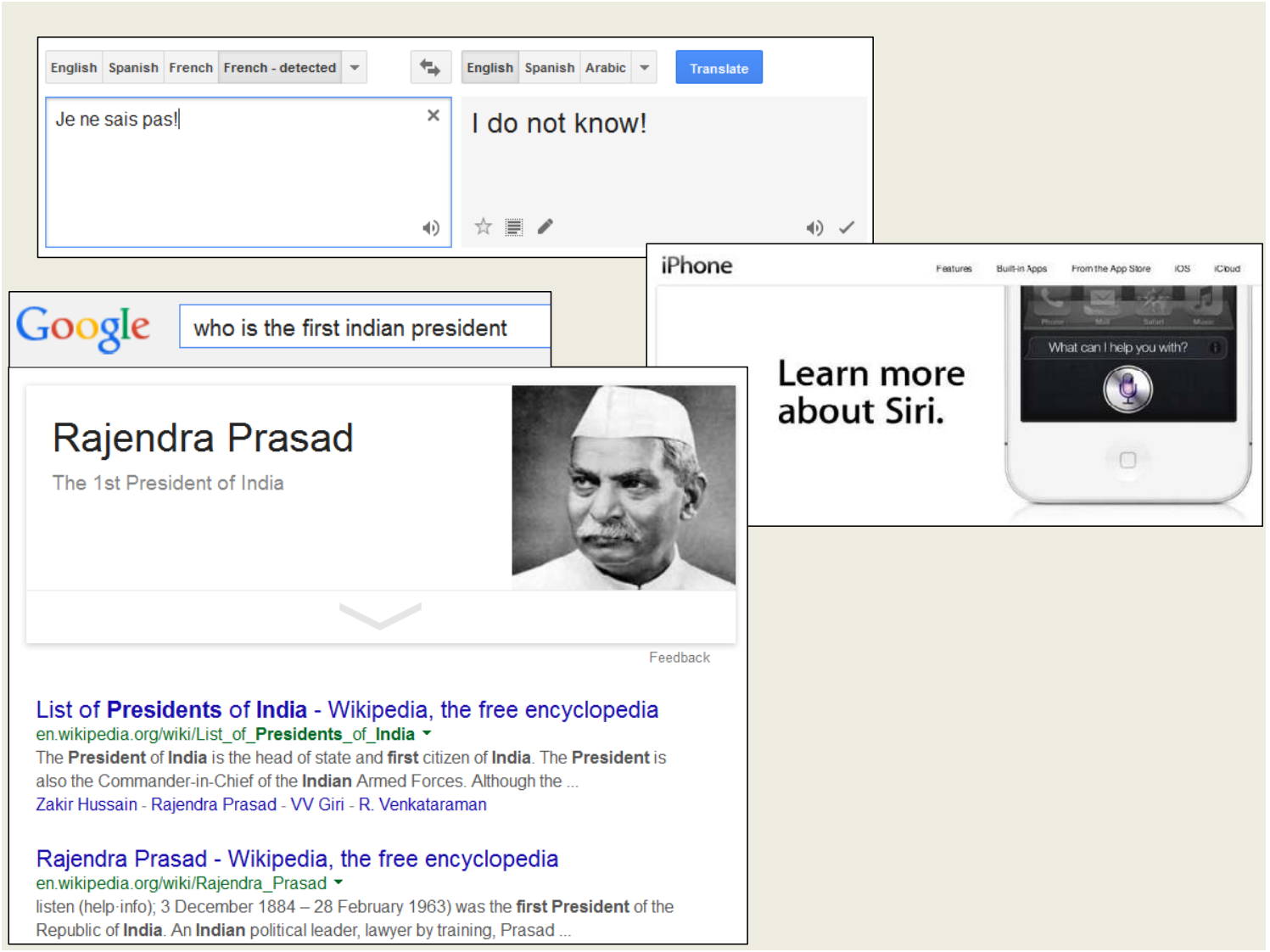
Foundations of Natural Language Processing
Lecture 1
Introduction

Alex Lascarides
(Slides based on those of Philipp Koehn, Alex Lascarides, Sharon Goldwater)

15 January 2019

School of informatics

What is Natural Language Processing?



What is Natural Language Processing?

Applications

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

Core technologies

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

This course

NLP is a big field! We focus mainly on core ideas and methods needed for technologies in the second column (and eventually for applications).

- Linguistic facts and issues
- Computational models and algorithms

More advanced methods and specific application areas covered in 4th/5th year courses:

- Natural Language Processing 2
- Machine Translation
- Text Technologies
- Automatic Speech Recognition

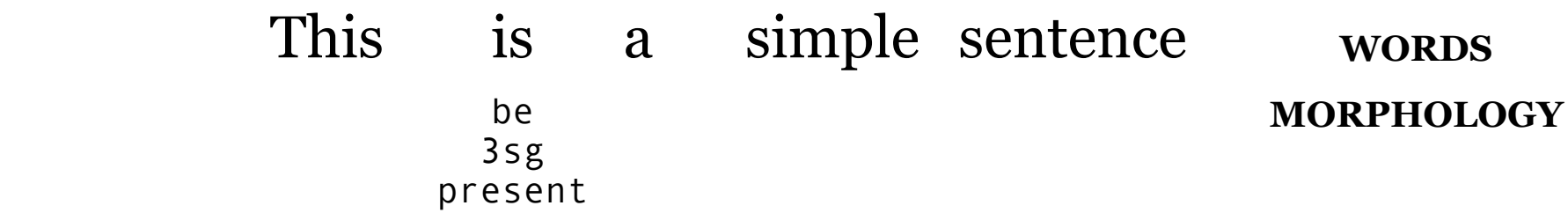
What does an NLP system need to "know"?

- Language consists of many levels of structure
- Humans fluently integrate all of these in producing/understanding language
- Ideally, so would a computer!

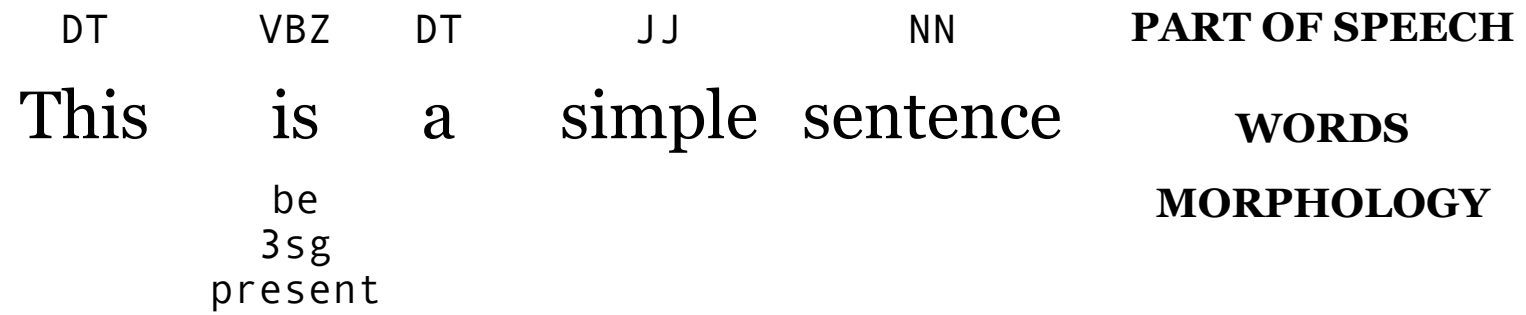
Words

This is a simple sentence **WORDS**

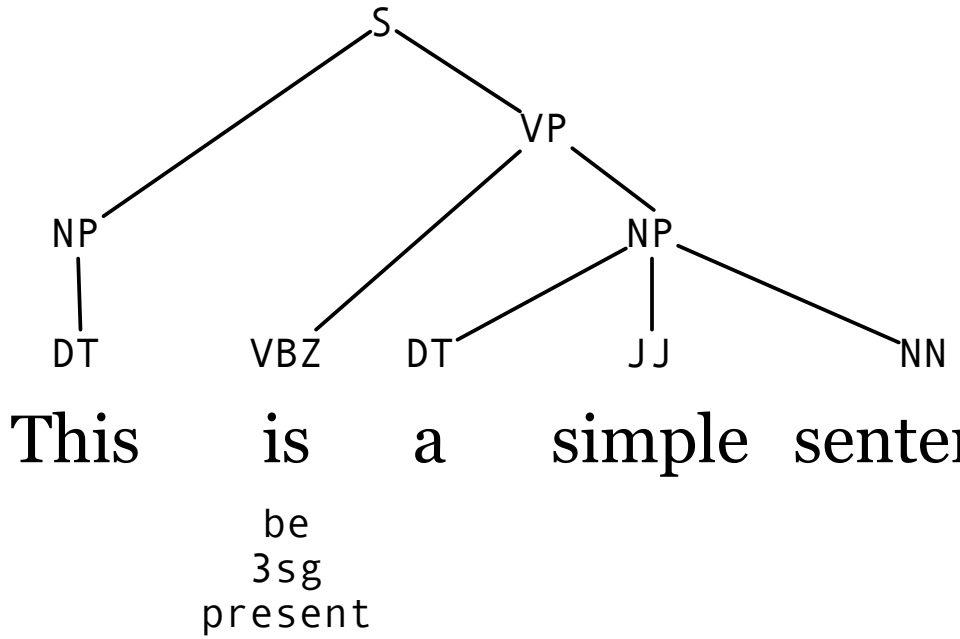
Morphology



Parts of Speech



Syntax



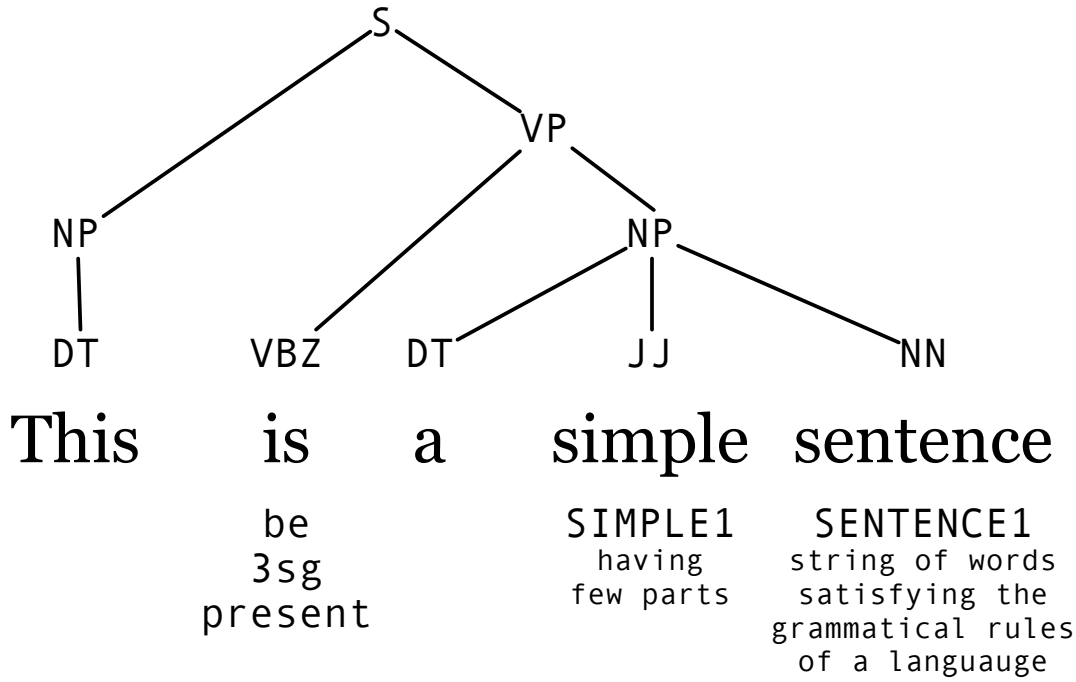
SYNTAX

PART OF SPEECH

WORDS

MORPHOLOGY

Semantics



SYNTAX

PART OF SPEECH

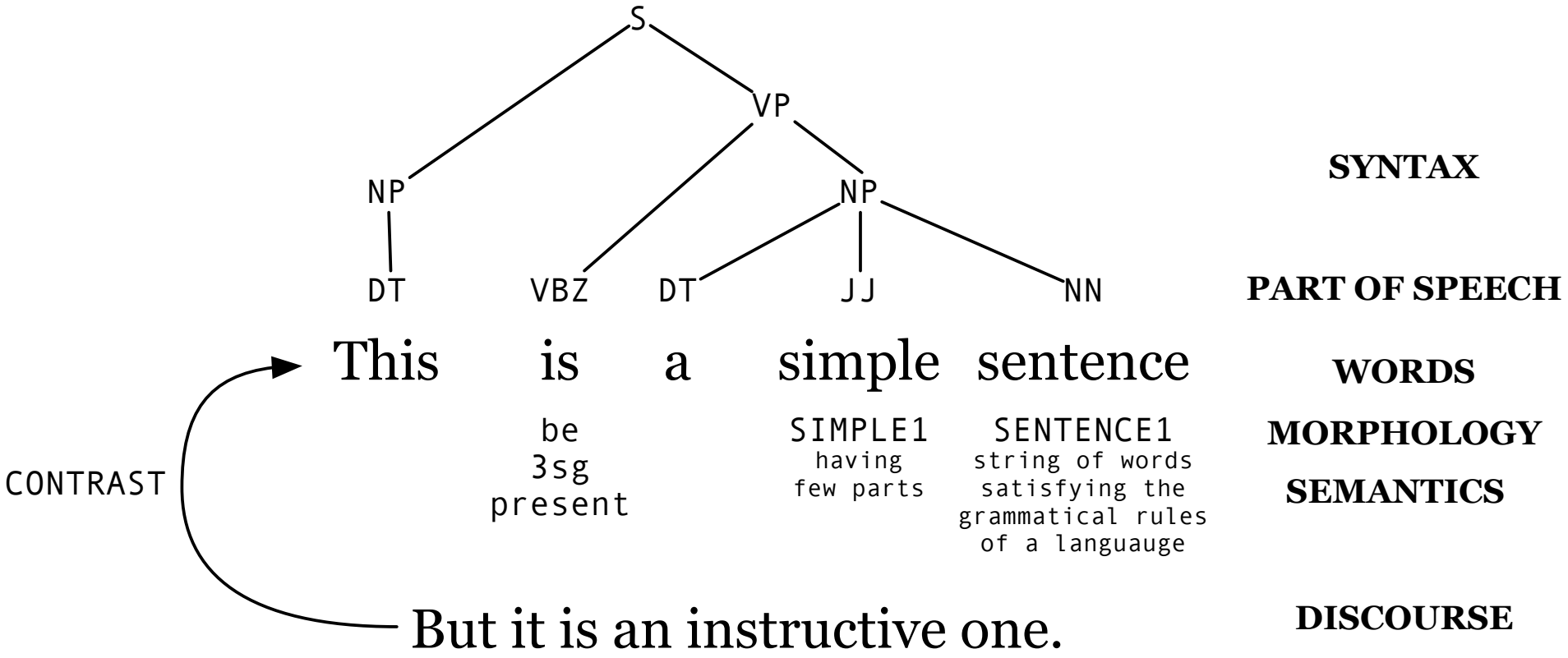
WORDS

MORPHOLOGY

SEMANTICS

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

Discourse


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

Why is NLP hard?

1. **Ambiguity** at many levels:

- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Quantifier scope: Every child loves some movie
- Multiple: I saw her duck
- Reference: John dropped the goblet onto the glass table and it broke.
- Discourse: The meeting is cancelled. Nicholas isn't coming to the office today.

How can we model ambiguity, and choose the correct analysis in context?

Ambiguity

Inf2a started to discuss 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

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** (cf. Inf2b).
- Normally requires **lots of data** about any particular phenomenon.

Why is NLP hard?

2. **Sparse data** due to **Zipf's Law**.

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

Word Counts

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

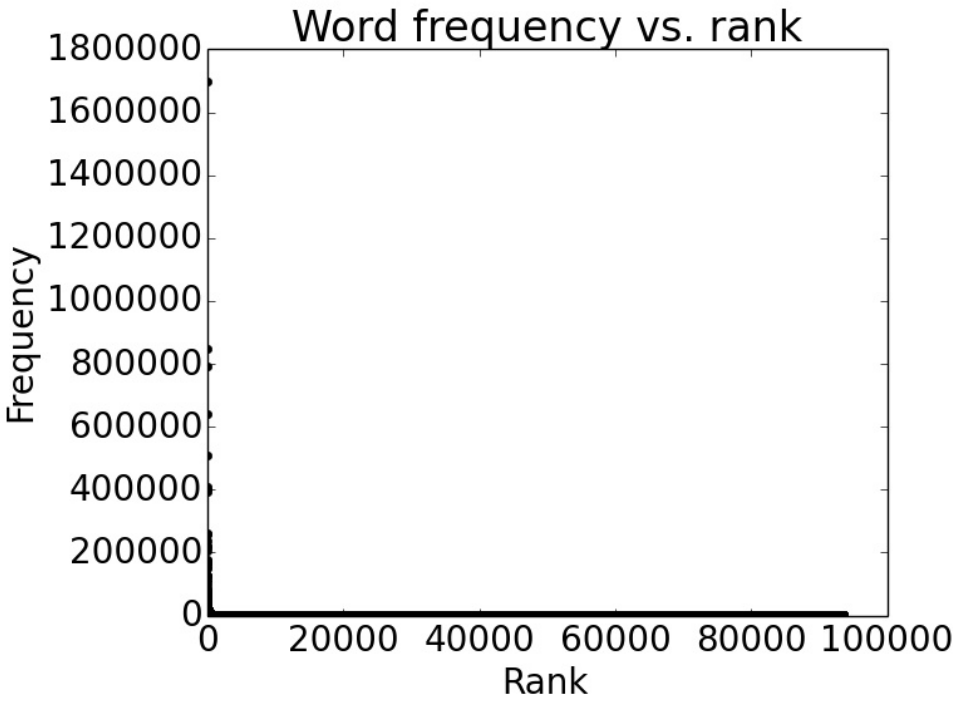
Word Counts

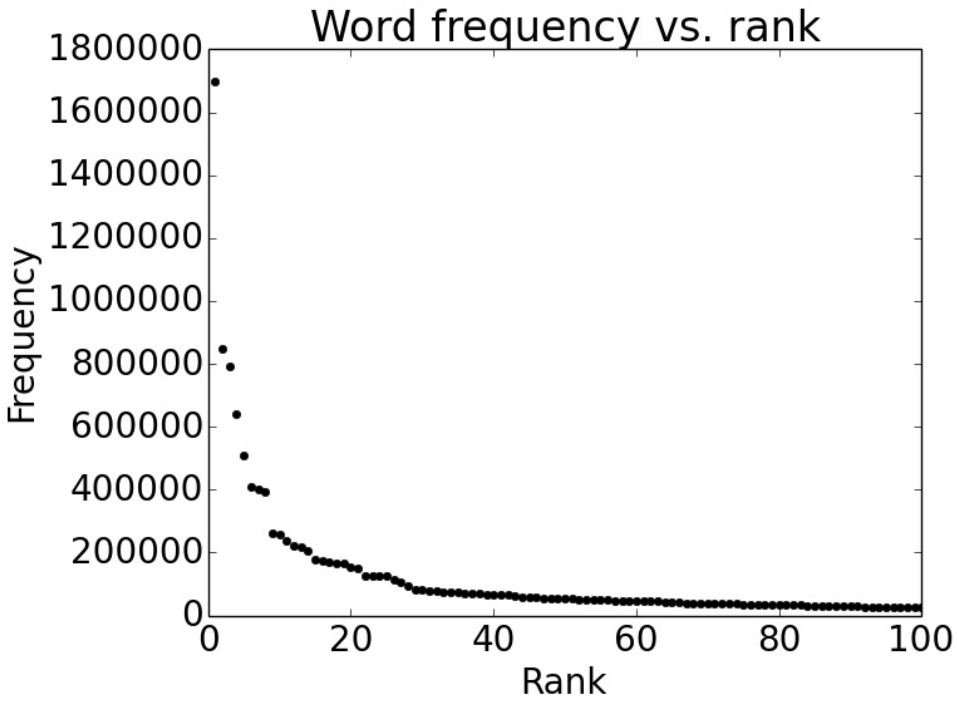
But also, out of 93638 distinct word types, 36231 occur only once.
Examples:

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

Plotting word frequencies

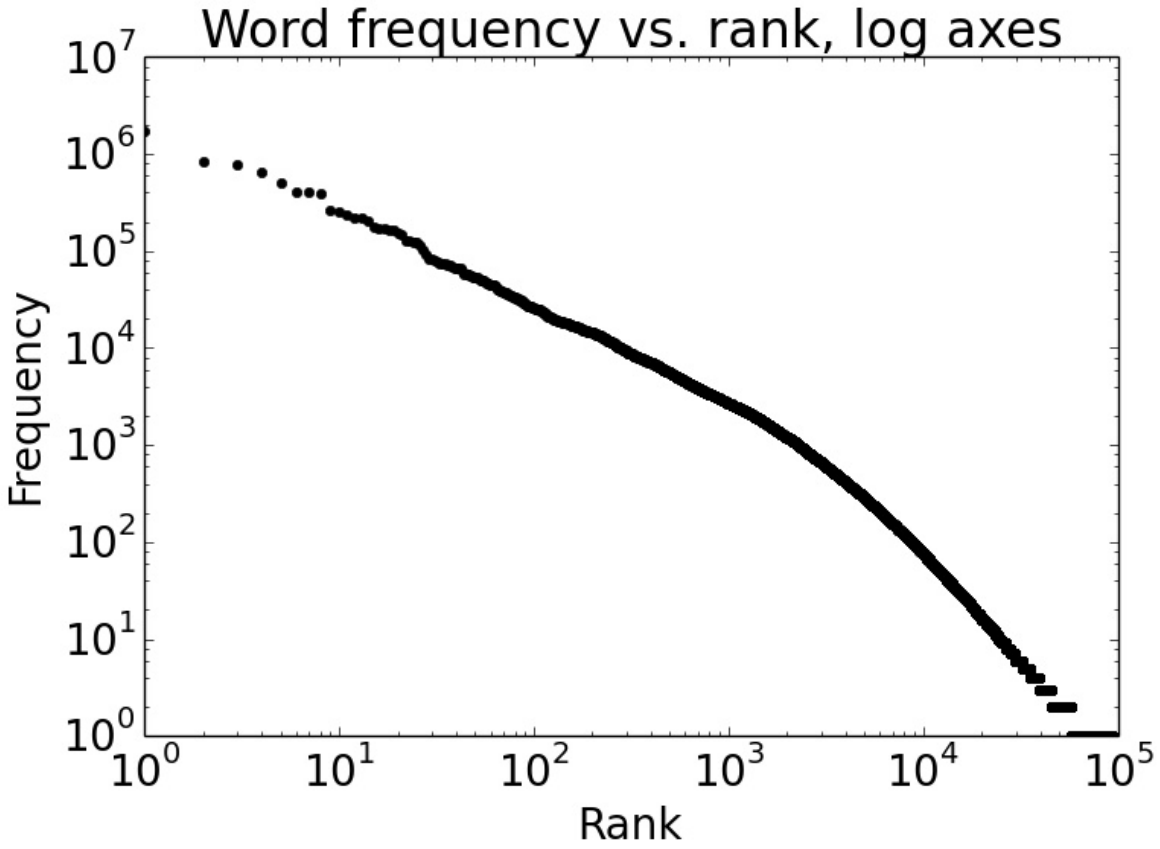
Order words by frequency. What is the frequency of nth ranked word?

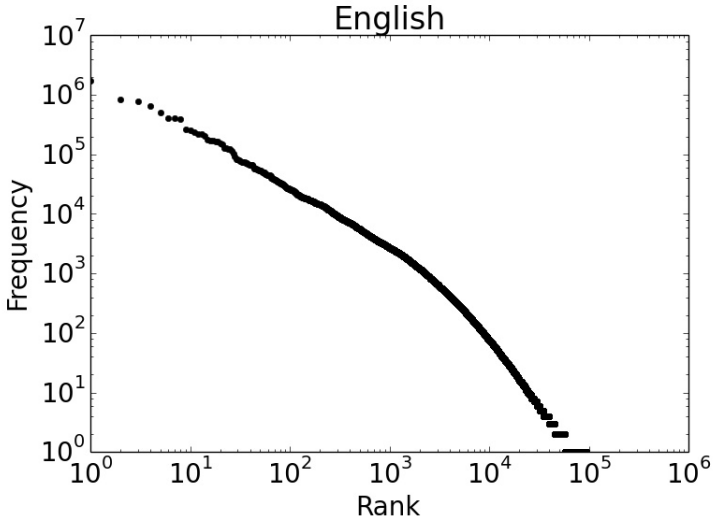


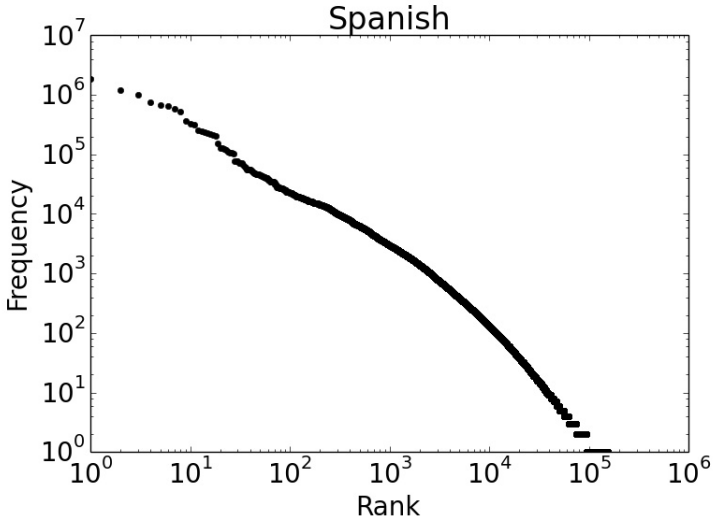


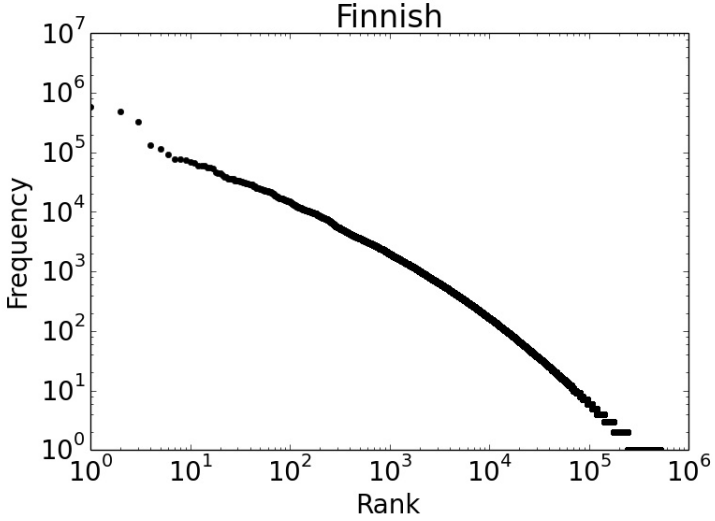
Rescaling the axes

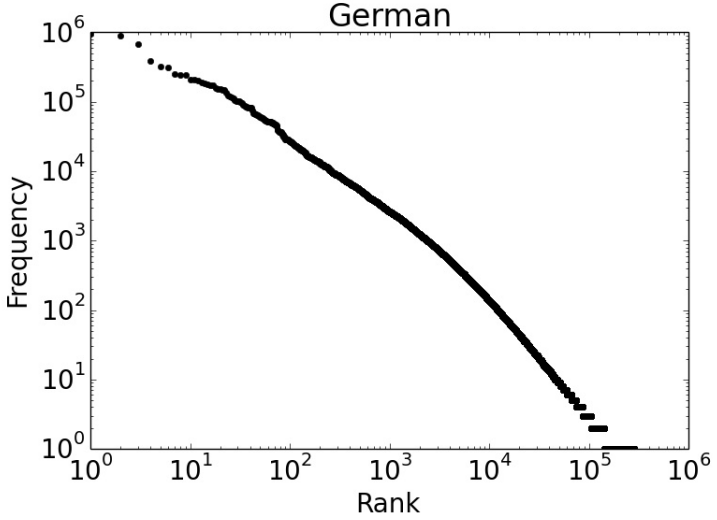
To really see what's going on, use logarithmic axes:




English
Frequency
Rank


Spanish
Frequency
Rank


Finnish
Frequency
Rank


German
Frequency
Rank

Zipf's law

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? $fr = k \quad \Rightarrow \quad f = \frac{k}{r} \quad \Rightarrow \quad \log f = \log k - \log r$

Implications of Zipf's Law

- 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 (e.g., syntactic rules in a CFG).
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen during training.

Why is NLP hard?

3. **Variation**

- Suppose we train a part of speech tagger on the Wall Street Journal:

  Mr./NNP Vinken/NNP is/VBZ chairman/NN of/IN Elsevier/NNP
  N.V./NNP ,/, the/DT Dutch/NNP publishing/VBG group/NN ./.

- What will happen if we try to use this tagger for social media??

  ikr smh he asked fir yo last name

Twitter example due to Noah Smith

Why is NLP hard?

4. **Expressivity**

- 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

Why is NLP hard?

5 and 6. **Context dependence** and **Unknown representation**

- 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?

That is, in the limit NLP is hard because *AI* is hard

- In particular, we've made remarkably little progress on the Knowledge Representation problem...