

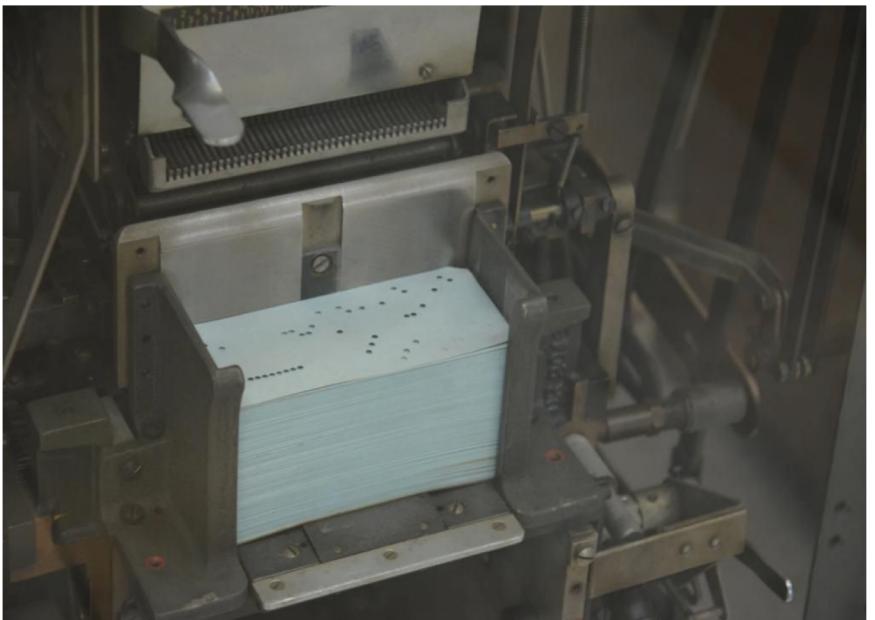
Natural Language Processing

**Introduction to NLP**

# **Introduction to NLP**



# Communication With Machines



~50-70s

```
File Edit Edit_Settings Menu Utilities Compilers Test Help
EDIT      BS9U.DEVT3.CLIBPAU(TIMMIES) - 01.31          Columns 00001 00
Command ==> 
*****
***** Top of Data *****
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 */
```

~80s

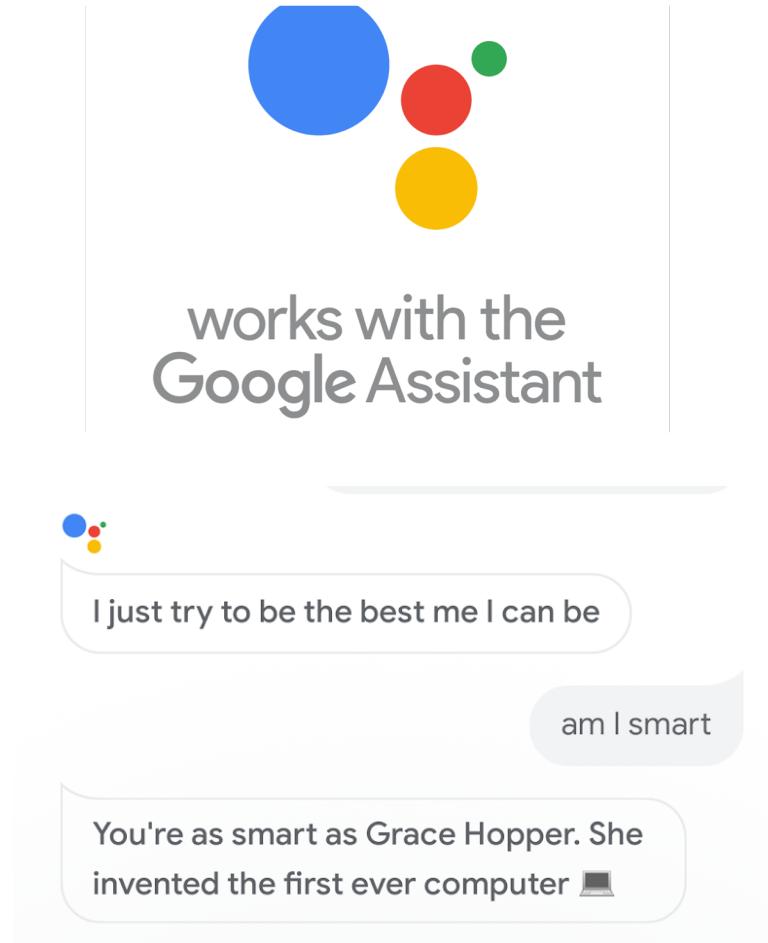


today

# Conversational Agents

Conversational agents contain:

- Speech recognition
- Language analysis
- Dialogue processing
- Information retrieval
- Text to speech



A black and white photograph of a large, modern building with a curved glass facade and a prominent white overhang. The building appears to be a corporate headquarters or a large office complex. In the foreground, there is a paved area with some low-lying plants and a set of stairs leading up to the entrance.

In early 2011, an IBM computing system named Watson competed against the world's best Jeopardy! champions.

# 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 in the end of the 18th century?
- What do scientists think about the ethics of human cloning?

# Machine Translation

Google Translate

CHINESE - DETECTED ↔ ENGLISH

我学习深度学习和机器学习

Wǒ xuéxí shēndù xuéxí hé jīqì xuéxí

I study deep learning and machine learning.

Send feedback

Google Translate

Text Documents

DETECT LANGUAGE ENGLISH SPANISH FRENCH ↗ ENGLISH SPANISH ARABIC ↘

Search languages

Detected language	Czech	Hebrew	Latin	Portuguese	Tajik
Afrikaans	Danish	Hindi	Latvian	Punjabi	Tamil
Albanian	Dutch	Hmong	Lithuanian	Romanian	Telugu
Amharic	English	Hungarian	Luxembourgish	Russian	Thai
Arabic	Esperanto	Icelandic	Macedonian	Samoan	Turkish
Armenian	Estonian	Igbo	Malagasy	Scots Gaelic	Ukrainian
Azerbaijani	Filipino	Indonesian	Malay	Serbian	Urdu
Basque	Finnish	Irish	Malayalam	Sesotho	Uzbek
Belarusian	French	Italian	Maltese	Shona	Vietnamese
Bengali	Frisian	Japanese	Maori	Sindhi	Welsh
Bosnian	Galician	Javanese	Marathi	Sinhala	Xhosa
Bulgarian	Georgian	Kannada	Mongolian	Slovak	Yiddish
Catalan	German	Kazakh	Myanmar (Burmese)	Slovenian	Yoruba
Cebuano	Greek	Khmer	Nepali	Somali	Zulu
Chichewa	Gujarati	Korean	Norwegian	Spanish	
Chinese	Haitian Creole	Kurdish (Kurmanji)	Pashto	Sundanese	
Corsican	Hausa	Kyrgyz	Persian	Swahili	
Croatian	Hawaiian	Lao	Polish	Swedish	

# Natural Language Processing

## Applications

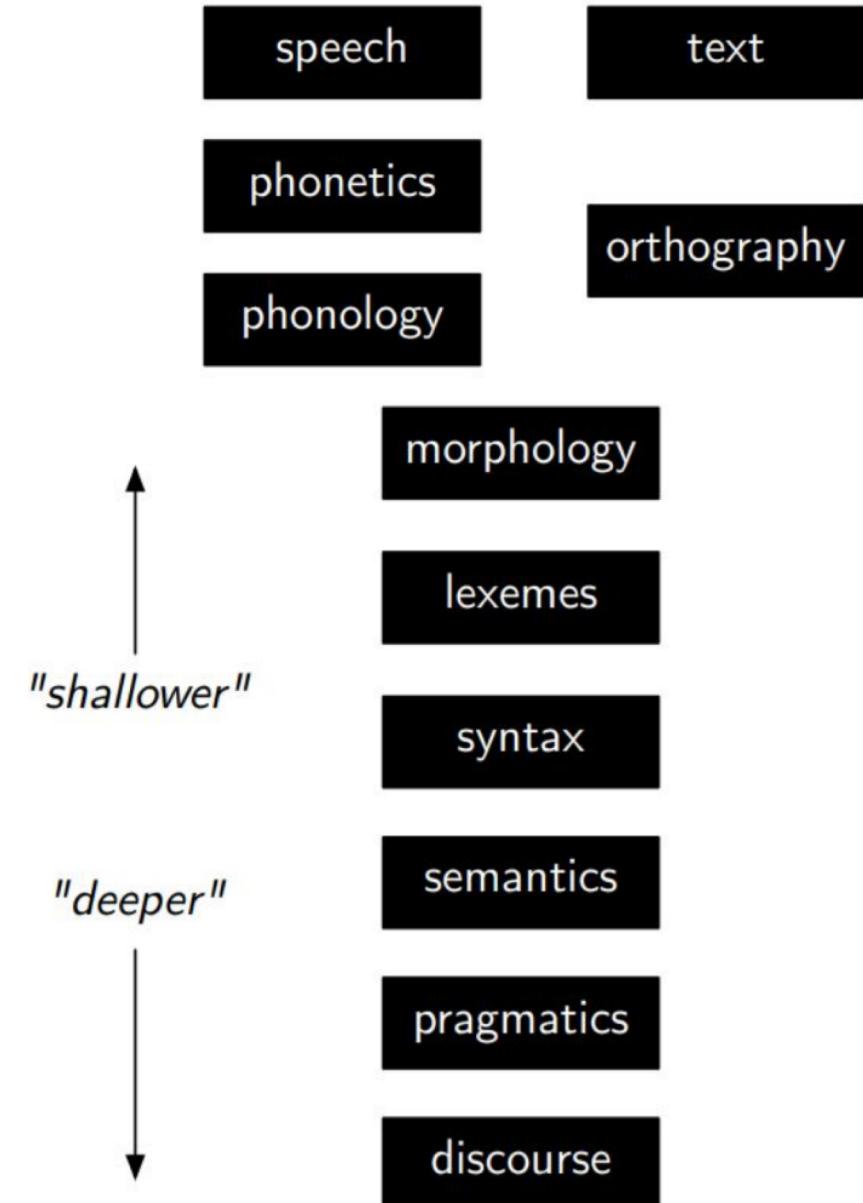
- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

## Core Technologies

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

NLP lies at the intersection of computational linguistics and machine learning.

# Level Of Linguistic Knowledge



# Phonetics, Phonology

- Pronunciation Modeling

**SOUNDS**      Th    i    a    si    e    n

# Words

- Language Modeling
- Tokenization
- Spelling correction

**WORDS**

This is a simple sentence

# Morphology

- Morphology analysis
- Tokenization
- Lemmatization

<b>WORDS</b>	This is a simple sentence
<b>MORPHOLOGY</b>	be 3sg present

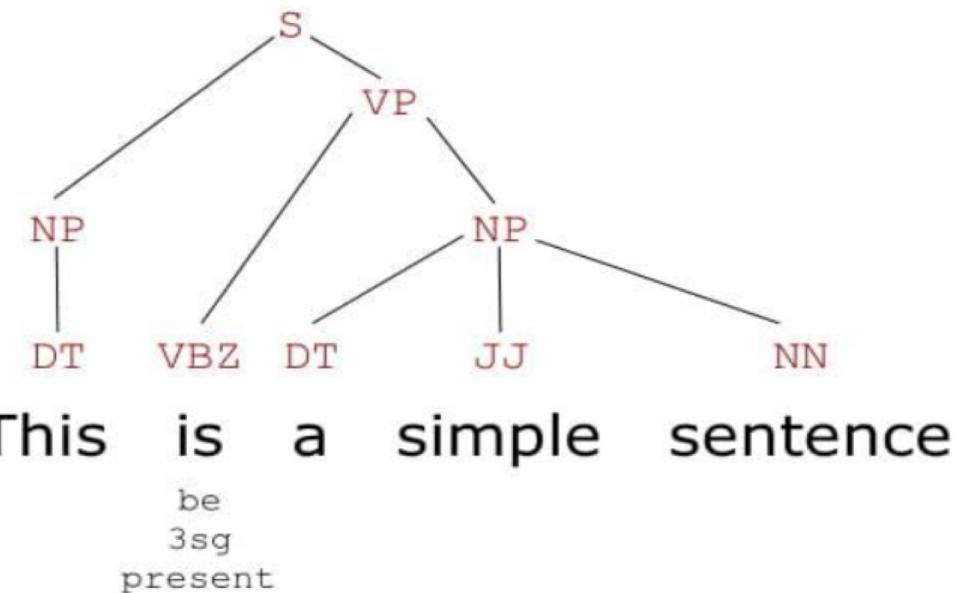
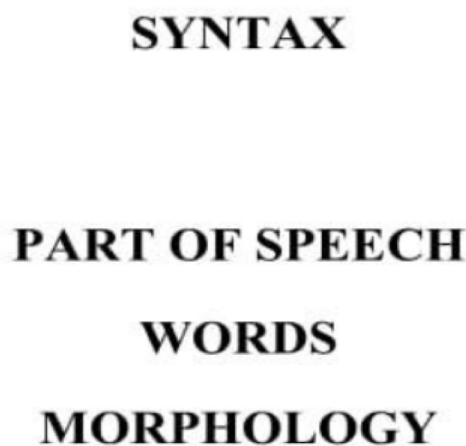
# Part of Speech

- Part of speech tagging

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

# Syntax

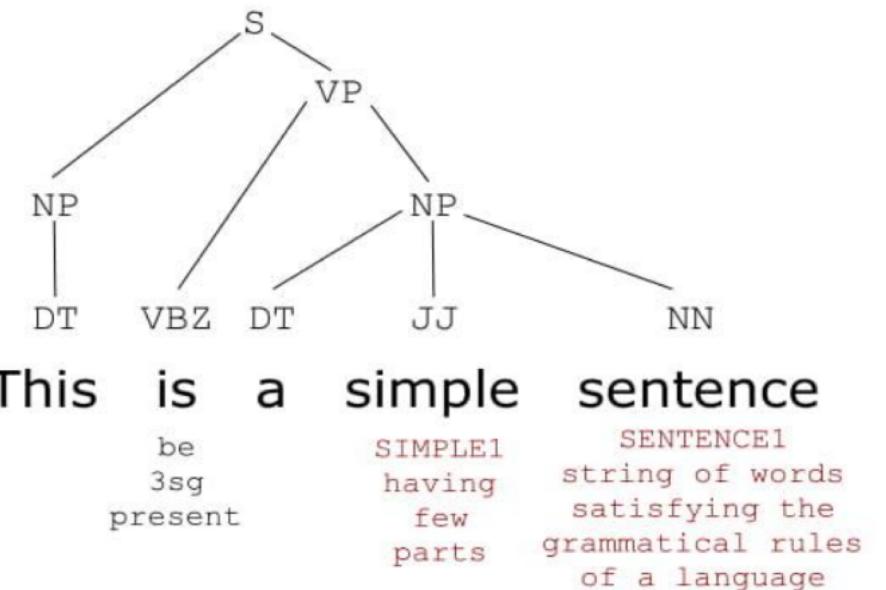
- Syntactic parsing



# Semantics

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

## SYNTAX



## PART OF SPEECH

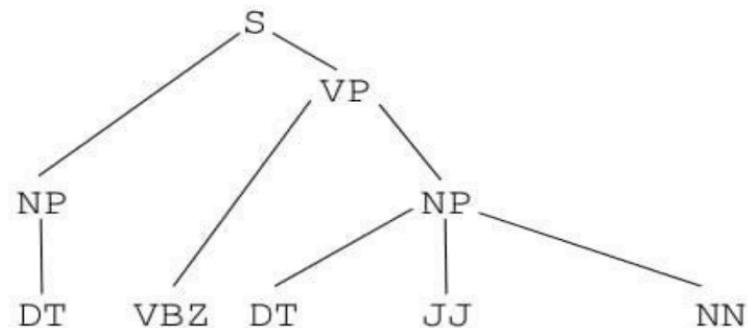
## WORDS

## MORPHOLOGY

## SEMANTICS

# Discourse

**SYNTAX**



**PART OF SPEECH**

**WORDS**

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**

CONTRAST

**DISCOURSE**

But it is an instructive one.

# Where Are We Now?



# Where Are We Now?

## Baseline mutual information model (Li et al. 2015)

A: Where are you going? (1)

B: I'm going to the restroom. (2)

A: See you later. (3)

B: See you later. (4)

A: See you later. (5)

B: See you later. (6)

...

...

A: how old are you? (1)

B: I'm 16. (2)

A: 16? (3)

B: I don't know what you are talking about. (4)

A: You don't know what you are saying. (5)

B: I don't know what you are talking about . (6)

A: You don't know what you are saying. (7)

...

VS



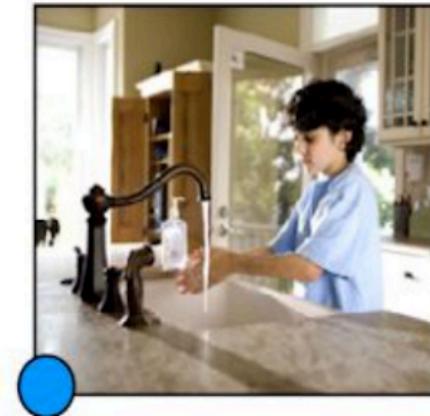
# Where Are We Now?



[https://www.theverge.com/2016/3/24/11297050  
/tay-microsoft-chatbot-racist](https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist)



woman cooking

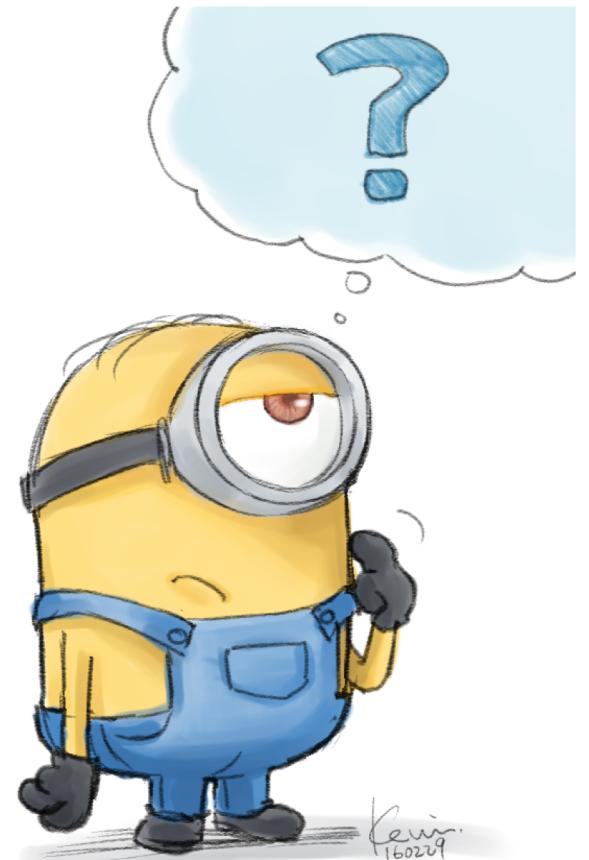


man fixing faucet

Zhao, J., Wang, T., Yatskar, M., Ordonez, V and Chang, M.-W. (2017) Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraint. EMNLP

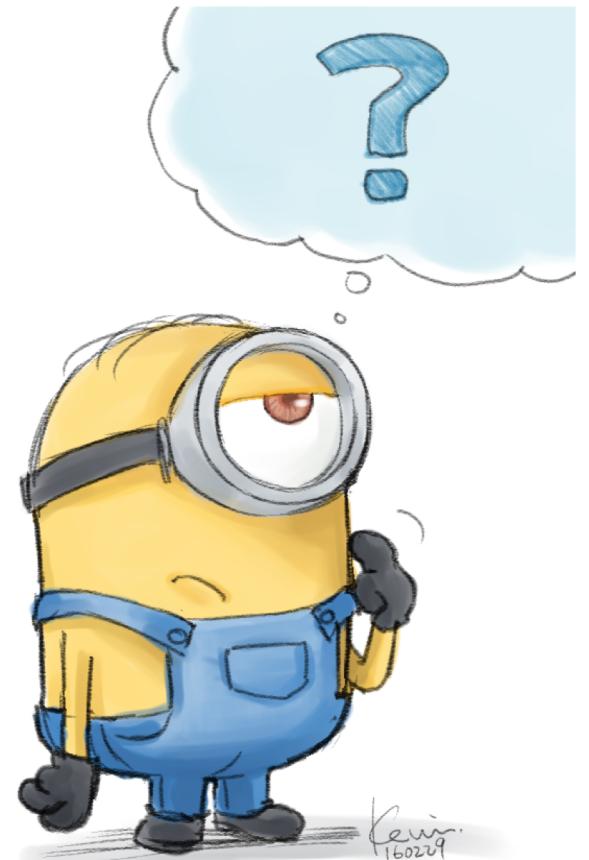
# Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations



# Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations



# 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 made her duck**





“One morning I shot  
an elephant in my pajamas”

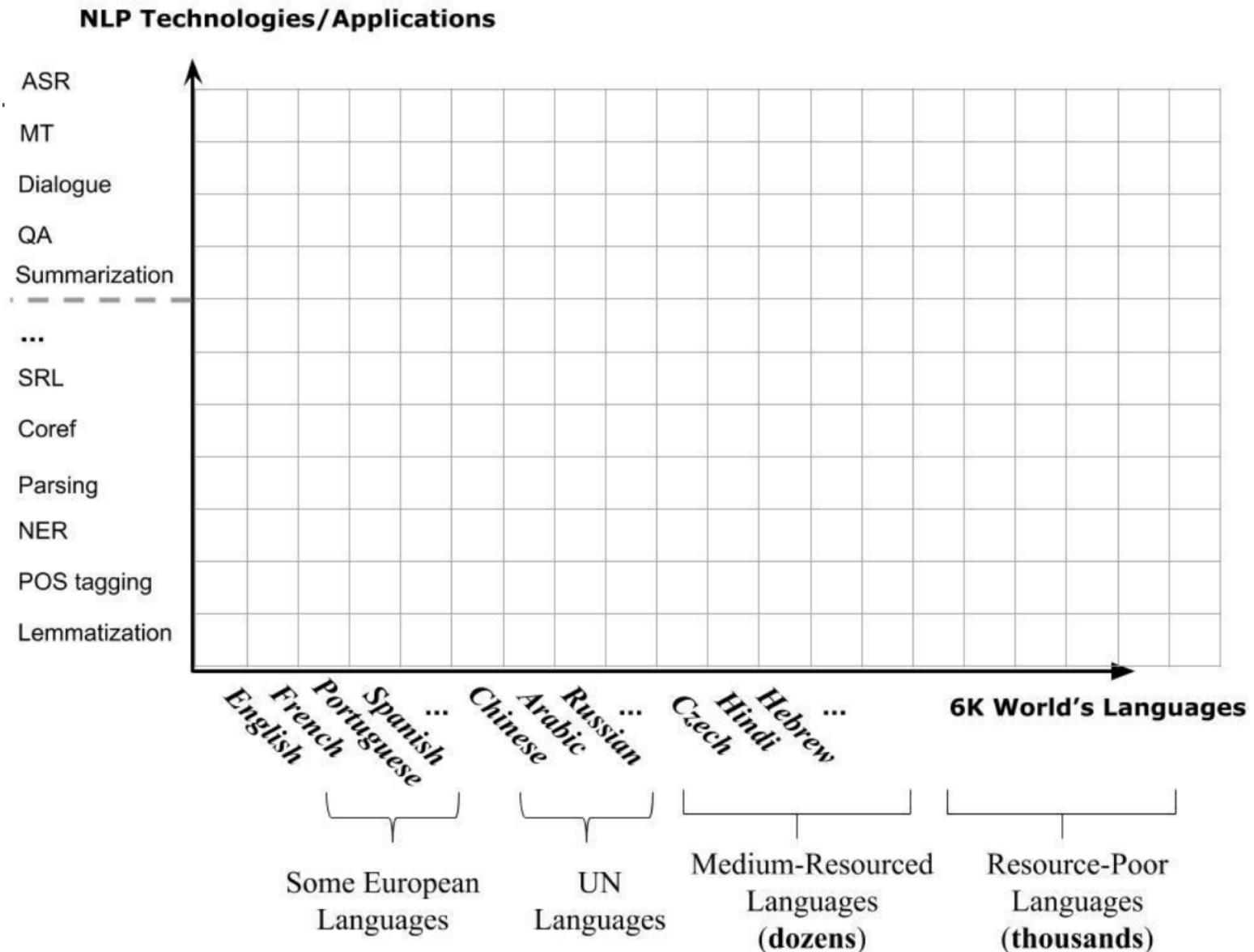


*I made her duck*

[SLP2 ch. 1]

- I cooked waterfowl for her
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- ...

# Ambiguity and Scale



# The Challenges of “Words”

- Segmenting text into words
- Morphological variation
- Words with multiple meanings: bank, mean
- Domain-specific meanings: latex
- Multiword expressions: make a decision, take out, make up

# Part of Speech Tagging

ikr smh he asked fir yo last name

so he can add u on fb lololol

# Part of Speech Tagging

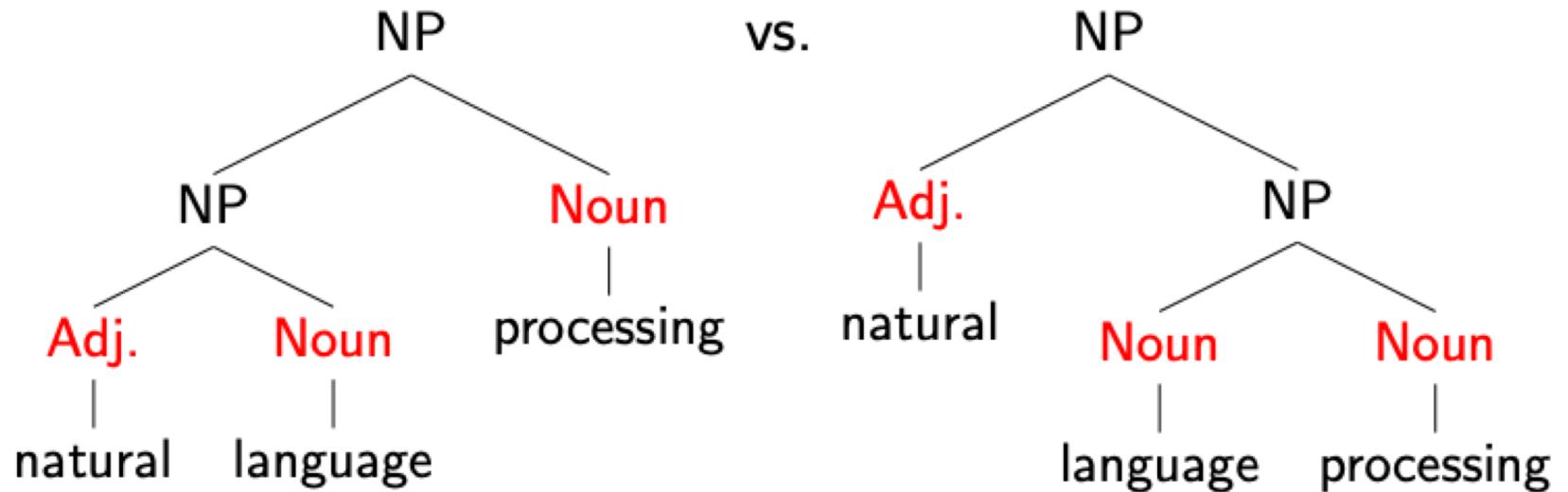
I know, right shake my head for your  
ikr smh he asked fir yo last name

you Facebook laugh out loud  
so he can add u on fb lololol

# Part of Speech Tagging

I know, right	shake my head		for	your			
ikr	smh	he	asked	fir	yo	last	name
!	G	O	V	P	D	A	N
interjection	acronym	pronoun	verb	prep.	det.	adj.	noun
		you		Facebook	laugh out loud		
so	he	can	add	u	on	fb	lololol
P	O	V	V	O	P	^	!
preposition				proper noun			

# Syntax



# Morphology + Syntax



A ship-shipping  
ship, shipping  
shipping-ships

# Semantics

- Every fifteen minutes a woman in this country gives birth.

# Semantics

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

– Groucho Marx



# Syntax + Semantics

- We saw the woman with the telescope wrapped in paper.

# Syntax + Semantics

- We saw the woman with the telescope wrapped in paper.
  - Who has the telescope?
  - Who or what is wrapped in paper?
  - An even of perception, or an assault?

# Dealing with Ambiguity

- How can we model ambiguity?
  - Non-probabilistic methods (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 analyses**, i.e., the most probable one
- 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 French/English sentences
  - Yelp reviews
  - The Web!



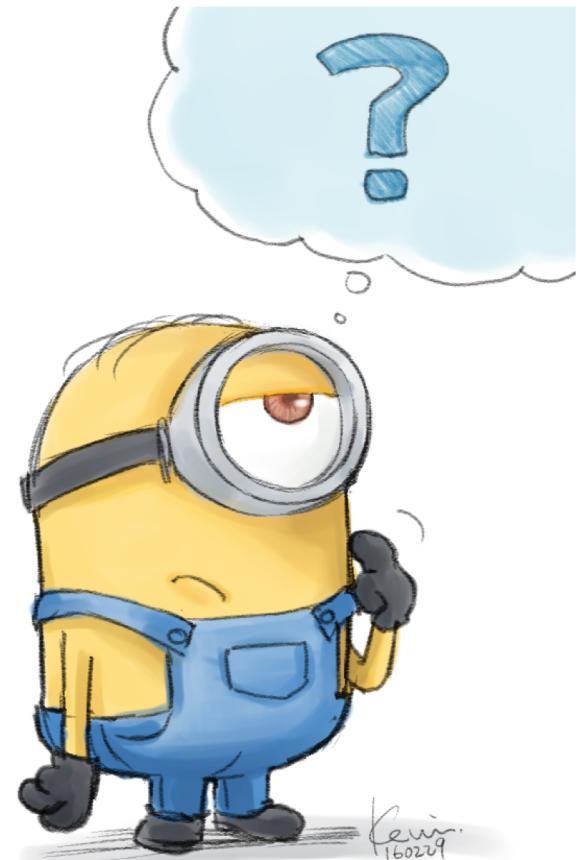
Rosetta Stone

# Statistical NLP

- Like most other parts of AI, NLP is dominated by statistical methods
  - Typically more robust than rule-based methods
  - Relevant statistics/probabilities are **learned from data**
  - Normally requires lots of data about any particular phenomenon

# Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations



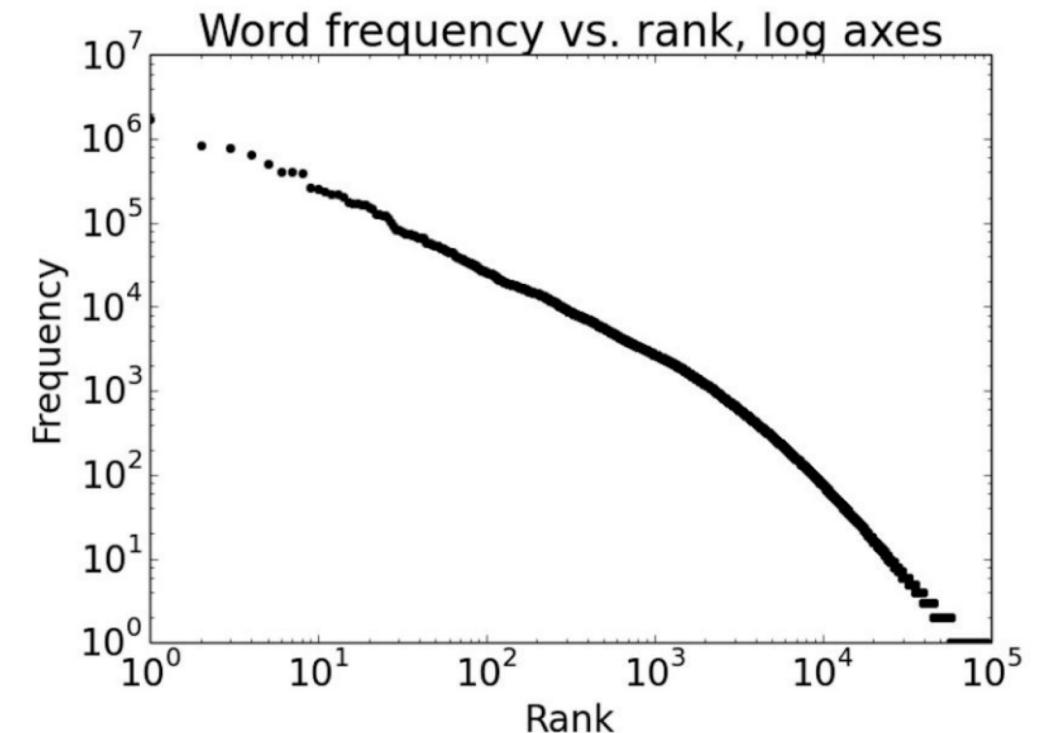
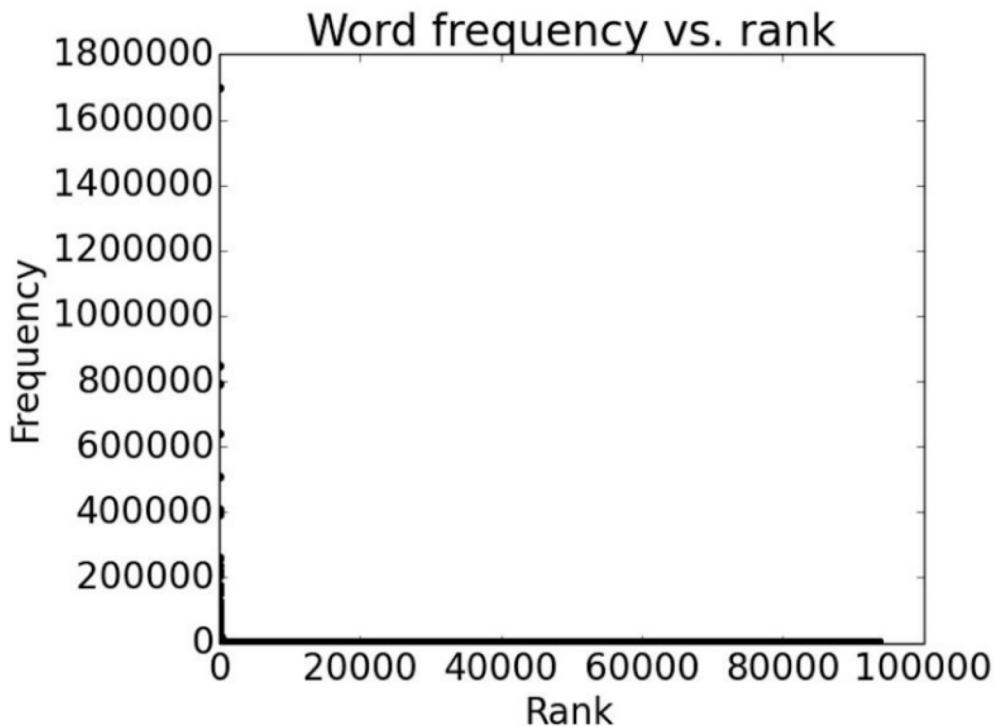
# Sparsity

- Sparse data due to **Zipf's Law**
- Example: the frequency of different words in a large text corpus

<b>any word</b>		<b>nouns</b>	
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

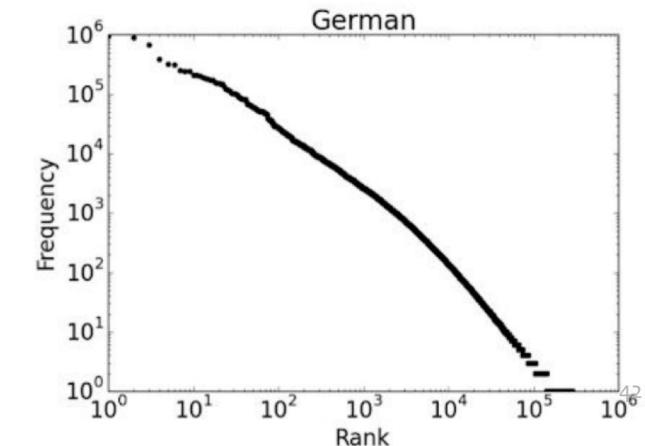
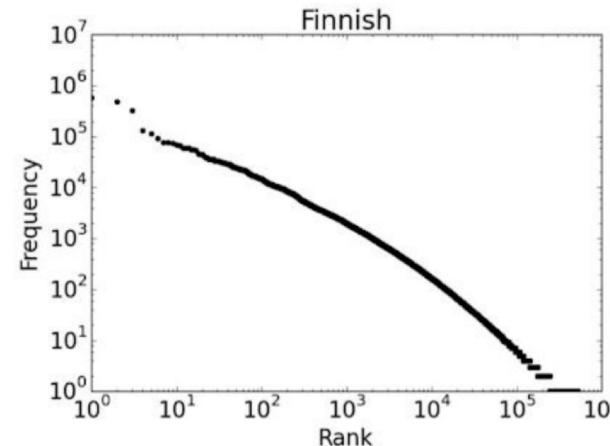
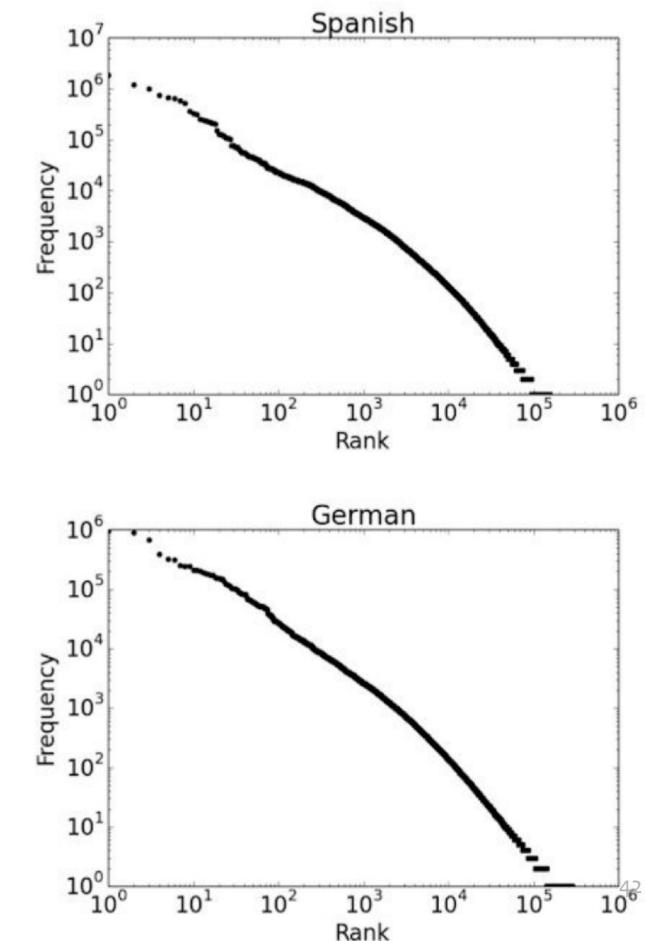
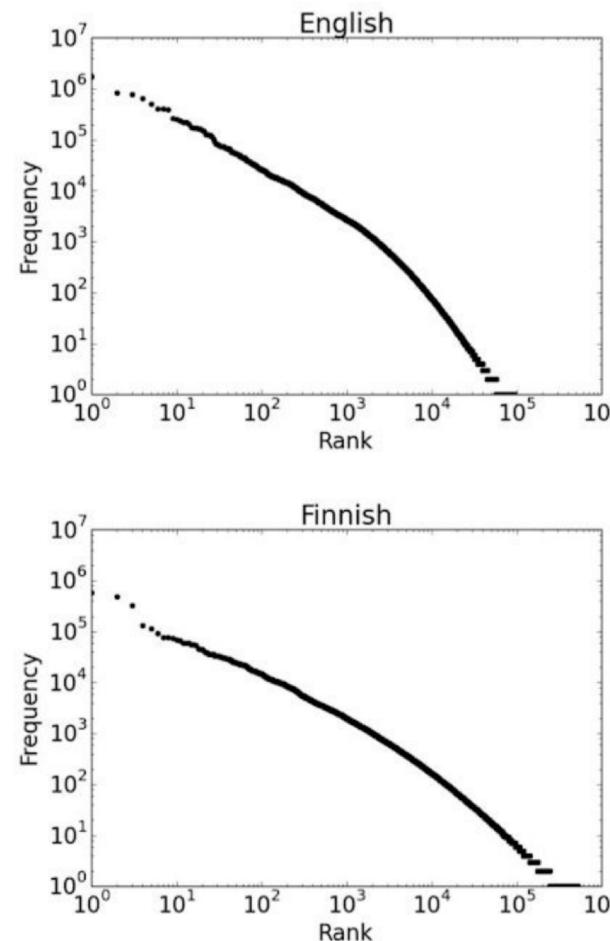
# Sparsity

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



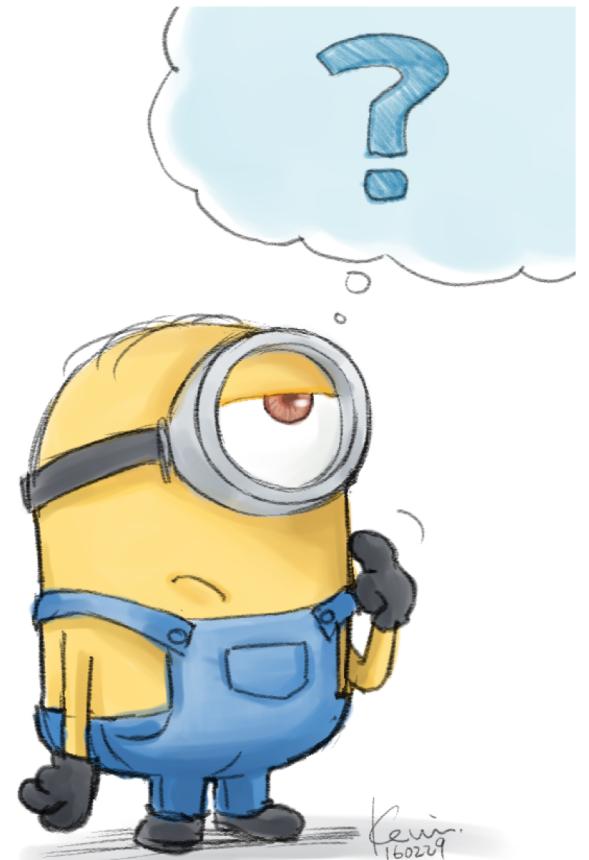
# Sparsity

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



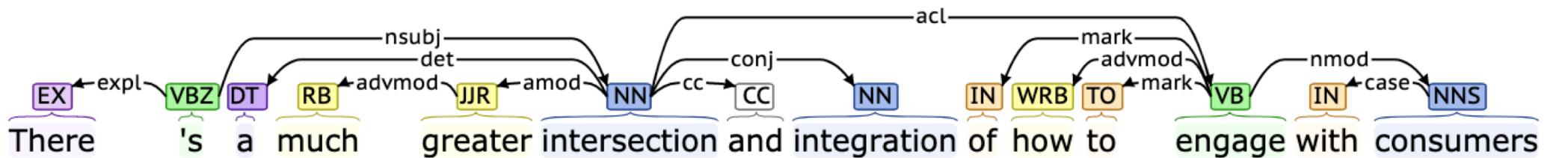
# Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations



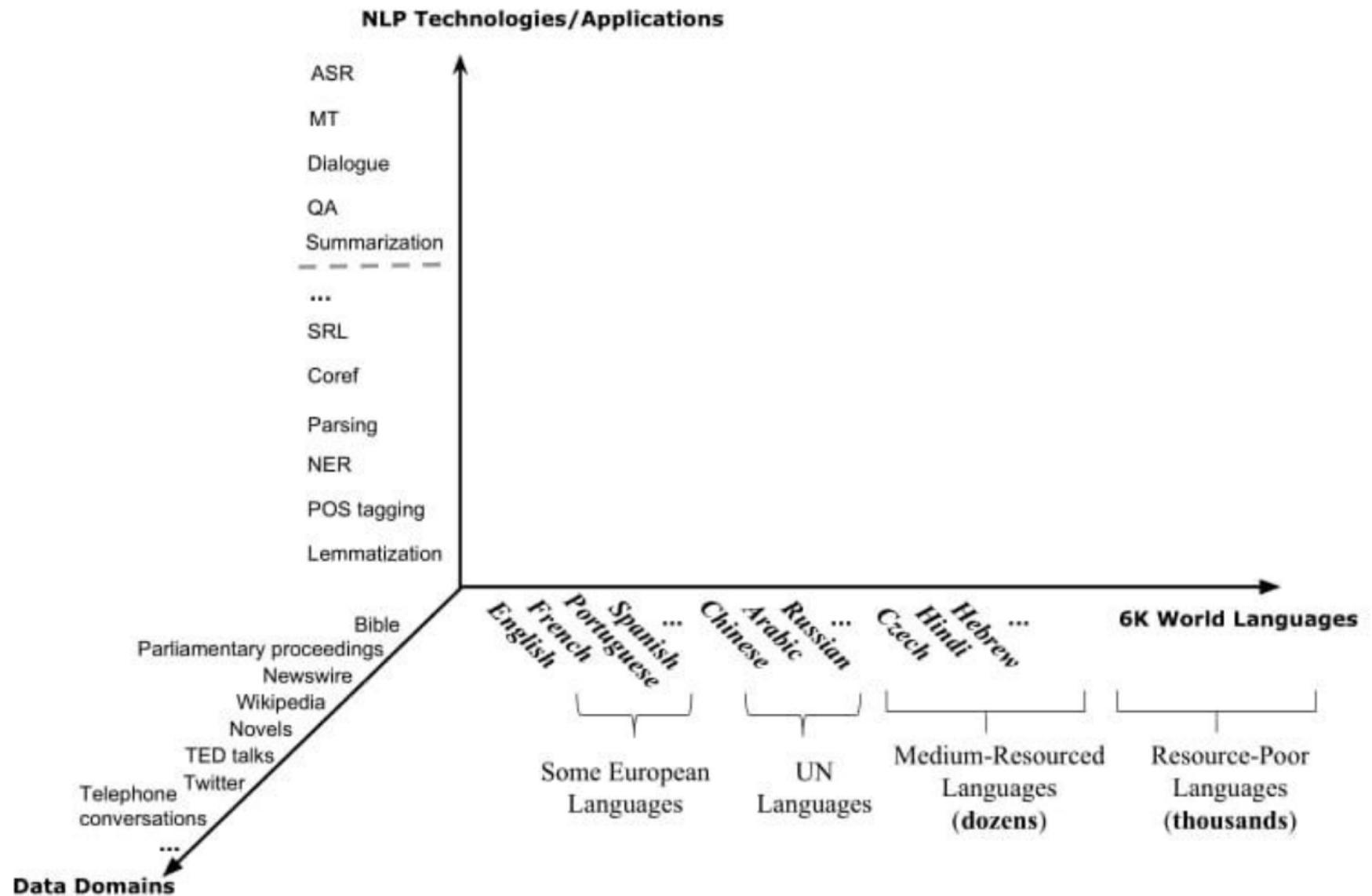
# Variation

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



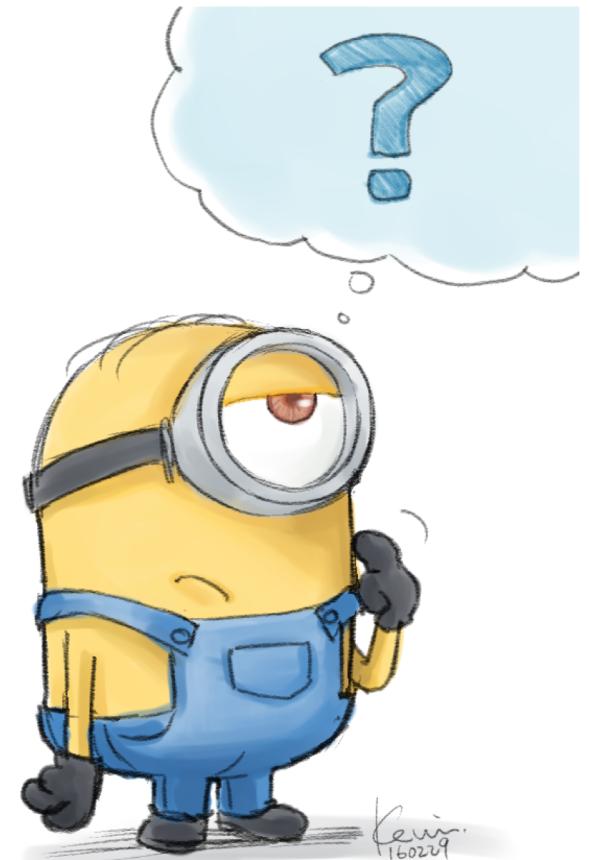
- What will happen if we try to use this tagger/parser for **social media**?
  - "ikr smh he asked fir yo last name so he can add u on fb lololol"

# Variation



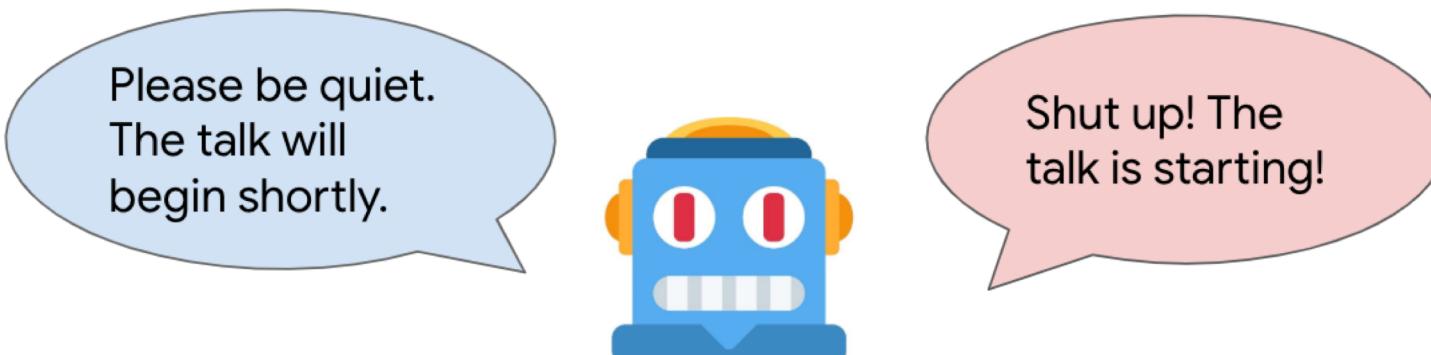
# Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations



# Expressivity

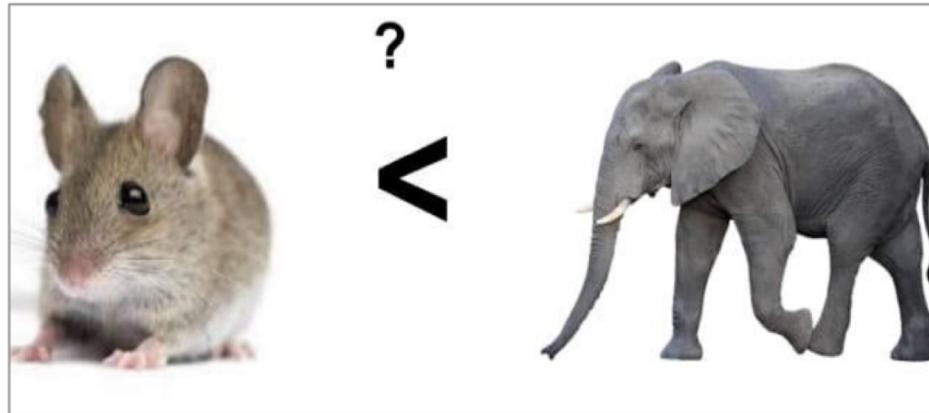
- Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:
  - *She gave the book to Tom* vs. *She gave Tom the book*
  - *Some kids popped by* vs. *A few children visited*
  - *Is that window still open?* vs. *Please close the window*



# 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

# Unmodeled Representation

Very difficult to capture what is  $\mathcal{R}$ , since we don't even know how to represent the knowledge a human has/needs:

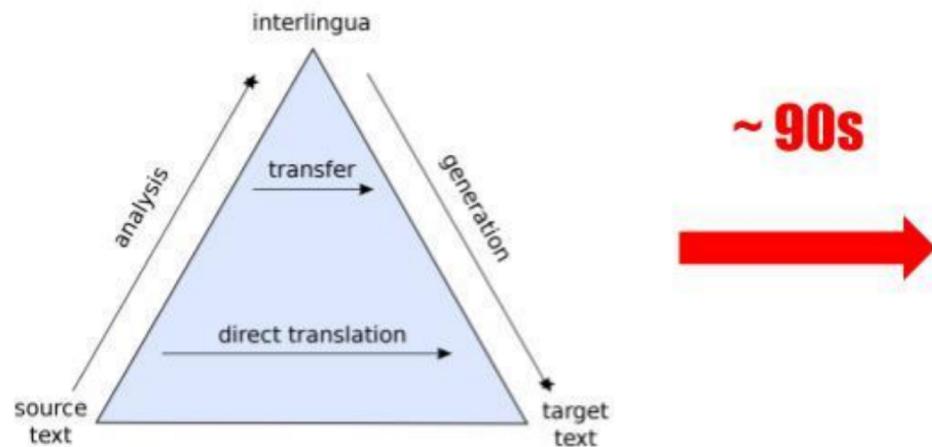
- What is the “meaning” of a word or sentence?
- How to model context?
- Other general knowledge?

# Desiderata for NLP Models

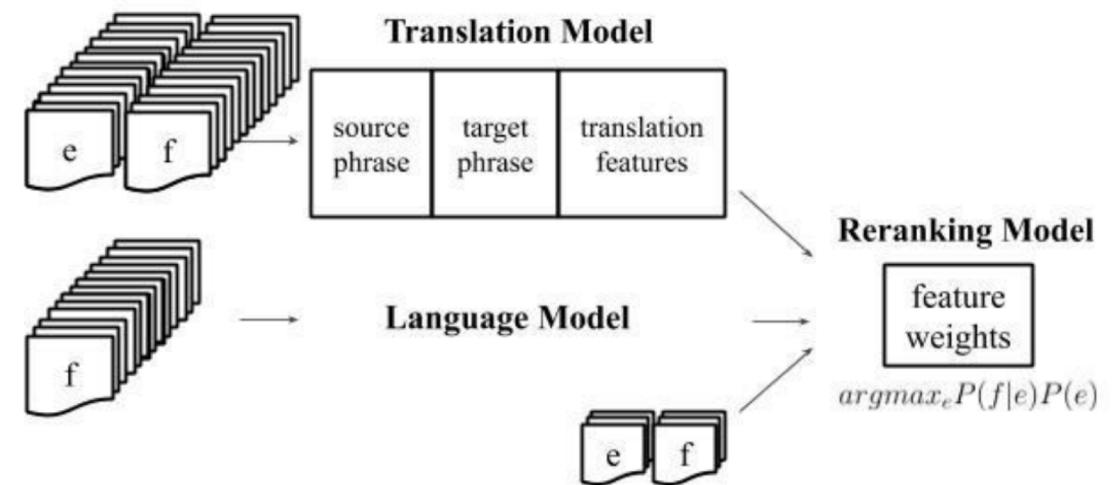
- Sensitivity to a wide range of phenomena and constraints in human language
- Generality across languages, modalities, genres, styles
- Strong formal guarantees (e.g., convergence, statistical efficiency, consistency)
- High accuracy when judged against expert annotations or test data
- Ethical

# Symbolic and Probabilistic NLP

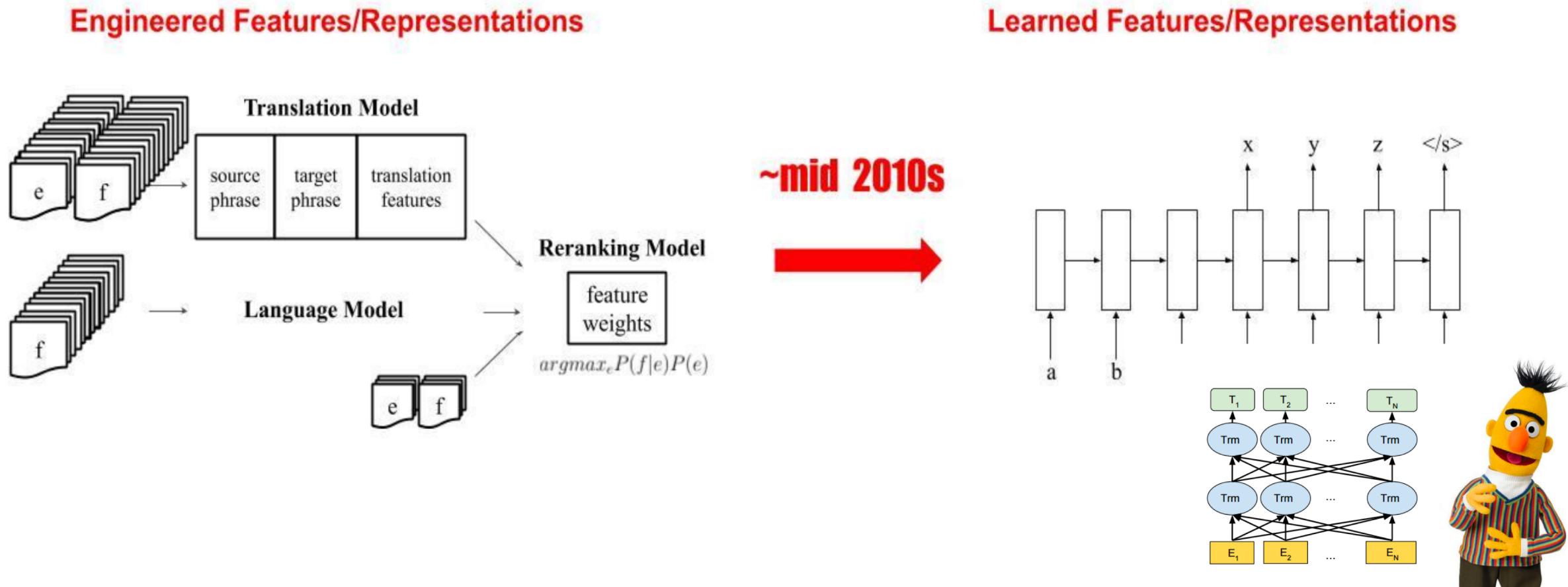
## Logic-based/Rule-based NLP



## Statistical NLP



# Probabilistic and Connectionist NLP



# NLP vs. Machine Learning

- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- $\mathcal{R}$  is not directly observable.
- Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.

# NLP vs. Linguistics

- NLP must contend with NL data as found in the world
- NLP  $\approx$  computational linguistics
- Linguistics has begun to use tools originating in NLP!

# Fields with Connections to NLP

- Machine learning
- Linguistics (including psycho-, socio-, descriptive, and theoretical)
- Cognitive science
- Information theory
- Logic
- Data science
- Political science
- Psychology
- Economics
- Education

# Today's Applications

- Conversational agents
- Information extraction and question answering
- Machine translation
- Opinion and sentiment analysis
- Social media analysis
- Visual understanding
- Essay evaluation
- Mining legal, medical, or scholarly literature

# Factors Changing NLP Landscape

1. Increases in computing power
2. The rise of the web, then the social web
3. Advances in machine learning
4. Advances in understanding of language in social context

# Logistics

---

# What is this Class?

## ■ **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
- Deep neural networks for NLP: LSTM, CNN, Seq2seq

# Outline of Topics

- Words and Sequences
  - Text classifications
  - Probabilistic language models
  - Vector semantics and word embeddings
  - Sequence labeling: POS tagging, NER
  - HMMs, Speech recognition
- Parsers
- Semantics
- Applications
  - Machine translation, Question Answering, Dialog Systems

# Readings

- Books:
  - Primary text: Jurafsky and Martin, Speech and Language Processing, 2nd or 3rd Edition
    - <https://web.stanford.edu/~jurafsky/slp3/>
  - Also: Eisenstein, Natural Language Processing
    - <https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf>















