Computational Linguistics I

CMSC 723 / LING 723 / INST 725

Marine Carpuat

What is language?

Wikipedia:

"Language is the ability to acquire and use complex systems of communication, particularly the human ability to do so, and a language is any specific example of such a system. The scientific study of language is called linguistics."

- Computational Linguistics (CL)
  - The science of doing what linguists do with language, but using computers
- Natural Language Processing (NLP)
  - The engineering discipline of doing what people do with language, but using computers
- Speech/Language/Text processing
- Human Language Technology

NLP State of the Art

Still a challenging problem!

AI's Language Problem

"Machines that truly understand language would be incredibly useful. But we don't know how to build them."

MIT Technology Review

Will Knight, Aug 9, 2016

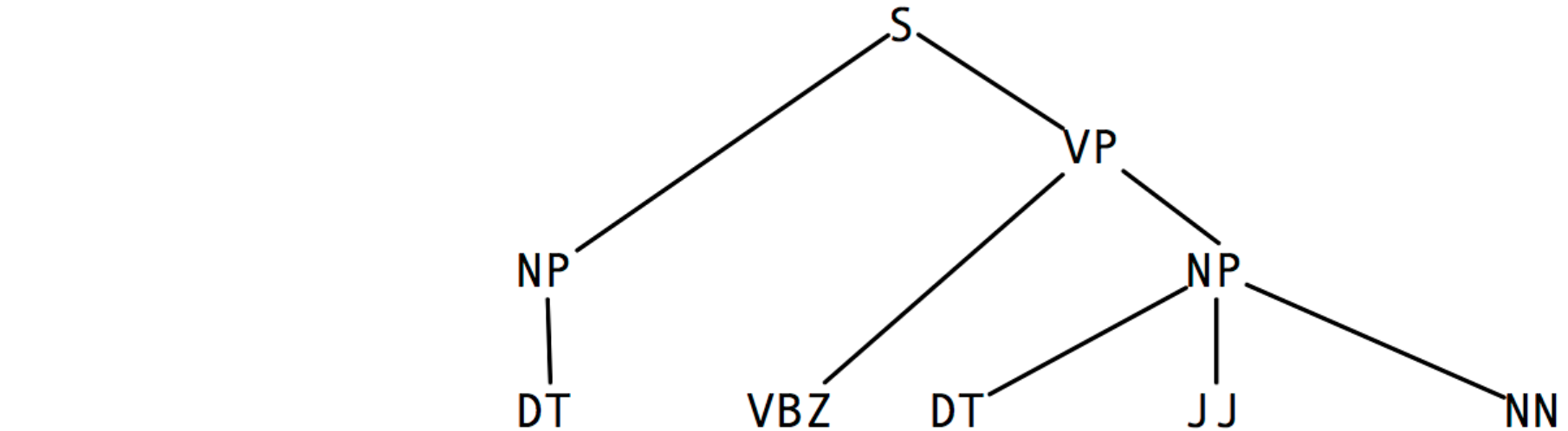
Many useful applications already exist

What does an NLP system need to “know”?

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

This is a simple sentence

Example from Nathan Schneider



SYNTAX

PART OF SPEECH

This is a simple sentence

WORDS

MORPHOLOGY

be
3sg
present

SIMPLE1
having
few parts

SENTENCE1
string of words
satisfying the
grammatical rules
of a language

SEMANTICS

CONTRAST

But it is an instructive one.

DISCOURSE

Why is NLP hard?

Ambiguity

At the word level

- Part of speech
  - [V Duck]!
  - [N Duck] is delicious for dinner.
- Word sense
  - I went to the bank to deposit my check.
  - I went to the bank to look out at the river

Ambiguity

At the syntactic level

- PP Attachment ambiguity
  - I saw the man on the hill with the telescope
- Structural ambiguity
  - I cooked her duck
  - Visiting relatives can be annoying
  - Time flies like an arrow

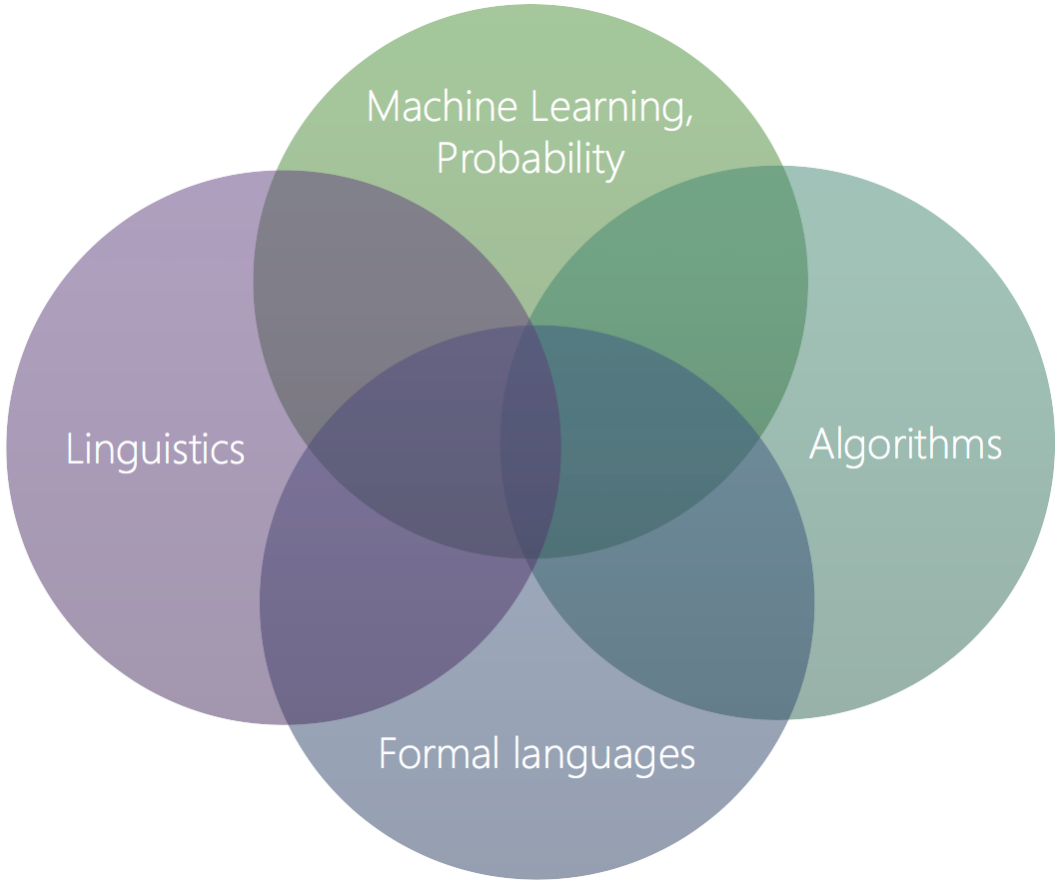
Ambiguity

- Quantifier scope
  - Everyone on the island speaks two languages.

- Hard cases require world knowledge, understanding of speaker goals
  - The city council denied the demonstrators the permit because they advocated violence
  - The city council denied the demonstrators the permit because they feared violence

Ambiguity

- NLP challenge: how can we model ambiguity, and choose the correct analysis in context?
- Approach: learn from data



Word counts

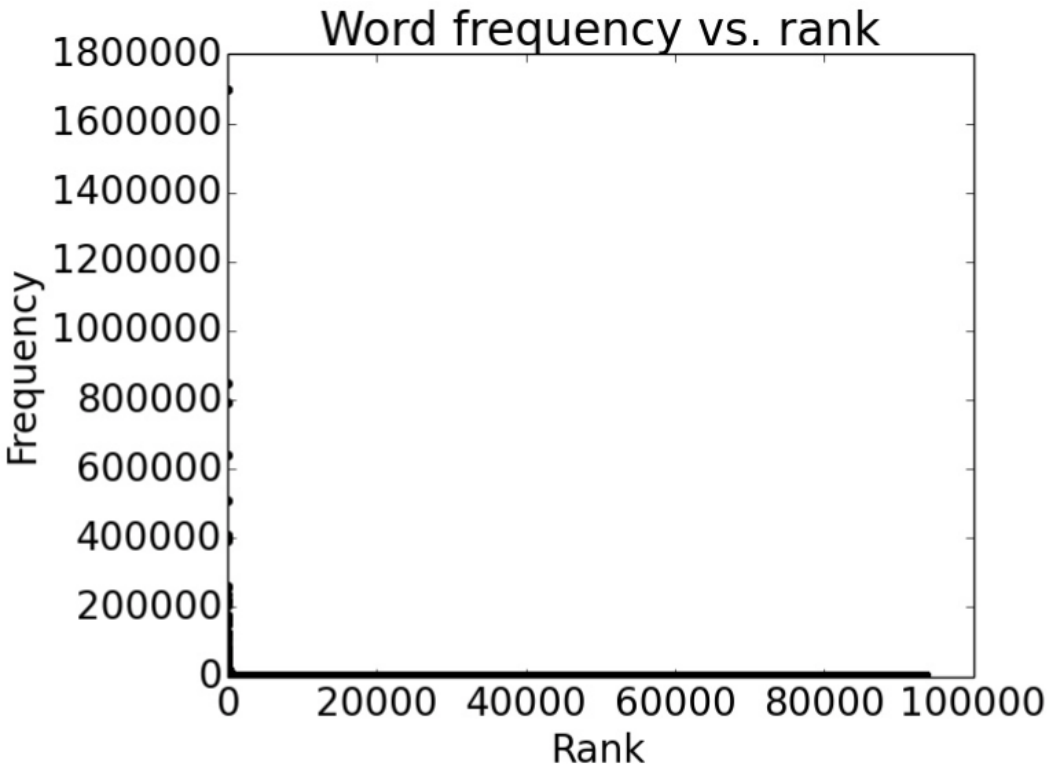
- Most frequent words in the English Europarl **corpus**
- (out of 24M word **tokens**)

any word		**nouns**	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States

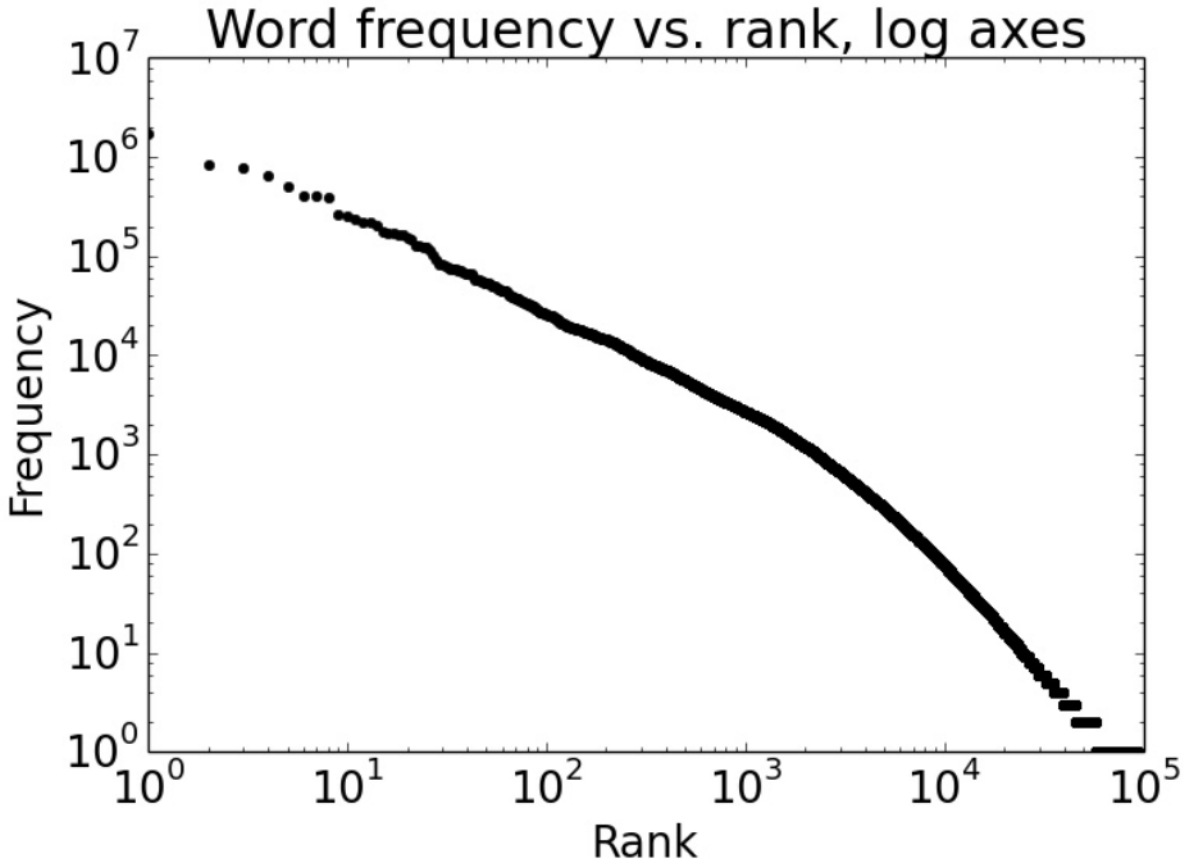
Word counts

- But also, out of the 93,638 distinct words (word **types**), 36,231 occur only once
    - cornflakes, mathematicians, fuzziness, jumbling
    - pseudo-rapporteur, lobby-ridden, perfunctorily,
    - Lycketoft, UNCITRAL, H-0695
    - policyfor, Commissioneris, 145.95, 27a

Plotting word frequencies



Plotting word frequencies (with log-log axes)



Zipf's law

$$f \times r \approx k$$

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

Zipf's law: implications

- Even in a very large corpus, there will be a lot of infrequent words
- The same holds for many other levels of linguistic structure
- Core NLP challenge: we need to estimate probabilities or to be able to make predictions for things we have rarely or never seen

Variation and Expressivity

- The same meaning can be expressed with different forms

  - I saw the man
  - The man was seen by me

  - She needed to make a quick decision in that situation
  - The scenario required her to make a split-second judgment

6,800 living languages

600 with written tradition

100 spoken by 95% of population

Social Impact

- NLP experiments and applications can have a direct effect on individual users' lives

- Some issues
  - Privacy
  - Exclusion
  - Overgeneralization
  - Dual-use problems

[Hovy & Spruit ACL 2016]

Today

- Levels of linguistic analysis in NLP
  - Morphology, syntax, semantics, discourse
- Why is NLP hard?
  - Ambiguity
  - Sparse data
    - Zipf's law, corpus, word types and tokens
    - Variation and expressivity
  - Social Impact

Course Logistics

http://www.cs.umd.edu/class/fall2017/cmsc723/

Before next class

- Read the syllabus
- Make sure you have access to piazza
- Get started on homework 1 – due Thursday Sep 7 by 12pm.