What is (E)NLP?

Nathan Schneider ~ 9 January 2025

Some slides adapted from Sharon Goldwater, Philipp Koehn, Alex Lascarides

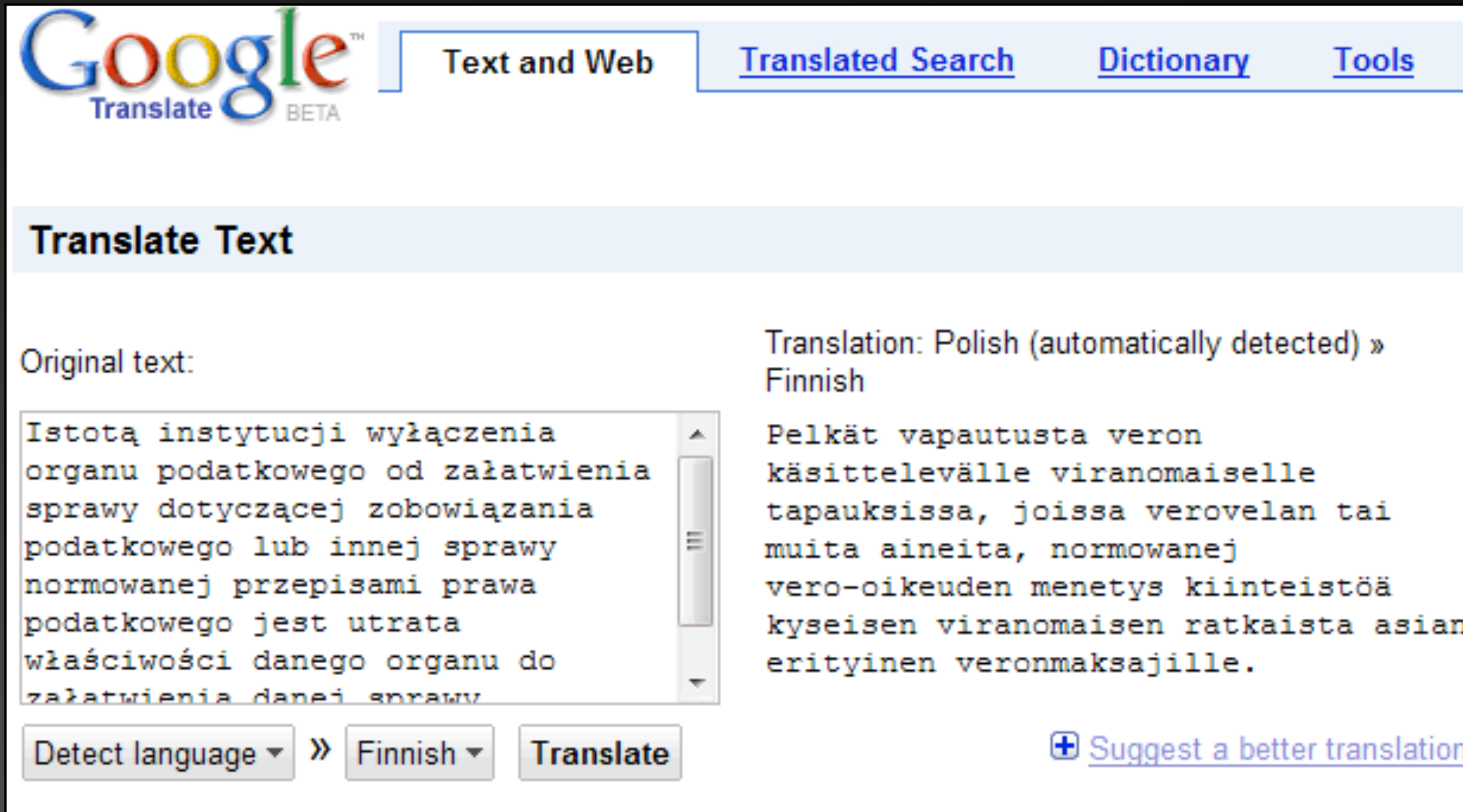
https://people.cs.georgetown.edu/nschneid/cosc5402/

What do YOU think?

- Team up with a partner you don't already know.
- Take 5 min. to discuss:
  - What have you heard lately about NLP & AI?
  - What do you expect to learn in this course?

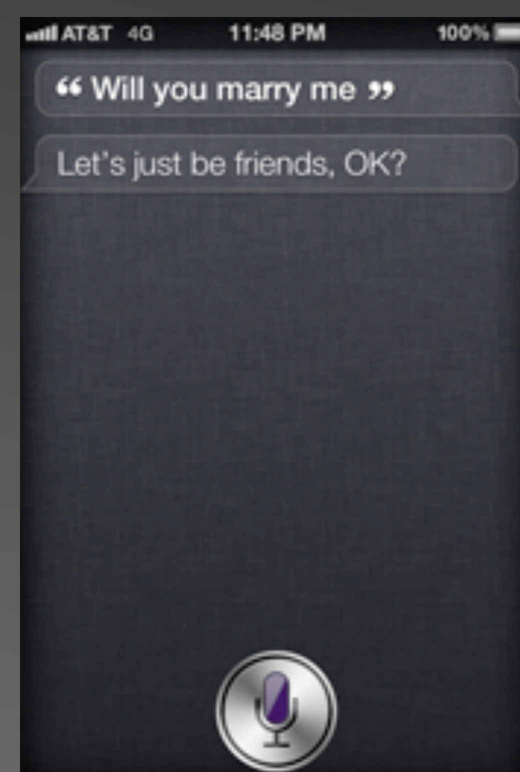
Introductions

- Say your name, program, year, language background
- and what you discussed with your partner


Google Translate BETA
Text and Web
Translated Search
Dictionary
Tools
Translate Text
Original text:
Istotą instytucji wyłączenia organu podatkowego od załatwienia sprawy dotyczącej zobowiązania podatkowego lub innej sprawy normowanej przepisami prawa podatkowego jest utrata właściwości danego organu do
Translation: Polish (automatically detected) » Finnish
Pelkät vapautusta veron käsittelevälle viranomaiselle tapauksissa, joissa verovelan tai muita aineita, normowanej vero-oikeuden menetys kiinteistöä kyseisen viranomaisen ratkaista asian erityinen veronmaksajille.
Detect language
Finnish
Translate
Suggest a better translation


$18,200
$21,440
$5,600
WATSON
contempt 97%
contemn 14%
Despised Icon 10%


Conduciendo en 64
0.0
Menú
NE
GARMIN nüvi


AT&T 4G 11:48 PM 100%
“ Will you marry me ”
Let's just be friends, OK?

Applications & Core Tasks

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

- Language modeling/text generation
- Part-of-speech tagging
- Syntactic parsing
- Named entity recognition
- Coreference resolution
- Word sense disambiguation
- Semantic role labeling
- …

NLP as a Field

- NLP lies at the intersection of **computational linguistics** and **artificial intelligence**.

- NLP is (to various degrees) informed by linguistics, but with practical/engineering rather than purely scientific aims.

- Processing **speech** (i.e., the acoustic signal) is separate.

This Course

- NLP is a big field! This course focuses mainly on **foundational** ideas and methods to answer the question: “How can we formulate computation for natural language?”
  - Linguistic facts and issues
  - Computational models and algorithms, especially using data (“empirical”)
  - More emphasis on representations and tasks than applications

What are your goals?

Why are you here? Perhaps you want to:

- work at a company that uses NLP (perhaps as the sole language expert among engineers)
- use NLP tools for research in linguistics (or other domains where text data is important: social sciences, humanities, medicine, …)
- conduct research in NLP (or IR, MT, etc.)

Linguistic Structure

- An important insight of linguistics is that language consists of many levels of structure
- Humans fluently integrate all of these in producing/ understanding language
- Ideally, so would a computer!

Linguistic Structure

This is a simple sentence **WORDS**

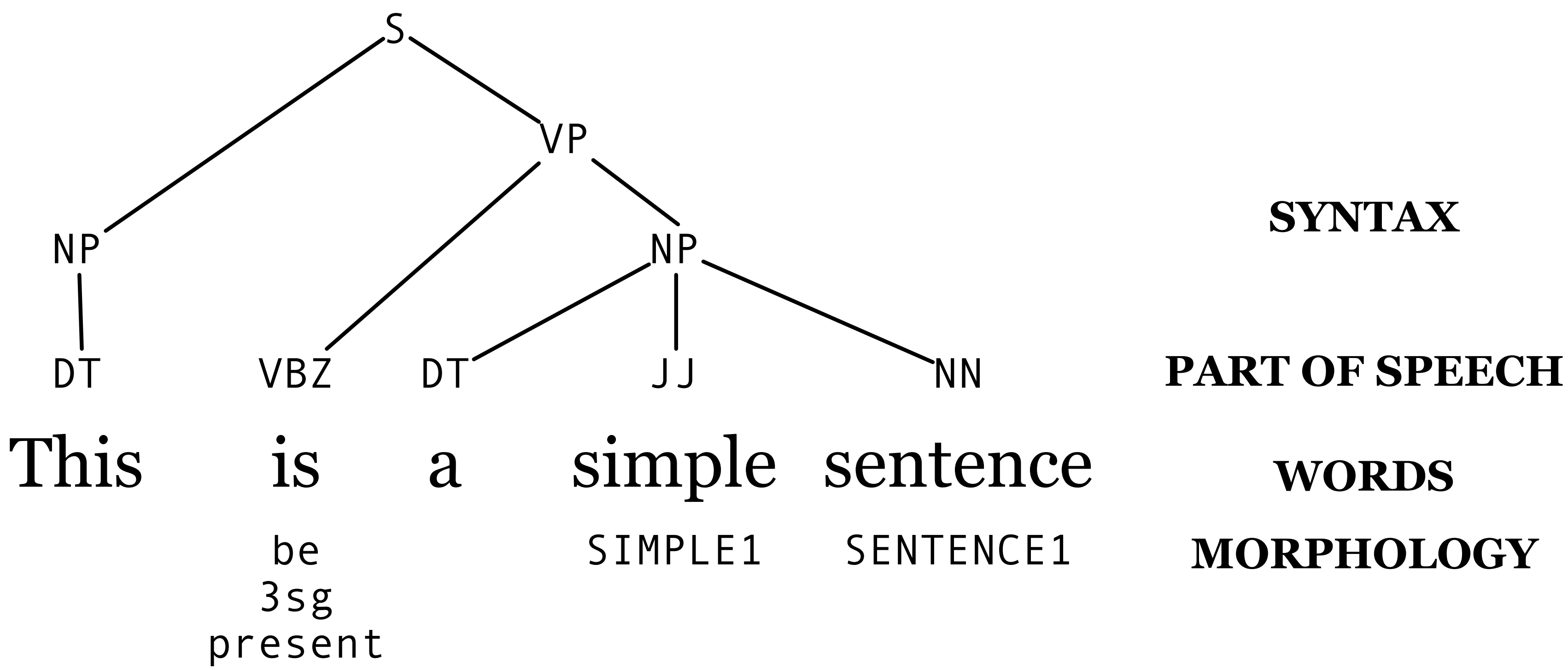
Linguistic Structure

This	is	a	simple	sentence	**WORDS**
	be 3sg present		SIMPLE1	SENTENCE1	**MORPHOLOGY**

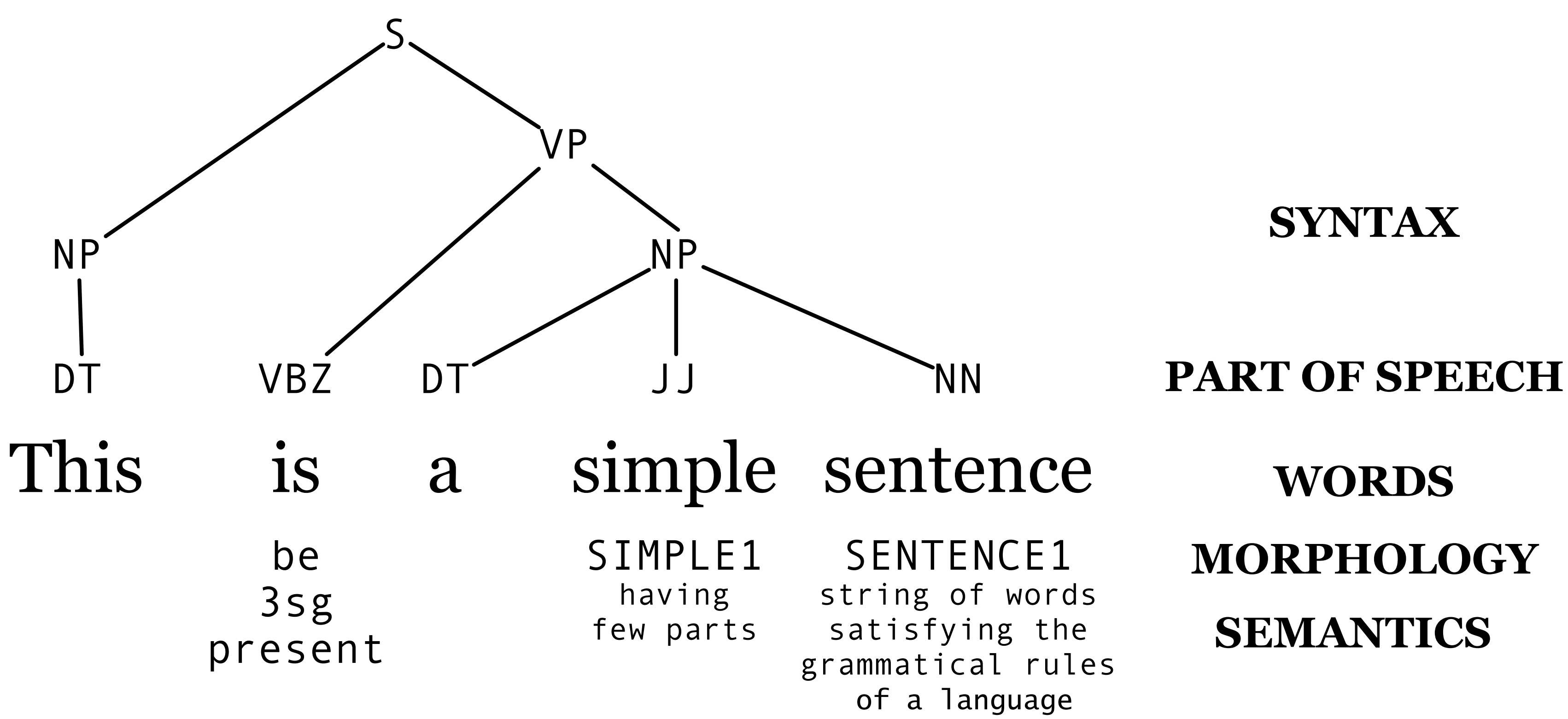
Linguistic Structure

DT	VBZ	DT	JJ	NN	**PART OF SPEECH**
This	is	a	simple	sentence	**WORDS**
	be 3sg present		SIMPLE1	SENTENCE1	**MORPHOLOGY**

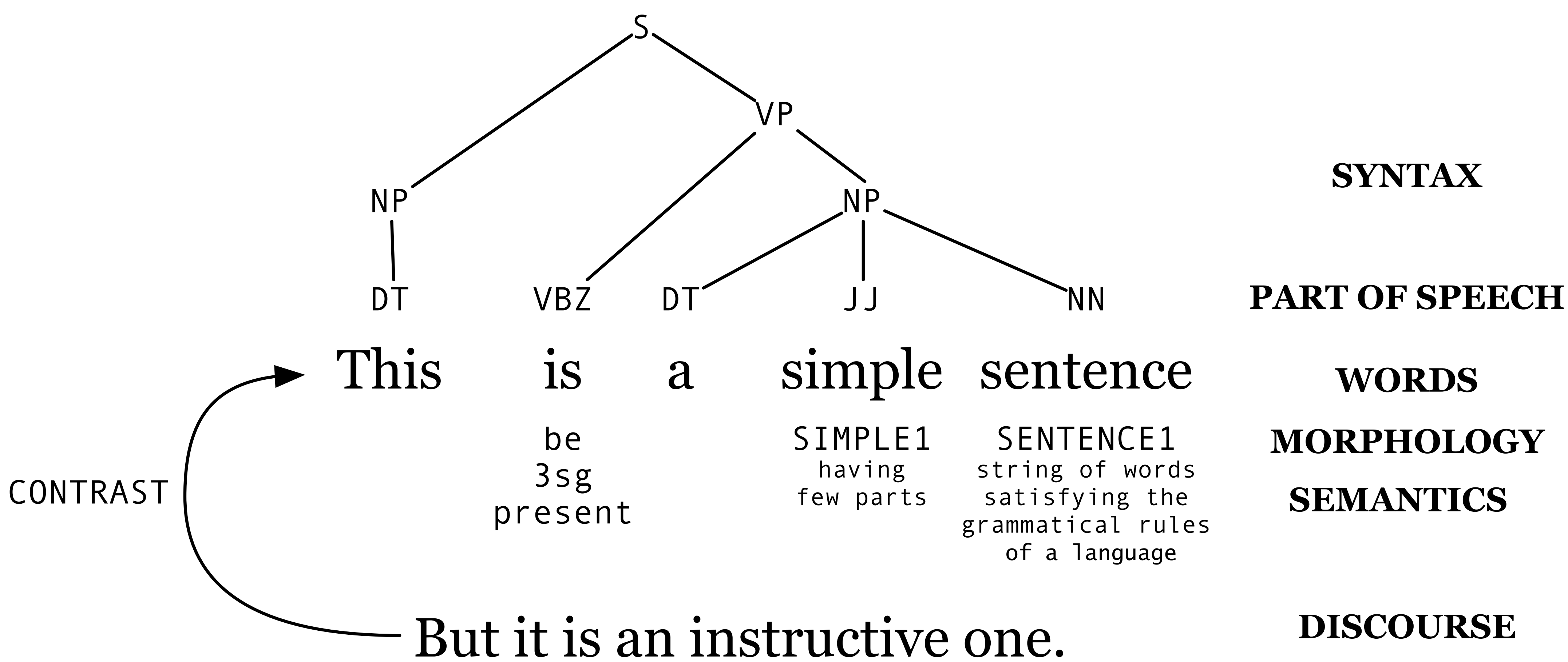
Linguistic Structure



Linguistic Structure



Linguistic Structure



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

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

Ambiguity

What can we do about 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*.

But the "best" analysis is only good if our probabilities are accurate. Where do they come from?

Statistical NLP

Like most other parts of AI, NLP 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.

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 "word" is a string of letters separated by spaces (a great oversimplification, we'll return to this issue...)

Word Counts

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

any word

Frequency	Token
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	Token
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

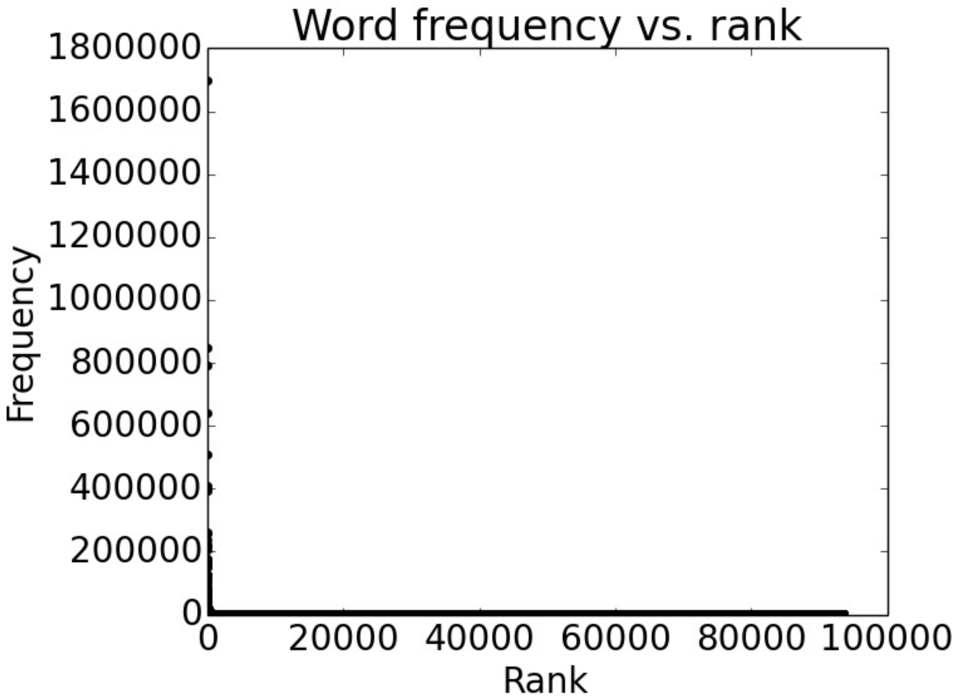
Word Counts

But also, out of 93,638 distinct words (**word types**), 36,231 occur only once. Examples:

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

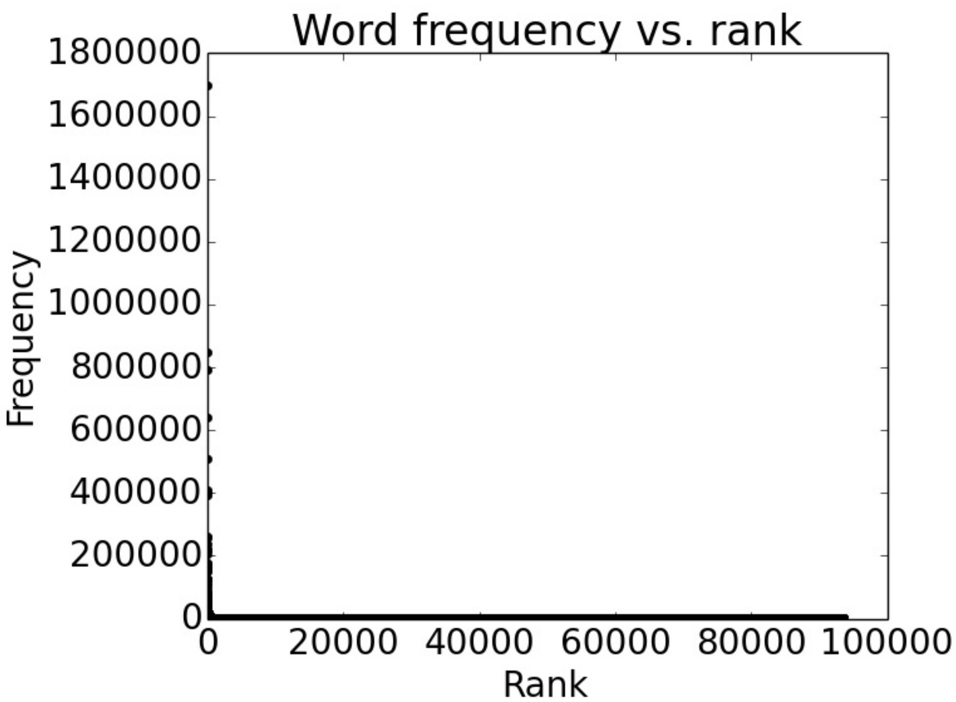
Plotting word frequencies

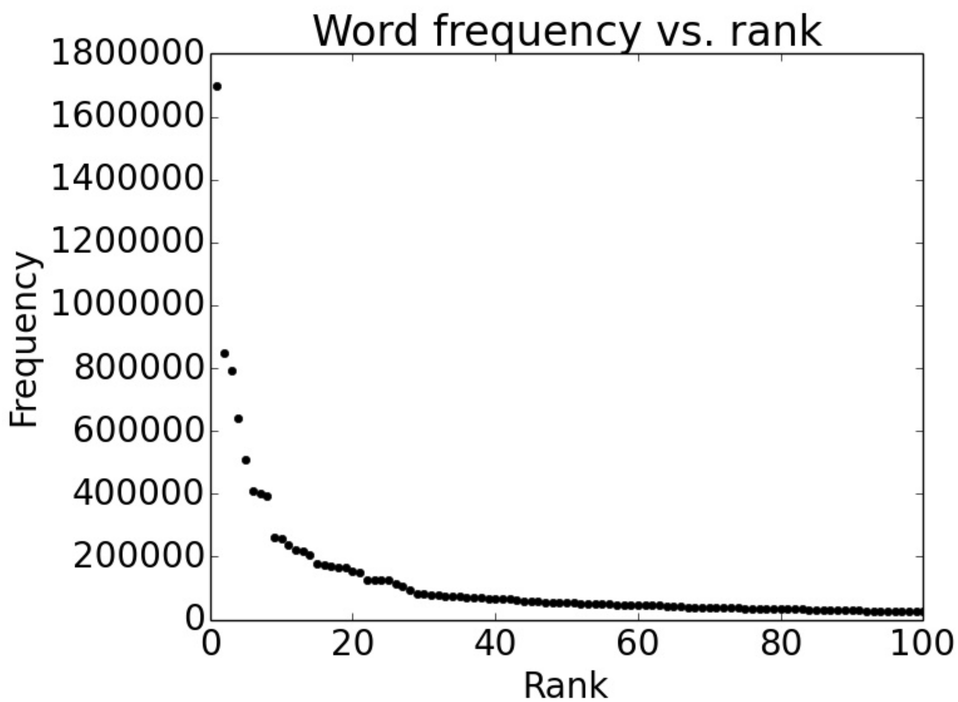
Order words by frequency. What is the frequency of nth ranked word?



Plotting word frequencies

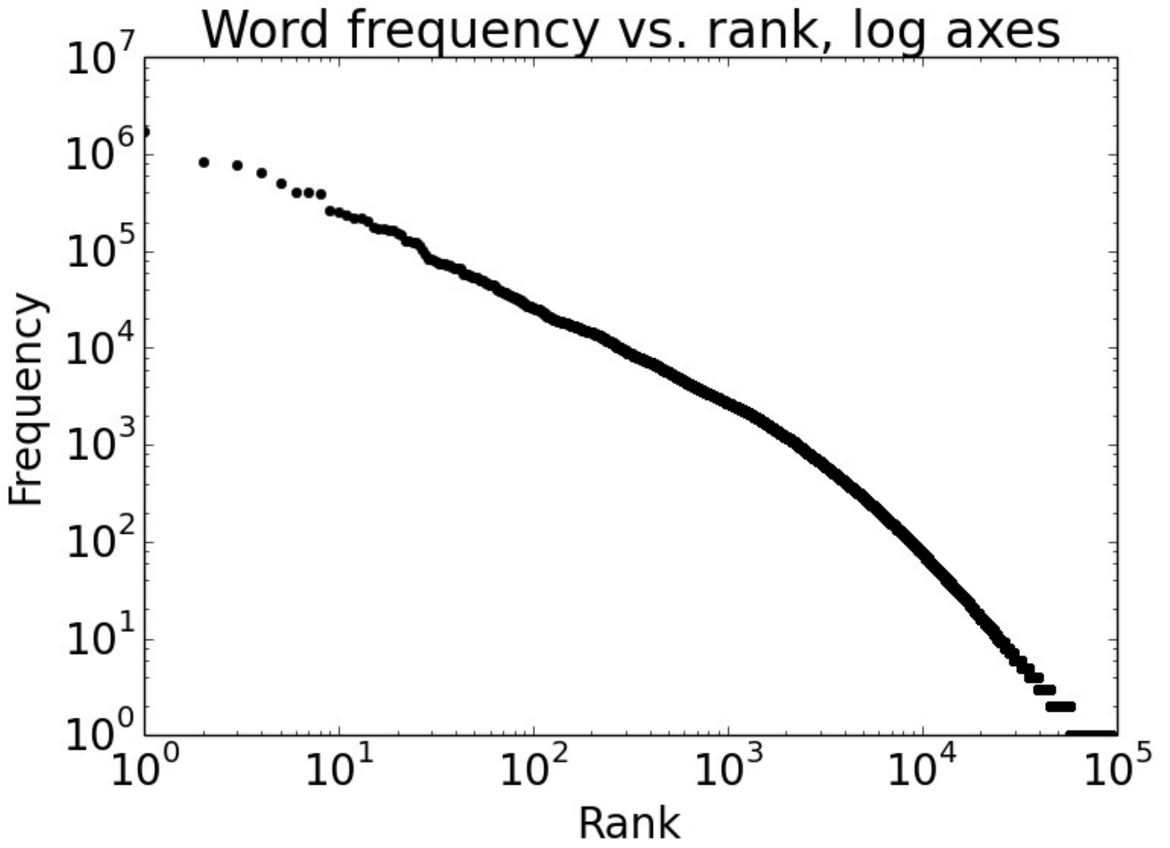
Order words by frequency. What is the frequency of nth ranked word?

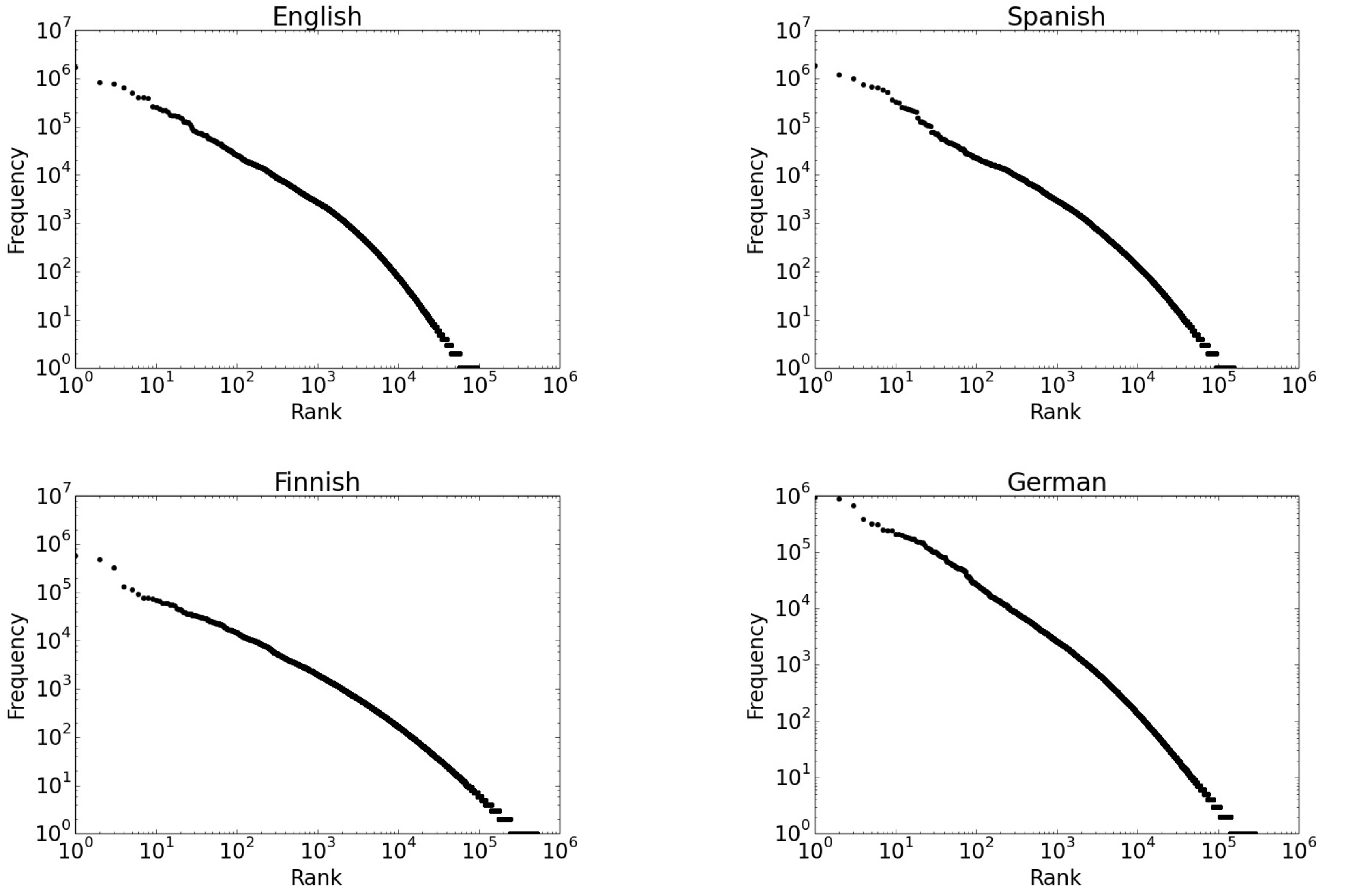




Rescaling the axes

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




English
Spanish
Finnish
German
Frequency
Rank
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7

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

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

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.

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 ./.

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?

Organization of Topics

- **Introduction, N-grams**: Some basics of text processing, linguistics, and probabilistic models of word sequences.
- Classification, Lexical Semantics with Classical Approaches
- Sequential Prediction, Part-of-Speech Tagging with Classical Approaches, Annotation
- Word Embeddings and Neural Networks
- Hierarchical Sentence Structure: Syntax, Grammars, and Parsing
- Neural Text Generation and Large Language Models

Backgrounds

- This course has enrollment from multiple programs!:
  - Linguistics
  - Computer Science
  - possibly: Data Science; …
- This means that there will be a diversity of backgrounds and skills, which is a fantastic opportunity for you to learn from fellow students.
- It also requires a bit of care to make sure the course is valuable for everyone.

What’s NOT in this course

- Formal language theory
- Computational morphology
- Compositional semantics
- Speech/signal processing, phonetics, phonology

Related Courses

- https://gucl.georgetown.edu/courses.php

Course Organization

- Instructor: **Nathan Schneider**
- TAs: **Xiulin Yang** + (starting after next week) **Blake Wang**
- Lectures: TuTh 11:00–12:15 ET, Walsh 497
- Office hours: stay tuned for times.
- Website: for syllabus, schedule (lecture slides/readings/assignments): https://people.cs.georgetown.edu/nschneid/cosc5402/
  - Make sure to read the syllabus!
  - No hard-copy textbook; readings will be posted online.
- We will also use Canvas for communication, submitting assignments.

Action Items

- All assignments will be linked from the schedule page on the course website.

- As a sort of pretest to make sure you are ready for this course, you have 1 week to do A0 (due before the start of class 1 week from today).
  - **It should not be hard or take very long**; if it takes you a long time you should consider a different course to practice Python skills.

- Canvas site is up. Submit assignment answers there.

- Registration:
  - The course is currently full. There is no waitlist anymore. If somebody drops before next Friday, it presumably would leave room for another student who meets the restrictions to add.