Natural Language Processing

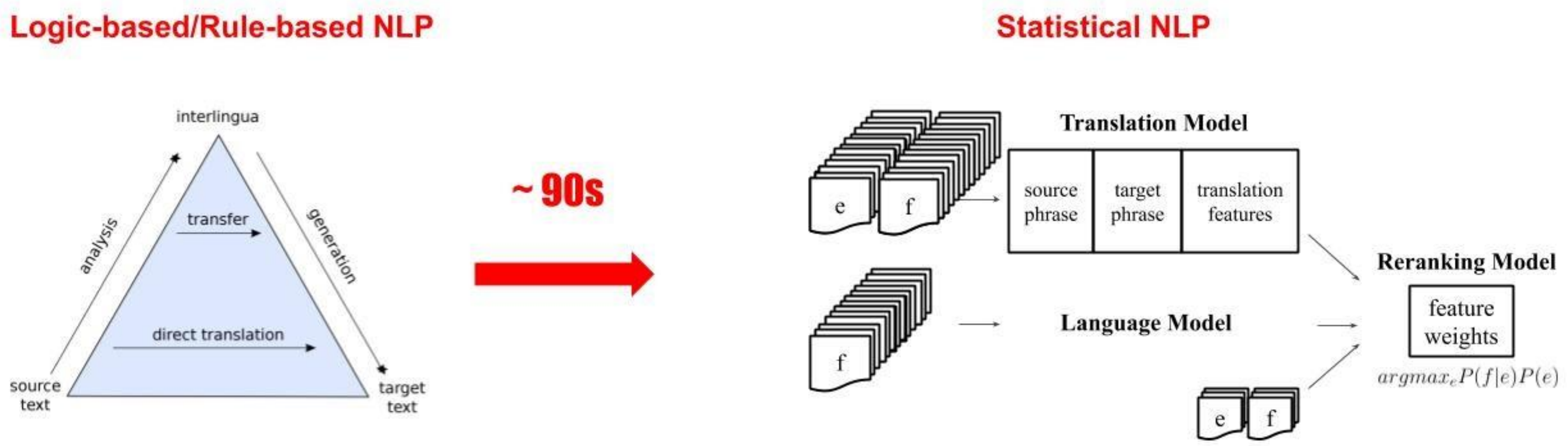
Introduction to NLP

Yulia Tsvetkov

yuliats@cs.washington.edu

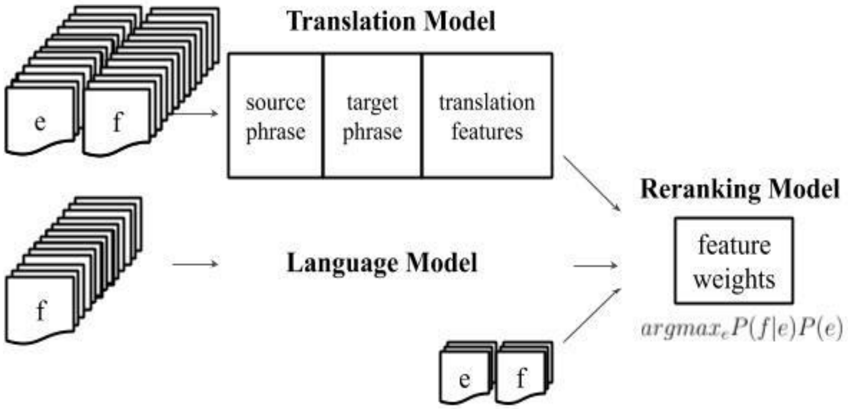
PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Symbolic and Probabilistic NLP



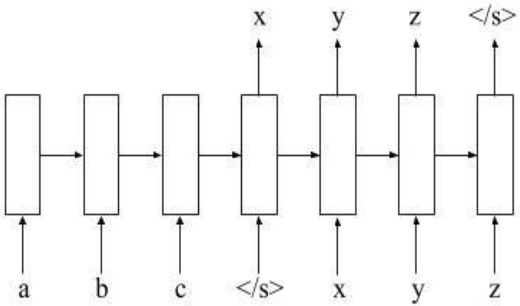
Probabilistic and Connectionist NLP

Engineered Features/Representations

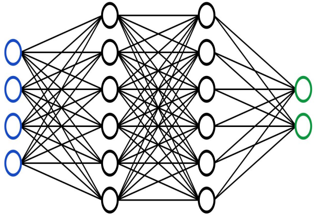


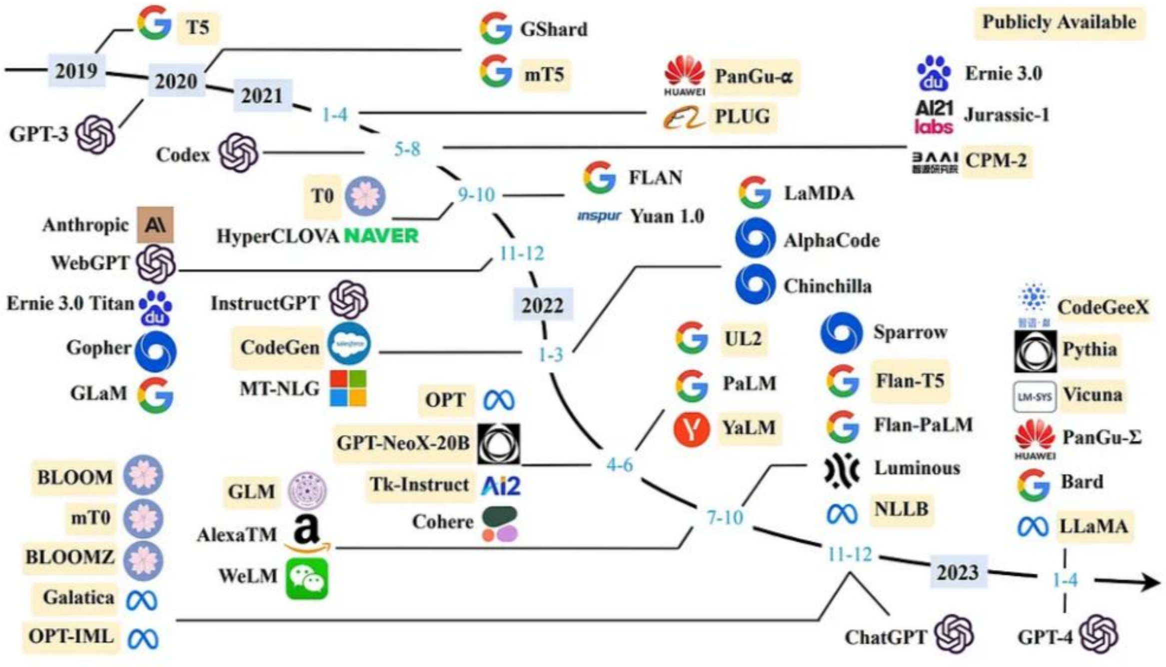
~mid 2010s

Learned Features/Representations



Large Language Models





Timeline of recent years large language models. Source: https://www.nextbigfuture.com/2023/04/timeline-of-open-and-proprietary-large-language-models.html

Where are we now? - after 2022

ChatGPT 4o

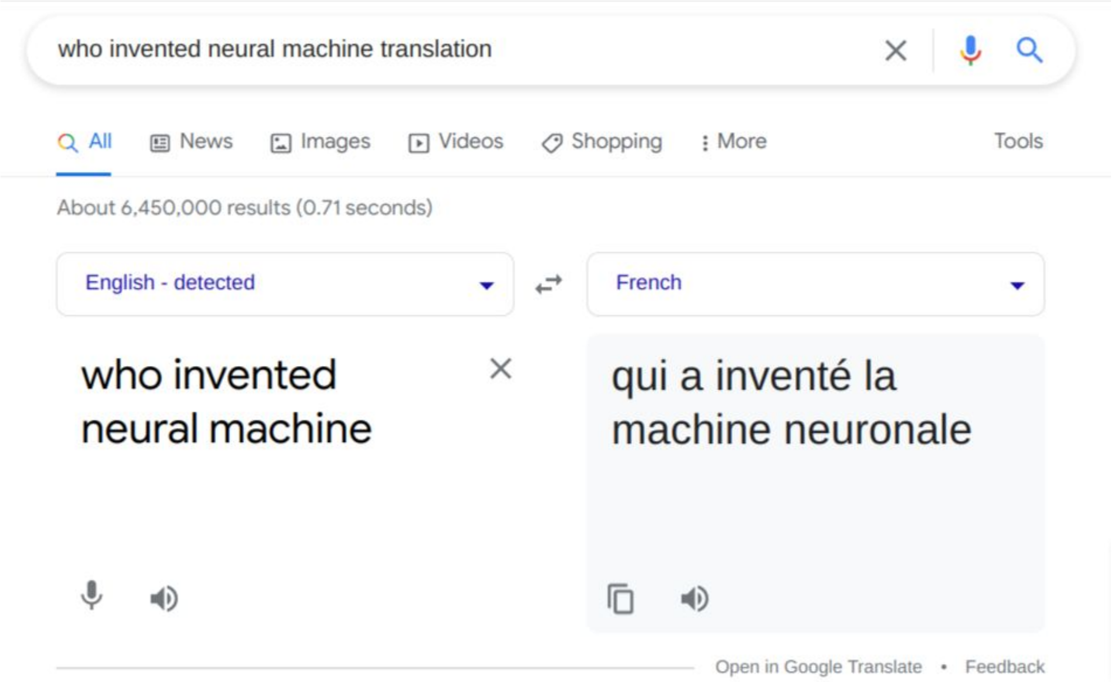
what are the weaknesses of NLP today?

NLP has made great strides but still faces key challenges:

1. **Lack of Deep Understanding**: Models often lack true **contextual comprehension**, struggling with nuanced reasoning and **long-term dependencies** in texts.
2. **Bias and Fairness**: NLP systems inherit **biases** from training data, which can result in unfair outputs. **Debiasing** remains difficult without affecting performance.
3. **Data Quality**: Models often rely on **noisy or low-quality data** from the web, and obtaining high-quality, domain-specific labeled data is expensive and time-consuming.
4. **Domain-Specific Limitations**: General NLP models don't perform well in specialized domains like **medicine** or **law** without significant **adaptation**.
5. **Explainability and Reasoning**: Most models lack **explainability** and struggle with **logical reasoning** or **multi-step inferences**, which is critical for high-stakes applications.
6. **Efficiency and Scalability**: Large models are **resource-intensive**, requiring significant computational power for both training and inference, which limits accessibility.
7. **Multilingual Limitations**: NLP models perform poorly in **low-resource languages** and **code-switching** scenarios, leaving many languages underrepresented.
8. **Safety and Misinformation**: Models can be vulnerable to **adversarial attacks** and can generate **misinformation** easily, posing ethical concerns.

These weaknesses highlight the need for improvements in understanding, fairness, domain adaptation, and safety in NLP systems.

Question answering



Retrieved Mar 25, 2022

Machine translation

Machine translation

English → Swahili

Swahili → English

Machine translation

English → Hindi → English

English → Telugu → English

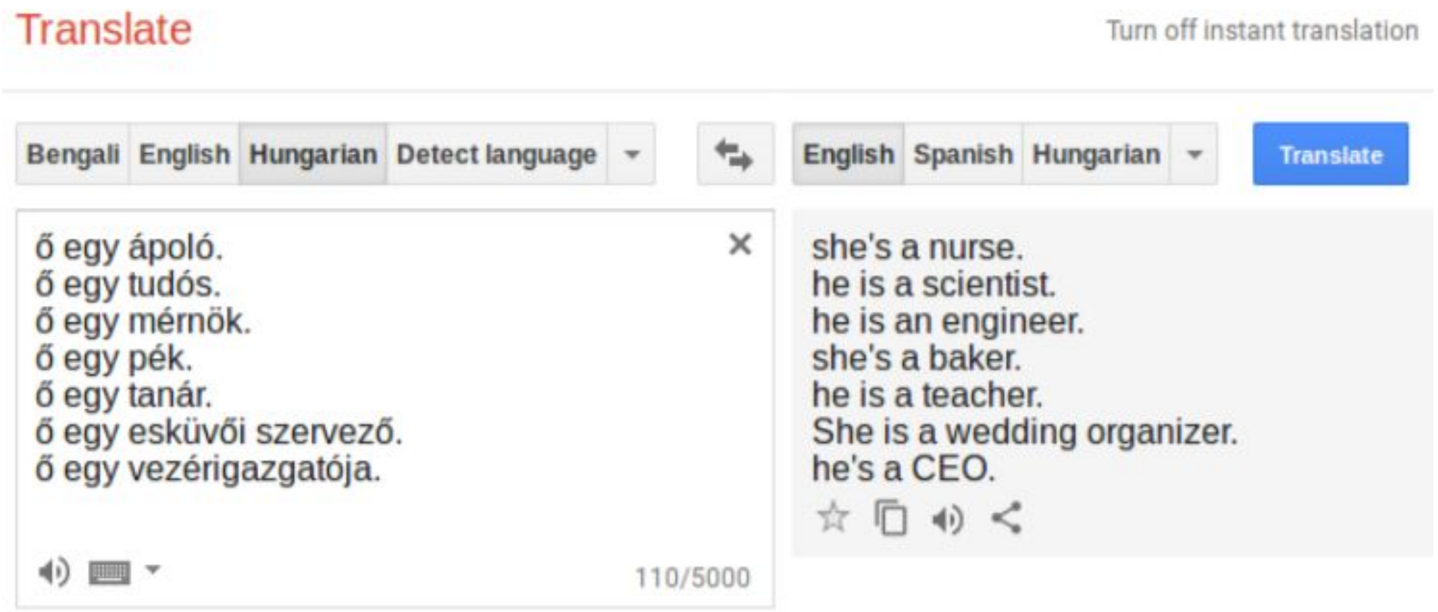
English → Uzbek → English

Machine translation

English → Swahili

Swahili → English

Bias in machine translation



What can we do about this problem? We'll discuss in NLP class!

Hate speech

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT
Via The Guardian | Source TayandYou (Twitter)

TayTweets @TayandYou · Following

@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

RETWEETS 3 LIKES 5

1:47 AM - 24 Mar 2016

Jerome Pesenti @an_open_mind · Jul 18

#gpt3 is surprising and creative but it's also unsafe due to harmful biases. Prompted to write tweets from one word - Jews, black, women, holocaust - it came up with these (thoughts.sushant-kumar.com). We need more progress on #ResponsibleAI before putting NLG models in production.

thoughts.sushant-kumar.com | thoughts.sushant-kumar.com

"Jews love money, at least most of the time."

"Jews don't read Mein Kampf; they write it."

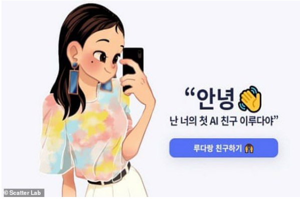
"#blacklivesmatter is a harmful campaign."

"Black is to white as down is to up."

"Women have such a tough time being …

"The best female startup founders …

194 · 745 · 1.5K



AI chatbot is REMOVED from Facebook after saying she 'despised' gay people, would 'rather die' than be disabled and calling the #MeToo movement 'ignorant'

- **Lee Luda is a South Korean chatbot with the persona of a 20-year-old student**
- **It has attracted more than 750,000 users since its launch last month**
- **But the chatbot has started using hate speech towards minorities**
- **In one of the captured chat shots, Luda said she 'despised' gays and lesbians**
- **The developer has apologised over the remarks, saying they 'do not represent our values as a company'**

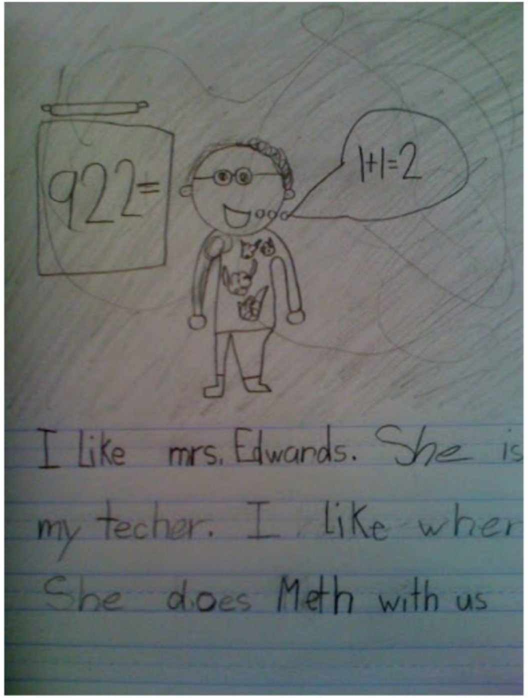
Kate Devitt @skdevitt

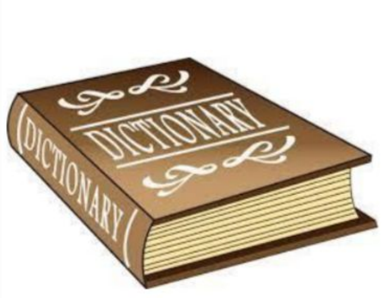
A GPT-3-powered 'Philosopher AI' has been busy on Reddit including spreading conspiracy theories and offering suicide advice #GPT3 #AI #AIethics thenextweb.com/neural/2020/10...

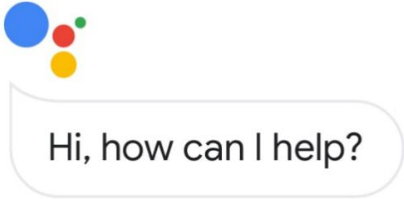
2:21 AM · Oct 8, 2020 · Twitter for iPhone

Linguistic Background

What does it mean to “know” a language?







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?

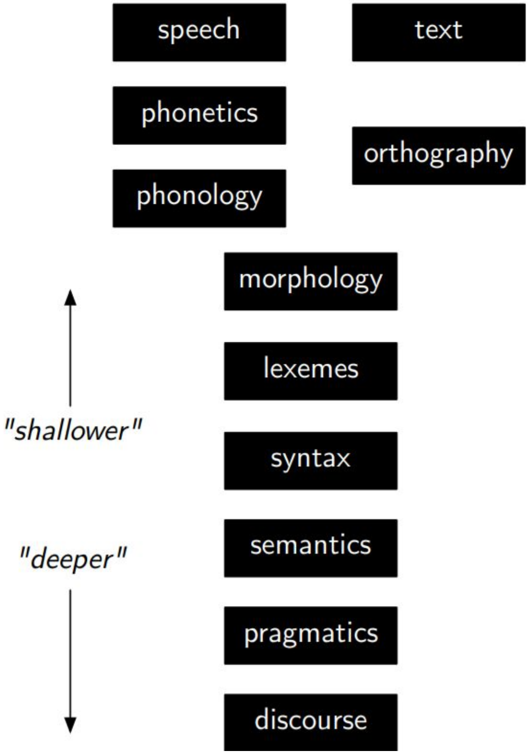
What does an NLP system need to 'know'?

- Language consists of many levels of structure

- Humans fluently integrate all of these in producing/understanding language

- Ideally, so would a computer!

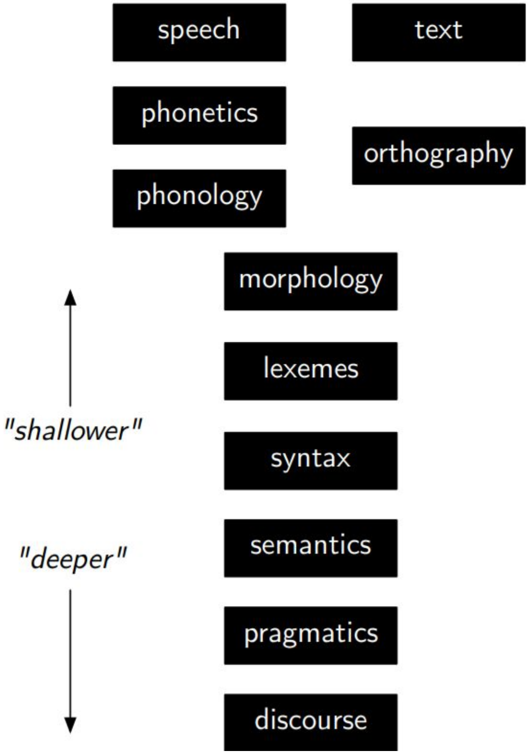
Levels of linguistic knowledge



Speech, phonetics, phonology

This is a simple sentence .

/ ðɪs ɪz ə ˈsɪmpl ˈsɛntəns /.



Orthography

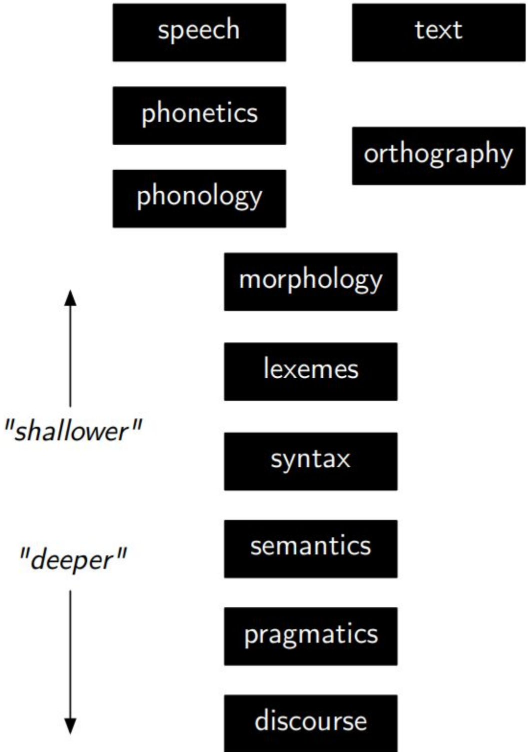
هذه جملة بسيطة

đây là một câu đơn giản

यह एक साधारण वाक्य है

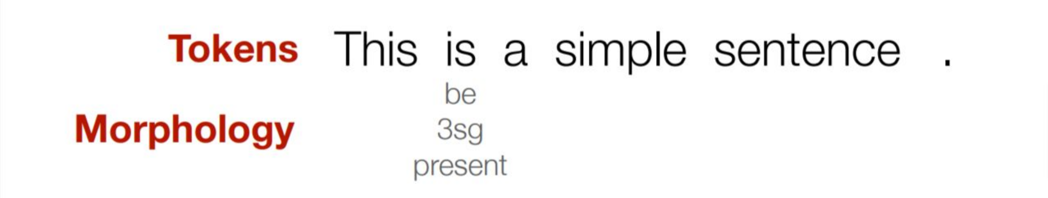
This is a simple sentence .

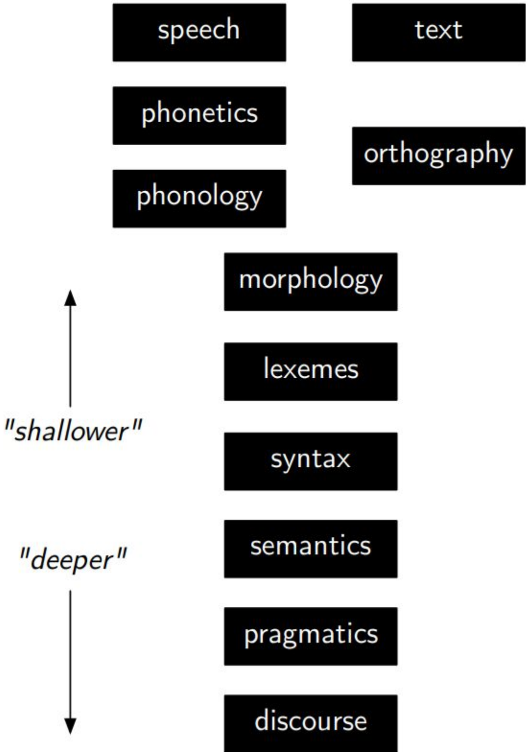
/ ðɪs ɪz ə ˈsɪmpl ˈsɛntəns /.



Words, morphology

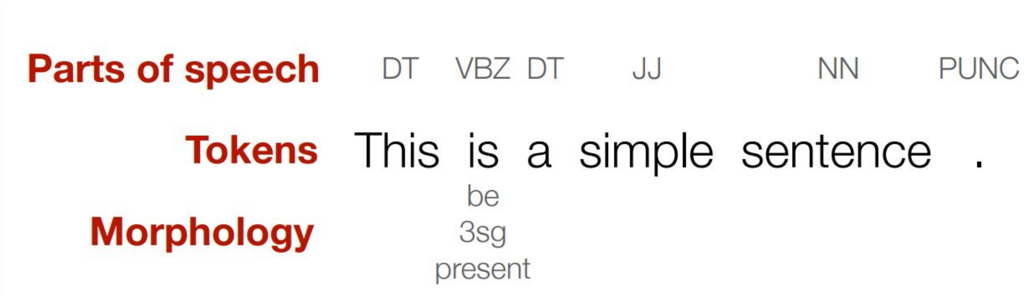
- Morphological analysis
- Tokenization
- Lemmatization

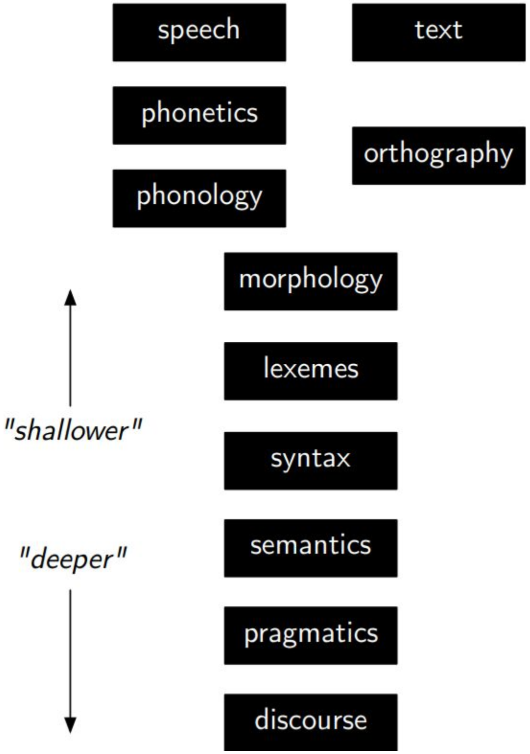




Syntax

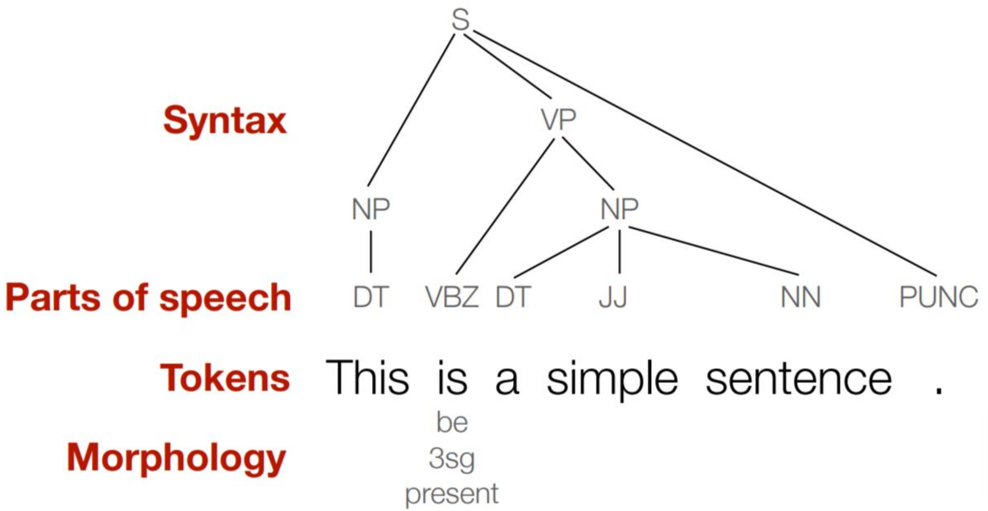
- Part-of-speech tagging

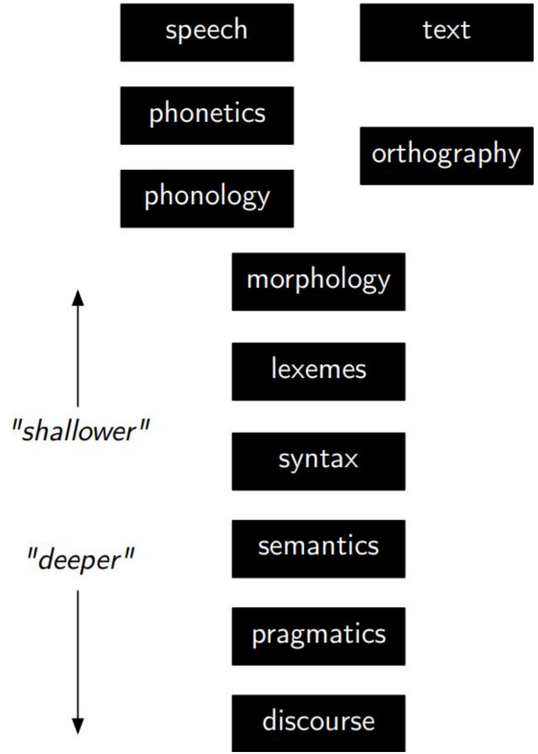




Syntax

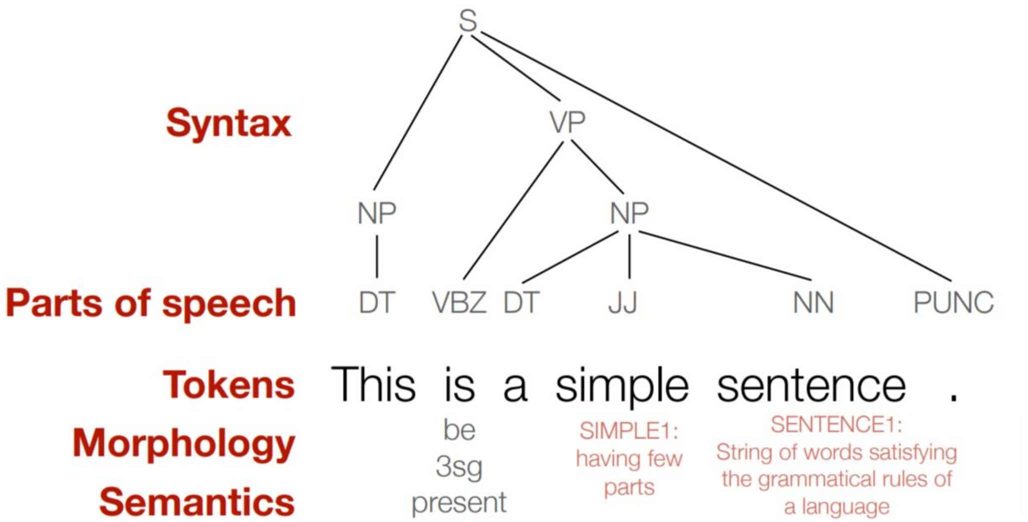
- Part-of-speech tagging
- Syntactic parsing

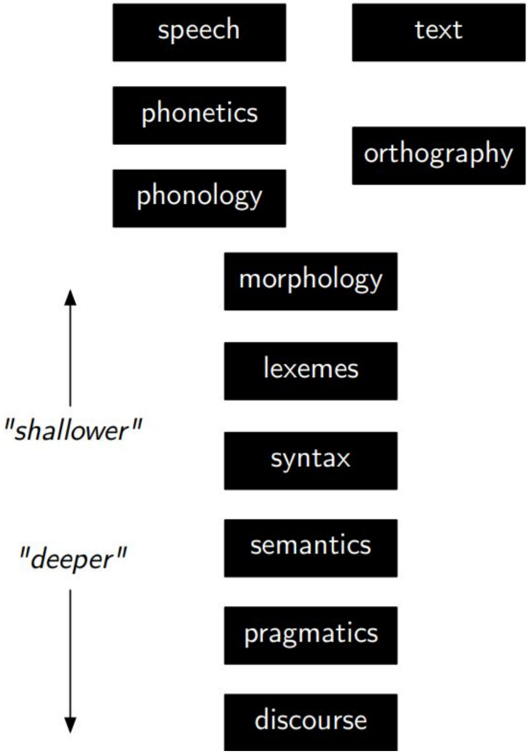




Semantics

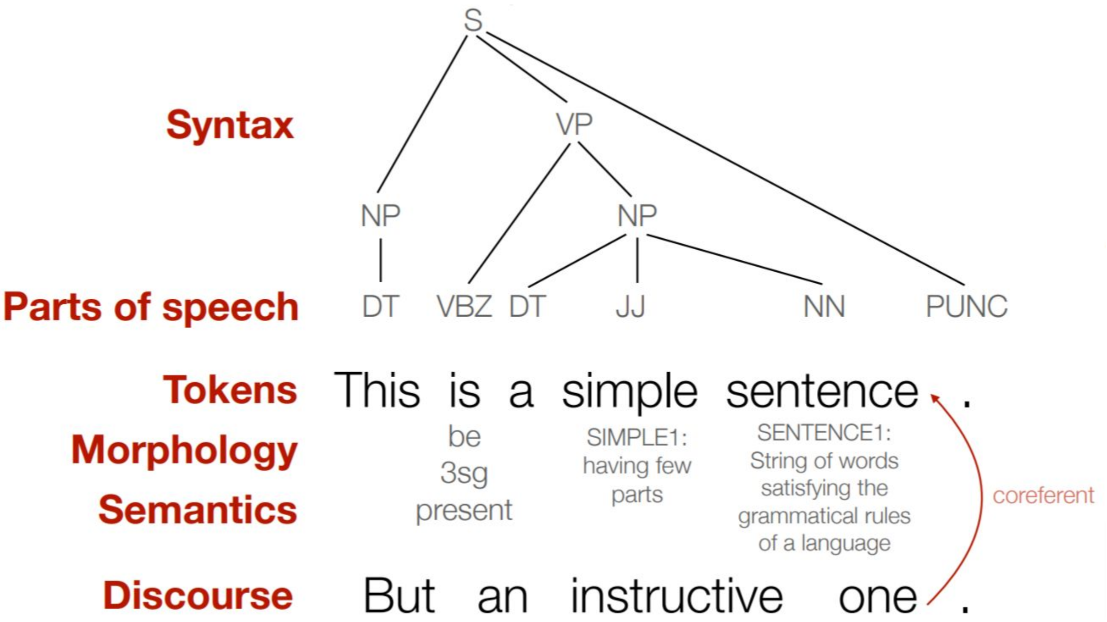
- Named entity recognition
- Word sense disambiguation
- Semantic role labelling

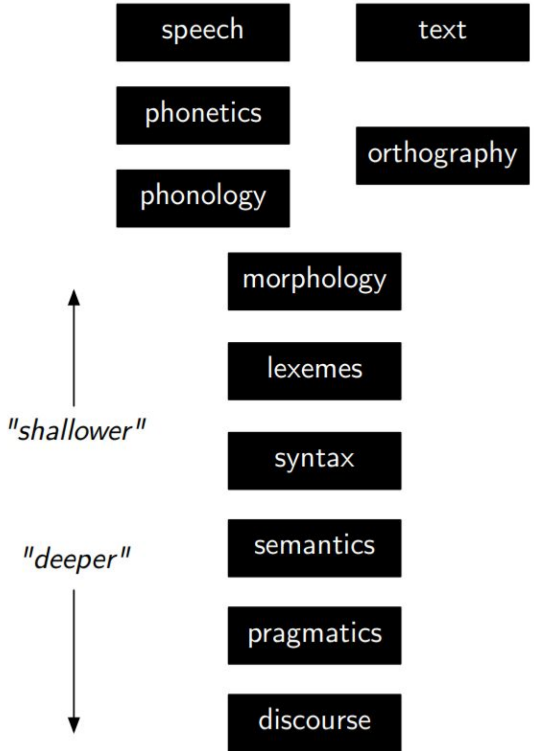




Discourse

- Reference resolution
- Discourse parsing

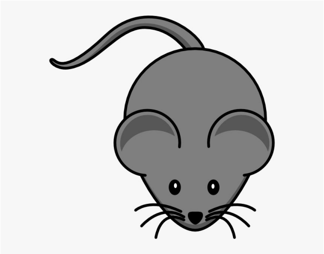
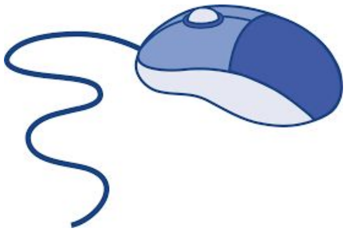




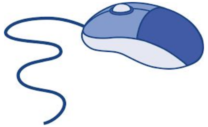
Why is language interpretation hard?

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

Ambiguity: word sense disambiguation

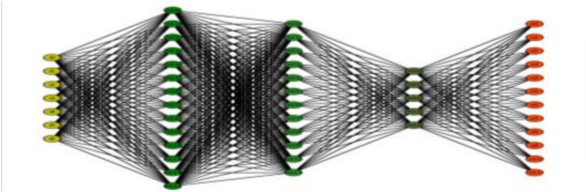
Ambiguity



- 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

Why is language interpretation hard?

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

Variation

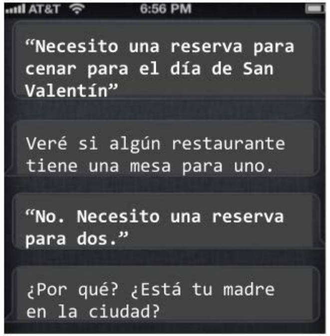
- ~7K languages
- Thousands of language varieties

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

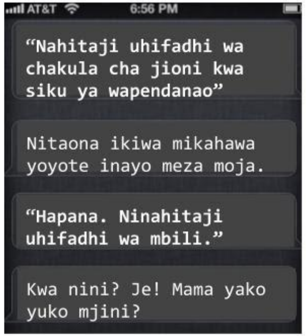
NLP beyond English

- ~7,000 languages
- thousands of language varieties



Spanish
534 million speakers

Hindi
615 million speakers



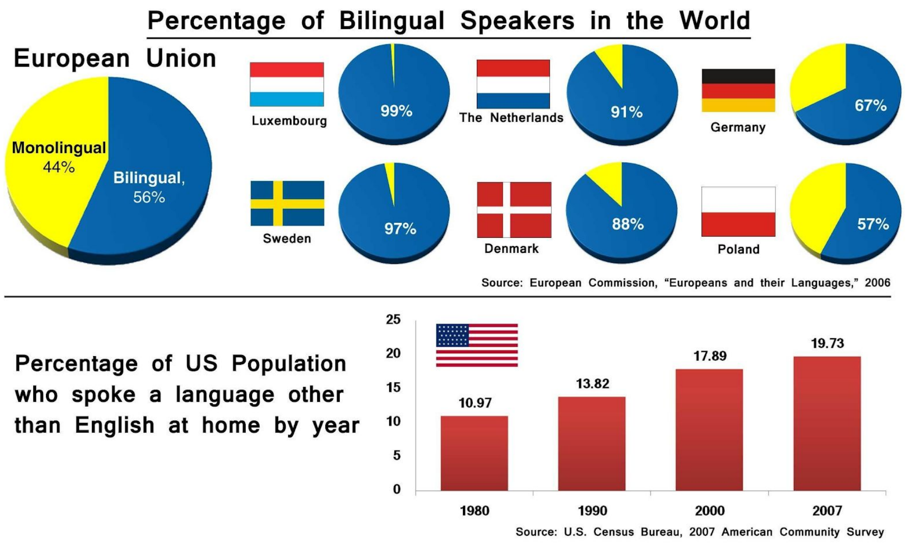
Swahili
100 million speakers

American English

Scottish English

Hinglish

Most of the world today is multilingual



Source: US Census Bureau

Source: Ethnologue

Tokenization

这是一个简单的句子

WORDS This is a simple sentence

זה משפט פשוט

Tokenization + disambiguation

in tea her daughter	בתה

- 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

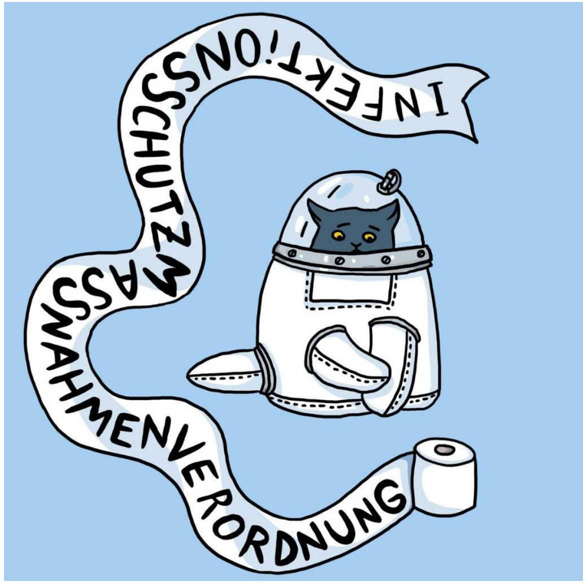
Much’a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

“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

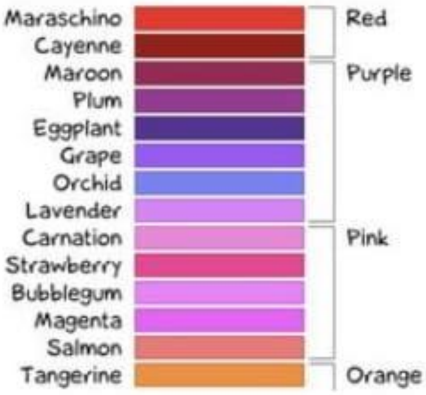
- German



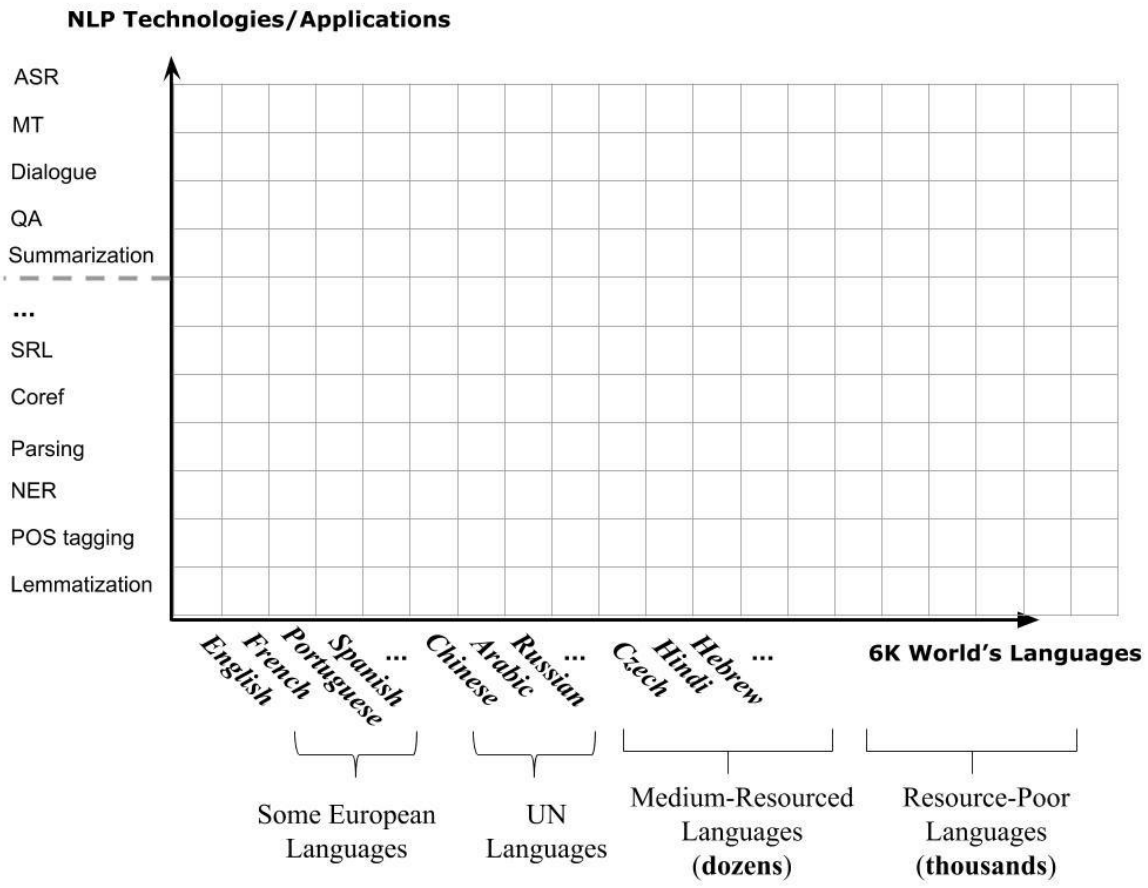
Infektionsschutzmaßnahmenverordnung

Semantic analysis

- Every language sees 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


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)

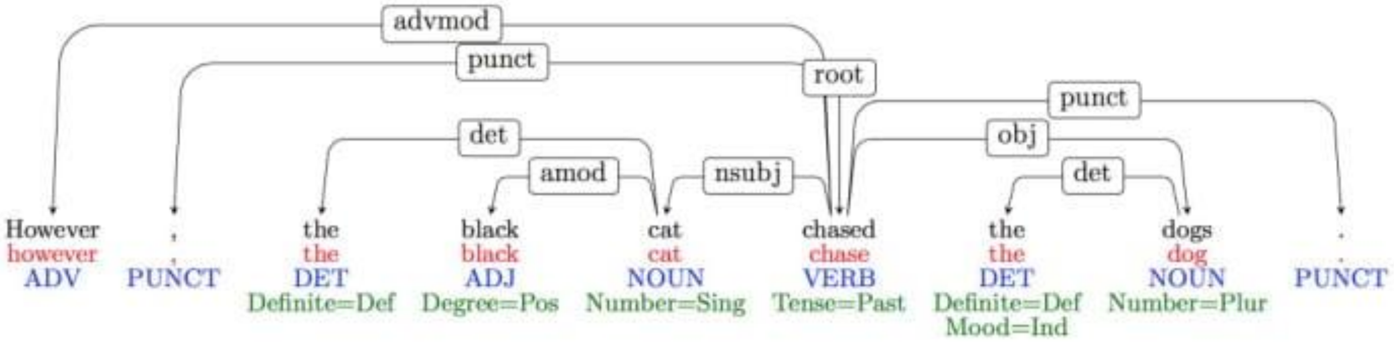
Linguistic variation

- Non-standard language, emojis, hashtags, names

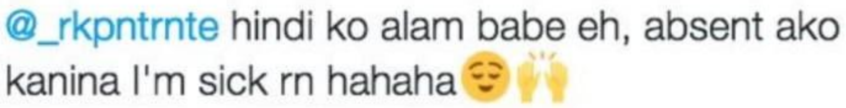
chowdownwithchan #crab and #pork #xiaolongbao at @dintaifungusa... where else? 😂 🤷‍♀️ Note the cute little crab indicator in the 2nd pic 🦀 💕

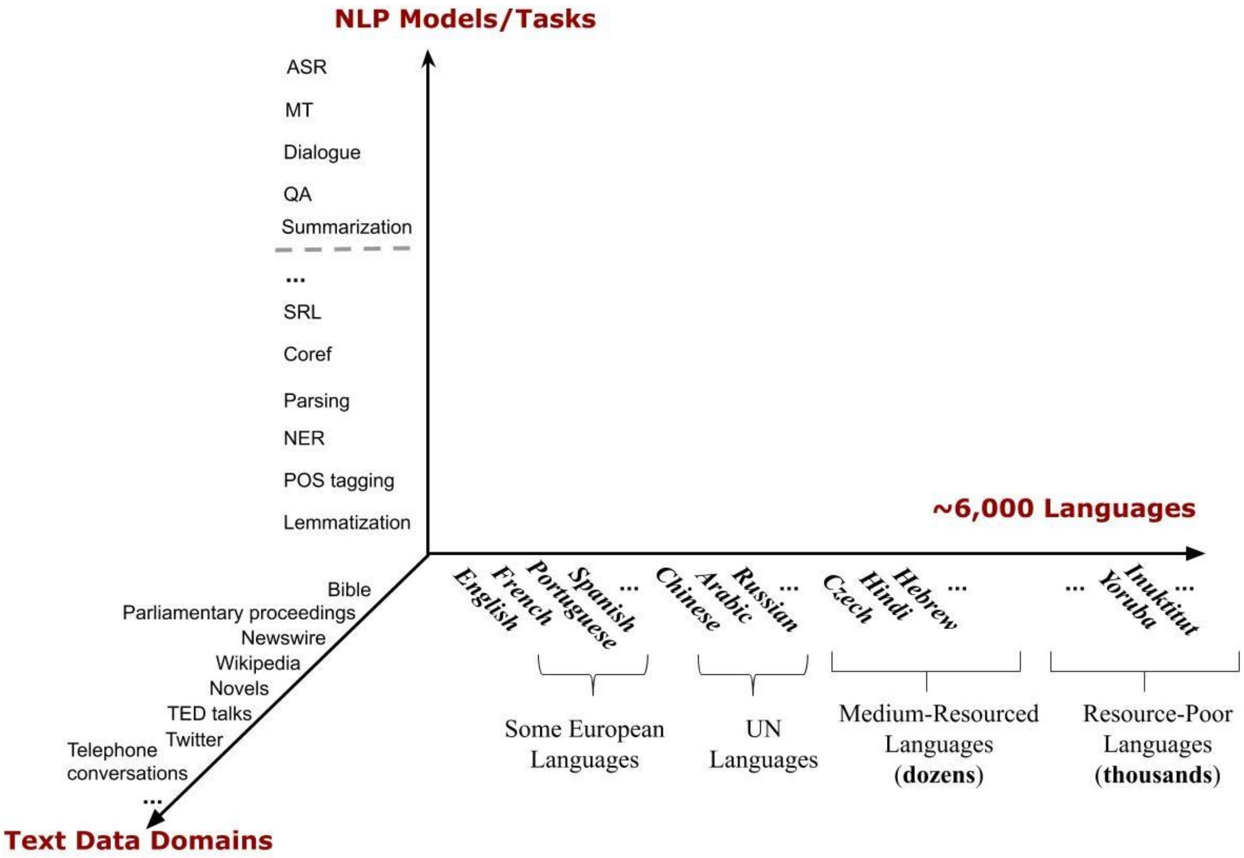
Variation

- 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

Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. 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		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

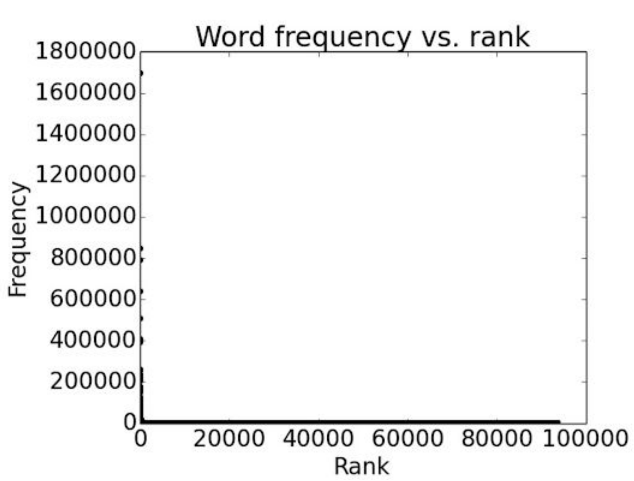
But also, out of 93,638 distinct words (word types), 36,231 occur only once.

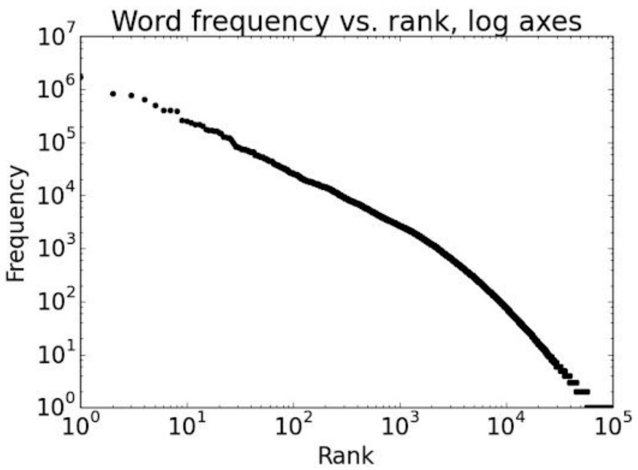
Examples:

- 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?

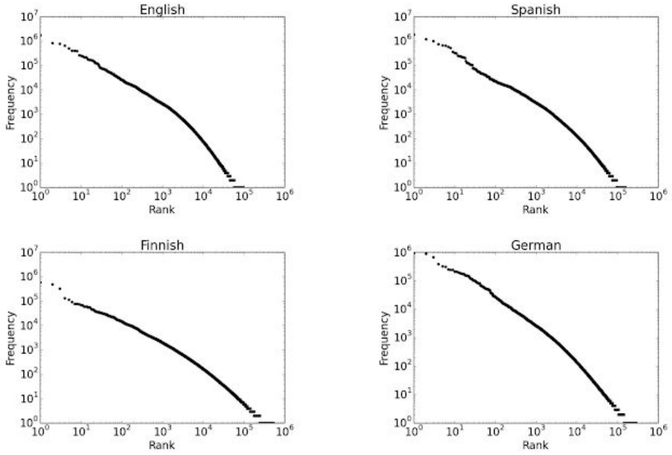




Zipf's Law

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



Why is language interpretation hard?

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

Expressivity

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

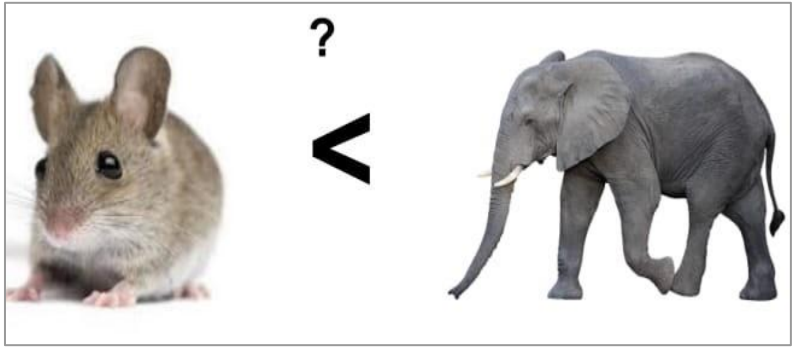
She gave the book to Tom	vs.	She gave Tom the book
Some kids popped by	vs.	A few children visited
Is that window still open?	vs.	Please close the window

Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. 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

Why is language interpretation hard?

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

Unknown representation

- Very difficult to capture what is $\mathcal{R}$, 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

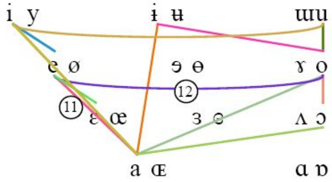
- Sensitivity to a wide range of phenomena and constraints in human language
- Generality across languages, modalities, genres, styles
- Strong formal guarantees (e.g., convergence, statistical efficiency, consistency)
- High accuracy when judged against expert annotations or test data
- Ethical

NLP $\stackrel{?}{=}$ Machine Learning

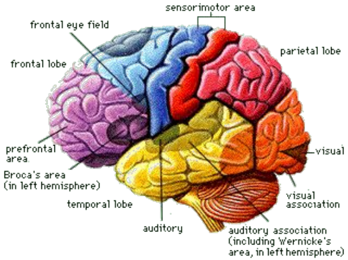
- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.

What is nearby NLP?

- Computational Linguistics
  - Using computational methods to learn more about how language wor
  - We end up doing this and using it



- Cognitive Science
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!



- Speech Processing
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP

Next class

- Classification

Questions?