Natural Language Processing

Introduction to NLP

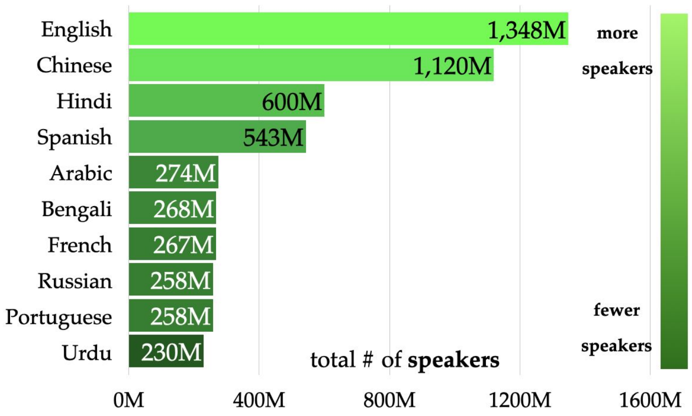
Sofia Serrano
sofias6@cs.washington.edu

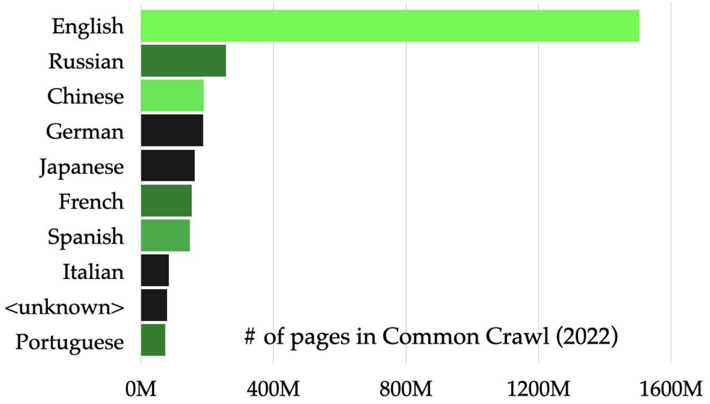
Credit to Yulia Tsvetkov and Noah Smith for slides

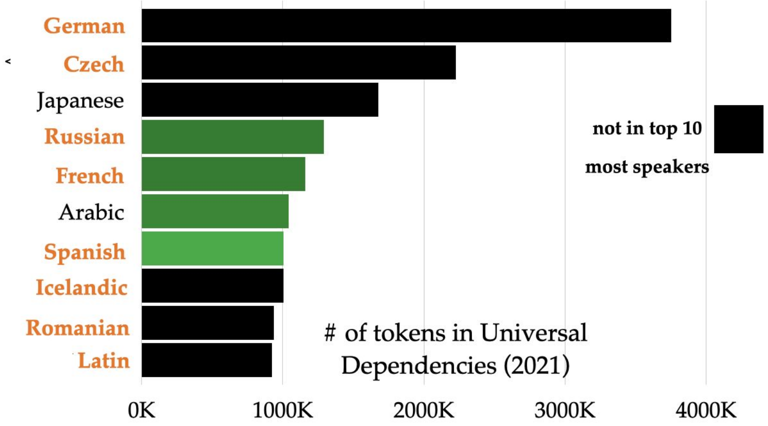
Announcements

- Academic Integrity Form is out on Canvas
- A1 is out on GitLab
  - Don't see it? Reply to this thread on Ed with your NetID: https://edstem.org/us/courses/32306/discussion/2365366
- Access to lecture recordings
  - No @cs.washington.edu google account? Click through to (request) access any lecture recording sooner rather than later so that we can give you access
- Make sure you can access the course machines
  - *(if connecting from off campus)* Run Husky OnNet VPN OR first ssh into an attu machine
  - `ssh yourNetID@nlpg00.cs.washington.edu` *(nlpg00-nlpg03)*
  - Not working?
    - Not a CSE major/no CSE account? Email ugrad-adviser@cs.washington.edu to request a CSE account (include your student ID number in the email) and CC Sofia
    - Still not working? Reply to this thread on Ed so that we can help troubleshoot: https://edstem.org/us/courses/32306/discussion/2368995

Following up on a question from last lecture



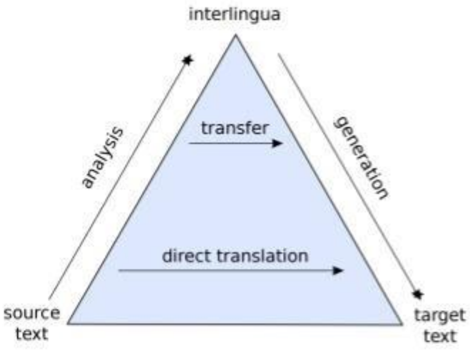




Credit to [Phoebe Mulcaire] for figures

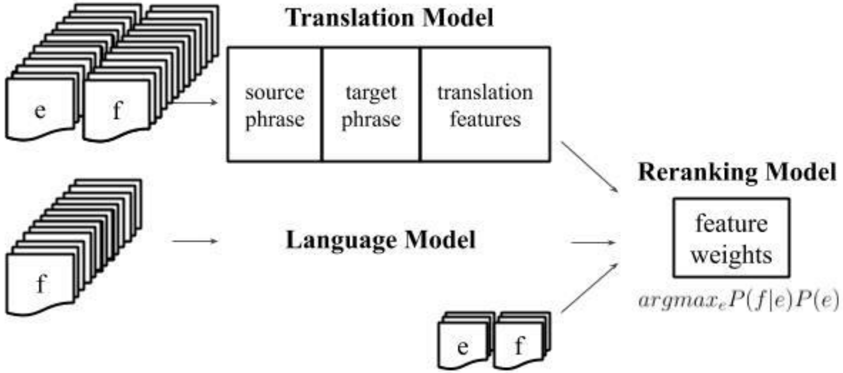
Symbolic and Probabilistic NLP

Logic-based/Rule-based NLP



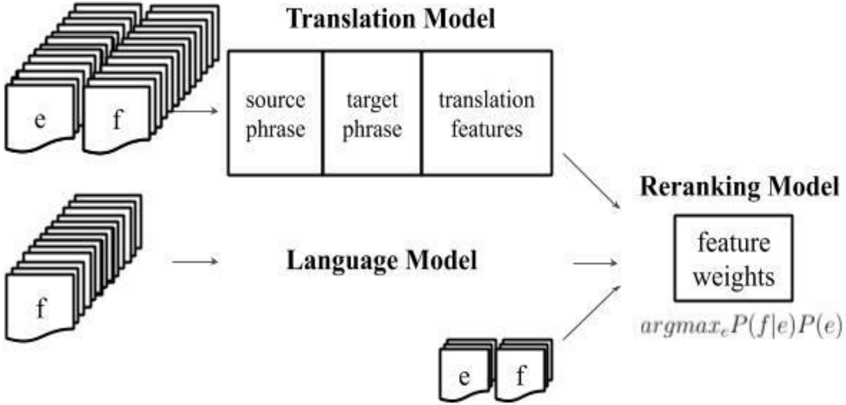
~ 90s

Statistical NLP



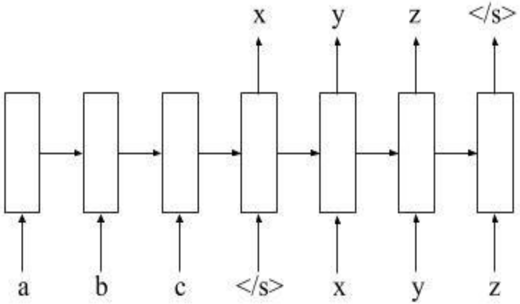
Probabilistic and Connectionist NLP

Engineered Features/Representations



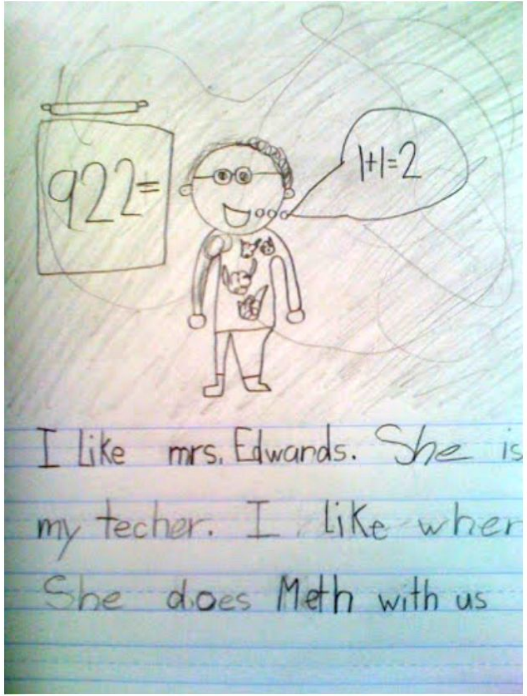
~mid 2010s

Learned Features/Representations



Linguistic Background

What does it mean to “know” a language?



(Thanks Canadian Internet Registration Authority!)

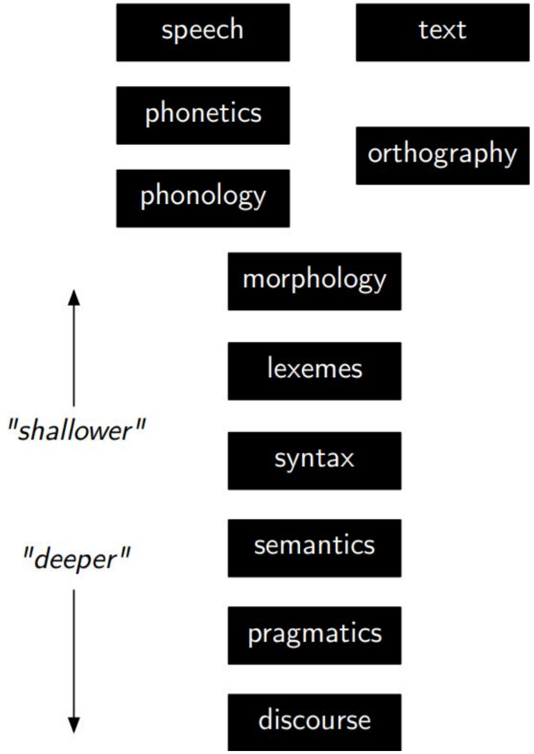
```
Last login: Mon Jan  9 08:08:57 2023 from 97.1
[sofias6@attu8 ~]$ wc myfile.txt
```

What do we need to “tell” a computer program so that it knows more English than `wc` or a dictionary, maybe even as much as a three-year-old, for example?

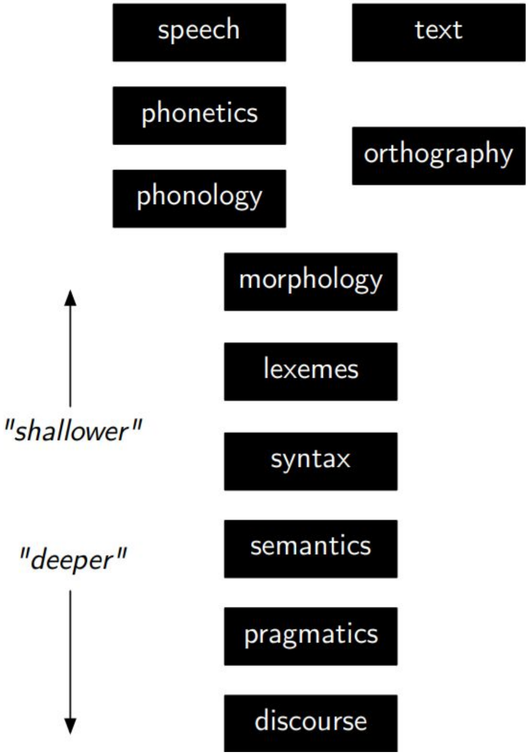
What does an NLP system need to 'know'?

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

Levels of linguistic knowledge



Speech, phonetics, phonology



Orthography

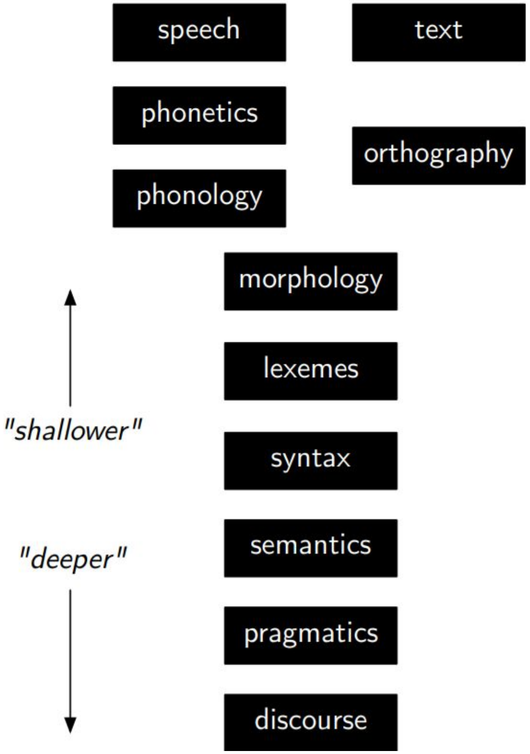
هذه جملة بسيطة

đây là một câu đơn giản

यह एक साधारण वाक्य है

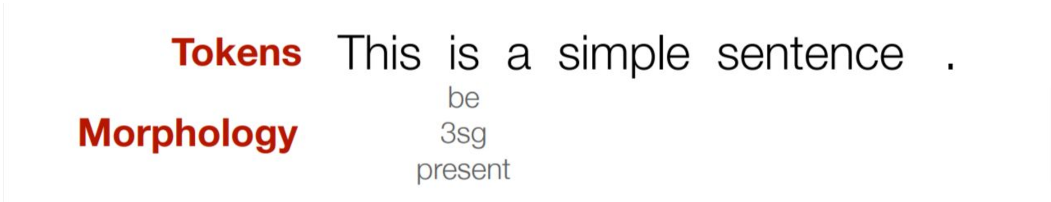
This is a simple sentence .

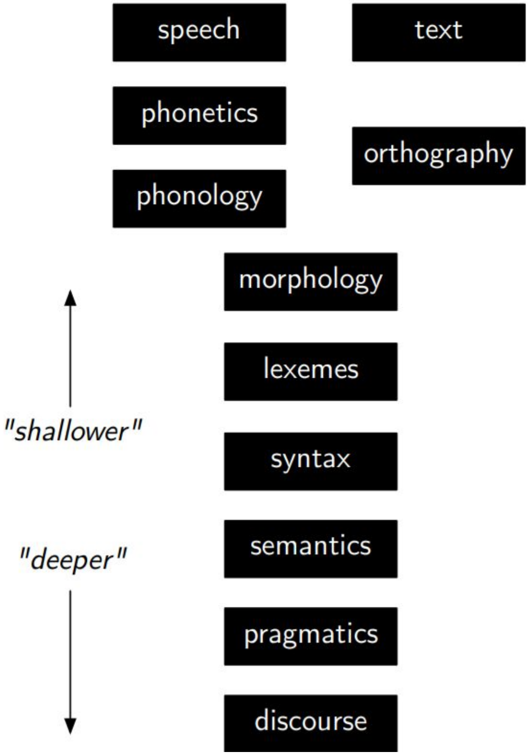
/ ðɪs ɪz ə ˈsɪmpl ˈsɛntəns /.



Words, morphology

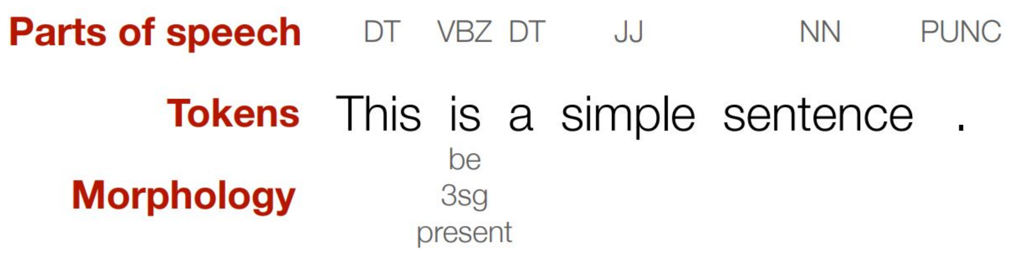
- Morphological analysis
- Tokenization
- Lemmatization

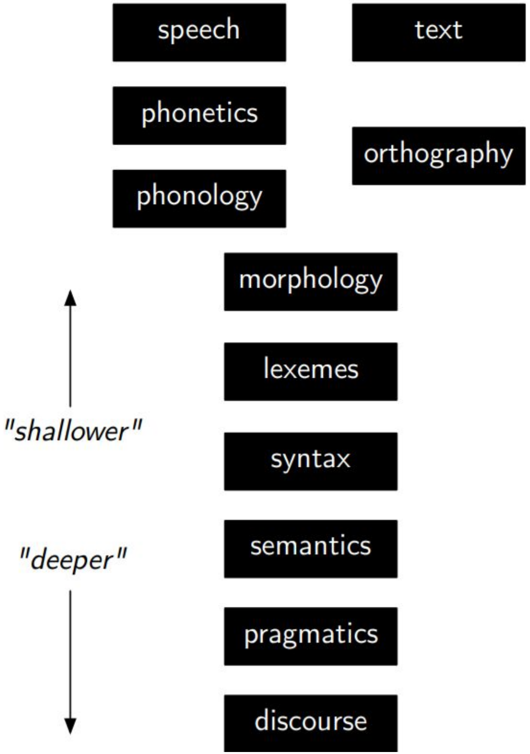




Syntax

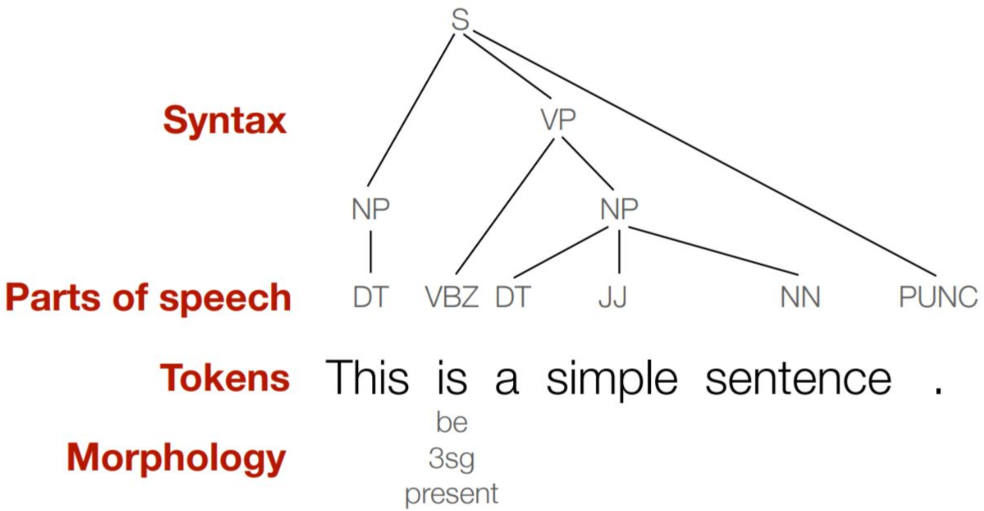
- Part-of-speech tagging

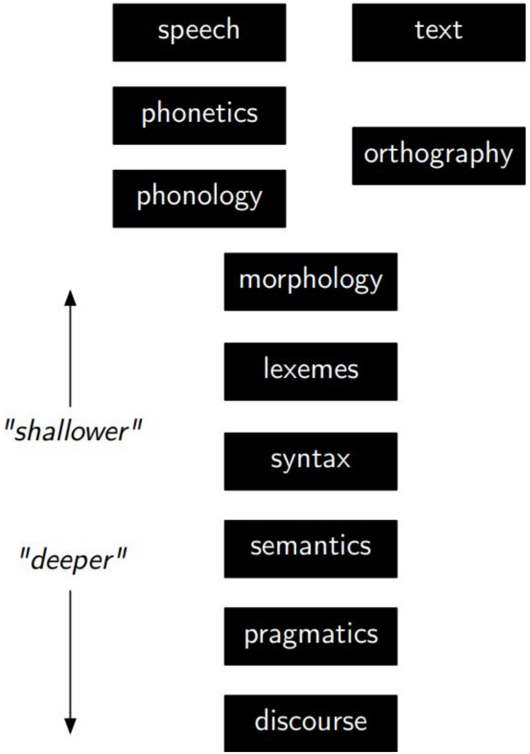




Syntax

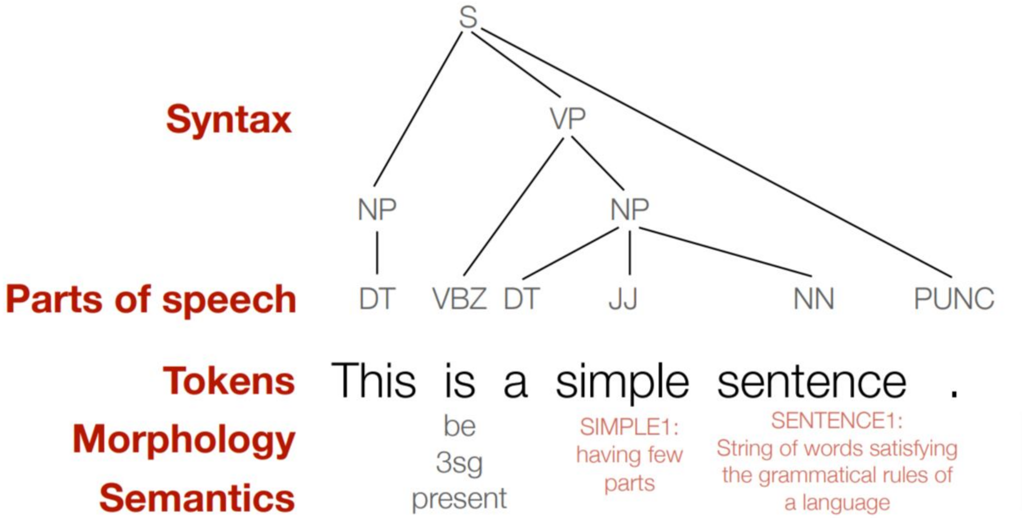
- Part-of-speech tagging
- Syntactic parsing

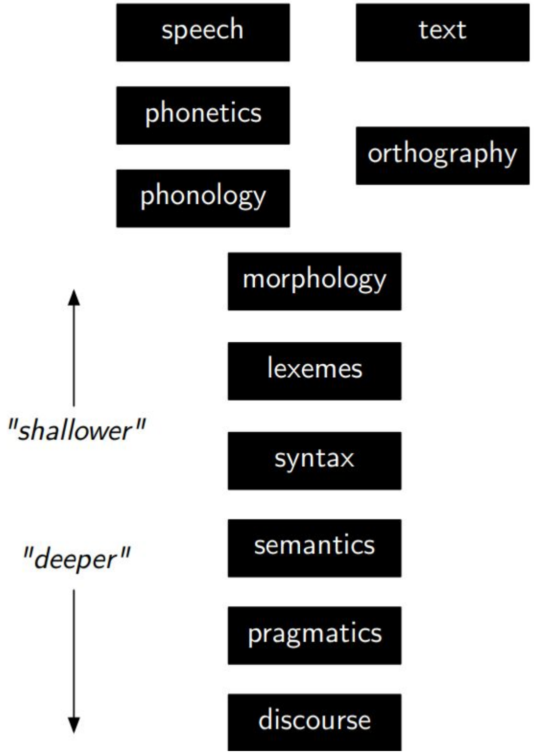




Semantics

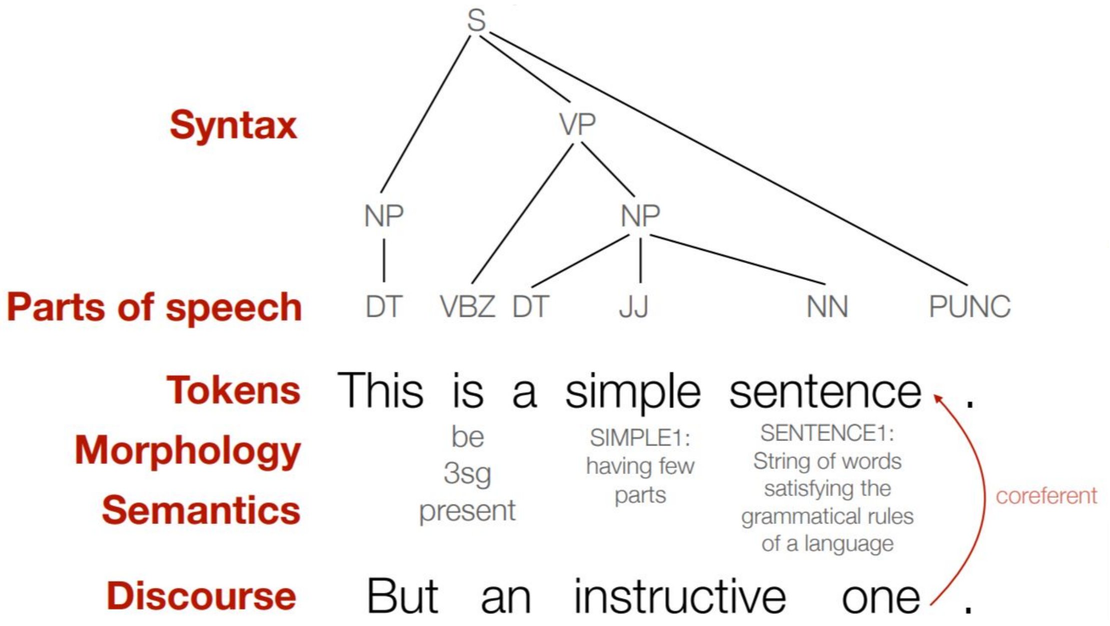
- Named entity recognition
- Word sense disambiguation
- Semantic role labeling

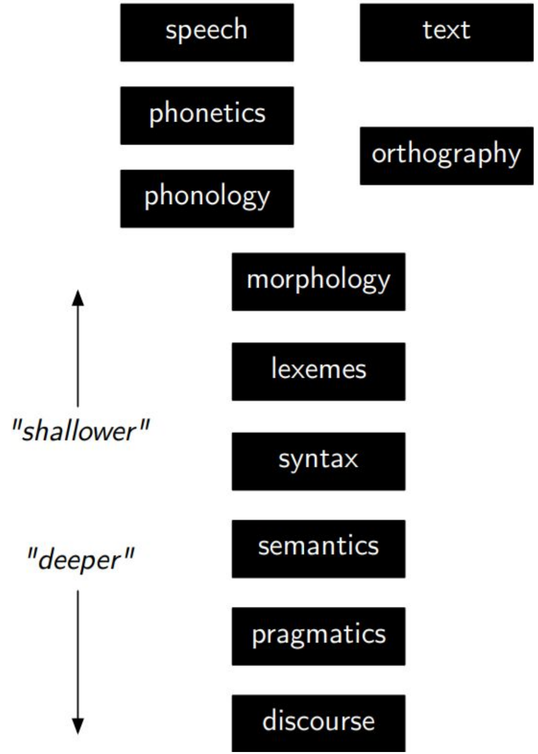




Discourse

- Reference resolution
- Discourse parsing



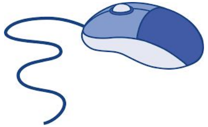


Linguistic challenges we'll need to deal with in designing NLP systems

What are some challenges for NLP systems?

1. Ambiguity
2. Variation
3. Sparsity
4. Expressivity
5. Unmodeled variables
6. Unknown representation $\mathcal{R}$

Ambiguity

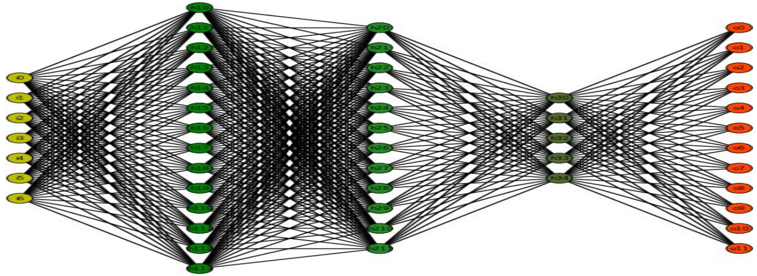



- Ambiguity at multiple levels:
  - Word senses: **bank** (finance or river?)
  - Part of speech: **chair** (noun or verb?)
  - Syntactic structure: **I can see a man with a telescope**
  - Multiple: **I saw her duck**



Dealing with ambiguity

- How can we model ambiguity and choose the correct analysis in context?
    - non-probabilistic methods (FSMs for morphology, CKY parsers for syntax) return *all possible analyses.*
    - probabilistic models (HMMs for part-of-speech tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return *the best possible analysis*, i.e., the most probable one according to the model
    - Neural networks, pretrained language models now provide end-to-end solutions



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

Corpora

- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
- Examples
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - Yelp reviews
  - The Web: billions of words of who knows what

What are some challenges for NLP systems?

1. Ambiguity
2. Variation
3. Sparsity
4. Expressivity
5. Unmodeled variables
6. Unknown representation $\mathcal{R}$

Scale

- ~7K languages
- Thousands of language varieties

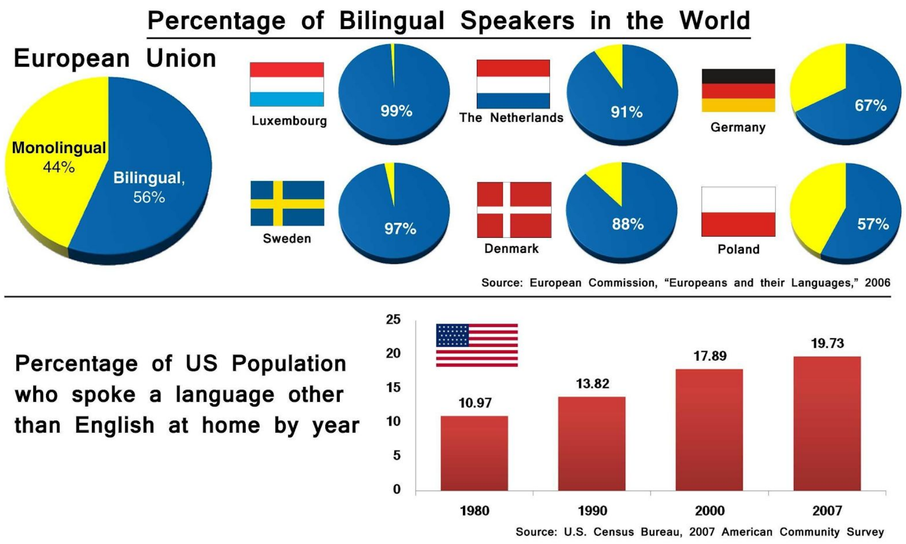
Englishes

Africa is a continent with a very high linguistic diversity: there are an estimated 1.5-2K African languages from 6 language families. 1.33 billion people

NLP beyond English

- ~7,000 languages
- thousands of language varieties

Most of the world today is multilingual

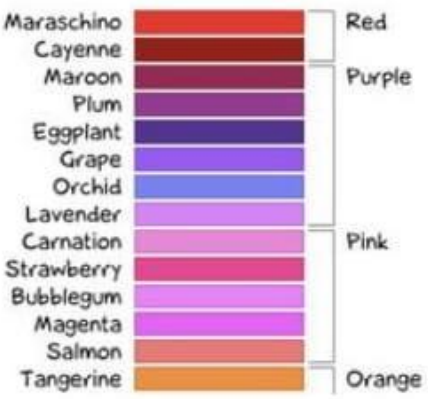


Source: US Census Bureau

Source: Ethnologue

Semantic analysis

- Every language represents the world in a different way
    - For example, it could depend on cultural or historical conditions



- Russian has very few words for colors, Japanese has hundreds
- Multiword expressions, e.g. happy as a clam, it’s raining cats and dogs or wake up and metaphors, e.g. love is a journey are very different across languages

Tokenization

这是一个简单的句子

WORDS

This is a simple sentence

זה משפט פשוט

Tokenization + disambiguation

in tea	בתה
her daughter	

- most of the vowels unspecified

in tea	בתה
in the tea	בהתה
that in tea	שבתה
that in the tea	שבהתה
and that in the tea	ושבהתה

ושבתה

and her saturday	ו+שבת+ה
and that in tea	ו+ש+ב+תה
and that her daughter	ו+ש+בת+ה

- most of the vowels unspecified
- particles, prepositions, the definite article, conjunctions attach to the words which follow them
- tokenization is highly ambiguous

Tokenization + morphological analysis

- Quechua

Much’ananayakapushasqakupuniñataqsunamá

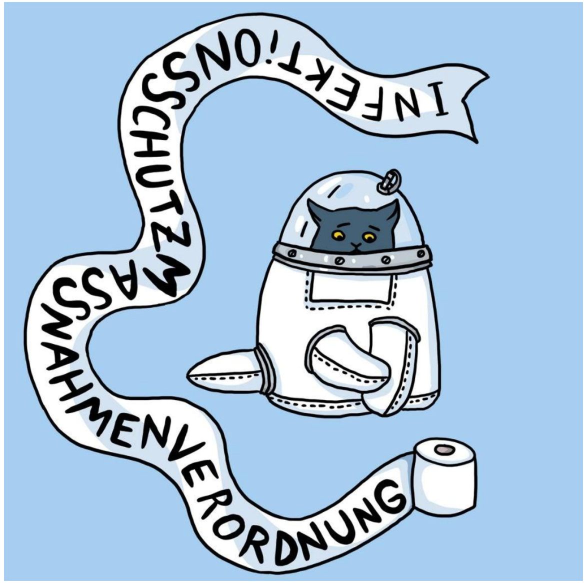
Much’a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

“So they really always have been kissing each other then”

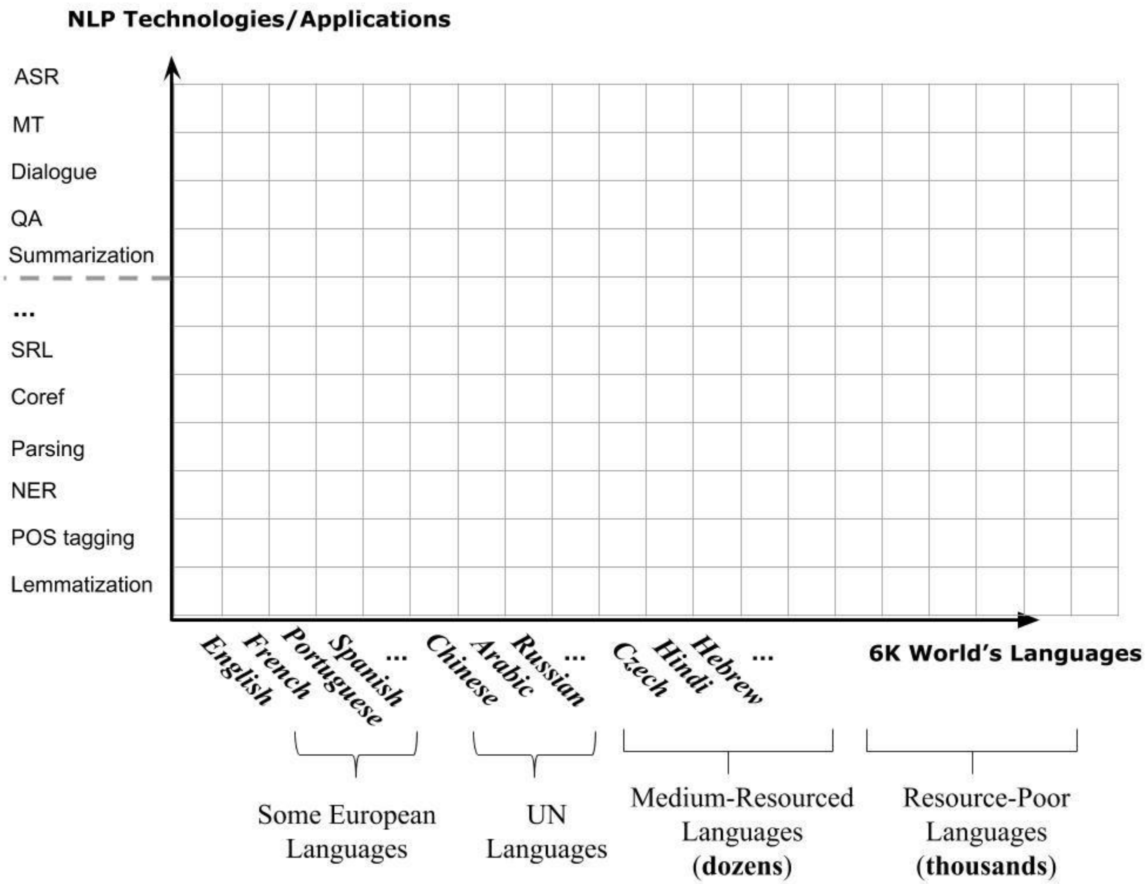
Much’a	to kiss
-na	expresses obligation, lost in translation
-naya	expresses desire
-ka	diminutive
-pu	reflexive (kiss *eachother*)
-sha	progressive (kiss*ing*)
-sqa	declaring something the speaker has not personally witnessed
-ku	3rd person plural (they kiss)
-puni	definitive (really*)
-ña	always
-taq	statement of contrast (...then)
-suna	expressing uncertainty (So...)
-má	expressing that the speaker is surprised

Tokenization + morphological analysis

- German



Infektionsschutzmaßnahmenverordnung


NLP Technologies/Applications
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
English
French
Portuguese
Spanish
...
Chinese
Arabic
Russian
...
Czech
Hindi
Hebrew
...
6K World's Languages
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)

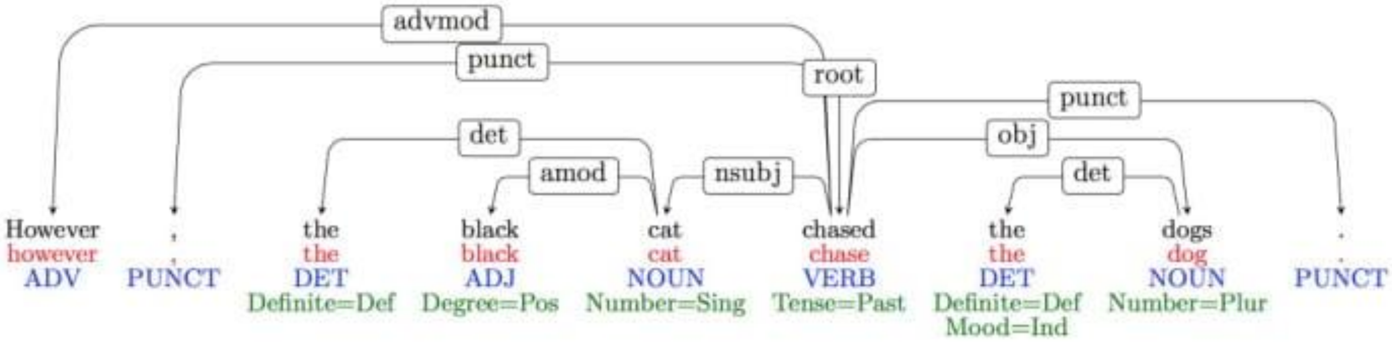
Linguistic variation

- Non-standard language, emojis, hashtags, names

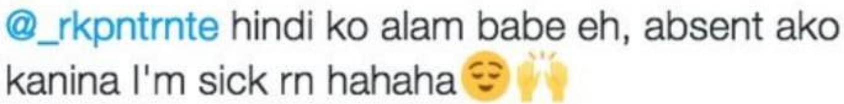
chowdownwithchan #crab and #pork #xiaolongbao at @dintaifungusa... where else? 😂🤷‍♀️ Note the cute little crab indicator in the 2nd pic 🦀💕

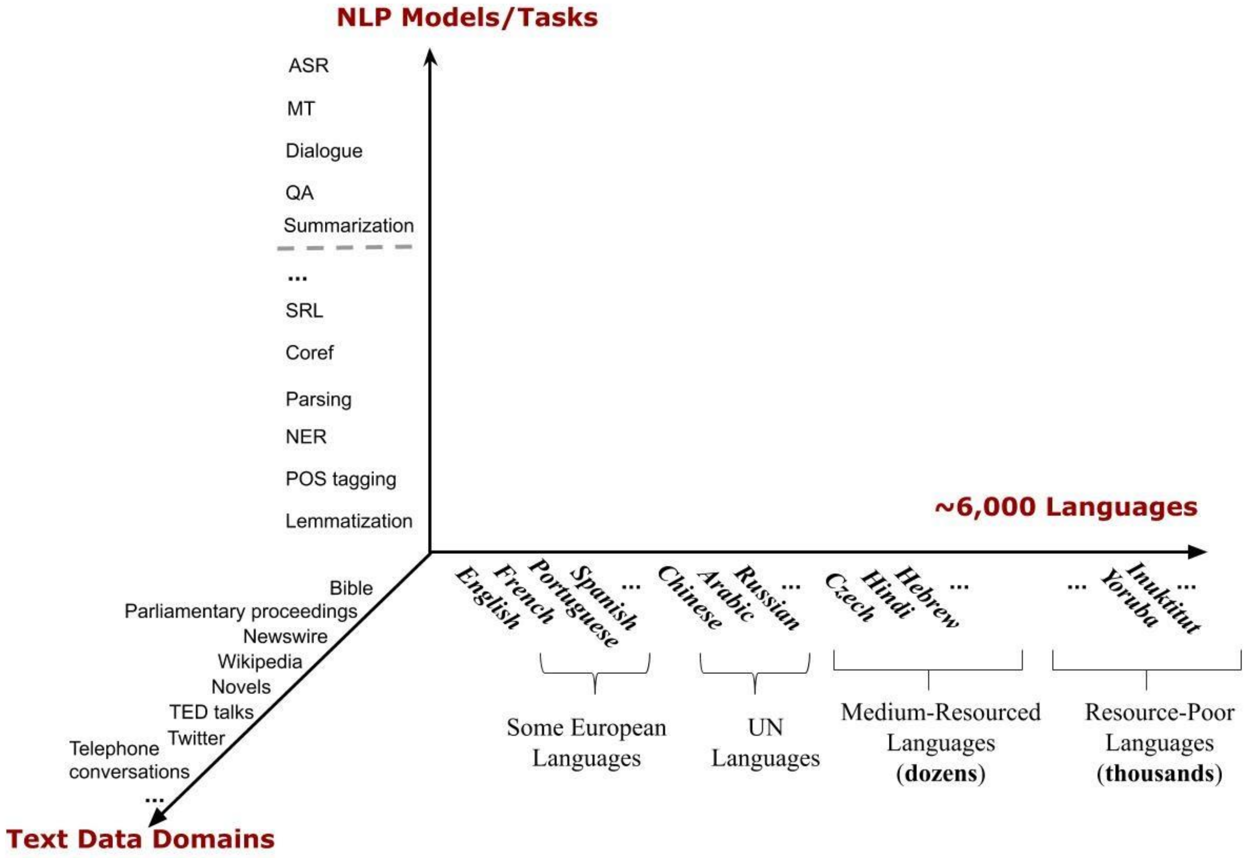
Variation

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



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




NLP Models/Tasks
ASR
MT
Dialogue
QA
Summarization
...
SRL
Coref
Parsing
NER
POS tagging
Lemmatization
~6,000 Languages
English
French
Portuguese
Spanish
...
Chinese
Arabic
Russian
...
Czech
Hindi
Hebrew
...
...
Yoruba
Inuktitut
...
Some European Languages
UN Languages
Medium-Resourced Languages (dozens)
Resource-Poor Languages (thousands)
Bible
Parliamentary proceedings
Newswire
Wikipedia
Novels
TED talks
Twitter
Telephone conversations
...
Text Data Domains

What are some challenges for NLP systems?

1. Ambiguity
2. Variation
3. Sparsity
4. Expressivity
5. Unmodeled variables
6. Unknown representation $\mathcal{R}$

Sparsity

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

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

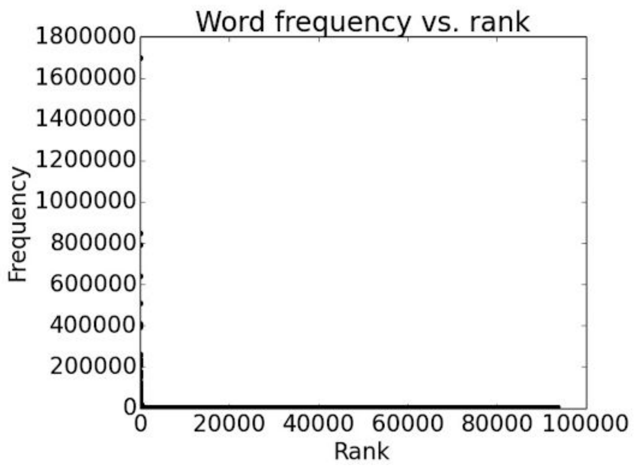
But also, out of 93,638 distinct words (word types), 36,231 occur only once.

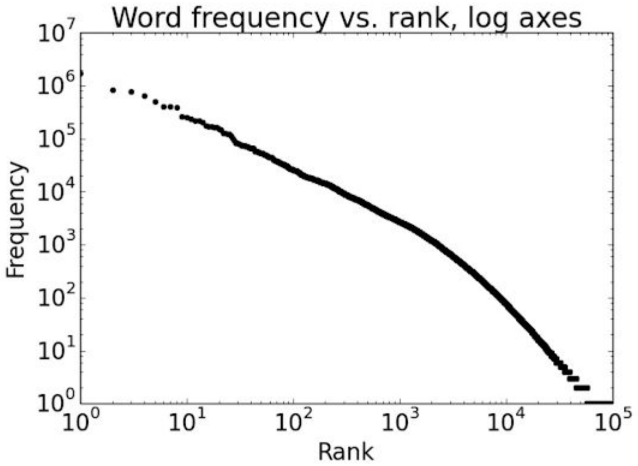
Examples:

- 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?

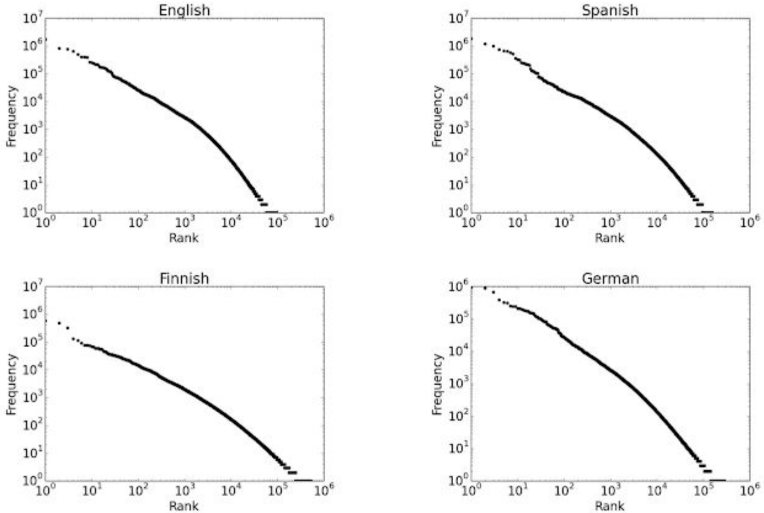




Zipf’s Law

Implications

- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen



What are some challenges for NLP systems?

1. Ambiguity
2. Variation
3. Sparsity
4. Expressivity
5. Unmodeled variables
6. Unknown representation $\mathcal{R}$

Expressivity

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

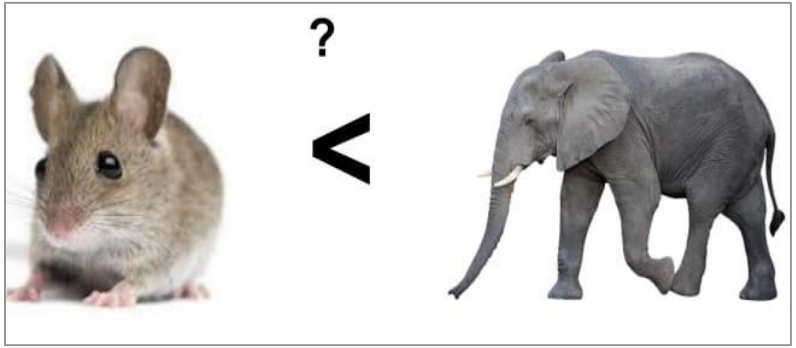
She gave the book to Tom	vs.	She gave Tom the book
Some kids popped by	vs.	A few children visited
Is that window still open?	vs.	Please close the window

What are some challenges for NLP systems?

1. Ambiguity
2. Variation
3. Sparsity
4. Expressivity
5. Unmodeled variables
6. Unknown representation $\mathcal{R}$

Unmodeled variables

“Drink this milk”



World knowledge

- I dropped the glass on the floor and it broke
- I dropped the hammer on the glass and it broke

What are some challenges for NLP systems?

1. Ambiguity
2. Variation
3. Sparsity
4. Expressivity
5. Unmodeled variables
6. Unknown representation $\mathcal{R}$

Unknown representation

- Very difficult to decide on a representation $\mathcal{R}$, 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?

Desiderata for NLP models

- Sensitivity to a wide range of phenomena and constraints in human language
- Generality across languages, modalities, genres, styles
- Computational efficiency at construction time and runtime
- Strong formal guarantees (e.g., convergence, statistical efficiency, consistency)
- High accuracy when judged against expert annotations and/or test data specific to a particular task
- Explainable to human users
- Ethical

Next class

- Text classification

Questions?