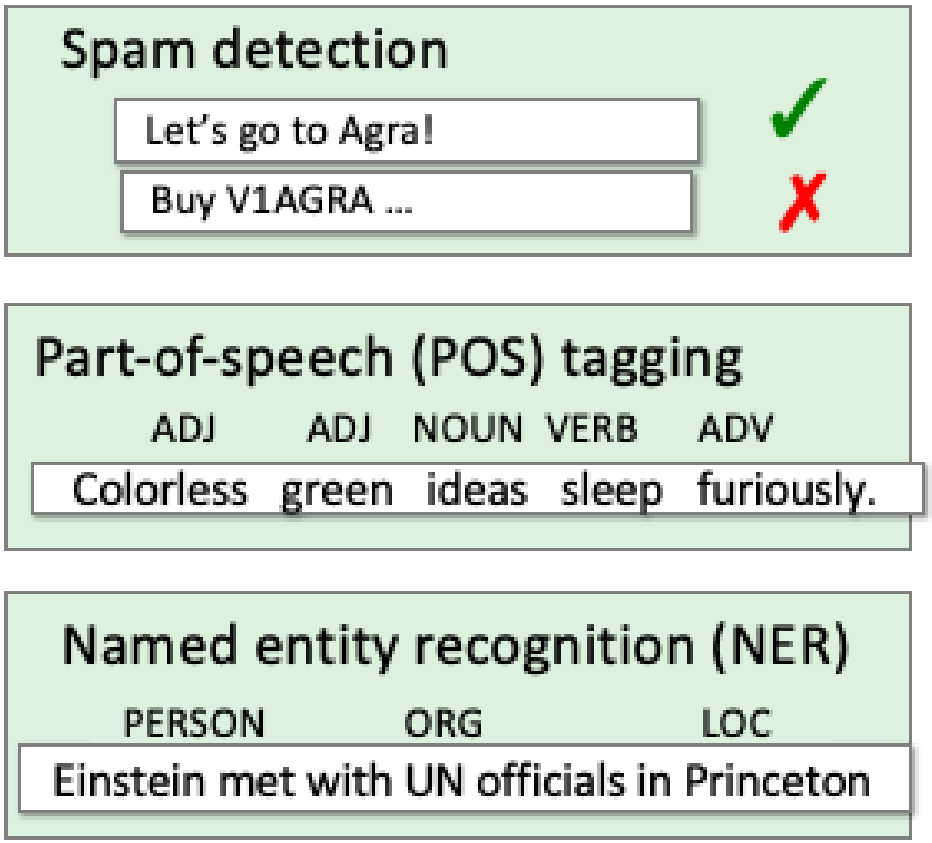




Watson `https:`

`//www.ibm.com/cloud/watson-natural-language-understanding`
`https://www.youtube.com/watch?v=lI-M7O_bRNg`

- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...


Spam detection
Let's go to Agra!
Buy V1AGRA ...
Part-of-speech (POS) tagging
ADJ ADJ NOUN VERB ADV
Colorless green ideas sleep furiously.
Named entity recognition (NER)
PERSON ORG LOC
Einstein met with UN officials in Princeton

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good

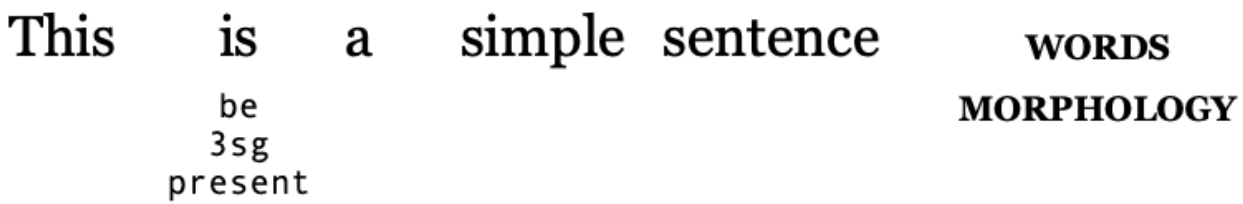
Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

- Language modelling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense disambiguation
- Semantic Role Labelling
- ...

This is a simple sentence WORDS

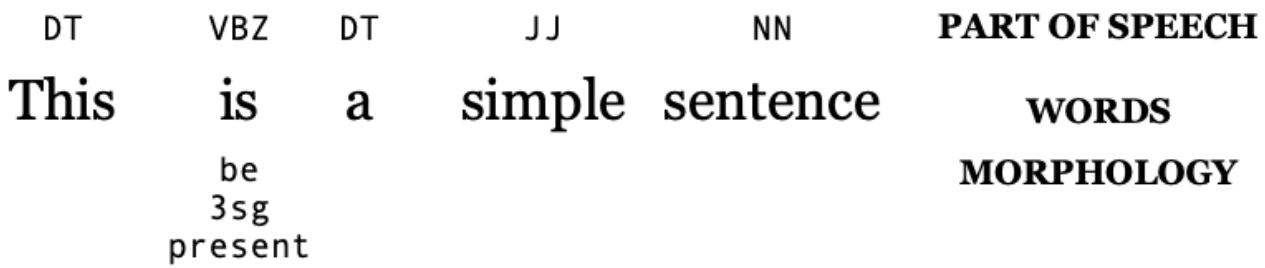

This is a simple sentence
WORDS
be
3sg
present
MORPHOLOGY

Parts of Speech
tagging

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	$
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	#
JJS	superlative adj	*wildest*	SYM	symbol	+,%, &	“	left quote	‘ or “
LS	list item marker	*1, 2, One*	TO	“to”	*to*	”	right quote	’ or ”
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

Figure 8.1 Penn Treebank part-of-speech tags (including punctuation).

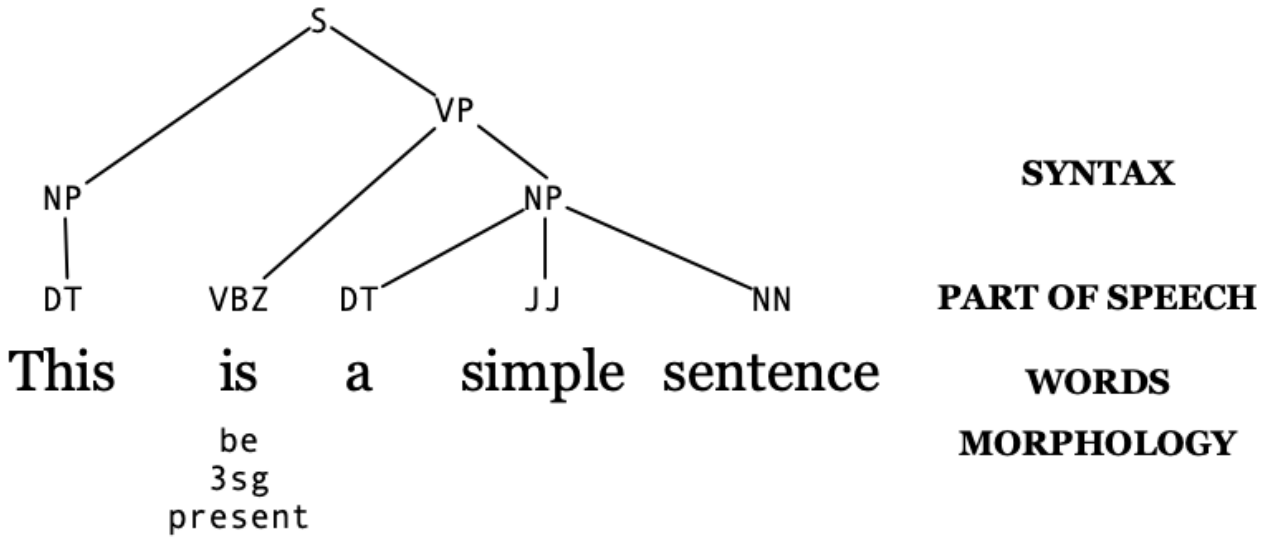
Parts of Speech
NLP

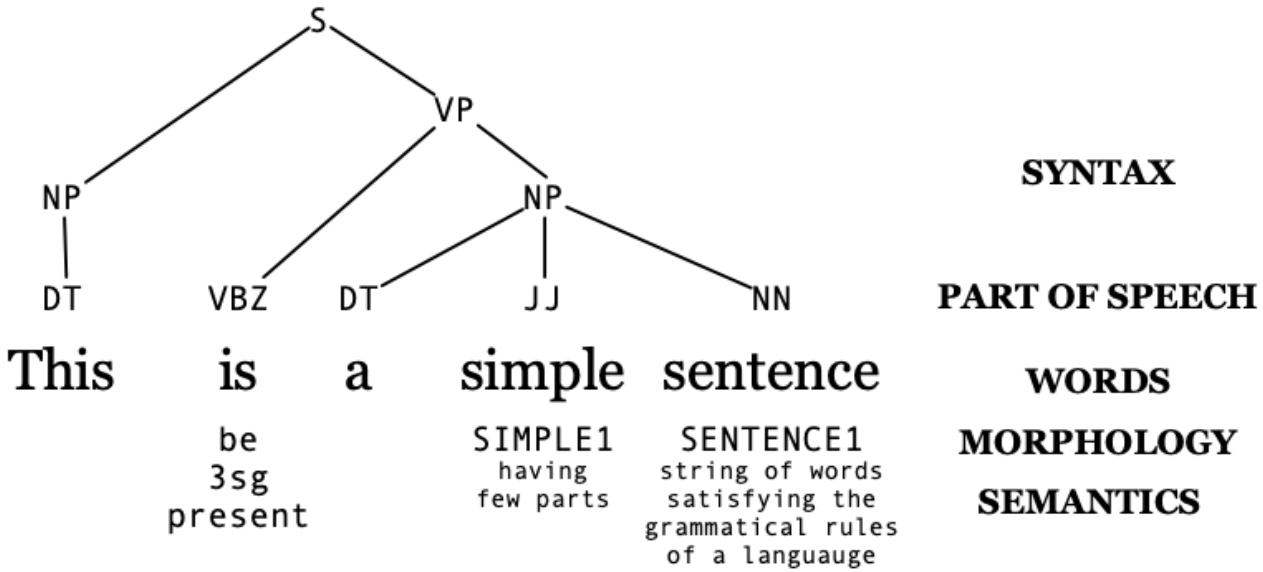


$$\mathrm{S} \longrightarrow \mathrm{NP} \quad \mathrm{VP}$$
$$\mathrm{NP} \longrightarrow \text{(Det)}\ \mathrm{N}_1$$
$$\mathrm{N}_1 \longrightarrow (\mathrm{AP}) \quad \mathrm{N}_1 \quad (\mathrm{PP})$$

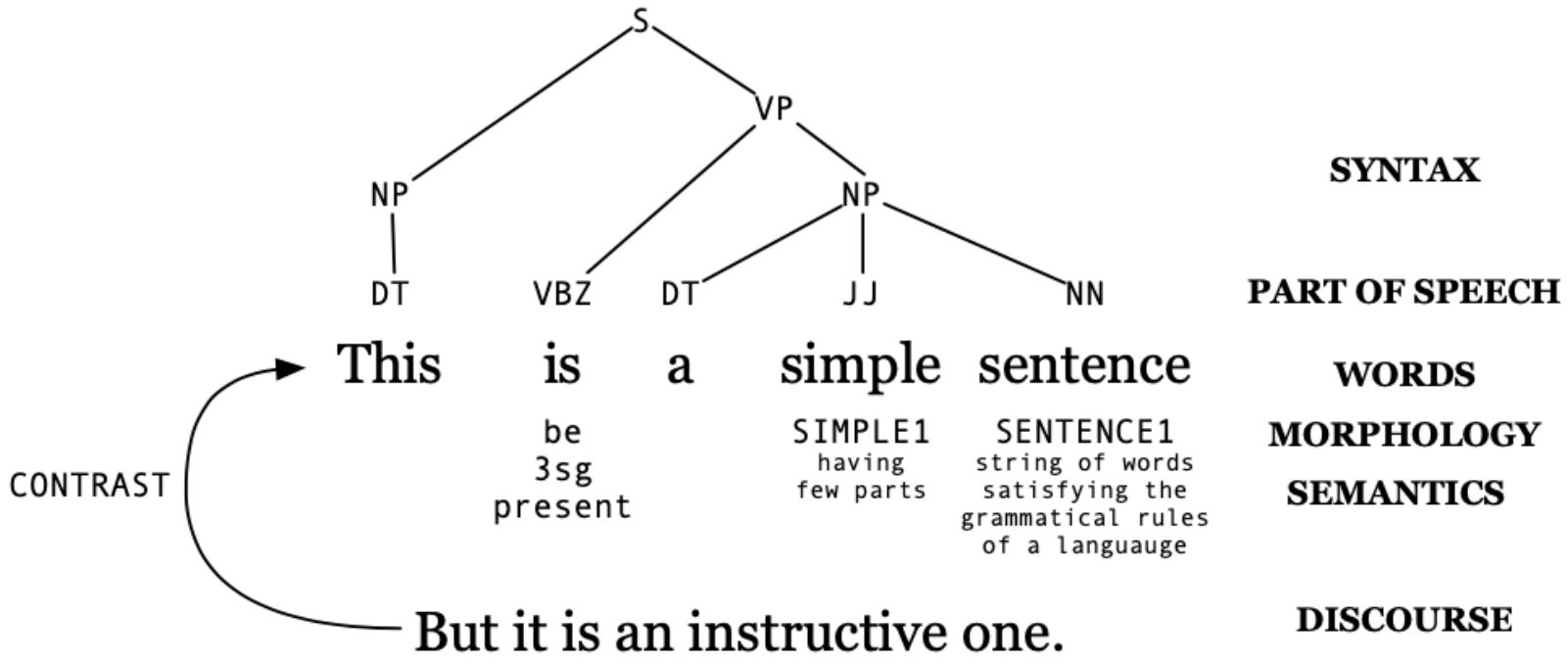
1. The first rule reads: A S (sentence) consists of a NP (noun phrase) followed by a VP (verb phrase).
2. The second rule reads: A noun phrase consists of an optional Det (determiner) followed by a N (noun).
3. The third rule means that a N (noun) can be preceded by an optional AP (adjective phrase) and followed by an optional PP (prepositional phrase). The round brackets indicate optional constituents.r

Syntax


S
VP
NP
NP
DT
VBZ
DT
JJ
NN
This
is
a
simple
sentence
be
3sg
present
SYNTAX
PART OF SPEECH
WORDS
MORPHOLOGY



$$\exists y(\textit{this\_dem}(x) \wedge \textit{be}(e, x, y) \wedge \textit{simple}(y) \wedge \textit{sentence}(y))$$


S
VP
NP
NP
DT
VBZ
DT
JJ
NN
This
is
a
simple
sentence
be
3sg
present
SIMPLE1
having
few parts
SENTENCE1
string of words
satisfying the
grammatical rules
of a languauge
CONTRAST
But it is an instructive one.
SYNTAX
PART OF SPEECH
WORDS
MORPHOLOGY
SEMANTICS
DISCOURSE

Ambiguity
Why NLP is hard?

- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Quantifier scope: Every child loves some movie
- Multiple: I saw her duck
- Reference: John dropped the goblet onto the glass table and it broke.

Ambiguity
Why NLP is hard?

Methods of dealing with ambiguity.

- non-probabilistic methods (FSMs for morphology, CKY parsers for syntax) return all possible analyses.
- probabilistic models (HMMs for POS tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return the best possible analysis, i.e., the most probable one according to the model.

This ”best” analysis is only good if our model’s probabilities are accurate. Where do they come from?

Statistical NLP
Why NLP is hard?

Like most other parts of AI, NLP today is dominated by statistical methods.

- Typically more robust than earlier rule-based methods.
- Relevant statistics/probabilities are learned from data.
- Normally requires lots of data about any particular phenomenon.

Sparse data due to Zipf's Law

Why NLP is hard?

- To illustrate, let's look at the frequencies of different words in a large text corpus.
- Relevant statistics/probabilities are learned from data.
- Assume a "word" is a string of letters separated by spaces (a great oversimplification, we'll return to this issue...)

Most frequent words (word types) in the English Europarl corpus (out of 24m word tokens)

any word	
Frequency	Type
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

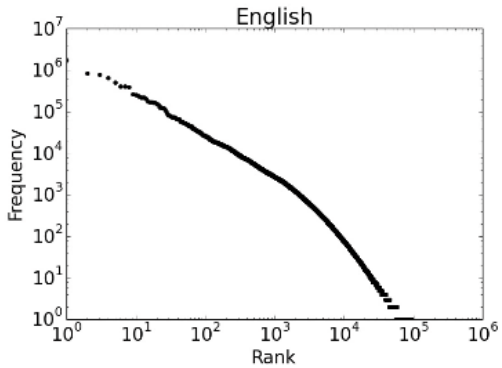
nouns	
Frequency	Type
124,598	European
104,325	Mr
92,195	Commission
66,781	President
62,867	Parliament
57,804	Union
53,683	report
53,547	Council
45,842	States

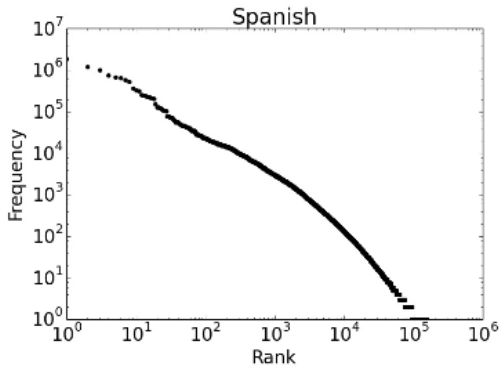
Sparse data due to Zipf's Law
Why NLP is hard?

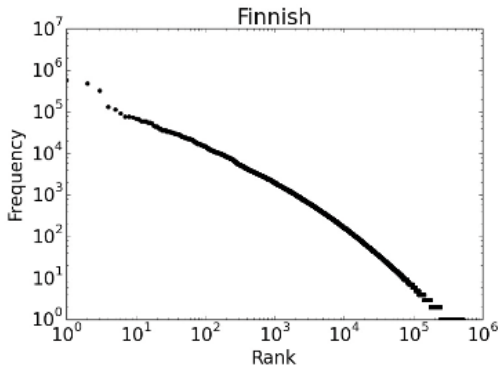
But also, out of 93638 distinct word types, 36231 occur only once.
Examples:

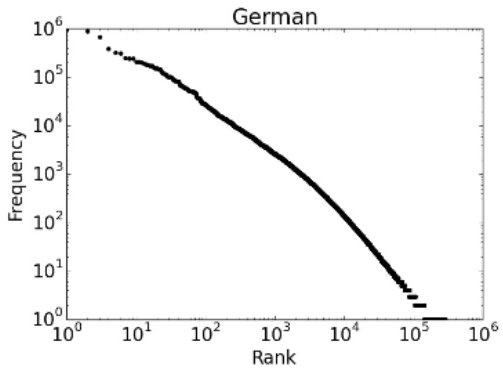
- cornflakes, mathematicians, fuzziness, jumbling
- pseudo-rapporteur, lobby-ridden, perfunctorily,
- Lycketoft, UNCITRAL, H-0695
- policyfor, Commissioneris, 145.95, 27a

Order words by frequency. What is the frequency of nth ranked word?









Zipf's law
Word frequencies

Summarizes the behaviour we just saw:

$$f \times r \approx k$$

- f = frequency of a word
- r = rank of a word (if sorted by frequency)
- k = a constant

Why a line in log-scales? $fr = k \Rightarrow \quad f = \frac{k}{r} \Rightarrow \log f = \log k - \log r$

- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words.
- In fact, the same holds for many other levels of linguistic structure
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen during training.

Expressivity
Why NLP is hard?

- Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:

  She gave the book to Tom *vs.* She gave Tom the book
  Some kids popped by *vs.* A few children visited
  Is that window still open? *vs.* Please close the window

Context dependence and Unknown representation

Why NLP is hard?

- Last example also shows that correct interpretation is context-dependent and often requires world knowledge.
- Very difficult to capture, since we don't even know how to represent the knowledge a human has/needs: What is the "meaning" of a word or sentence? How to model context? Other general knowledge?
- In particular, we've made remarkably little progress on the Knowledge Representation problem...

Other
Why NLP is hard?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad
the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

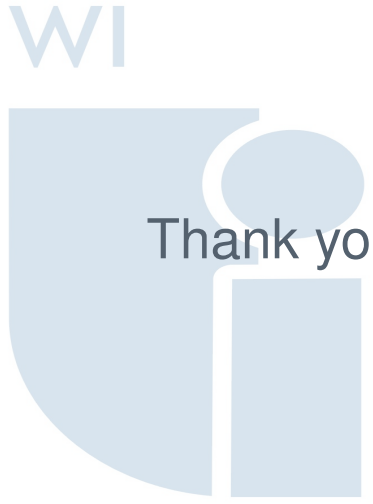
unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …
Let It Be was recorded …
… a mutation on the *for* gene …



Thank you for your attention