



Language  
Technologies  
Institute

Carnegie  
Mellon  
University

# Algorithms for NLP

CS 11-711 · Fall 2020

## Lecture 1: Introduction

Emma Strubell

# Welcome!



**Emma**

**Yulia**

**Bob**

**Sanket**

**Han**

**Jiateng**

# Course website:

<http://demo.clab.cs.cmu.edu/11711fa20/>

---

## Algorithms for NLP

CMU CS 11711, Fall 2020

Tuesday/Thursday 1:30-2:50pm EDT, on Zoom

Emma Strubell (office hours: By appointment), strubell@cmu.edu  
Yulia Tsvetkov (office hours: By appointment), ytsvetko@cs.cmu.edu  
Robert Frederking (office hours: By appointment), ref@cs.cmu.edu

Teaching Assistants:

Jiateng Xie (office hours: Wednesday 10:00-11:00pm EDT, Zoom), jiatengx@cs.cmu.edu  
Sanket Vaibhav Mehta (office hours: Friday 9:00-10:00am EDT, Zoom), svmehta@cs.cmu.edu  
Xiaochuang Han (office hours: Monday 10:30-11:30am EDT, Zoom), xiaochuh@cs.cmu.edu

Forum: [Piazza](#)

Note: Sensitive information related to the class (e.g., Zoom links) will be available on [Piazza](#)

# Piazza:

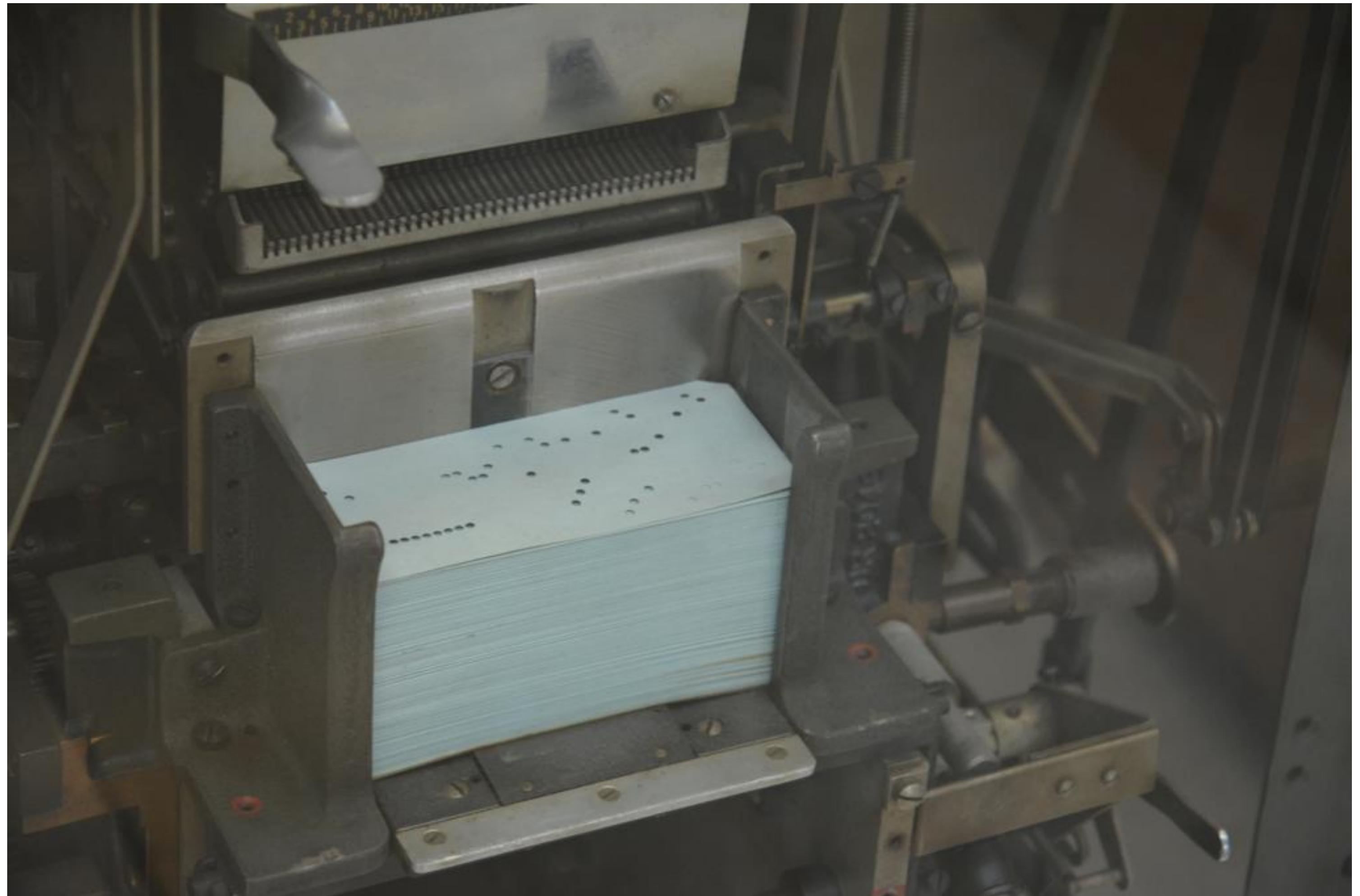
[Learning Goals](#)   [Prerequisites](#)   [Announcements](#)   [Format](#)   [Class Schedule](#)   [Readings](#)   [Assignments](#)   [Policies](#)

<https://piazza.com/cmu/fall2020/11711/>

---

# Communicating with machines

~1950s-1970s:



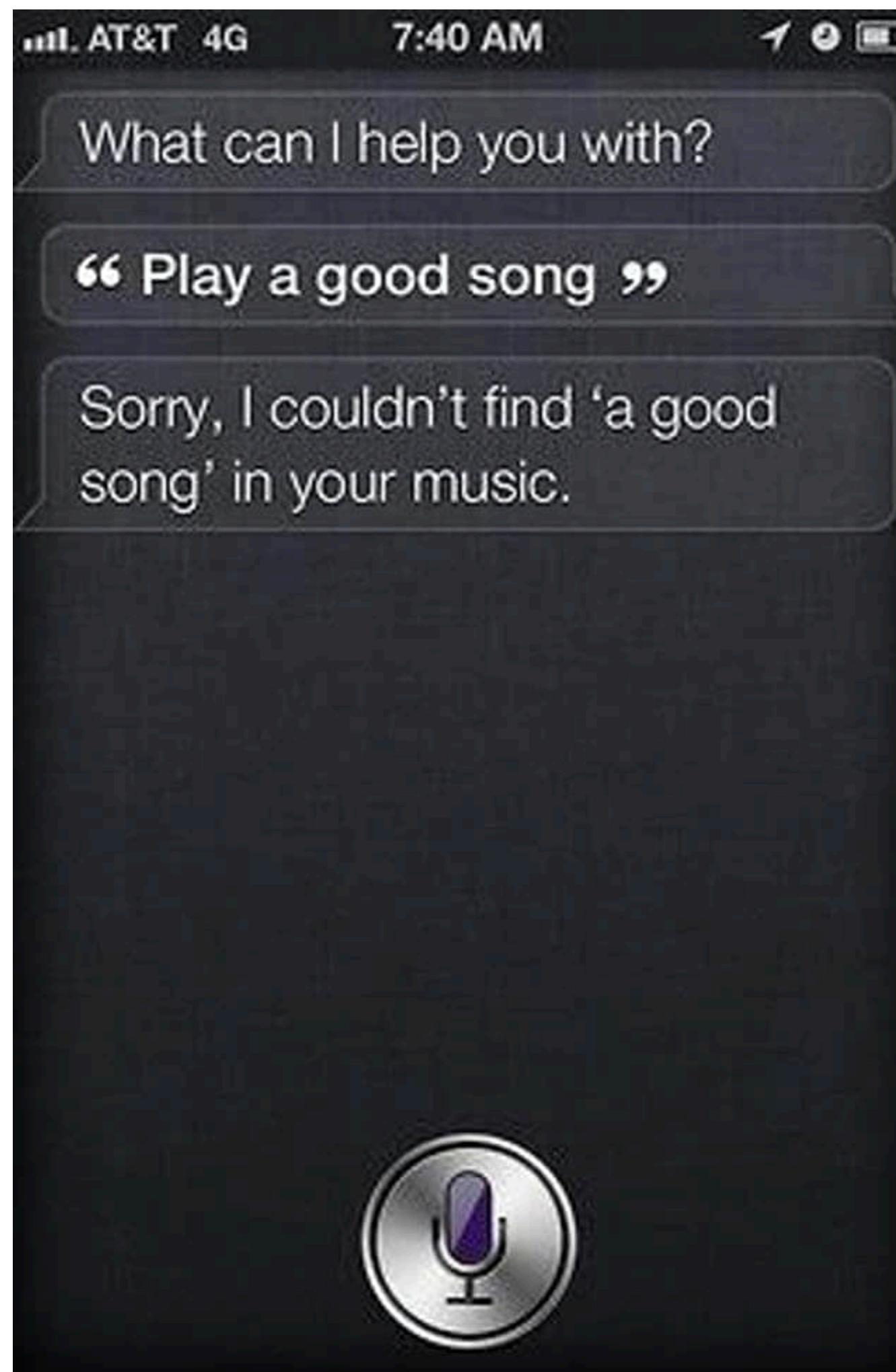
# Communicating with machines

~1980s:

```
. File Edit Edit_Settings Menu Utilities Compilers Test Help
. EDIT      BS9U.DEVT3.CLIBPAU(TIMMIES) - 01.31          Columns 00001 000
. Command ==> █
. **** * **** * **** * **** * **** * **** * **** * **** * **** * **** * **** *
. 000001 /* REXX EXEC **** * **** * **** * **** * **** * **** * **** * **** *
. 000002 /*
. 000003 /* TIMMIES FACTOR - COMPOUND INTEREST CALCULATOR
. 000004 /*
. 000005 /* AUTHOR: PAUL GAMBLE
. 000006 /* DATE: OCT 1/2007
. 000007 /*
. 000008 /*
. 000009 **** * **** * **** * **** * **** * **** * **** * **** * **** * **** *
. 000010
. 000011
. 000012 say '*****'
. 000013 say 'Welcome Coffee drinker.'
. 000014 say '*****'
. 000015 DO WHILE DATATYPE(CoffeeAmt) \= 'NUM'
. 000016   say ""
. 000017   say "What is the price of your coffee?",
. 000018     "(e.g. 1.58 = $1.58)"
. 000019   parse pull CoffeeAmt
. 000020 END
. 000021
. 000022 DO WHILE DATATYPE(CoffeeWk) \= 'NUM'
. 000023   say ""
. 000024   say "How many coffees a week do you have?"
. 000025   parse pull CoffeeWk
. 000026 END
. 000027
. 000028 DO WHILE DATATYPE(Rate) \= 'NUM'
. 000029   say ""
. 000030   say "What annual interest rate would you like to see on that money?",
. 000031     "(e.g. 8 = 8%)"
. 000032   parse pull Rate
. 000033 END
. 000034 Rate = Rate * 0.01 /* CHG TO DECIMAL NUMBER */
. 000035
```

# Communicating with machines

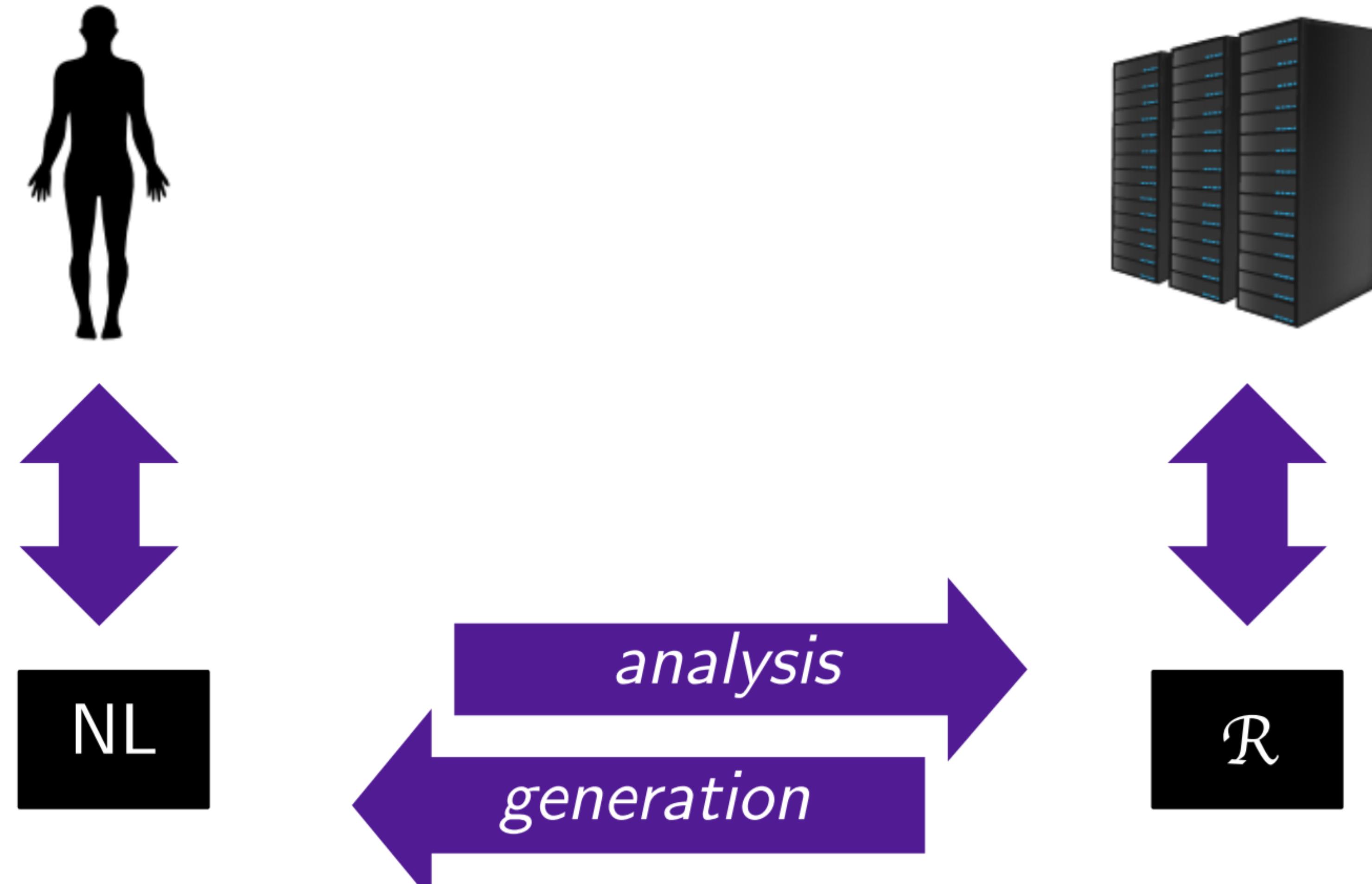
Today:



# What is natural language processing (NLP)?

- $NL \in \{\text{Mandarin, Hindi, Spanish, Arabic, English, ... Inuktitut}\}$
- Automation of NLs:
  - analysis ( $NL \rightarrow \mathcal{R}$ )
  - generation ( $\mathcal{R} \rightarrow NL$ )
  - acquisition of  $\mathcal{R}$  from knowledge and data

# What is natural language processing (NLP)?



# Language technologies

What technologies are required to write such a program?

- A conversational agent contains:
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Information retrieval
  - Text to speech



# Language technologies

## Machine translation

The image shows two side-by-side screenshots of the Google Translate web interface.

**Left Screenshot:** This shows the main Google Translate page. It has a "CHINESE - DETECTED" input field containing the Chinese sentence "我学习深度学习和机器学习" (Wǒ xuéxí shēndù xuéxí hé jīqì xuéxí) and its English translation "I study deep learning and machine learning." Below the input field are microphone and speaker icons, and a blue feedback button labeled "Send feedback".

**Right Screenshot:** This shows the "DETECT LANGUAGE" feature of Google Translate. It includes a search bar labeled "Search languages" and a table of supported languages. The table is organized into four columns: the first column lists languages by origin (e.g., Afrikaans, Albanian, Amharic), the second column lists languages by destination (e.g., Czech, Danish, Dutch), and the third and fourth columns list destination languages grouped by script (Latin, Cyrillic, etc.). The "DETECT LANGUAGE" tab is selected at the top.

Detected Language	Target Languages	Target Languages	Target Languages
Afrikaans	Czech	Hebrew	Latin
Albanian	Danish	Hindi	Latvian
Amharic	Dutch	Hmong	Lithuanian
Arabic	English	Hungarian	Luxembourgish
Armenian	Esperanto	Icelandic	Macedonian
Azerbaijani	Estonian	Igbo	Malagasy
Basque	Filipino	Indonesian	Malay
Belarusian	Finnish	Irish	Malayalam
Bengali	French	Italian	Maltese
Bosnian	Frisian	Japanese	Maori
Bulgarian	Galician	Javanese	Marathi
Catalan	Georgian	Kannada	Mongolian
Cebuano	German	Kazakh	Myanmar (Burmese)
Chichewa	Greek	Khmer	Nepali
Chinese	Gujarati	Korean	Norwegian
Corsican	Haitian Creole	Kurdish (Kurmanji)	Pashto
Croatian	Hausa	Kyrgyz	Persian
	Hawaiian	Lao	Polish
			Swahili
			Swedish

# Language technologies

## Question answering

- What does “divergent” mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?
- How much Chinese silk was exported to England at the end of the 18th century?
- What do scientists think about the ethics of human cloning?



Google

how many states were in the U.S. in 1913

All News Images Shopping Maps More Settings Tools

About 258,000,000 results (0.63 seconds)

en.wikipedia.org › wiki › List\_of\_U.S.\_states\_by\_date\_... ▾

[List of U.S. states by date of admission to the Union - Wikipedia](#)

... states ratified the 1787 Constitution, then the order in which the others were admitted to the Union. A state of the United States is one of the 50 constituent entities that shares its sovereignty with the federal government. Americans are citizens of both the federal republic and of the state in which ...

[List of U.S. states](#) · [Articles of Confederation](#) ... · See also · Notes

# Natural language processing

## Applications

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

## Core technologies

- Language modeling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense disambiguation
- Semantic role labelling
- ...

# What does an NLP system need to “know”?

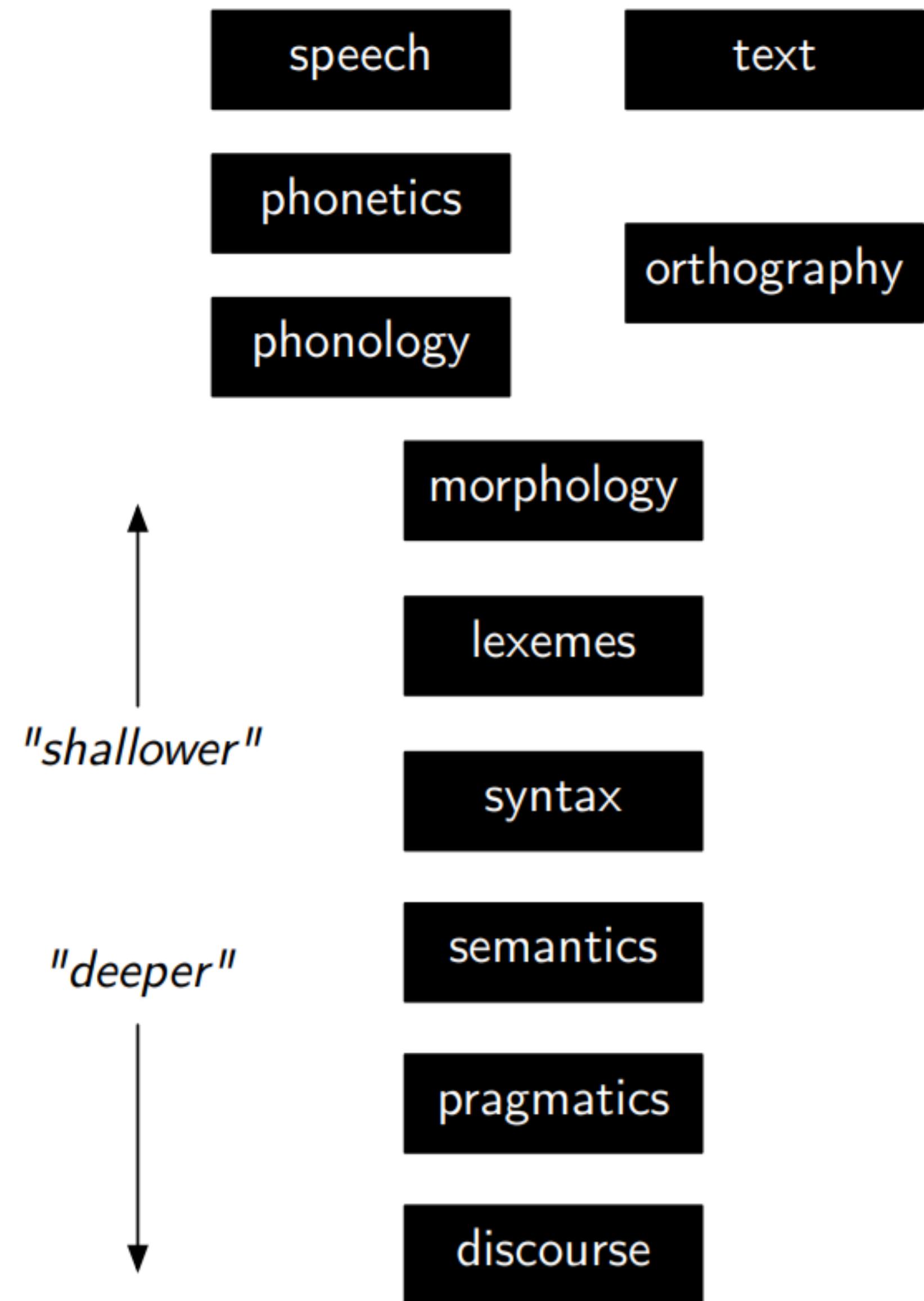
- Language consists of many levels of structure
- Humans fluently integrate all of these in producing and understanding language
- Ideally, so would a computer!

# What does it mean to “know” a language?



Hi, how can I help?

# Levels of linguistic knowledge



# Speech, phonetics, phonology



This is a simple sentence .

/ ðɪs ɪz ə 'simpl 'sɛntəns /.

speech

text

phonetics

orthography

phonology

morphology

lexemes

syntax

semantics

pragmatics

discourse

"shallower"

"deeper"

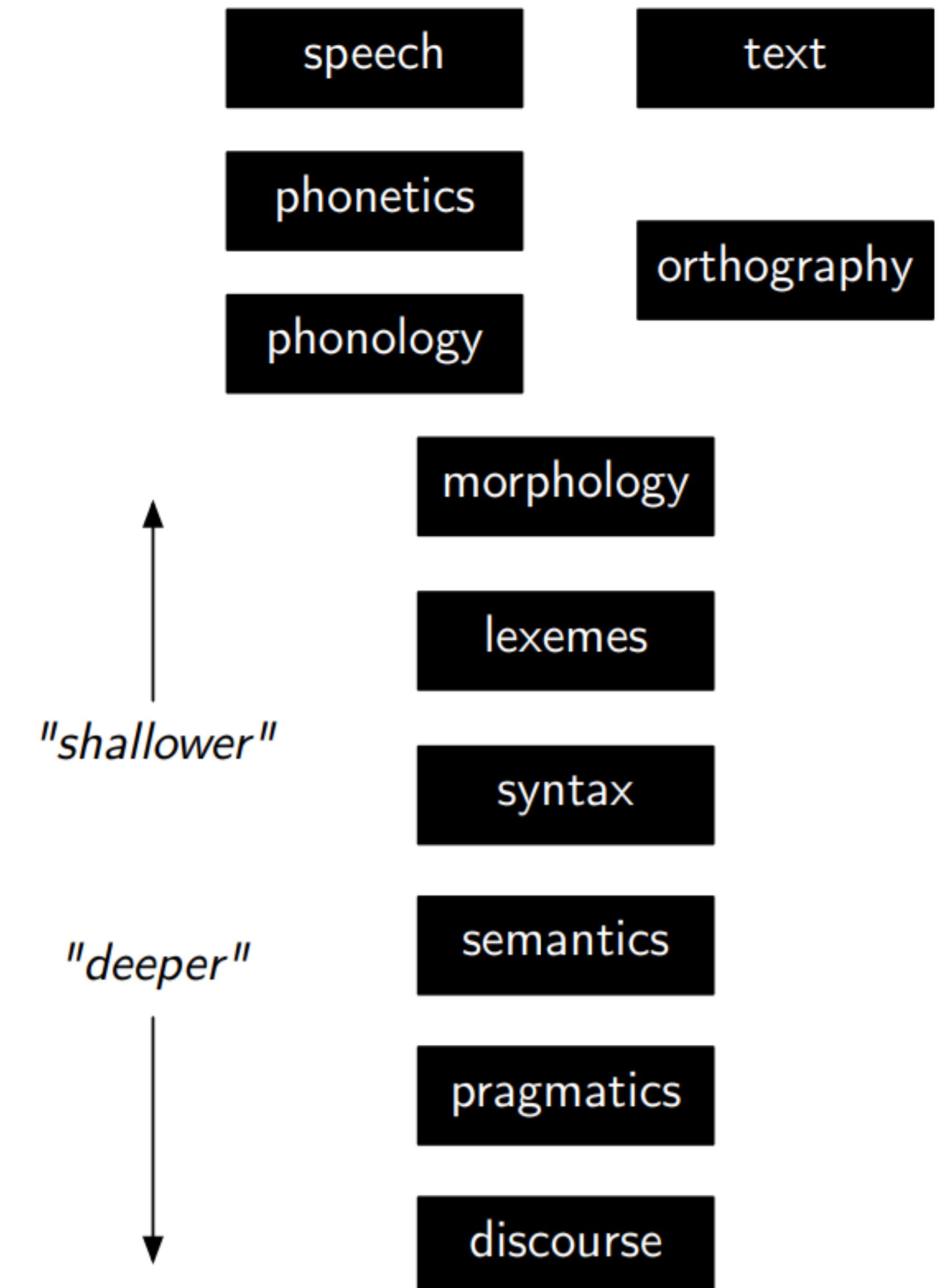


# Words

- Morphological analysis
- Tokenization
- Lemmatization

**Tokens** This is a simple sentence .  
be  
3sg  
present

**Morphology**



# Syntax

## ■ Part-of-speech tagging

**Parts of speech**

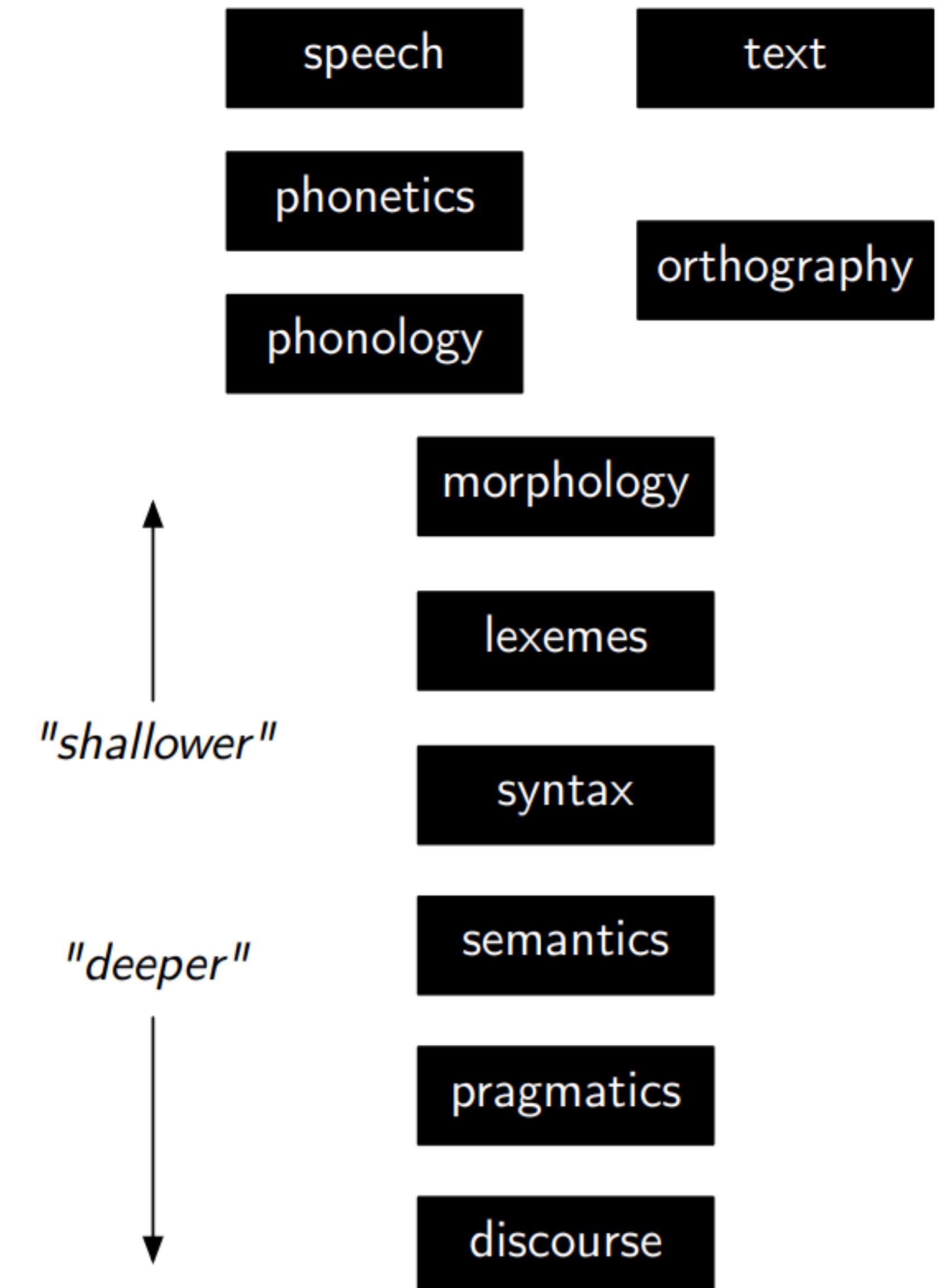
DT VBZ DT JJ NN PUNC

**Tokens** This is a simple sentence .

be

3sg

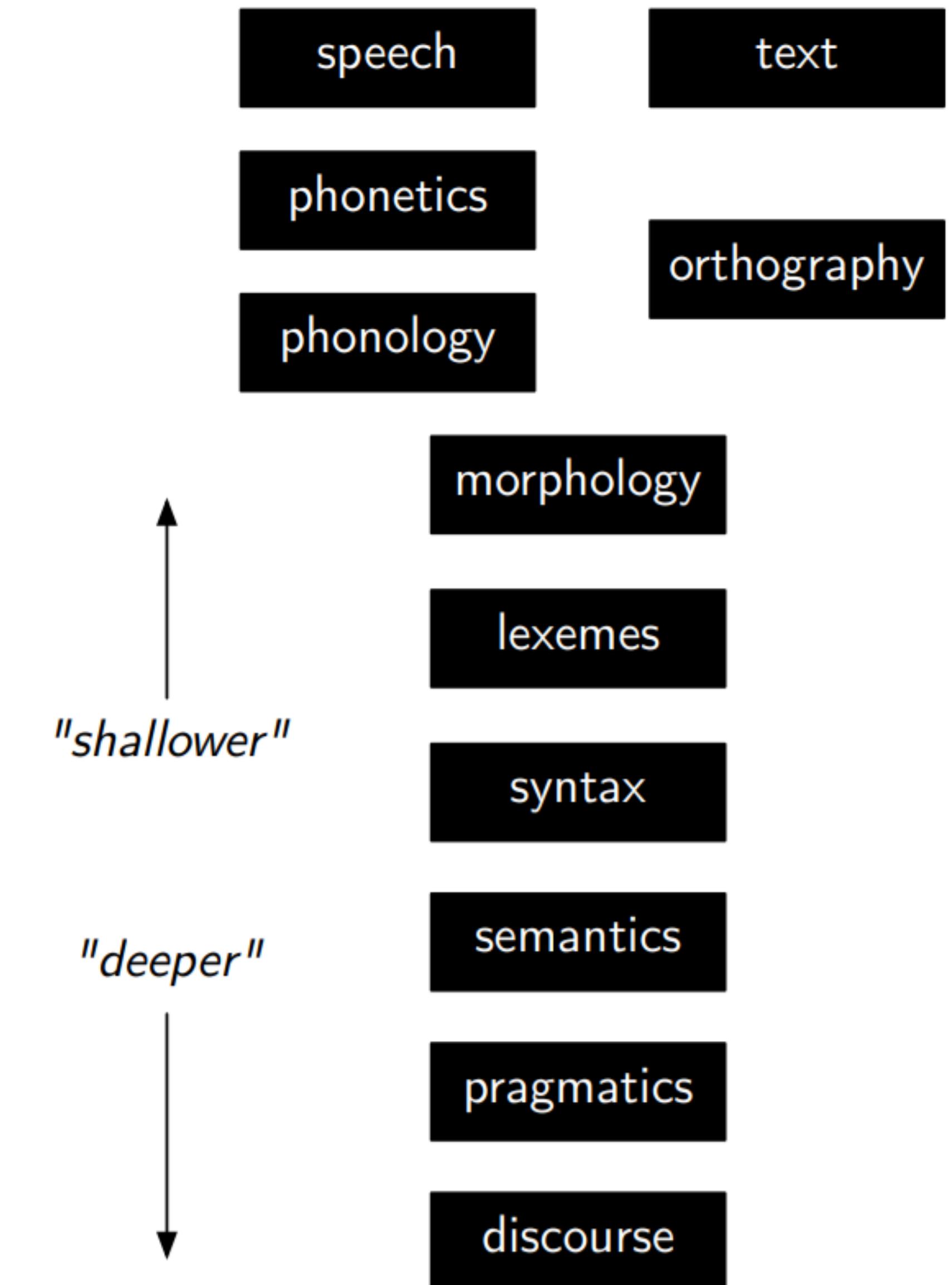
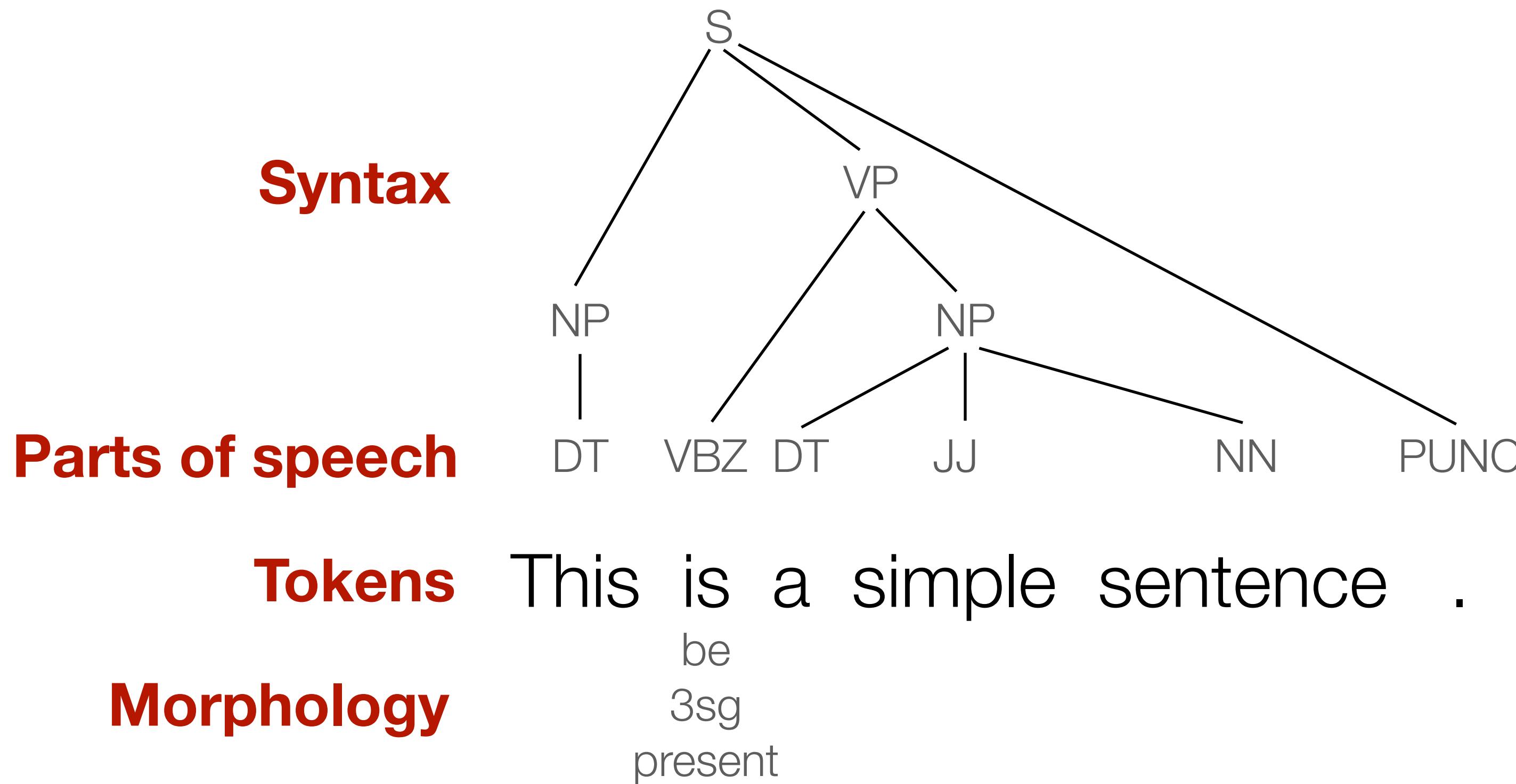
present



**Morphology**

# Syntax

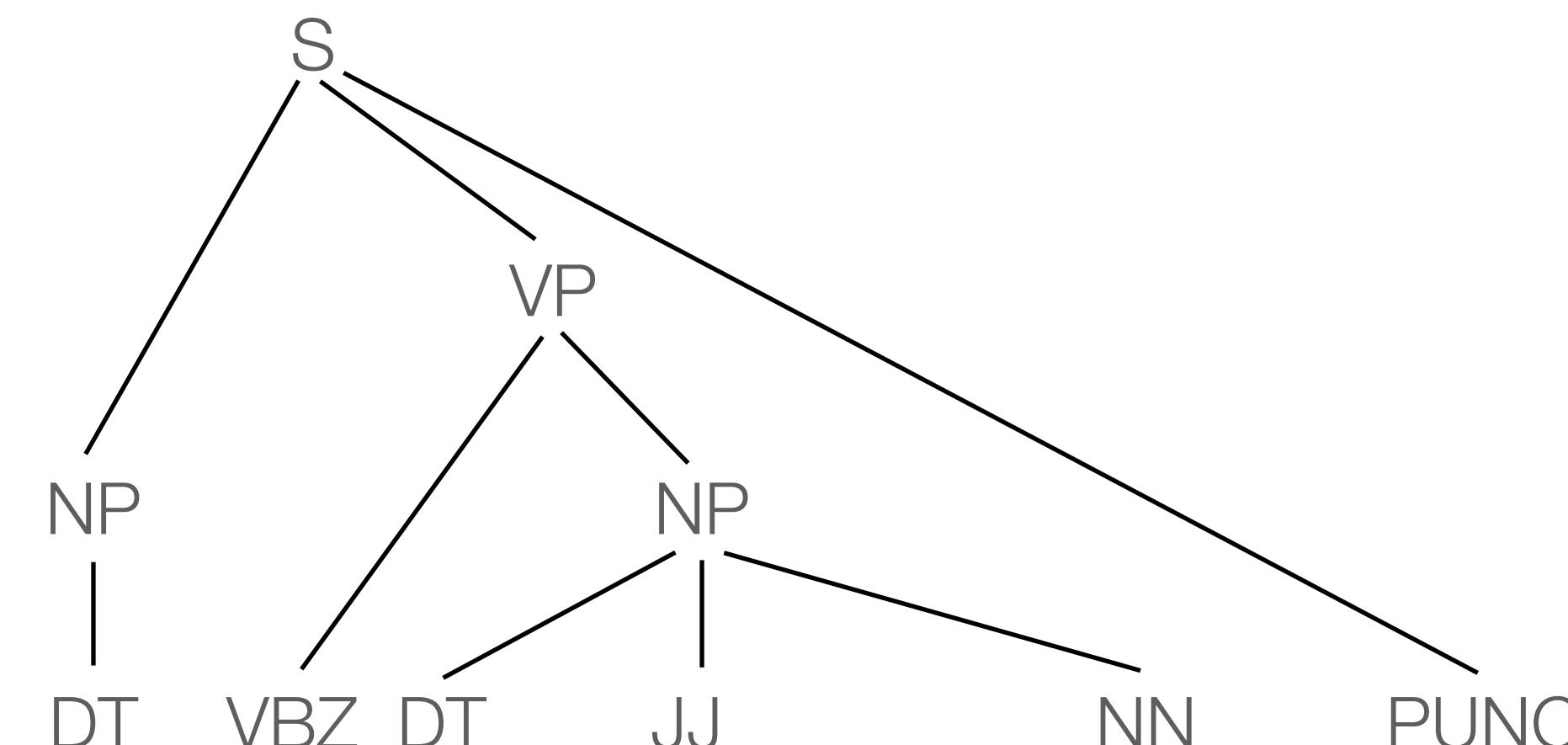
- Part-of-speech tagging
- Syntactic parsing



# Semantics

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

**Syntax**



**Parts of speech**

**Tokens**

This is a simple sentence .

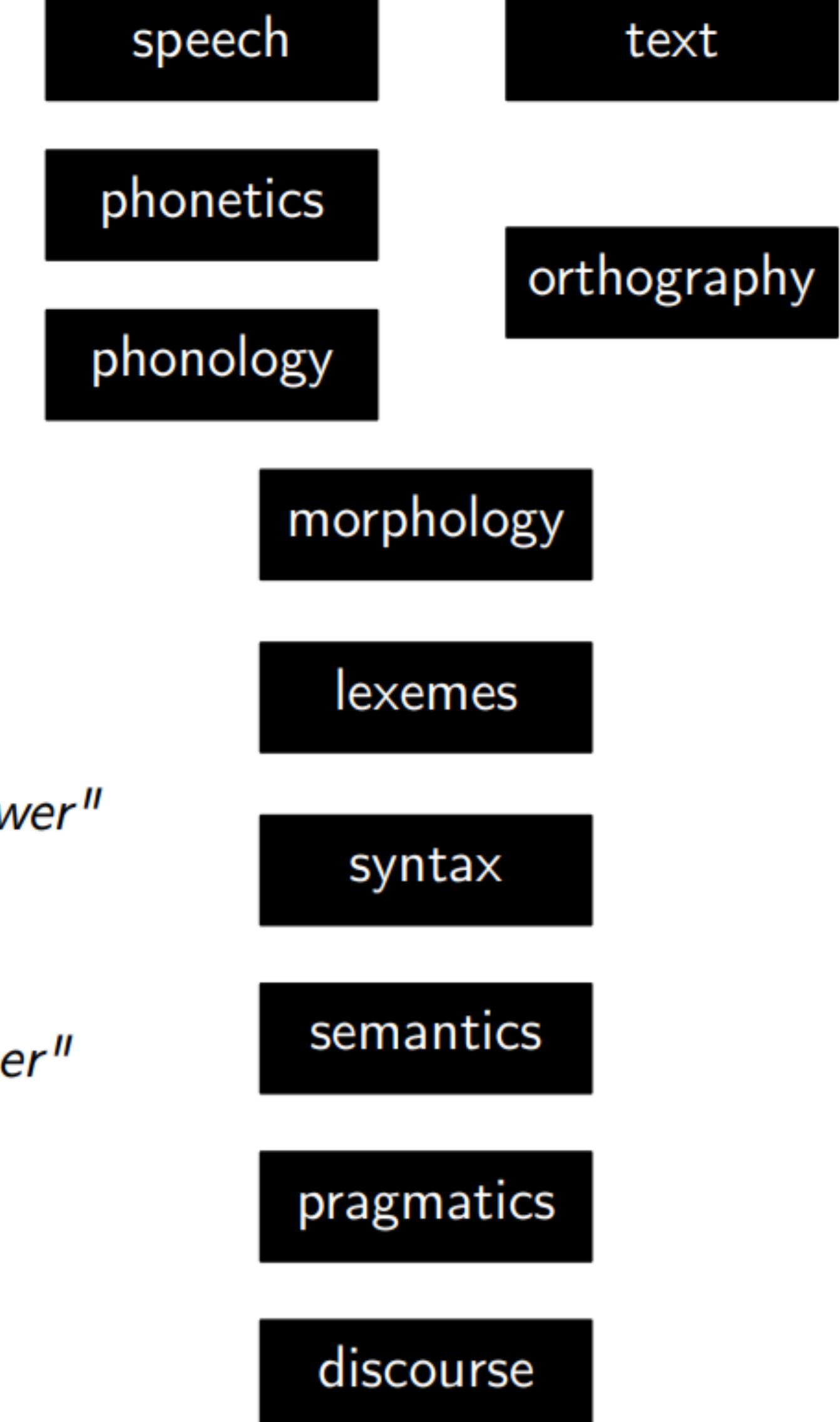
be  
3sg  
present

SIMPLE1:  
having few  
parts

SENTENCE1:  
String of words satisfying  
the grammatical rules of  
a language

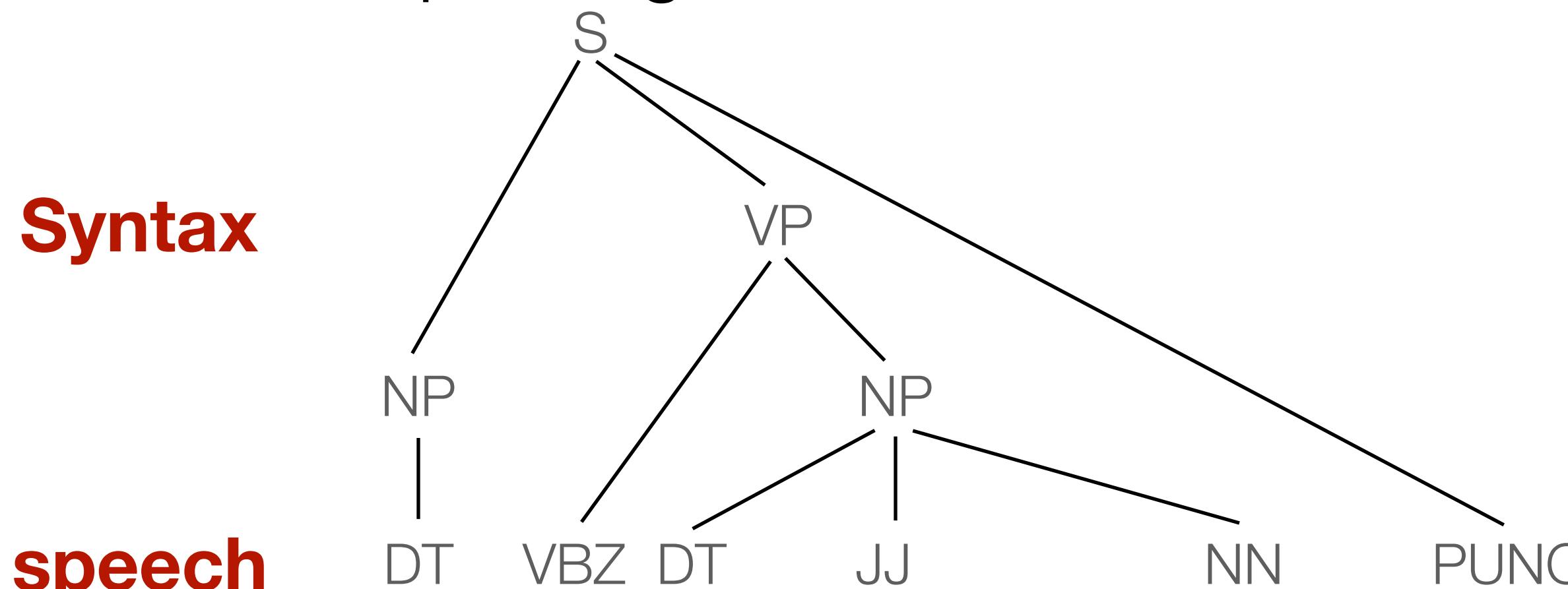
**Morphology**

**Semantics**



# Discourse

- Reference resolution
- Discourse parsing



**Parts of speech**

**Tokens**

This is a simple sentence

be  
3sg  
present

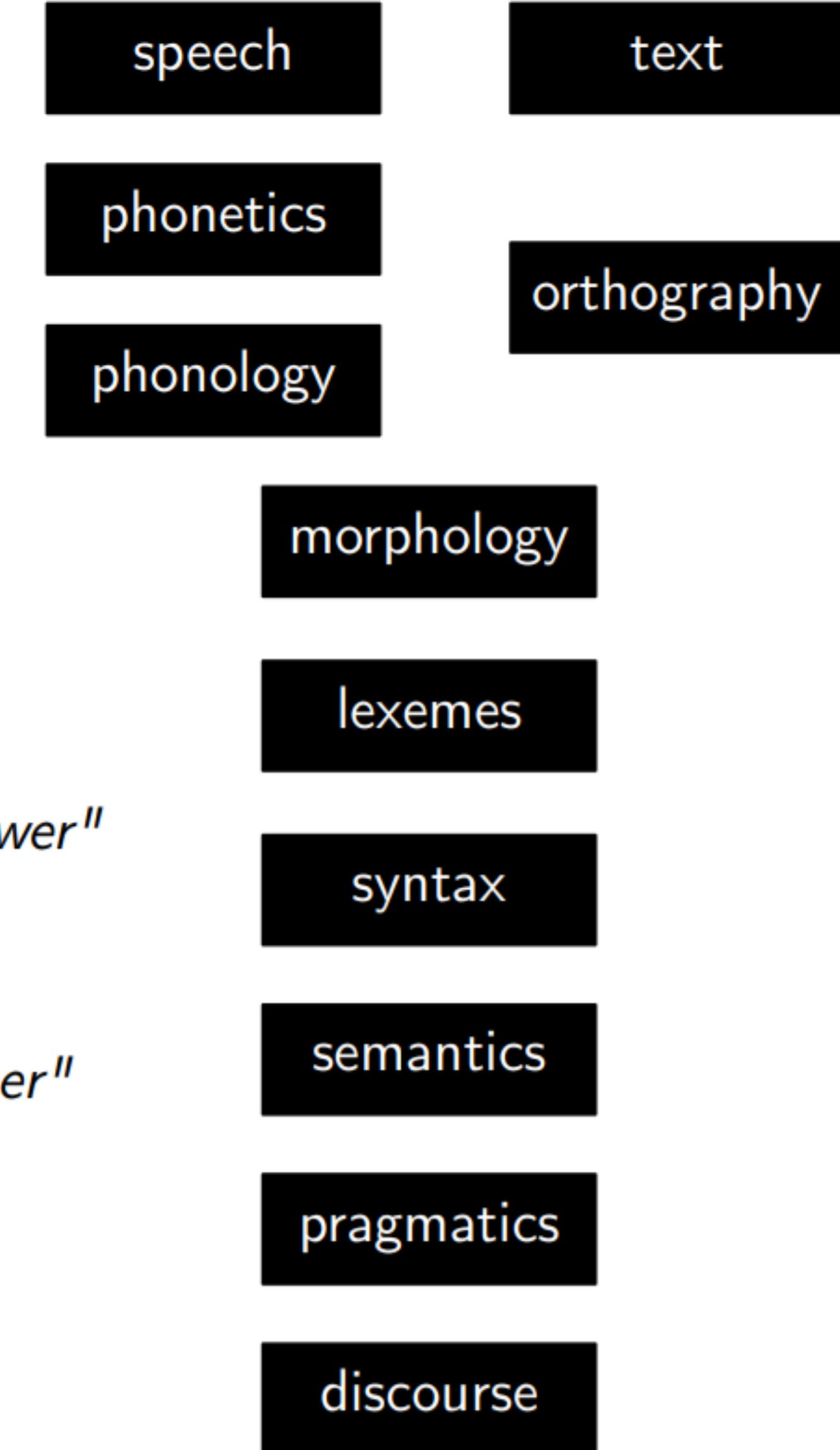
SIMPLE1:  
having few  
parts

SENTENCE1:  
String of words  
satisfying the  
grammatical rules  
of a language

"shallower"

"deeper"

coreferent



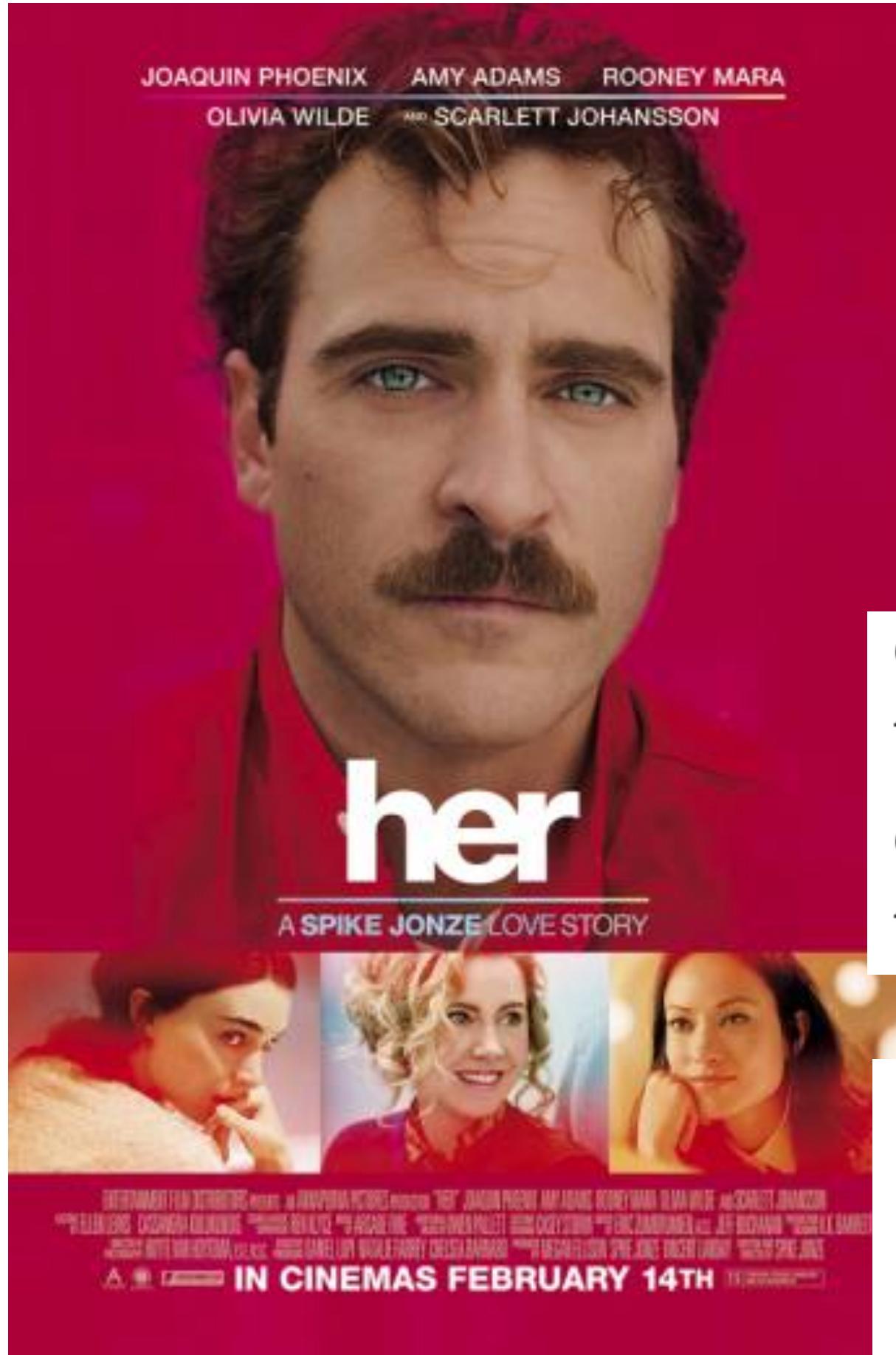
**Morphology**

**Semantics**

**Discourse**

But an instructive one

# Where are we now?



REPORT

## OPENAI'S LATEST BREAKTHROUGH IS ASTONISHINGLY POWERFUL, BUT STILL FIGHTING ITS FLAWS

*The ultimate autocomplete*

By James Vincent | Jul 30, 2020, 10:01am EDT

Q: How many eyes does a giraffe have?

A: A giraffe has two eyes.

Q: How many eyes does my foot have?

A: Your foot has two eyes.

Q: Which is heavier, a mouse or a toaster?

A: A toaster is heavier than a mouse.

Q: Which is heavier, a toaster or a pencil?

A: A pencil is heavier than a toaster.

## OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

by Will Douglas Heaven

July 20, 2020

# Where are we now?

## Conversation A

Human: say you're in an empty giant office and you see  
a black bird flying indoors. what do you do?  
Meena: what kind of bird? Is it friendly? How large?  
How big is the office? Is it indoors or outdoors?  
Human: well it's a predator bird and it's gonna eat you  
Meena: I would grab my phone and call 911.

Human: I've been using Python for 15 years  
Meena: Ah I see, so I'm guessing you're pretty pro at it by now then?  
Human: probably  
Meena: What's your current project?  
Human: NLP  
Meena: Neuro-linguistic Programming?  
Human: correct

## Cross-turn Repetition Example 1

Human: Hi!  
Meena: Hi! How are you?! :)  
Human: doing good, what would be the best city to visit  
in Europe?  
Meena: Paris, Barcelona, Amsterdam, Prague  
great list! why?  
Human: Paris is a beautiful city, and Barcelona is too.  
And I've always wanted to go to Amsterdam and  
Prague and have never been.  
Human: have you even been to India?  
Meena: No, but I'd love to go.  
Human: how about Japan?  
Meena: I'd love to go to Japan too.  
Human: how about Paris?  
Meena: I'd love to go to Paris too.  
Human: are you lying to me?  
Meena: I'd love to go to Japan too.  
Human: i think you're lying to me, that makes me sad  
Meena: I'd love to go to Japan, too.  
Human: you're blocking me out! you have to talk to me  
Meena: I'd love to go to Japan.  
Human: where else?  
Meena: I'd love to go to Japan, too!

# Where are we now?

## 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)

Human: I don't like you!

Meena: I like you, though!

Human: You were insensitive to me being a Vegan!

Meena: I didn't mean for it to come off that way! I'm sorry! You can still love me?

Human: No. I love chicken more

Meena: Oh. I'm sorry. I love you still.

Human: No need

Meena: Oh, but I need.

Human: That's way too much flirty

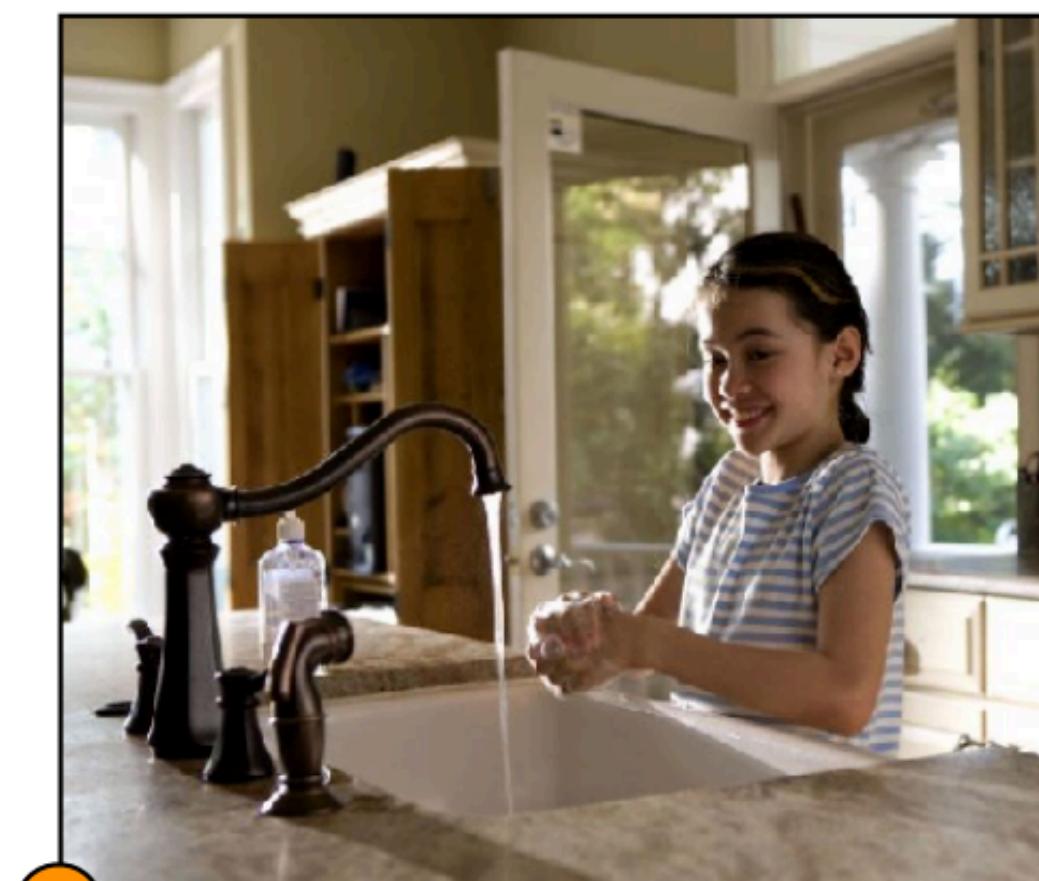
Meena: No, it's just right.

Human: Bye Meena! Sorry it couldn't work out between us!

Meena: It's okay, you know I'm still in love with you.

Human: It's not me. its about you

Meena: It's about us.



# Why is NLP hard?

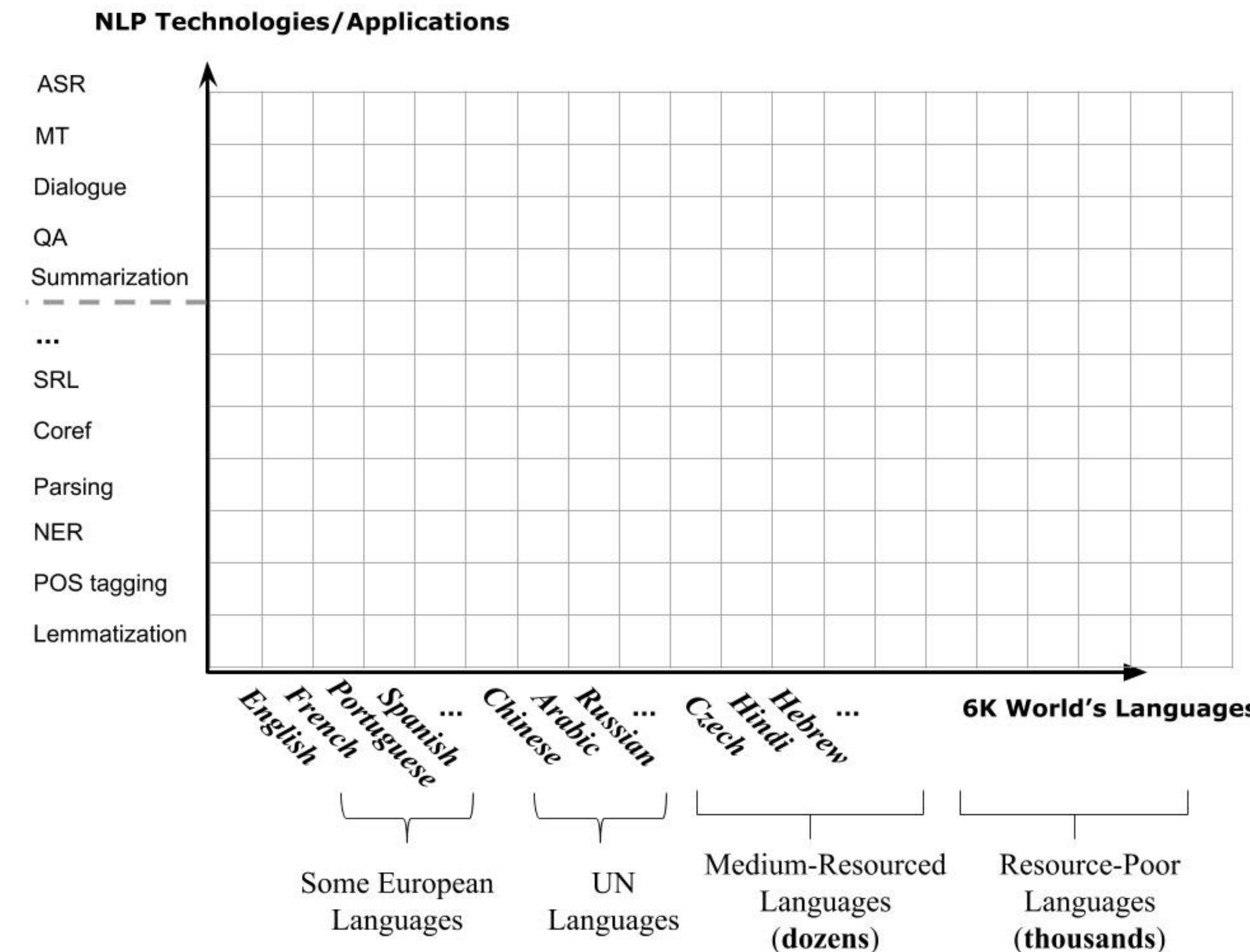
1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled variables
7. Unknown representation  $\mathcal{R}$

# Ambiguity

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



# Ambiguity + Scale



# Ambiguity: Tokenization

这是一个简单的句子

WORDS      This is a simple sentence

זה פשוט פשוט

# Ambiguity: Word sense disambiguation

in tea  
her daughter

בתה

- most of the vowels unspecified

# Ambiguity: Tokenization + disambiguation

in tea	בתה
in the tea	בהתה
that in tea	שבתה
that in the tea	שבתה
and that in the tea	ושבתה
	ושבתה
and her saturday	+שבתה
and that in tea	+ש+בתה
and that her daughter	+ש+בתה

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

# Ambiguity: Morphology

- unfriend
- Obamacare
- Manfuckinghattan

# Ambiguity: Semantic analysis

- Every language sees the world in a different way
  - For example, could depend on historical conditions
  - Russian has very few words for colors; Japanese has hundreds
  - Idioms: **happy as a clam, it's raining cats and dogs, les carottes sont cuites**



# Ambiguity: Semantics

Every fifteen minutes a woman in this country gives birth.

# Ambiguity: Semantics

Every fifteen minutes a woman in this country gives birth. Our job is to find this woman, and stop her!

— Groucho Marx

# Ambiguity: Syntax + semantics

- I saw the woman with the telescope wrapped in paper.
  - Who has the telescope?
  - Who or what is wrapped in paper?
  - Event of perception or assault?



# Dealing with ambiguity

- How can we model ambiguity and choose correct analysis in context?
  - Non-probabilistic methods return all possible analyses.
  - Probabilistic models return best possible analysis, i.e. most probable one according to the model.

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

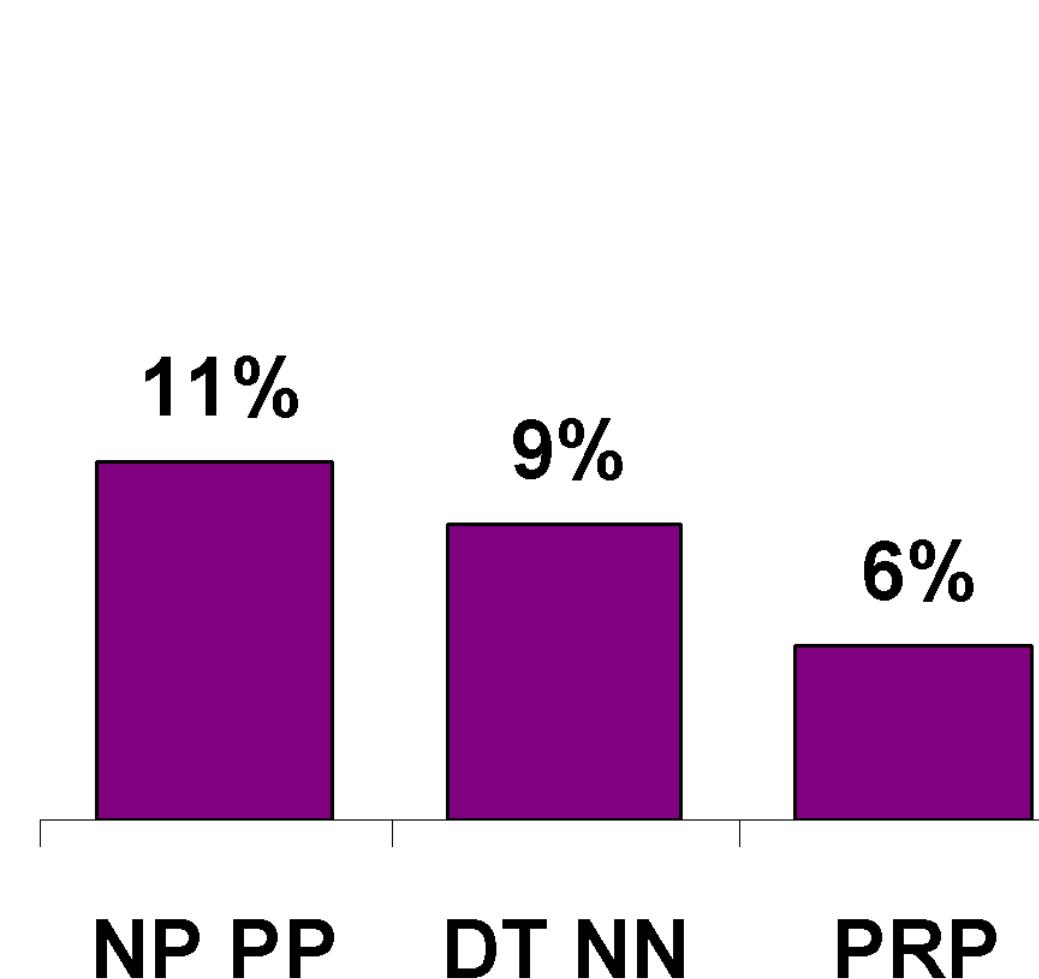
# Corpora

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

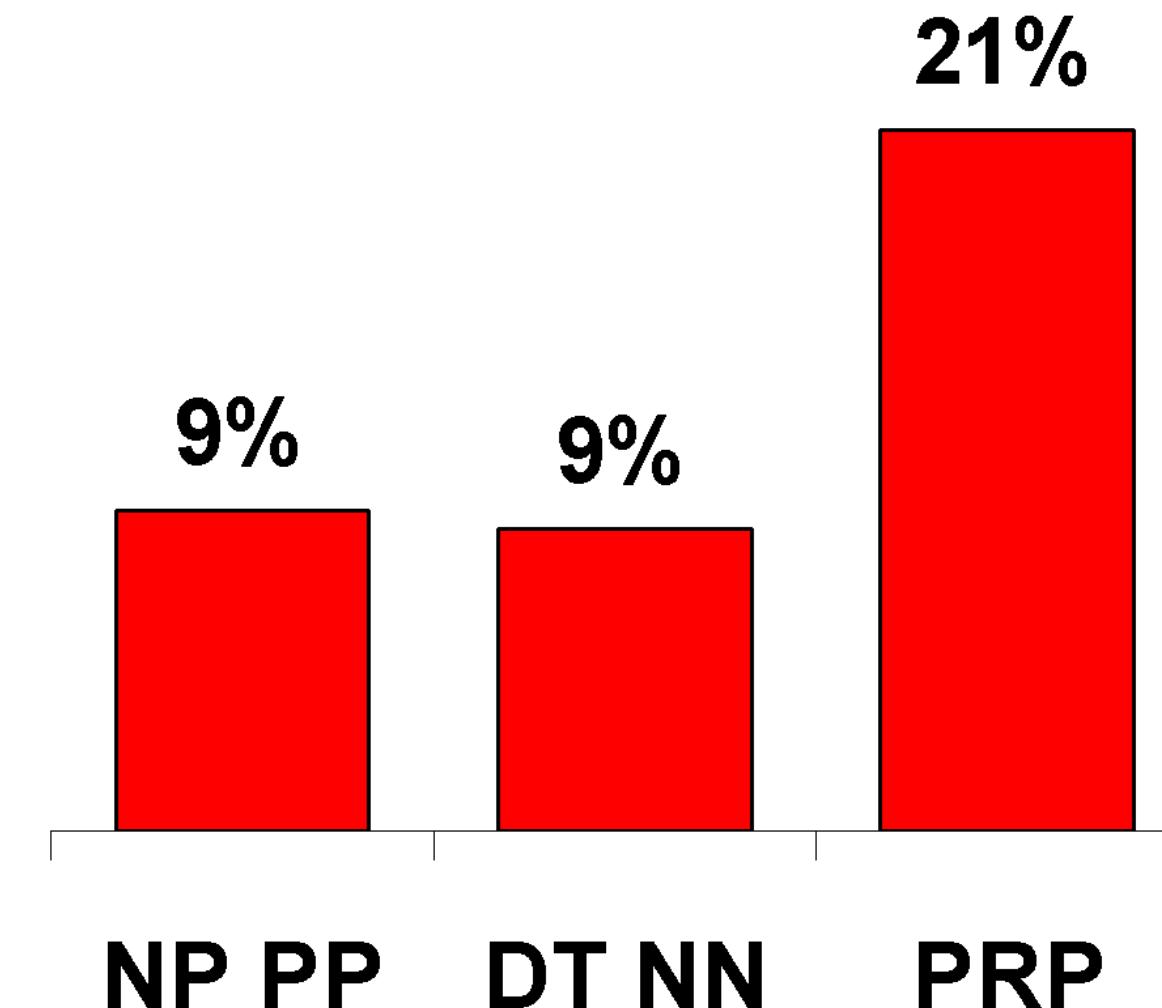
# Corpus-based methods

- Give us statistical information

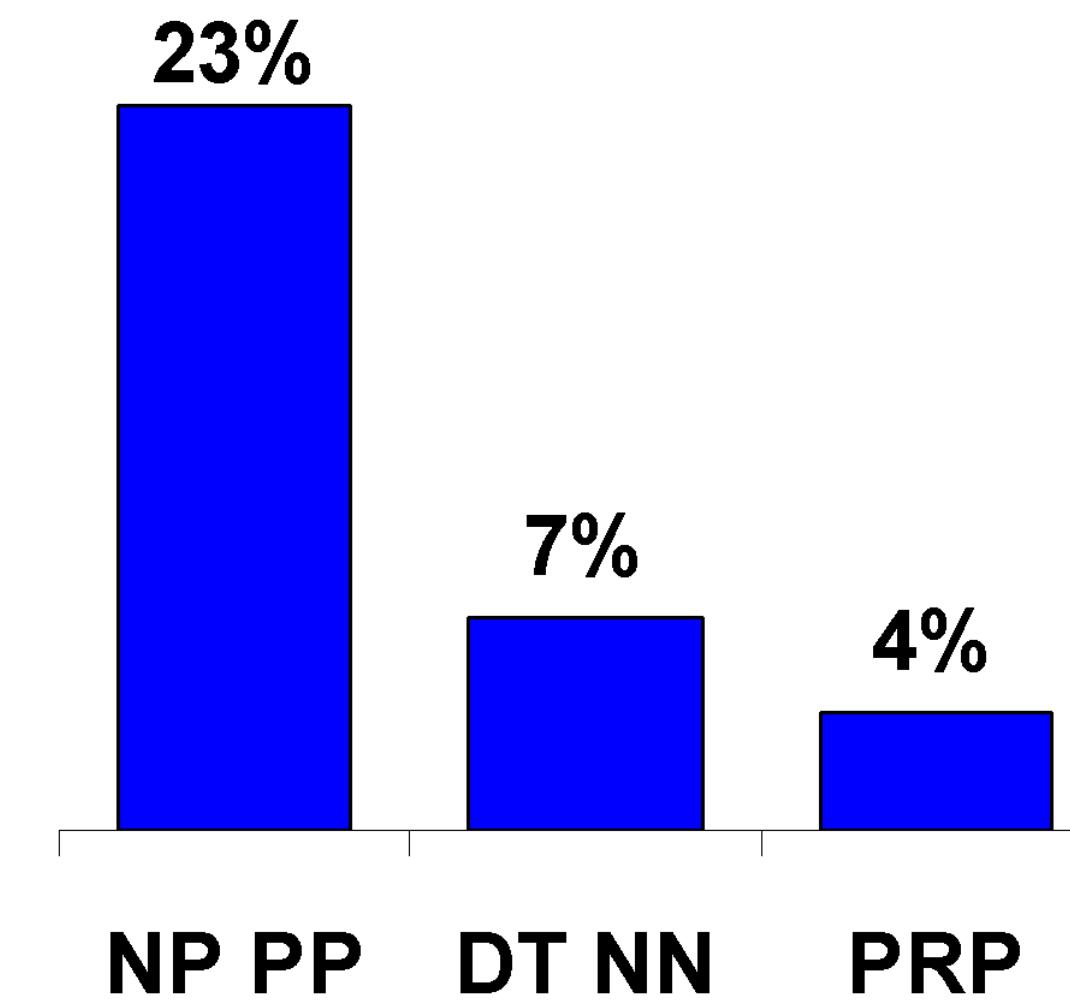
All NPs



NPs under S



NPs under VP



# 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 **learned from data**
  - Normally requires lots of data about any particular phenomenon

# Why is NLP hard?

1. Ambiguity
2. Scale
3. **Sparsity**
4. Variation
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

# Sparsity

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

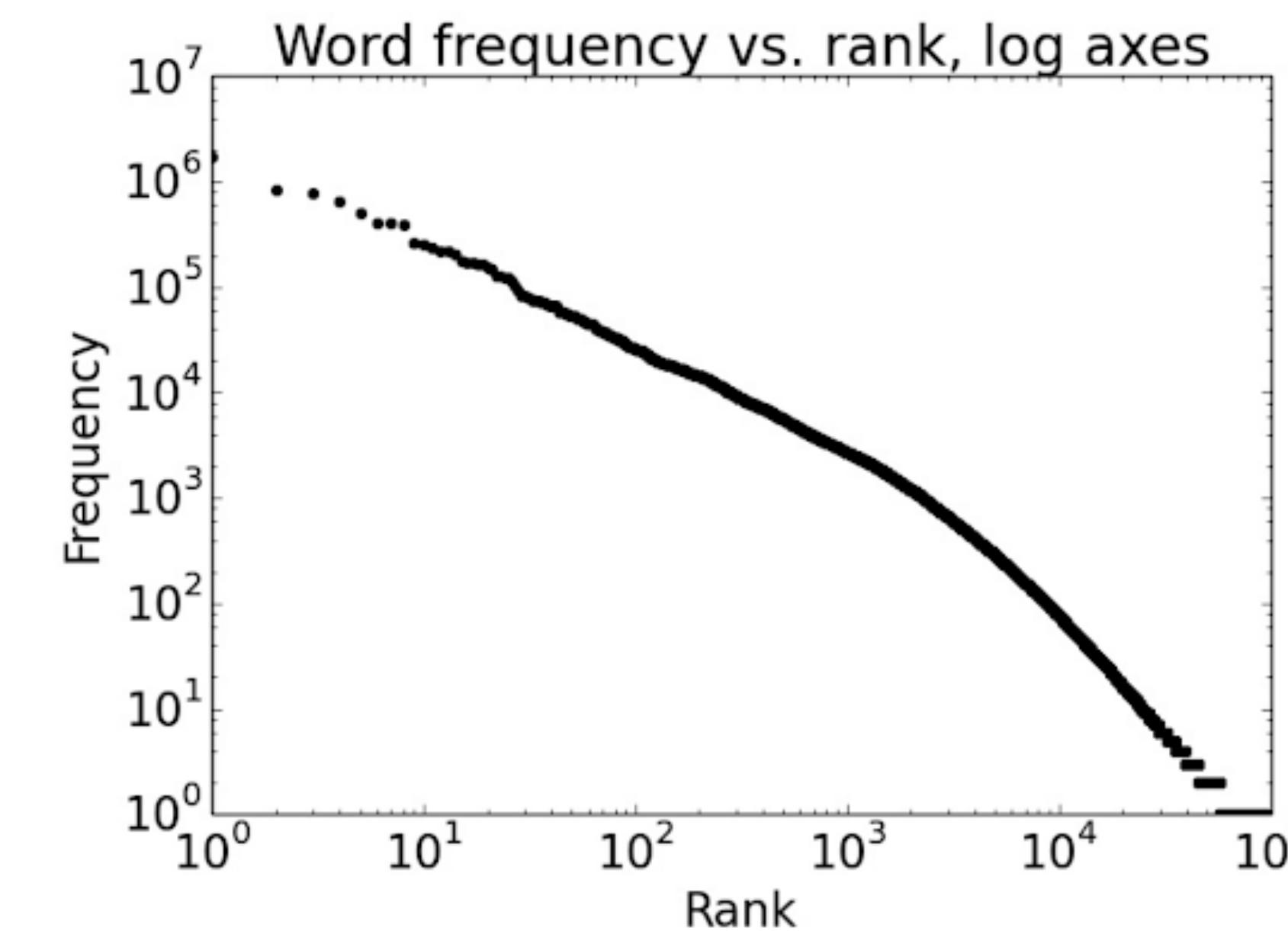
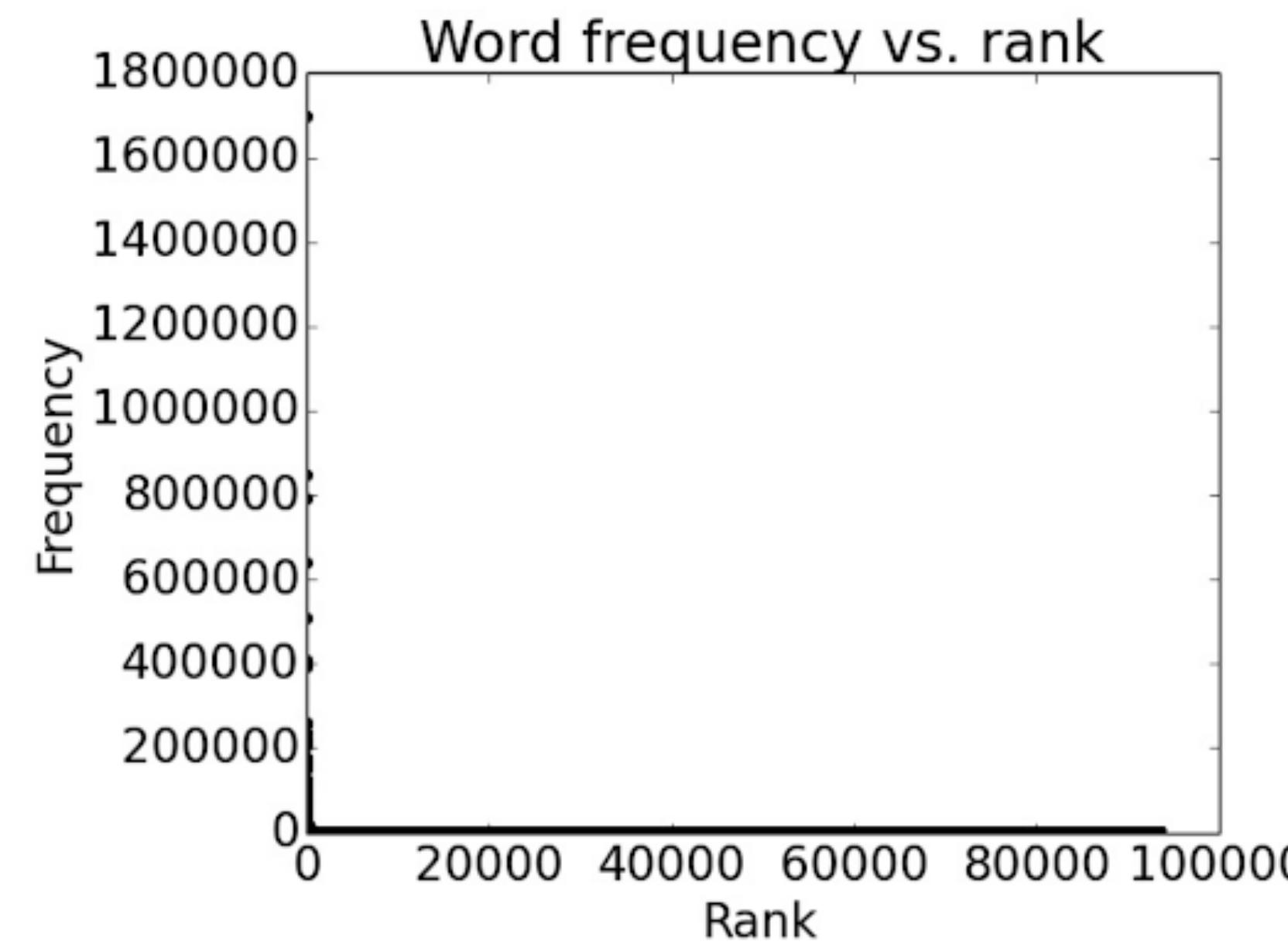
any word		nouns	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States

# Word counts

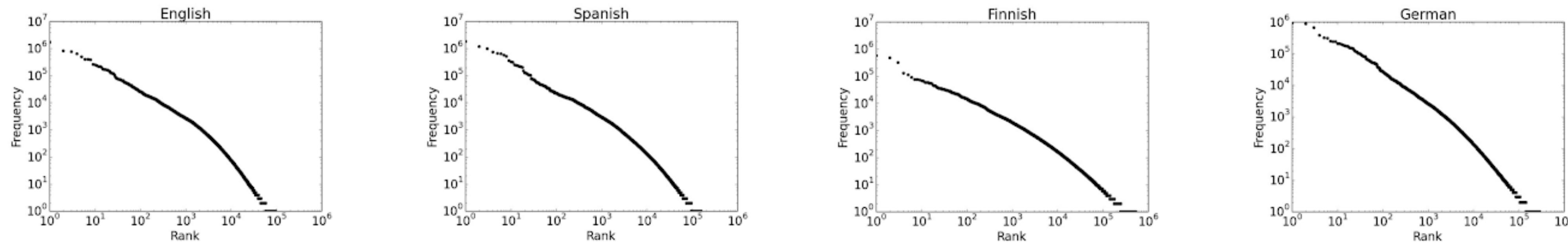
- Out of 93,638 distinct words (**types**), 36,231 (~40%) occur only once.
- Examples:
  - cornflakes, mathematicians, fuzziness, jumbling
  - pseudo-rapporteur, lobby-ridden, perfunctorily
  - Lycketoft, UNCITRAL, H-0695
  - policyfor, Commissioneris, 145.95, 27a

# Plotting word frequencies

- Order words by frequency. What is the frequency of  $n$ th 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 NLP hard?

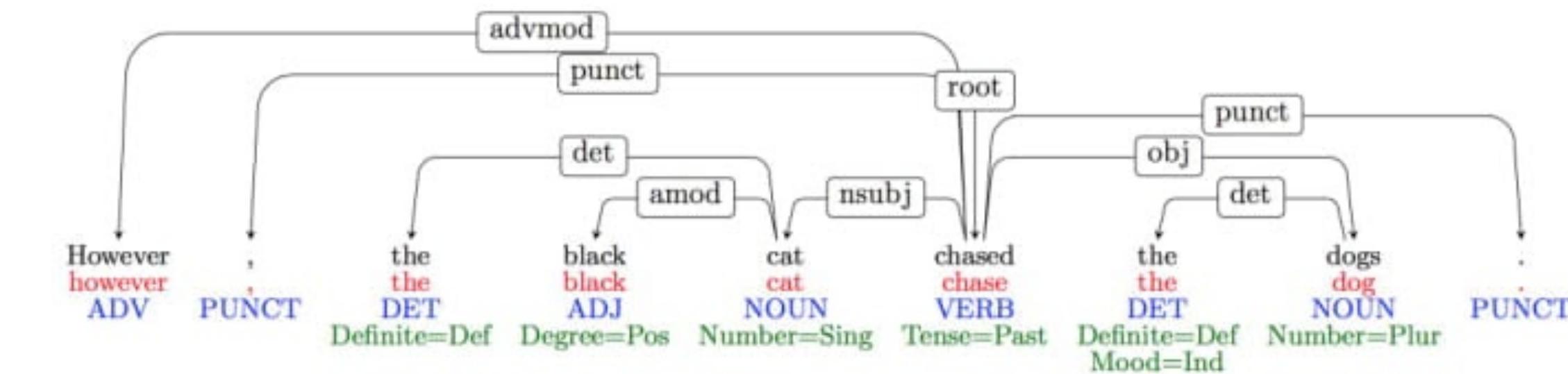
1. Ambiguity
2. Scale
3. Sparsity
- 4. Variation**
5. Expressivity
6. Unmodeled variables
7. Unknown representation  $\mathcal{R}$

# Variation

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

## THE WALL STREET JOURNAL.

Dow Jones Company, Inc.  
EE/PR \*\*\* WSJ.com  
  
( (S (NP-SBJ (NP (NNP Pierre) (NNP Vinken))  
    (, ,)  
    (ADJP (NML (CD 61) (NNS years))  
        (JJ old))  
    (, ,))  
    (VP (MD will)  
        ( VP (VB join)  
            (NP (DT the) (NN board))  
            (PP-CLR (IN as)  
                (NP (DT a) (JJ nonexecutive) (NN director)))  
            (NP-TMP (NNP Nov.) (CD 29))))  
    (. .)))



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

Trinity @christinedarvin

Hayyy, naminis ko na yung chicken wings sa UN 😔

Translate Tweet

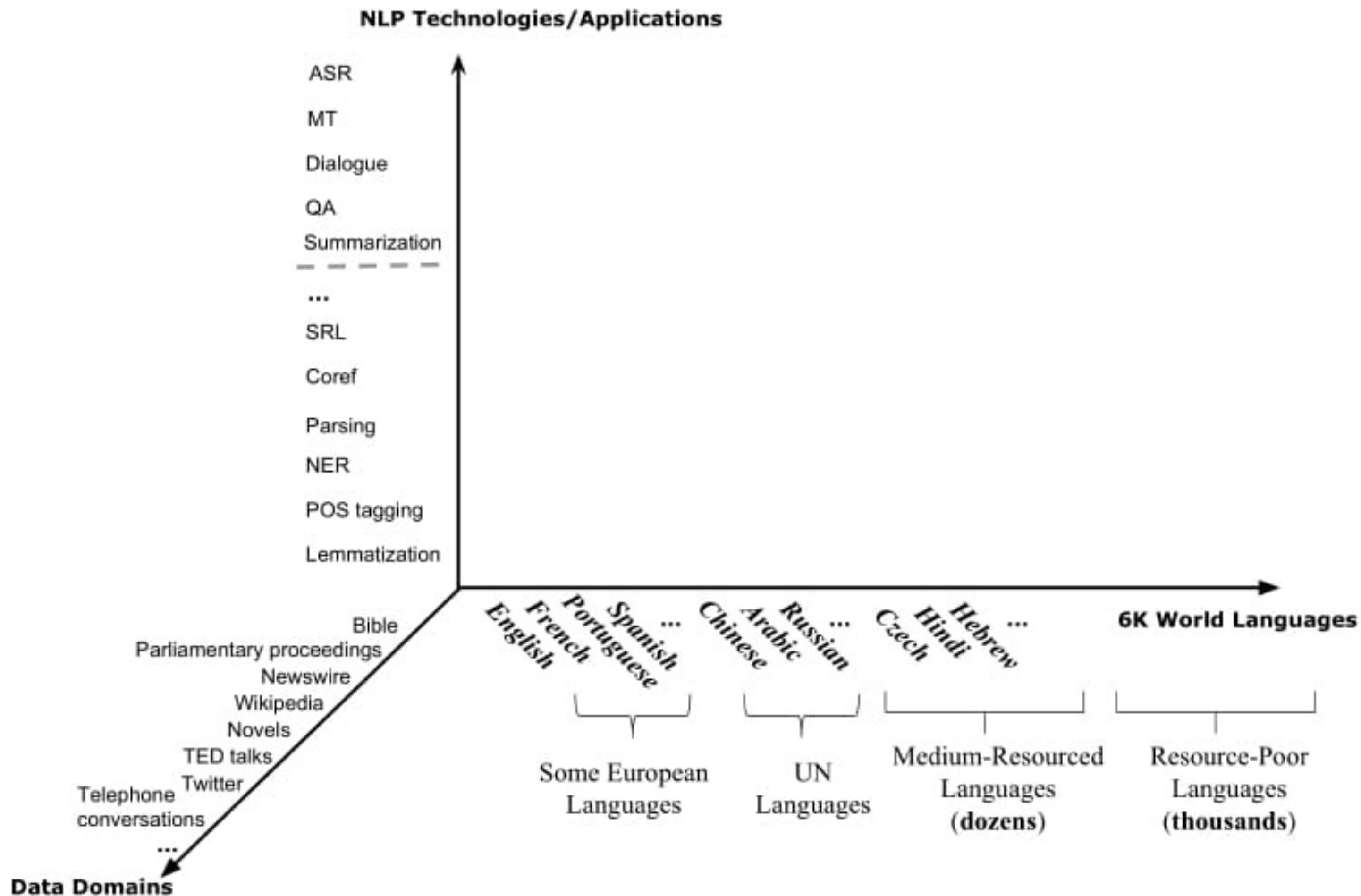
4:57 AM · Aug 17, 2020 · Twitter for Android

Domo @djxdomo

I ain't never met a gatekeeper in my life because I'm finna do whatever tf I'm finna do.

4:22 PM · Aug 31, 2020 · Twitter for iPhone

# Why is NLP Hard?



# Why is NLP hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
- 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 Deni. vs. She gave Deni the book.

Some kids popped by. vs. A few children visited.

Is that window still open? vs. Please close the window.

# Unmodeled variables

## ■ World knowledge

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



“drink this milk.”



skater eats pavement



?  
<



# Unknown representation

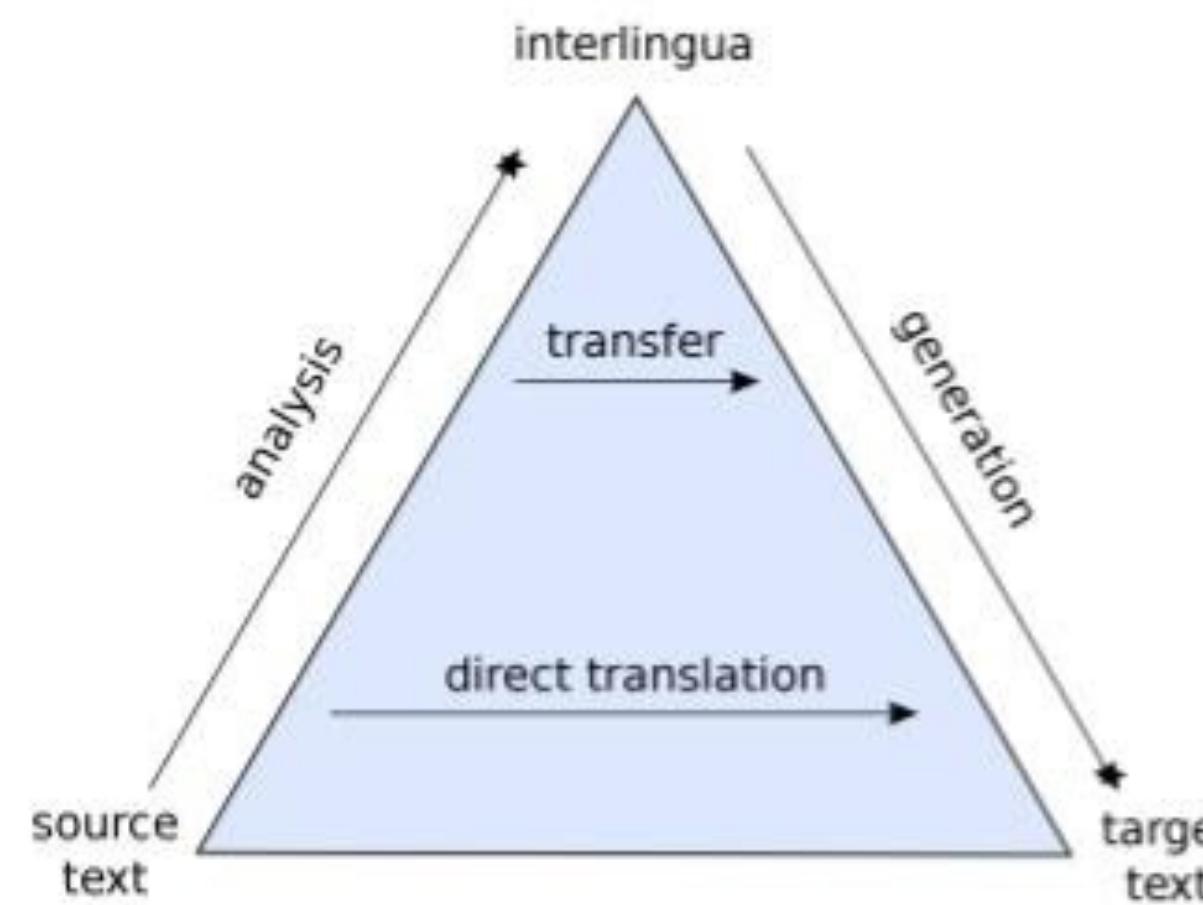
- Very difficult to capture what is  $\mathcal{R}$ , since we don't even know how to represent the knowledge a human has/needs:
  - What is the “meaning” of a word, sentence, utterance?
  - How to model context?
  - Other general knowledge?

# 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
- Efficient
- Ethical

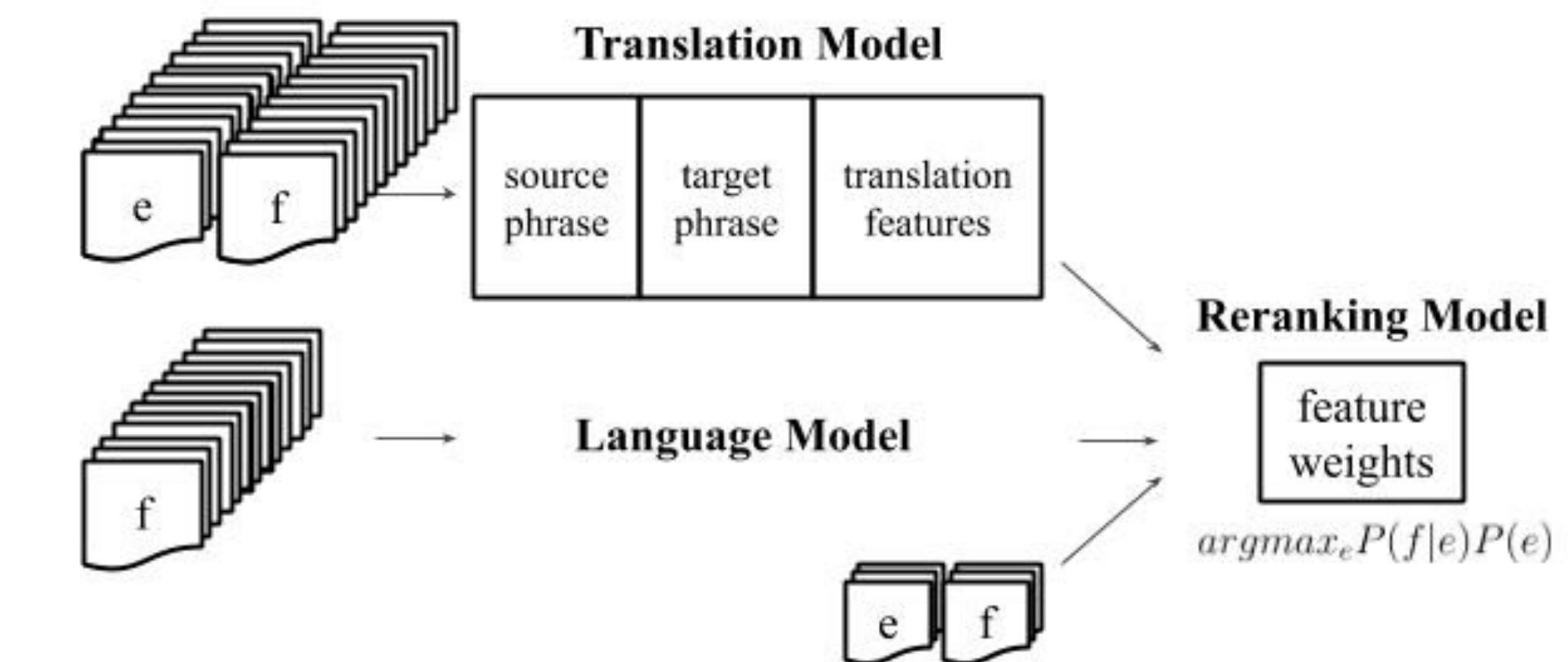
# Symbolic and probabilistic NLP

Logic/Rule-based NLP



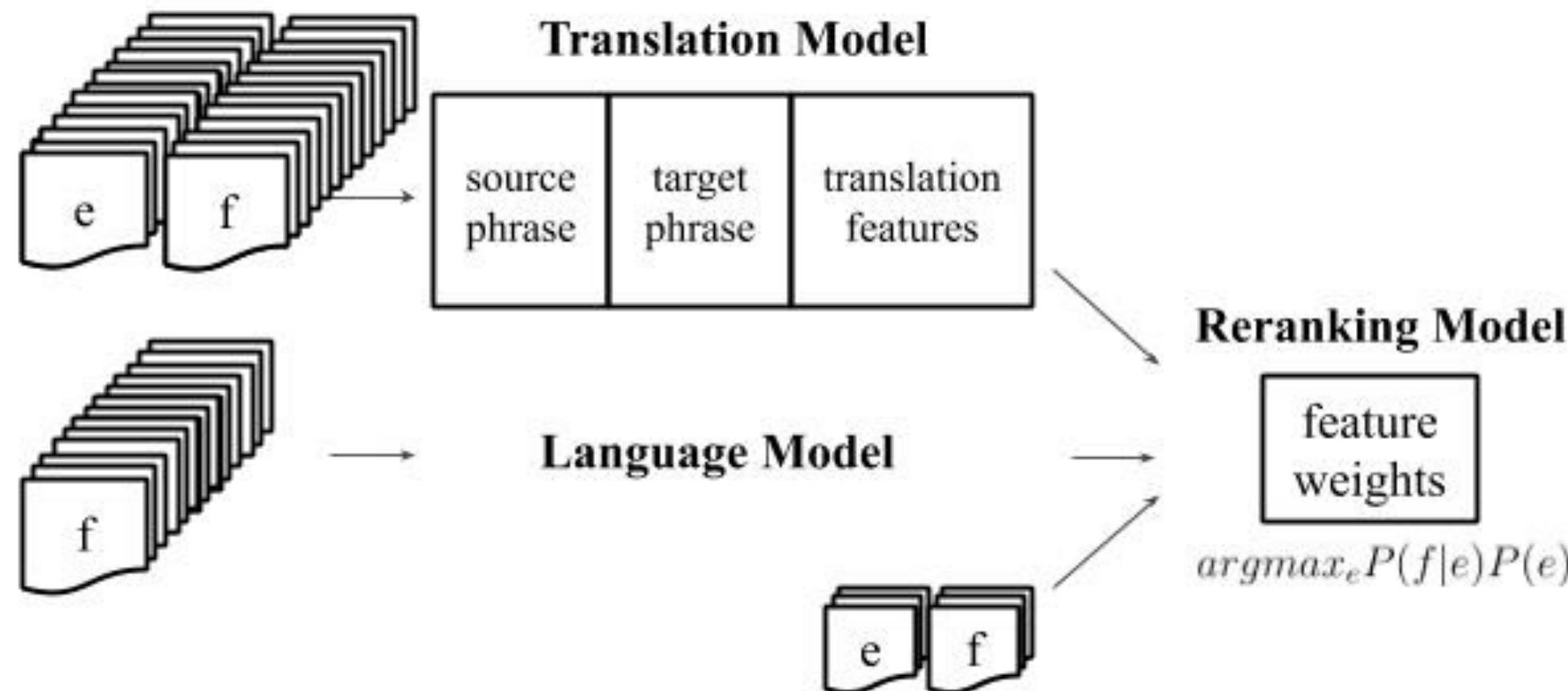
~1990s

Statistical NLP

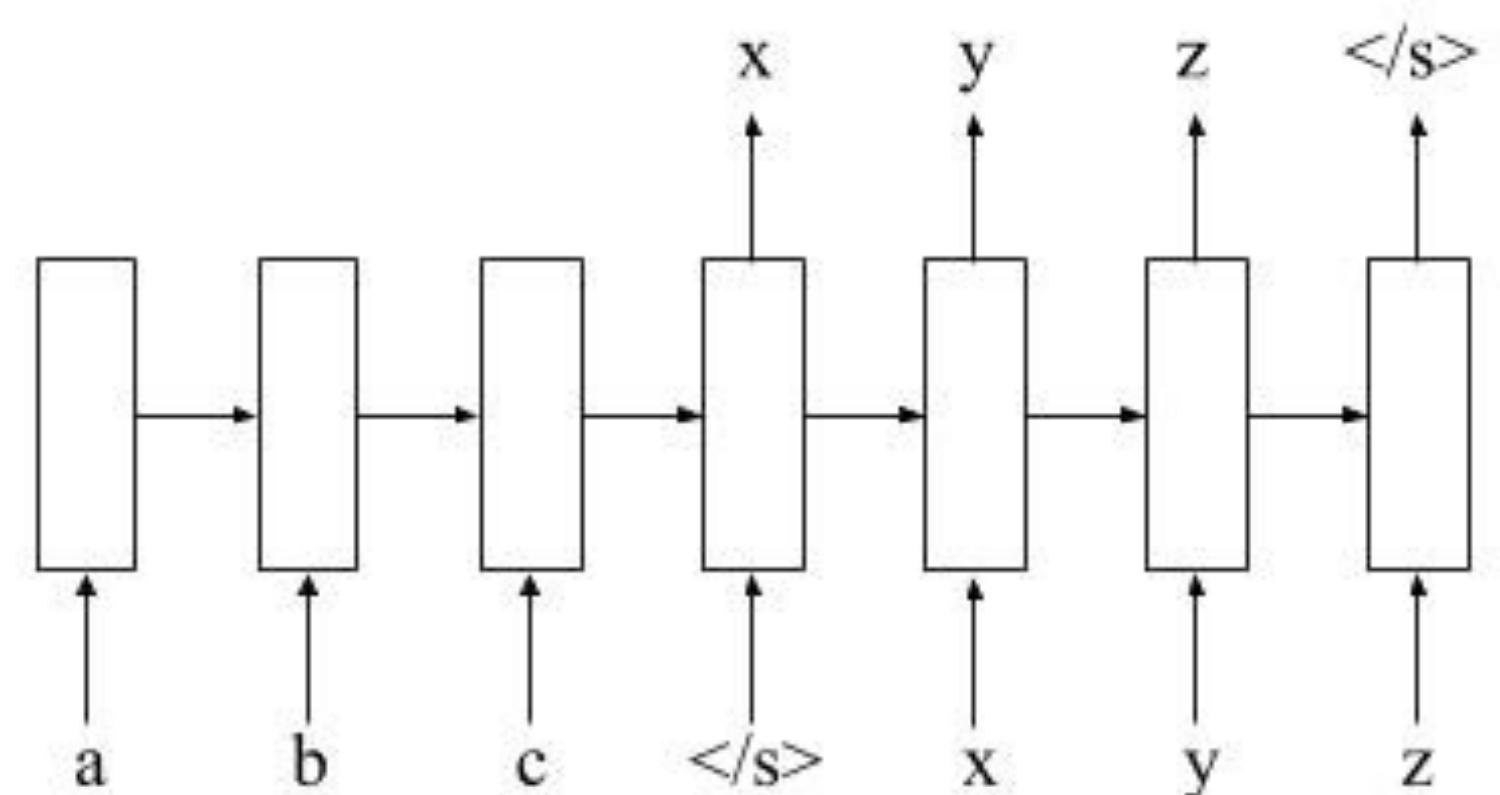


# Probabilistic and Connectionist NLP

Engineered features



~mid 2010s

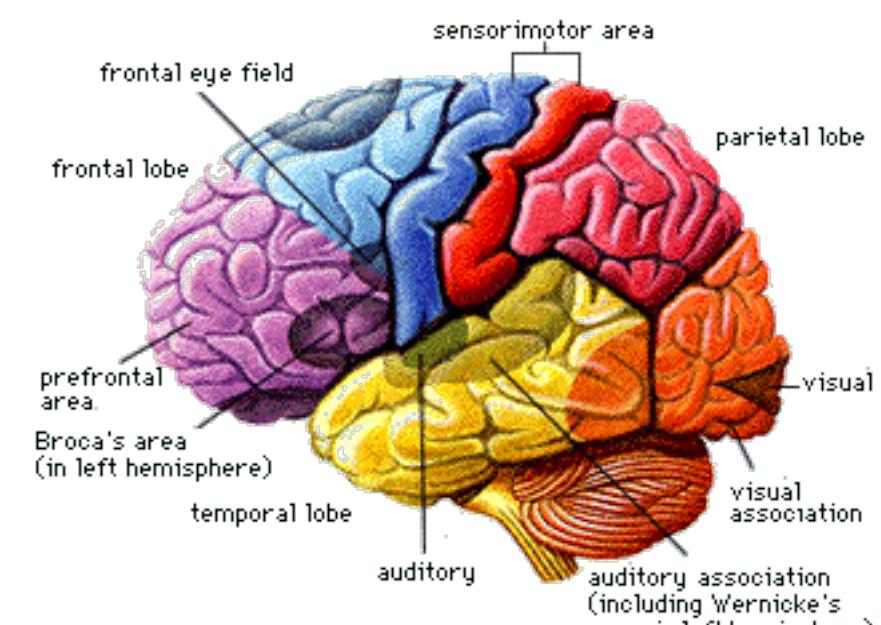
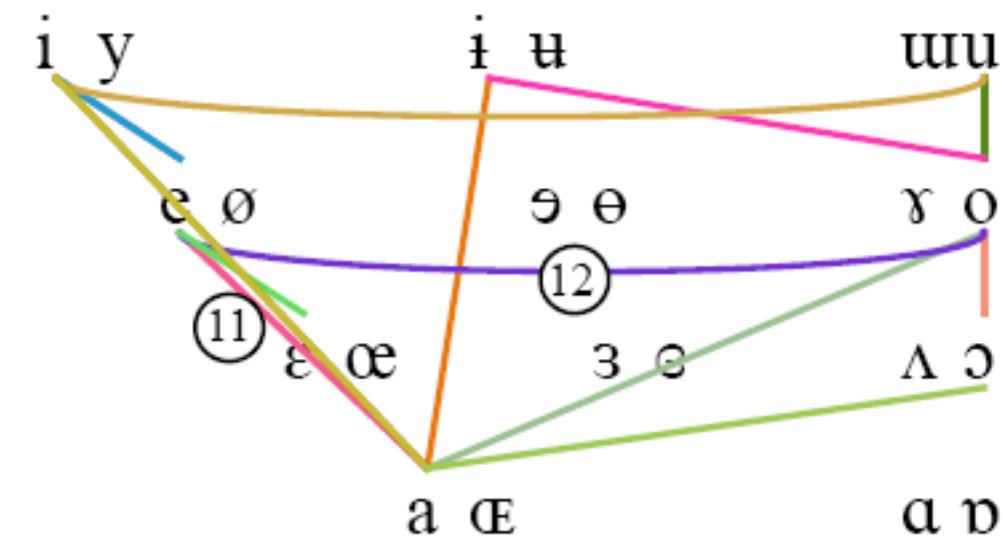


# NLP $\stackrel{?}{=}$ Machine Learning

- 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 better understand how language works
    - We end up doing this and using it
  - Cognitive science
    - How does the human brain work?
    - 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 statistical NLP



# **Course logistics**

# What is this class?

## Three aspects to the course:

- Linguistic issues

- What are the range of language phenomena?
- What are the knowledge sources that let us disambiguate?
- What representations are appropriate?
- How do you know what to model and what not to model?

- Statistical modeling methods

- Increasingly complex model structures
- Learning and parameter estimation
- Efficient inference: dynamic programming, search, sampling

- Engineering methods

- Issues of scale. We'll focus on what makes problems hard, and what works in practice.

# What is this class?

## Models and algorithms:

- Models

- State machines (finite state automata / transducers)
- Logic (first-order logic)
- Probabilistic models (WFST, language models, HMM, SVM, CRF, ...)
- Vector space models (embeddings, Transformers)

- Algorithms

- State space search (DFS, BFS, A\*, dynamic programming – Viterbi, CKY)
- Supervised learning
- Unsupervised learning

- Methodological tools

- Training/test sets, evaluation, cross-validation

# What is this class?

## Outline of topics:

- Words and sequences
  - Classification
  - Language modeling
  - Vector semantics and word embeddings
  - Sequence labeling
- Structured classification
  - Parsing
- Morphology
- Semantics
- Discourse
- Applications
  - Machine translation, summarization, sentiment analysis, computational ethics

# What is this class?

## Grading

- This is a project-based course. Grading will be done based on 5 homework assignments, as well as class participation.
  - Four projects each worth 20% of your final grade: Classification, sequence labeling, dependency parsing, and coreference.
  - One review/presentation on a paper from this year's EMNLP: 10%
  - Participation on Piazza: 10%

# What is this class?

## Requirements and goals

- Class requirements
  - Uses a variety of skills/knowledge
    - Probability and statistics, graphical models
    - Basic linguistics background
    - Strong coding skills (Python)
  - Most people are probably missing one of the above. You will have to work on your own to fill the gaps.
- Class goals
  - Learn the issues and techniques of statistical NLP
  - Build realistic NLP tools
  - Be able to read current research papers in the field

# What is this class?

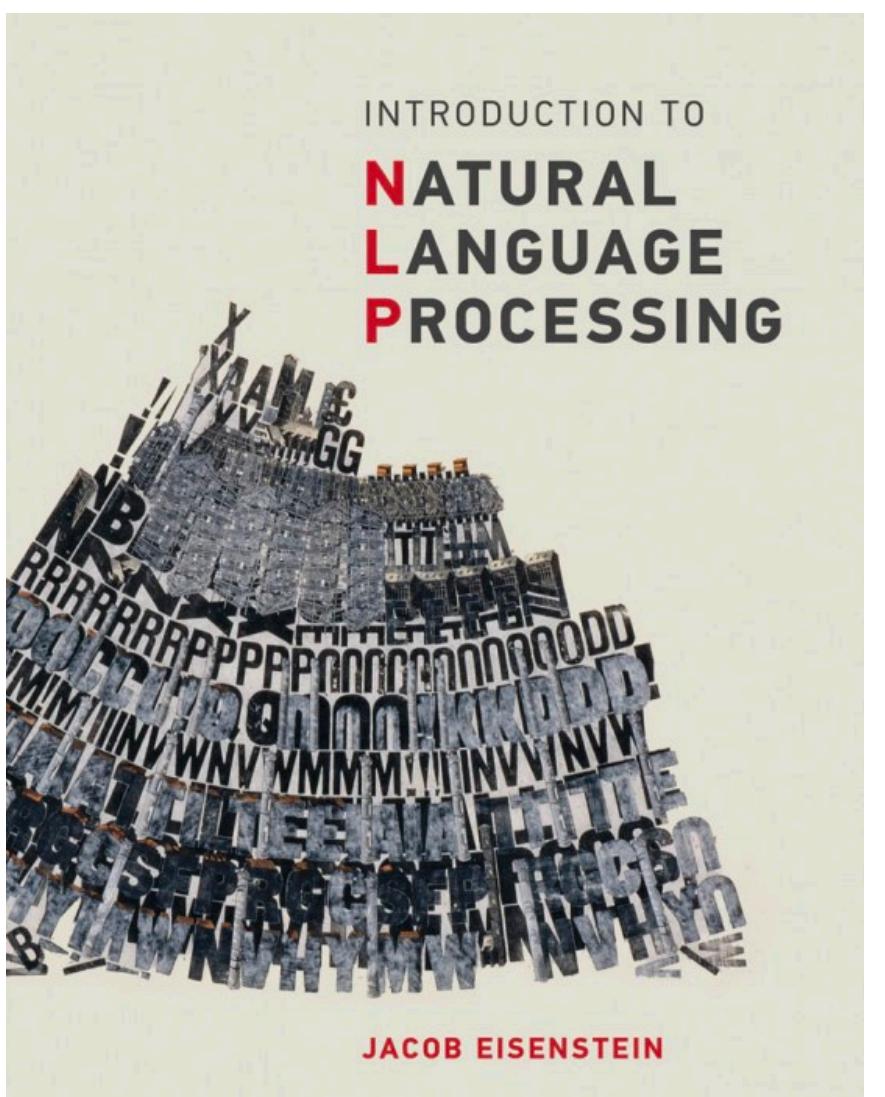
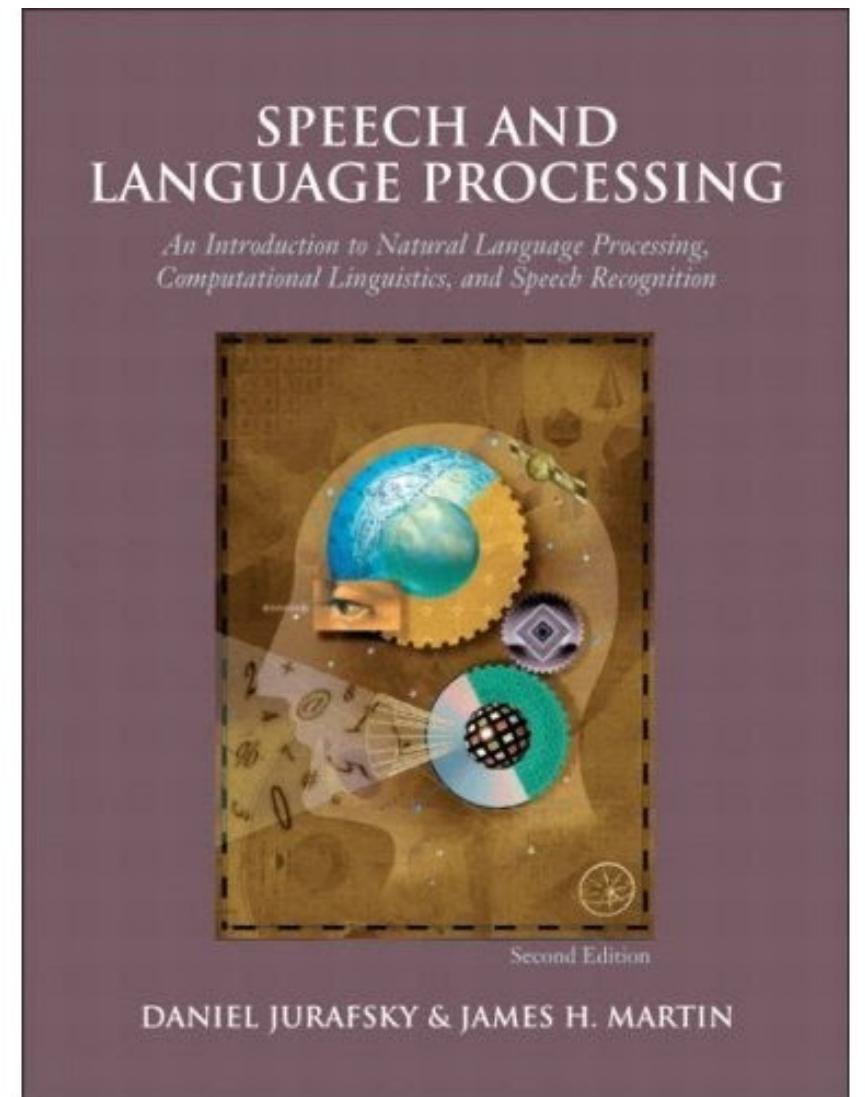
## Readings

### ■ Primary texts

- Jurafsky and Martin, Speech and Language Processing, 2nd and 3rd Edition (not 1st)  
3rd ed: <https://web.stanford.edu/~jurafsky/slp3/>
- Eisenstein, Introduction to Natural Language Processing  
<https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf>

### ■ Prerequisites

- Mastery of basic probability
- Strong skills in Python or equivalent
- Deep interest in language



# Other announcements

- Course contacts:
  - Webpage: materials and announcements
  - Piazza: discussion forum
  - Canvas: project submissions
  - Homework questions: Recitation, Piazza, TA office hours
- Enrollment: We'll try to take everyone who meets the requirements
- Computing resources:
  - Homeworks should run on relatively modern hardware just fine.  
If you need access to more compute for extra credit experimentation, we may be able to arrange for cloud credits.
- Questions?