
Advanced Natural Language Processing

Lecture 1

Introduction

Philipp Koehn

21 September 2011



Welcome to ANLP

- Lecturers: Frank Keller, Philipp Koehn
- Lectures:
 - Tuesdays, Wednesdays and Fridays, 9:00am
 - CMB, Seminar Room 2
- Three assignments (worth 30%) will be given out over the semester
Exam counts for 70% of the grade
- Up to date information online at
<http://www.inf.ed.ac.uk/teaching/courses/anlp/>
- Recommended book: "Speech and Language Processing", 2nd edition,
Jurafsky and Martin, 2008, Prentice Hall.

Applications

- Speech Recognition
- Machine Translation
- Information Retrieval
- Dialog
- Question Answering
- Information Extraction
- Summarization
- Textual Entailment
- Sentiment Analysis
- ...

Courses

- Speech Recognition → ASR
- Machine Translation → MT
- Information Retrieval → TTS
- Dialog → NLG
- Question Answering
- Information Extraction
- Summarization
- Textual Entailment
- Sentiment Analysis
- ...

Linguistics and Data

Quotes

It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

Noam Chomsky, 1969

Whenever I fire a linguist our system performance improves.

Frederick Jelinek, 1988

Conflicts?

- Scientist vs. engineer
- Explaining language vs. building applications
- Rationalist vs. empiricist
- Insight vs. data analysis

Language Processing

Linguistics

- words
- morphology
- parts of speech
- syntax
- senses
- semantics
- discourse

Methods

- finite state models
- hidden Markov models
- grammars and parsing
- statistical parsing
- sense disambiguation

Words

This is a simple sentence **WORDS**

Morphology

This is a simple sentence

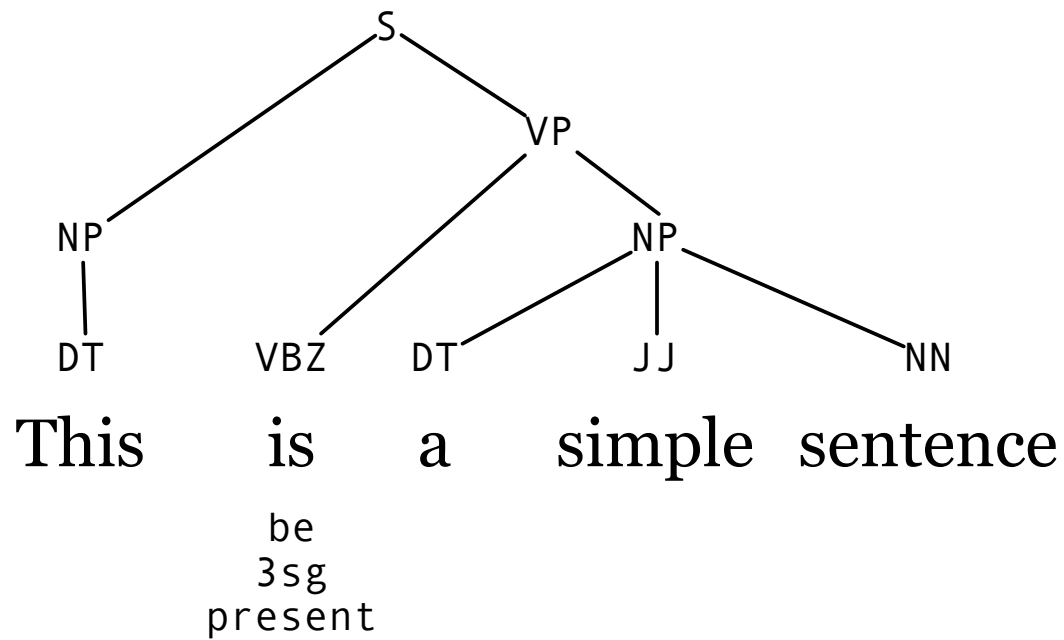
be
3sg
present

WORDS
MORPHOLOGY

Parts of Speech

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

Syntax



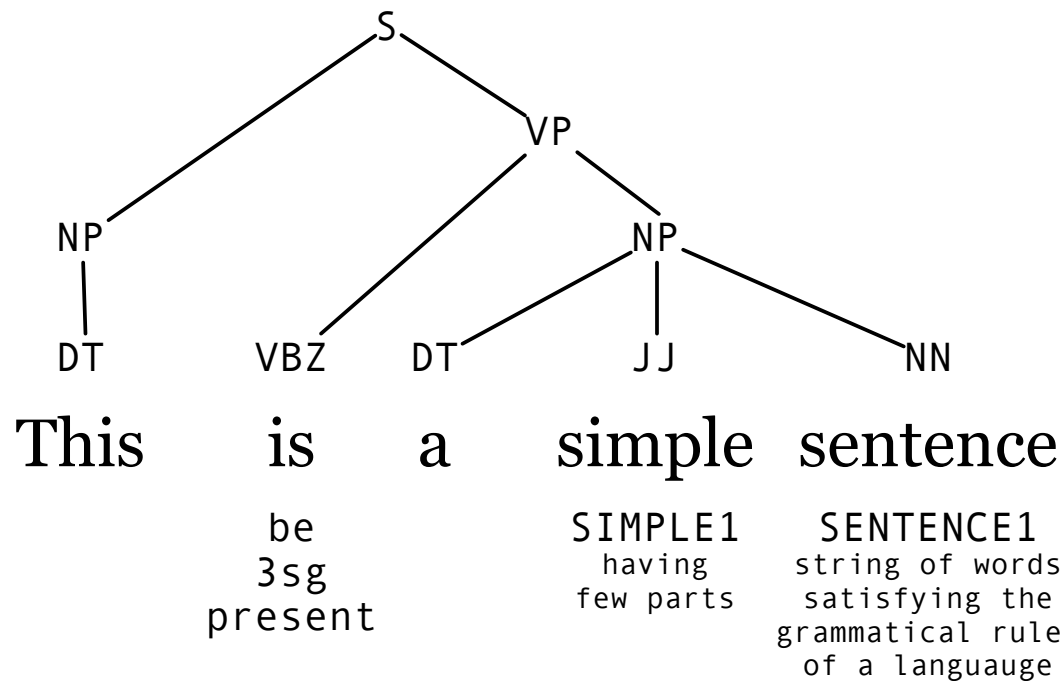
SYNTAX

PART OF SPEECH

WORDS

MORPHOLOGY

Semantics



SYNTAX

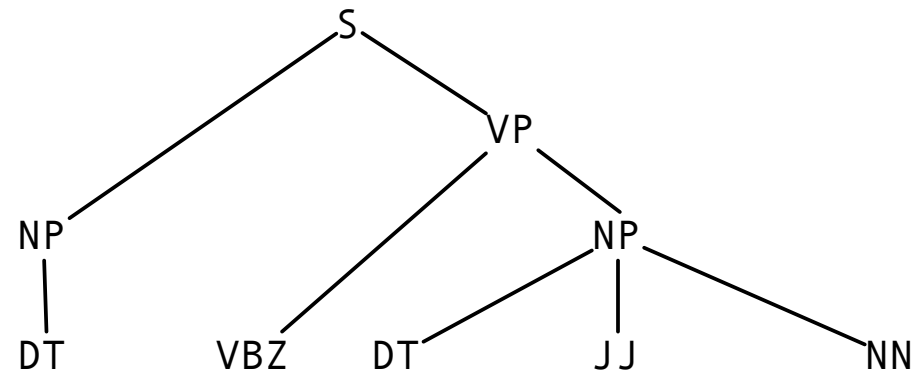
PART OF SPEECH

WORDS

MORPHOLOGY

SEMANTICS

Discourse



SYNTAX

PART OF SPEECH

This is a simple sentence

WORDS

be
3sg
present

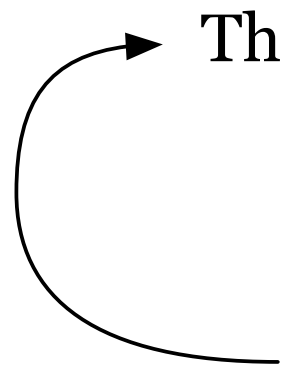
SIMPLE1
having
few parts

SENTENCE1
string of words
satisfying the
grammatical rules
of a language

MORPHOLOGY

SEMANTICS

CONTRAST



But it is an instructive one.

DISCOURSE

Why is Language Hard?

- Ambiguities on many levels
- Rules, but many exceptions
- No clear understand how humans process language
- Can we learn everything about language by automatic data analysis?

Data: Words

- Definition: strings of letters separated by spaces
- But how about:
 - punctuation: commas, periods, etc. typically separated (tokenization)
 - hyphens: [high-risk](#)
 - clitics: [Joe's](#)
 - compounds: [website](#), [Computerlinguistikvorlesung](#)
- And what if there are no spaces:

伦敦每日快报指出,两台记载黛安娜王妃一九九七年巴黎死亡车祸调查资料的手提电脑,被从前大都会警察总长的办公室里偷走.

Word Counts

Most frequent words in the English Europarl corpus

any word		nouns	
Frequency in text	Token	Frequency in text	Content word
1,929,379	the	129,851	European
1,297,736	,	110,072	Mr
956,902	.	98,073	commission
901,174	of	71,111	president
841,661	to	67,518	parliament
684,869	and	64,620	union
582,592	in	58,506	report
452,491	that	57,490	council
424,895	is	54,079	states
424,552	a	49,965	member

Word Counts

But also:

There is a large tail of words that occur only once.

33,447 words occur once, for instance

- cornflakes
- mathematicians
- Tazhikistan

Zipf's law

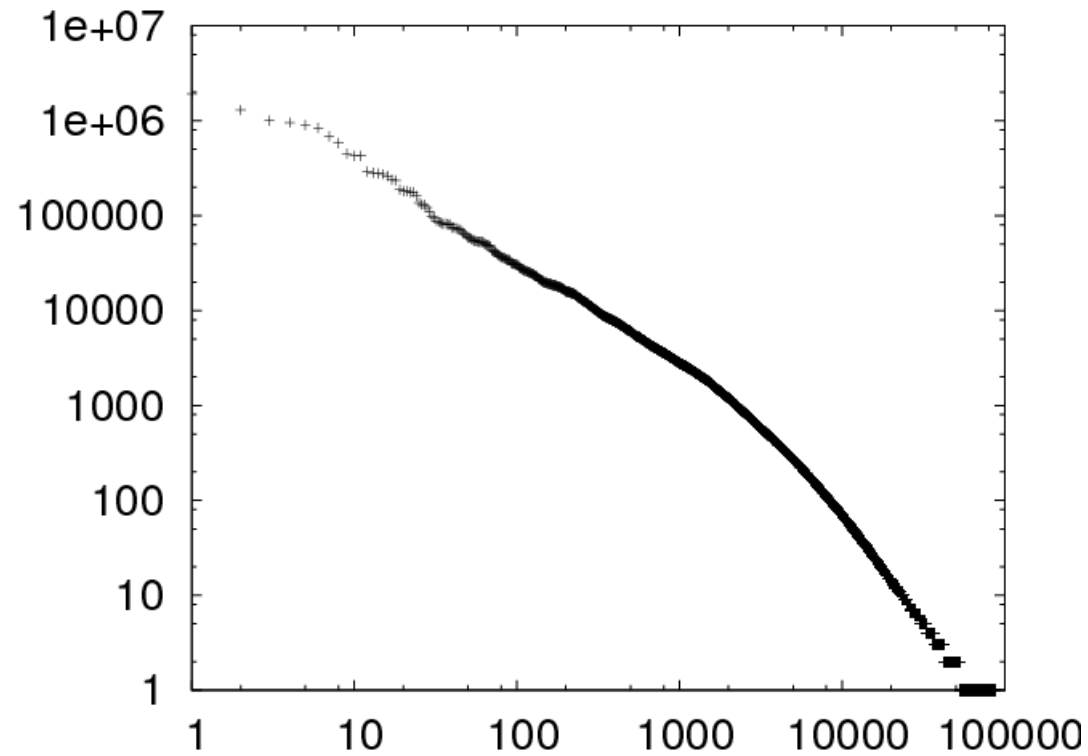
$$f \times r = k$$

f = frequency of a word

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

k = a constant

Zipf's law as a graph



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

Linguistics and Data

- Data
 - looking at real use of language in text
 - can learn a lot from empirical evidence
 - but: Zipf's law: there will be always instances that are rarely seen
- Linguistics
 - build a better understanding of language structure
 - linguistic analysis points to what is important
 - but: many ambiguities cannot be explained easily