Natural Language Processing

# Introduction to NLP

Instructor: Moushmi Dasgupta

Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.

# AGENDA

## Introduction to NLP and Business Applications

1.1 What is Language?

1.2 Building Blocks of Language

1.3 Why is NLP Challenging?

1.4 Machine Learning, Deep Learning, and NLP: An Overview

1.5 Approaches to NLP in Business Analytics

Pre-requisite:

1. Python programming
2. An understanding of Machine Learning
3. Invest in attending classroom sessions (Weekly 1 or 2 classes of 3+ hours duration)
4. Invest in yourself with1 hour of self study everyday

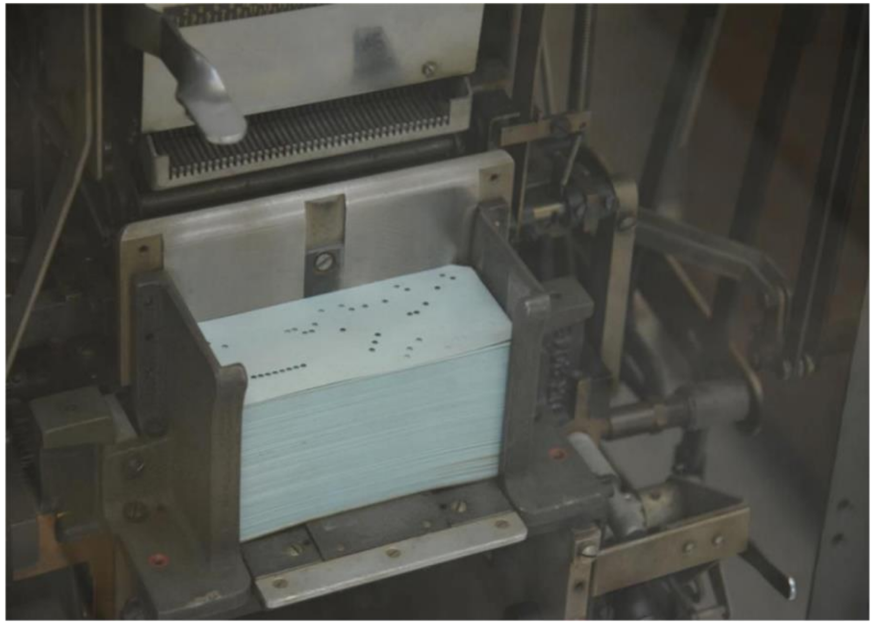
# Human Language

Google search reports that there are **7,151 living languages**

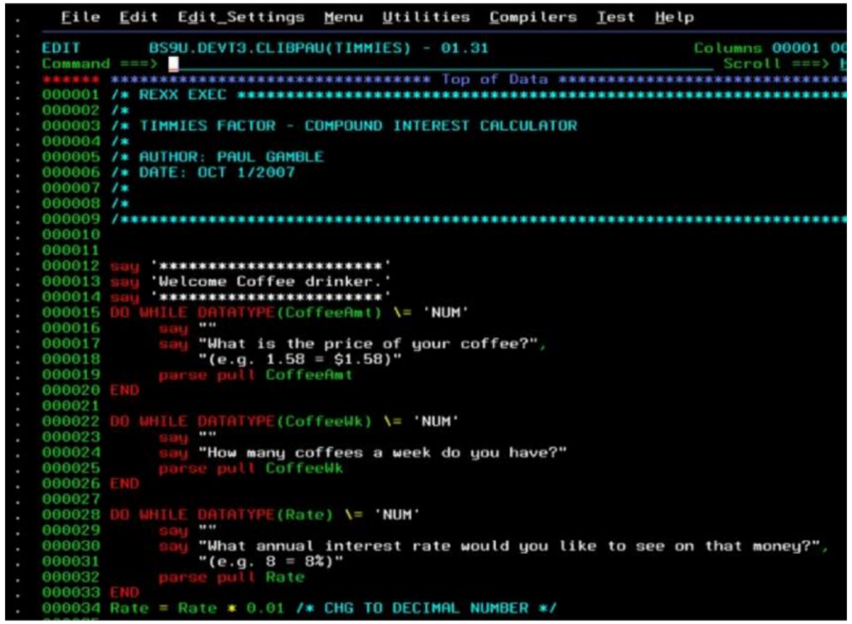
The system of sounds and writing that human beings use to express their thoughts, ideas and feelings

Language as understood by the machine learning algorithms

# Communication With Machines



~50-70s



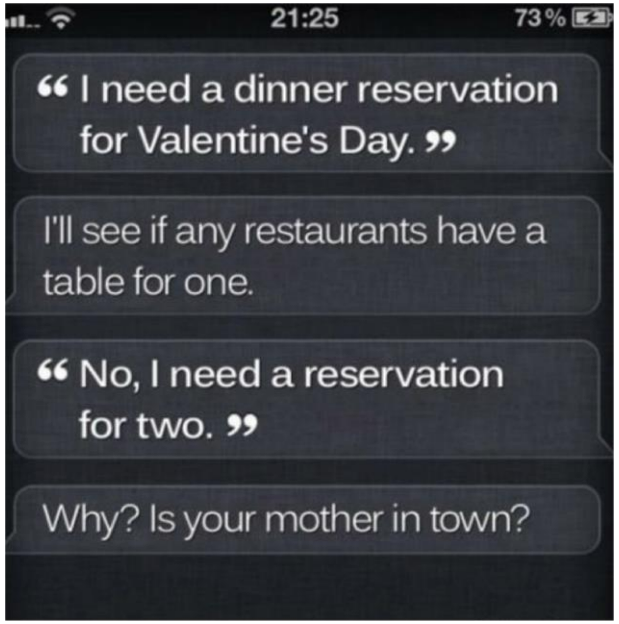
```
EDIT       BS9U.DEVT3.CLIBPAU(TIMMIES) - 01.31                Columns 00001 00
Command ===>                                                    Scroll ===> H
****** ***************************** Top of Data ******************************
/* REXX EXEC ******************************************************************
/*
/* TIMMIES FACTOR - COMPOUND INTEREST CALCULATOR
/*
/* AUTHOR: PAUL GAMBLE
/* DATE: OCT 1/2007
/*
/*
/******************************************************************************


say '***********************'
say 'Welcome Coffee drinker.'
say '***********************'
DO WHILE DATATYPE(CoffeeAmt) \= 'NUM'
     say ""
     say "What is the price of your coffee?",
         "(e.g. 1.58 = $1.58)"
     parse pull CoffeeAmt
END

DO WHILE DATATYPE(CoffeeWk) \= 'NUM'
     say ""
     say "How many coffees a week do you have?"
     parse pull CoffeeWk
END

DO WHILE DATATYPE(Rate) \= 'NUM'
     say ""
     say "What annual interest rate would you like to see on that money?",
         "(e.g. 8 = 8%)"
     parse pull Rate
END
Rate = Rate * 0.01 /* CHG TO DECIMAL NUMBER */
```

~80s



“ I need a dinner reservation for Valentine's Day. ”

I'll see if any restaurants have a table for one.

“ No, I need a reservation for two. ”

Why? Is your mother in town?

today

INDIAN INSTITUTE OF MANAGEMENT CALCUTTA

CMC Limited
Your IT Partner

amdocs

KANTAR TNS.

teradata.

Java

aws

Google Cloud

kubernetes

MICRO, SMALL & MEDIUM ENTERPRISES
सूक्ष्म, लघु एवं मध्यम उद्यम

Open for Innovation
KNIME

PySpark

spaCy

भारतीय प्रौद्योगिकी संस्थान कानपुर
INDIAN INSTITUTE OF TECHNOLOGY KANPUR

R

Apache Airflow

hadoop

RMIT UNIVERSITY

Keras

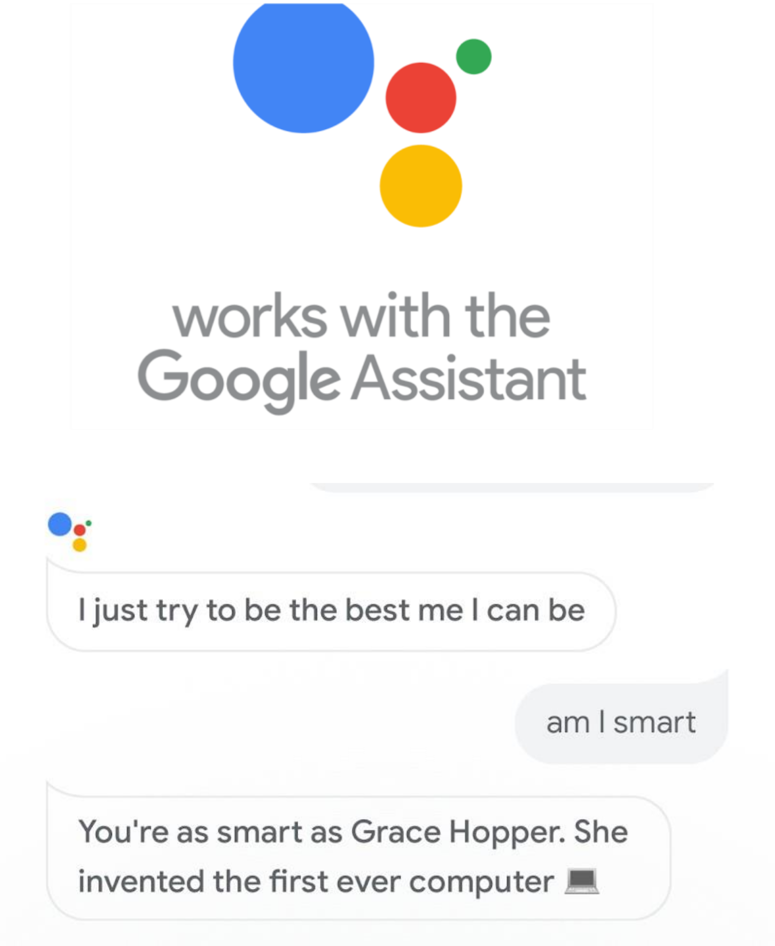
# Analytics Domain

https://www.linkedin.com/in/moushmi1234/

# Conversational Agents

Conversational agents contain:

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

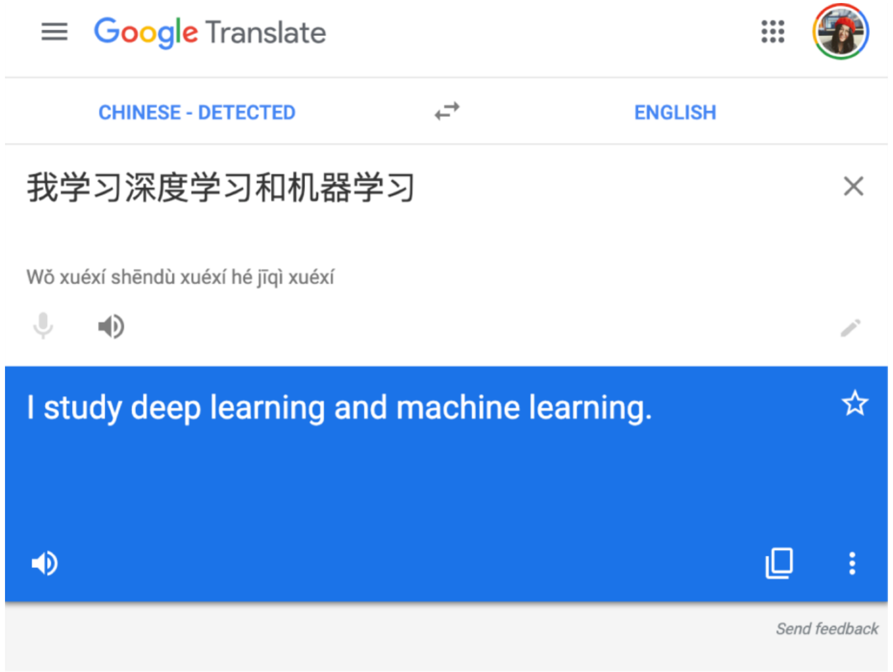


I just try to be the best me I can be

am I smart

You're as smart as Grace Hopper. She invented the first ever computer 💻

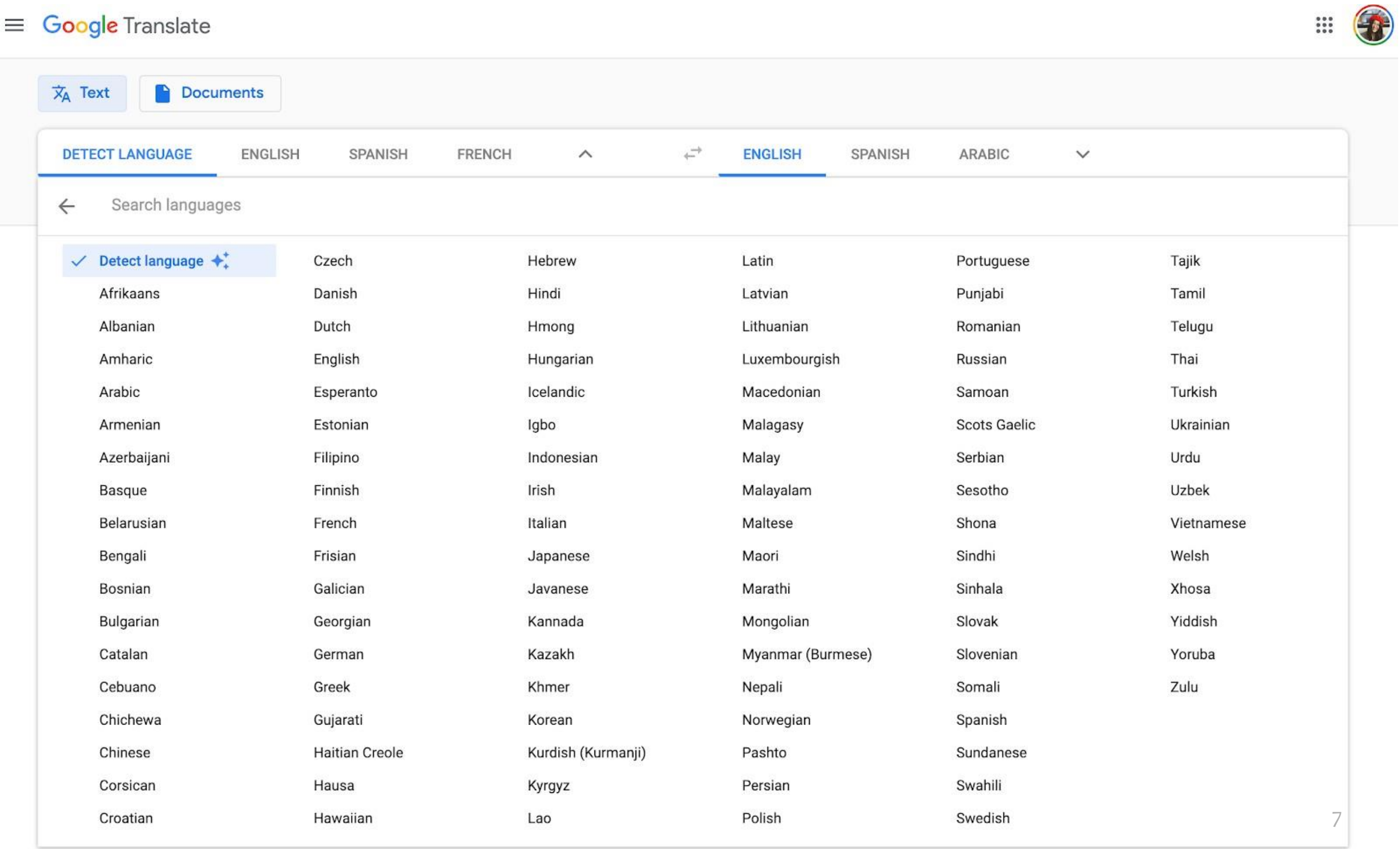
# Machine Translation



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.



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

# Natural Language Processing

## Applications

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

## Core Technologies

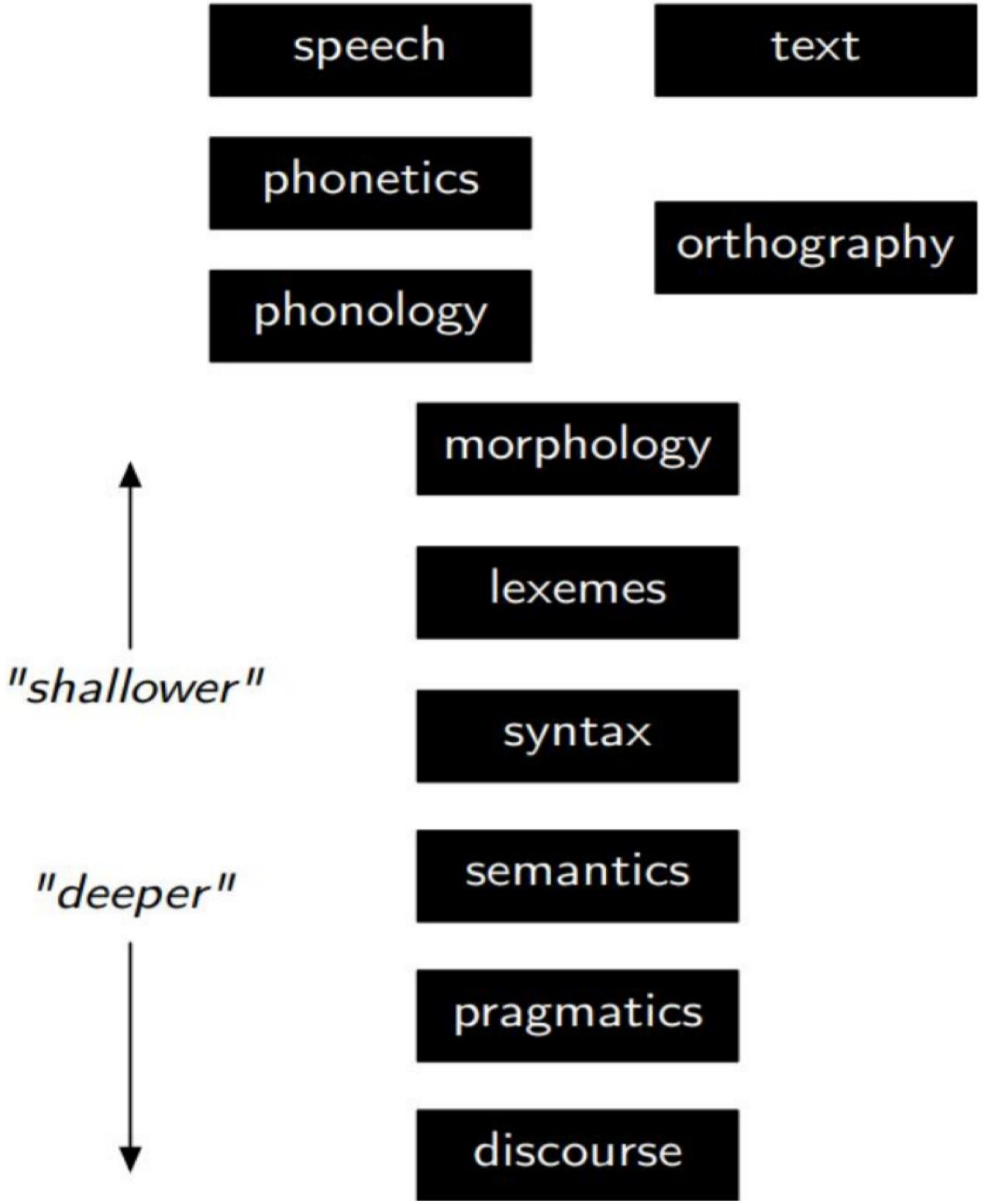
- Language modeling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Word sense disambiguation
- Semantic role labeling
- …

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

## Natural Language Processing (NLP) Examples

- Email filters.
- Smart assistants – Siri, Alexa, Google Assistant
- Search results
- Predictive text Analytics
- Language translation
- Digital phone calls
- Data analysis
- Text analytics

# Level Of Linguistic Knowledge


speech

text

phonetics

orthography

phonology

morphology

lexemes

"shallower"

syntax

semantics

"deeper"

pragmatics

discourse

# Phonetics, Phonology

- Pronunciation Modeling

**SOUNDS** Th i a si e n

Phonetics - the study of the sounds of human speech

Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.

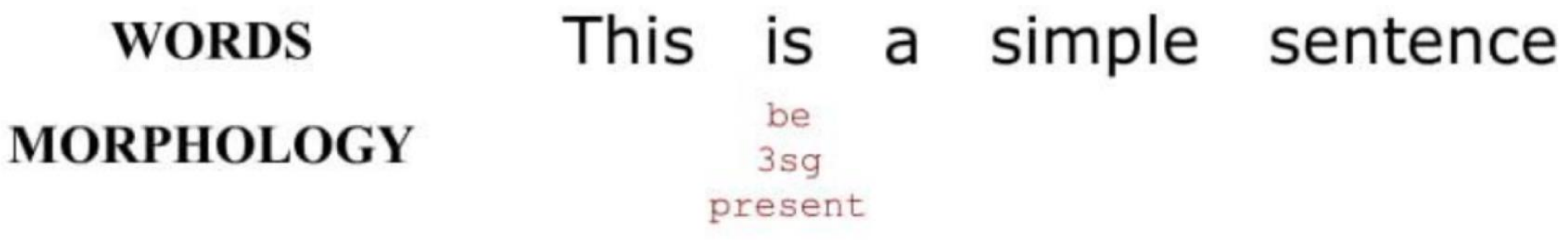
# Words

- Language Modeling
- Tokenization
- Spelling correction

WORDS This is a simple sentence

# Morphology

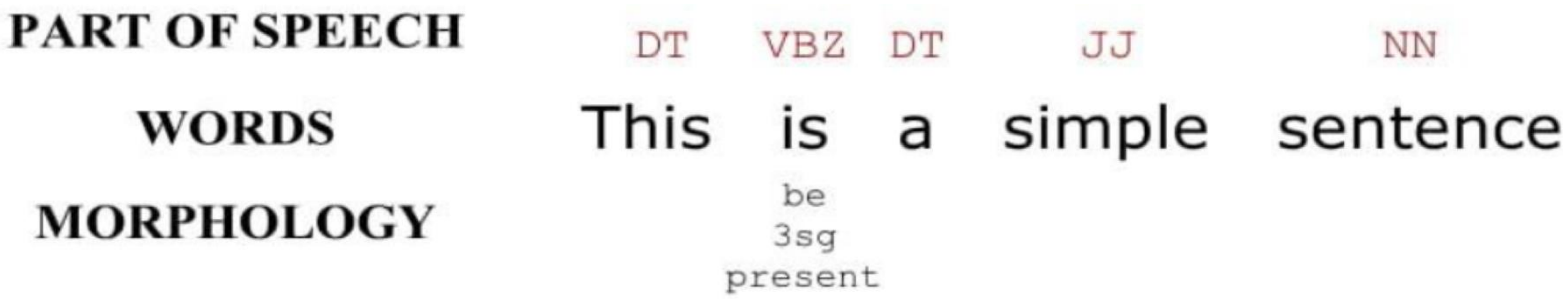
- Morphology analysis
- Tokenization
- Lemmatization



Morphology - the form of words, studied as a branch of linguistics

# Part of Speech

- Part of speech tagging

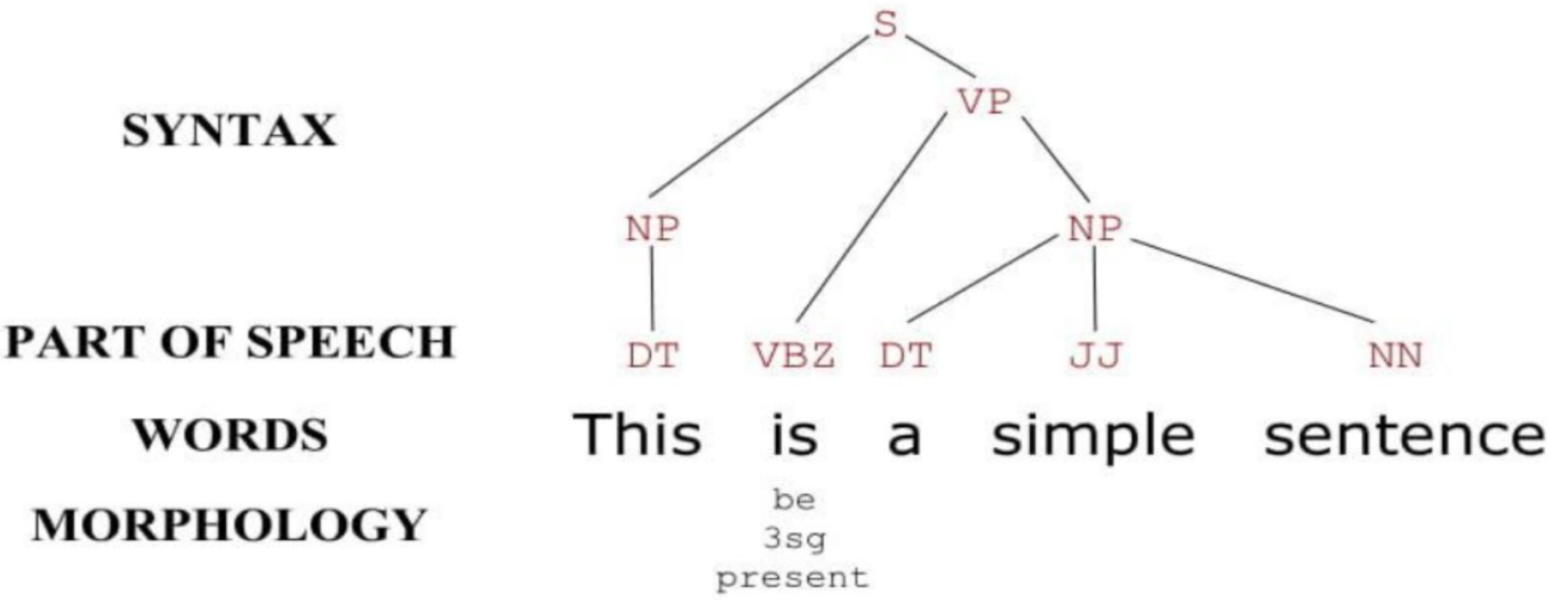


| | | | | | |
|---|---|---|---|---|---|
| PART OF SPEECH | DT | VBZ | DT | JJ | NN |
| WORDS | This | is | a | simple | sentence |
| MORPHOLOGY | | be<br>3sg<br>present | | | |

Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.

# Syntax

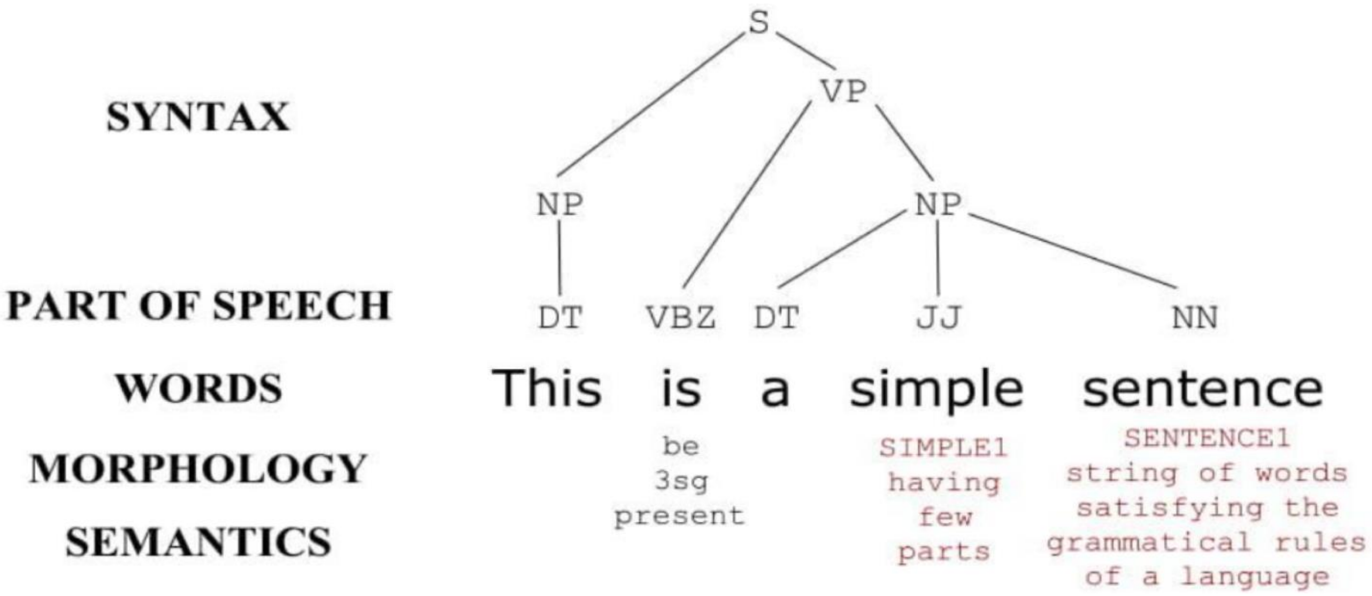
- Syntactic parsing



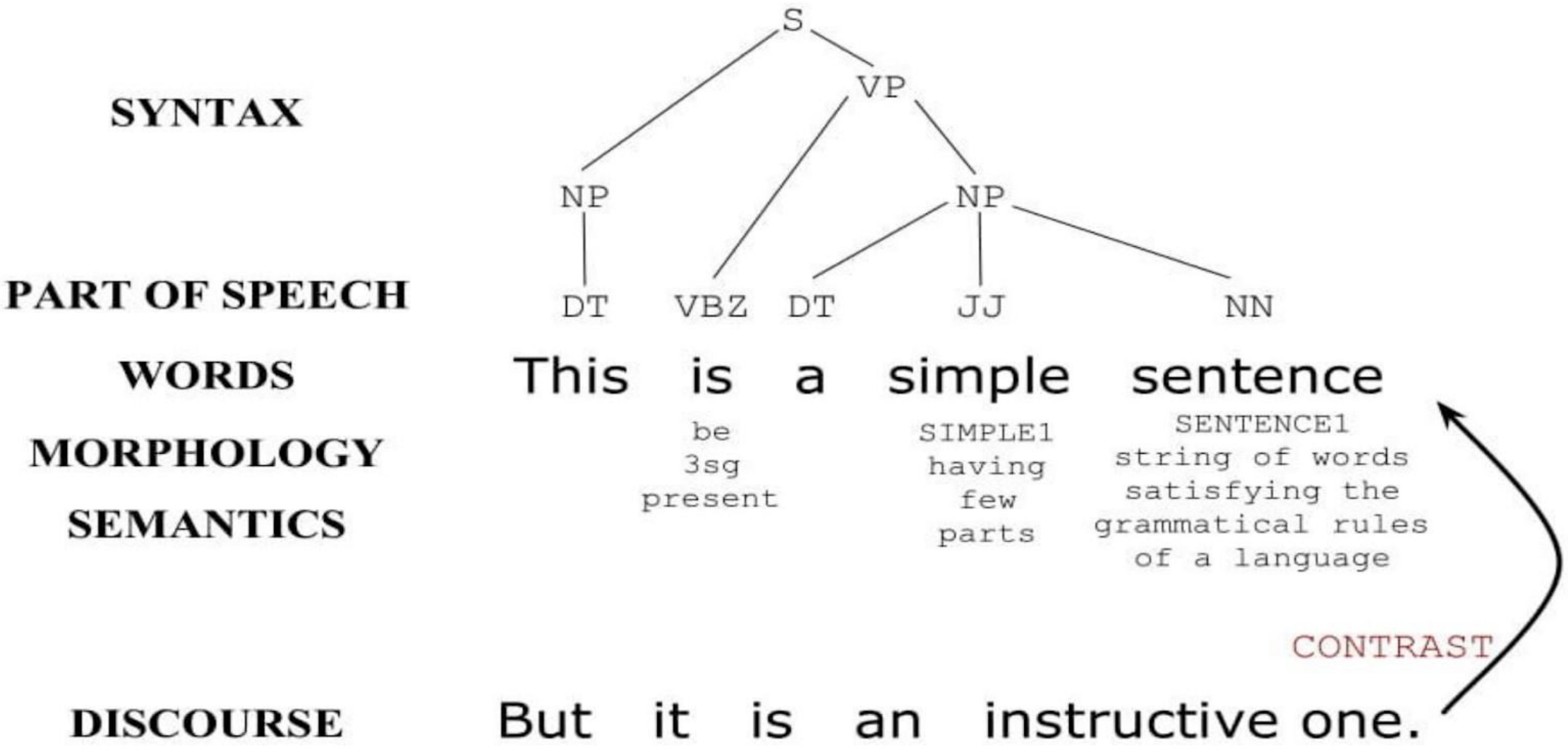
Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.

# Semantics

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



# Discourse



SYNTAX
PART OF SPEECH
WORDS
MORPHOLOGY
SEMANTICS
DISCOURSE
S
VP
NP
NP
DT
VBZ
DT
JJ
NN
This is a simple sentence
be
3sg
present
SIMPLE1
having
few
parts
SENTENCE1
string of words
satisfying the
grammatical rules
of a language
CONTRAST
But it is an instructive one.

# English Lexicon

A lexicon, word-hoard, wordbook, or word-stock is the vocabulary of a person, language, or branch of knowledge (such as nautical or medical). ...

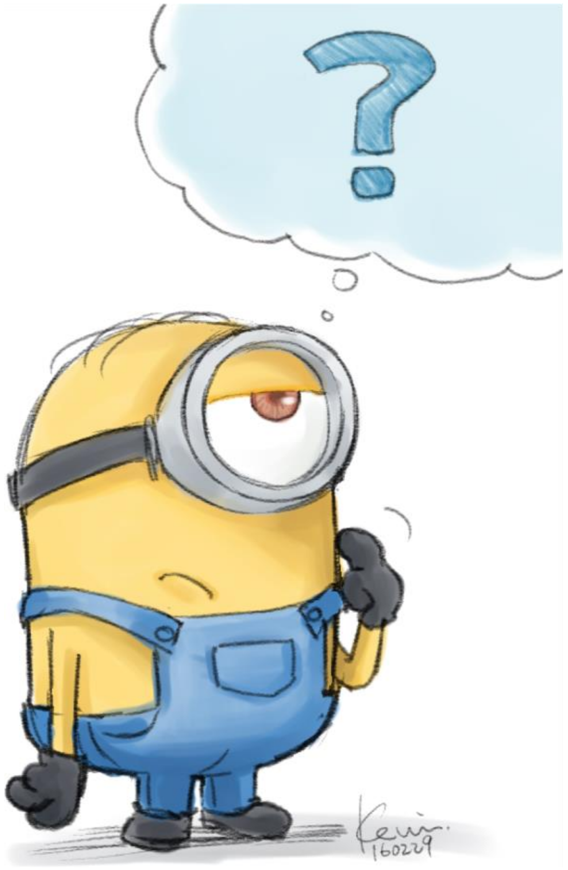
The word "lexicon" derives from the Greek λεξικόν (lexicon), neuter of λεξικός (lexikos) meaning
"of or for words."

**Train**

- Railway Station
- Platform
- Luggage
- Ticket collector
- Passengers
- Coach Number
- Berth

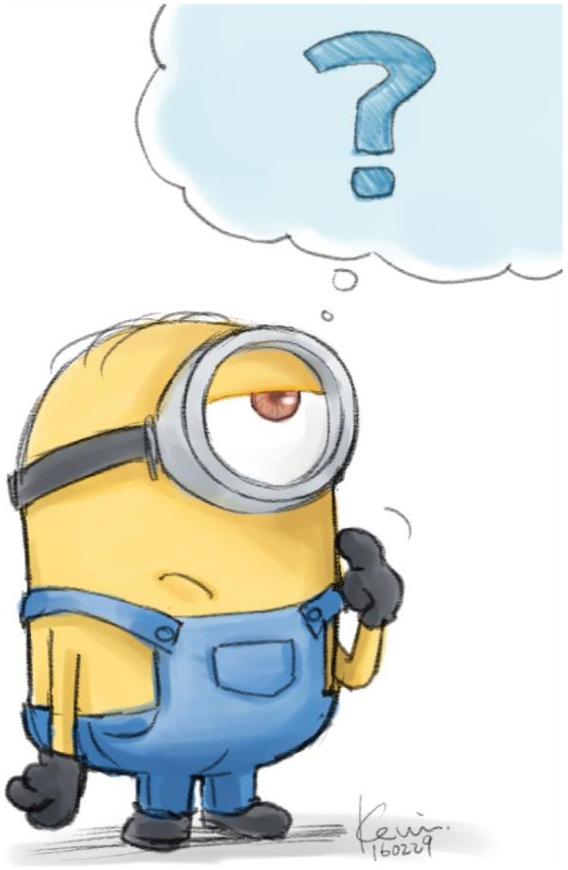
# Why NLP is Hard?

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



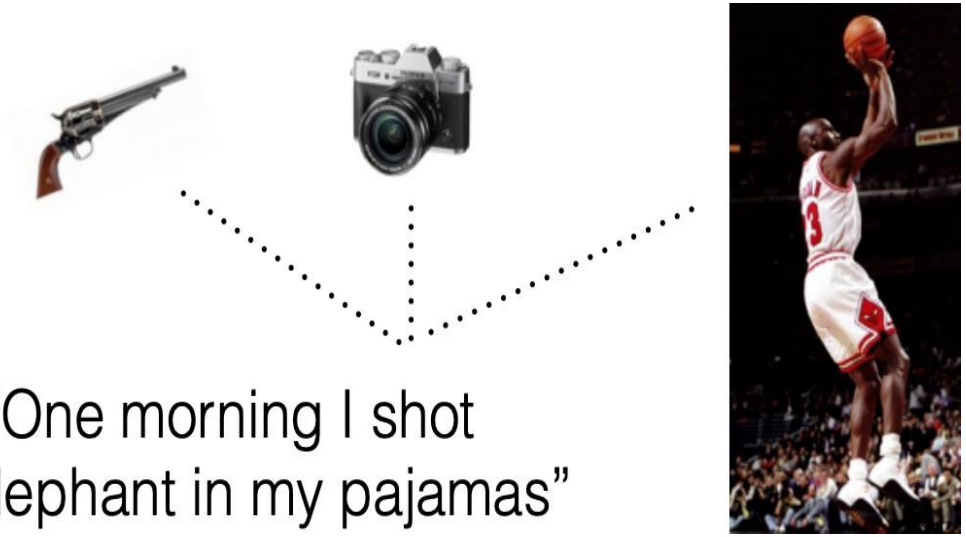
# Why NLP is Hard?

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
- …

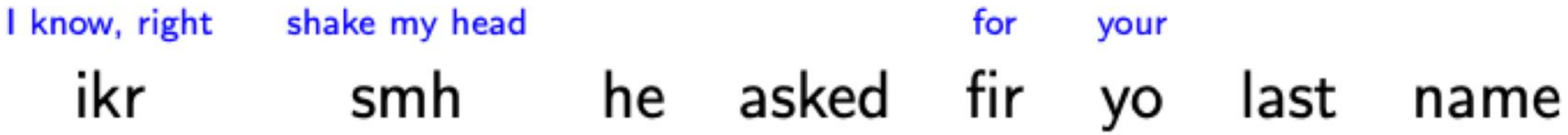
# Part of Speech Tagging

## Sentences with all 8 Parts of Speech

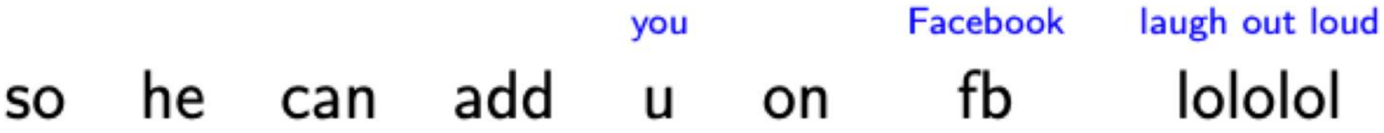
1. Noun – Tom lives in **New York**.
2. Pronoun – Did **she** find the book she was looking for?
3. Verb – I **reached** home.
4. Adverb – The tea is **too** hot.
5. Adjective – The movie was **amazing**.
6. Preposition – The candle was kept **under** the table.
7. Conjunction – I was at home all day, **but** I am feeling very tired.
8. Interjection – **Oh**! I forgot to turn off the stove.

It is **a process of converting a sentence to forms – list of words, list of tuples (where each tuple is having a form (word, tag))**. The tag in case of is a part-of-speech tag, and signifies whether the word is a noun, adjective, verb, and so on.

# Part of Speech Tagging

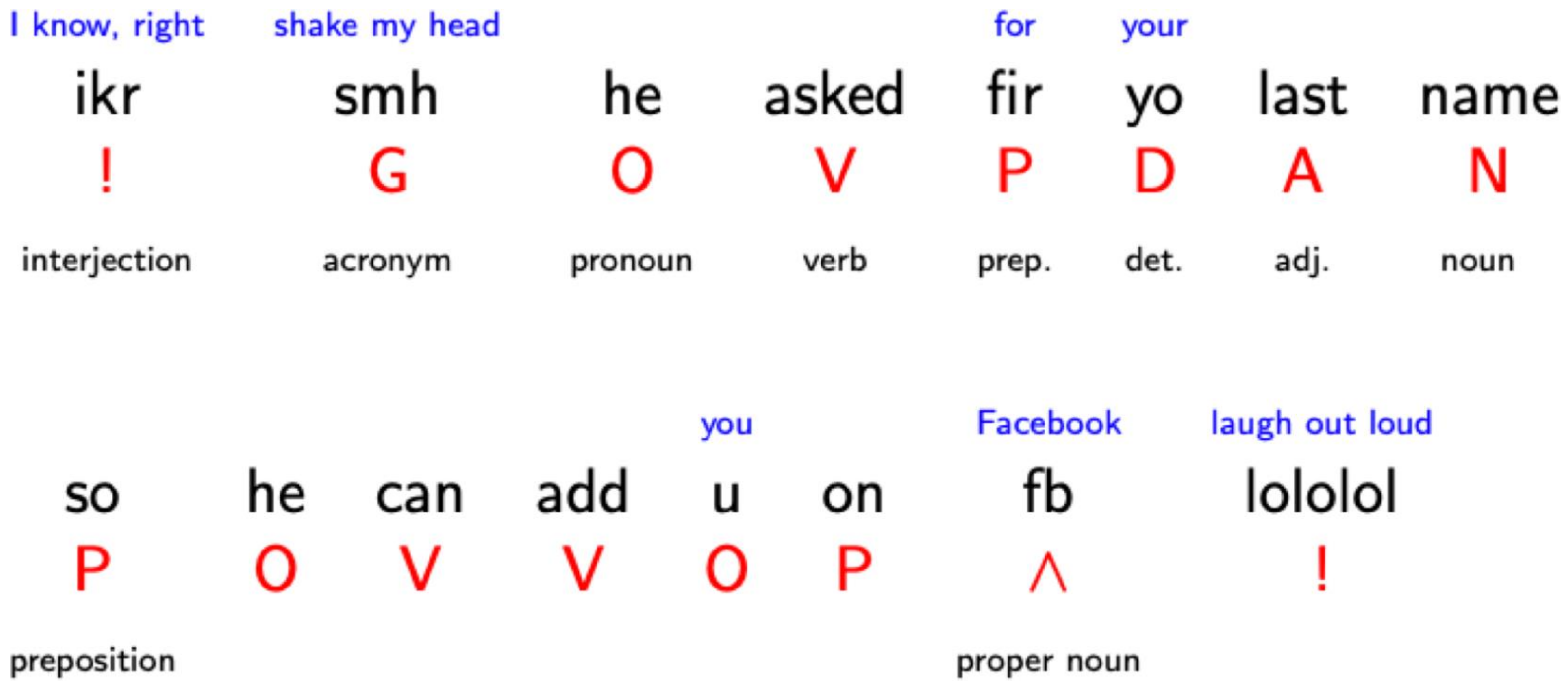


| | | | | | | | |
|---|---|---|---|---|---|---|---|
| I know, right | shake my head | | | for | your | | |
| ikr | smh | he | asked | fir | yo | last | name |



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

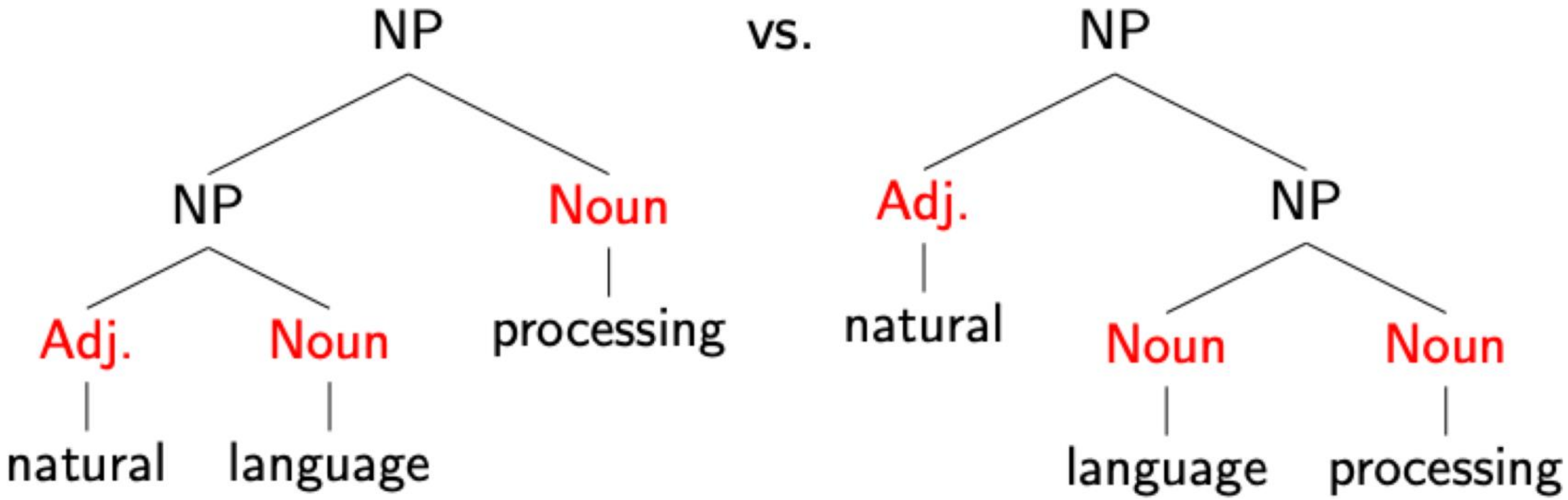
# Part of Speech Tagging



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 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


NP
NP
Noun
Adj.
Noun
natural
language
processing
vs.
NP
Adj.
NP
natural
Noun
Noun
language
processing

# Morphology + Syntax

A ship-shipping ship, shipping shipping-ships

Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.

# Semantics

377 people, equivalent to a jumbo jet crashing, die every day.

Our job is to find this jumbo jet and stop it!

# Syntax + Semantics

We saw the woman with the telescope wrapped in paper.

Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.

# 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

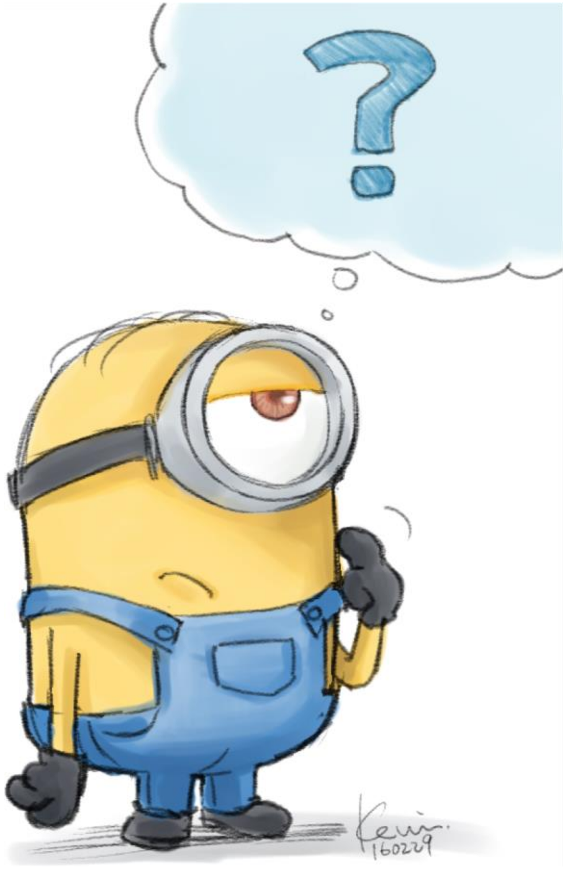
**Demotic, hieroglyphic and Greek**.

# 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

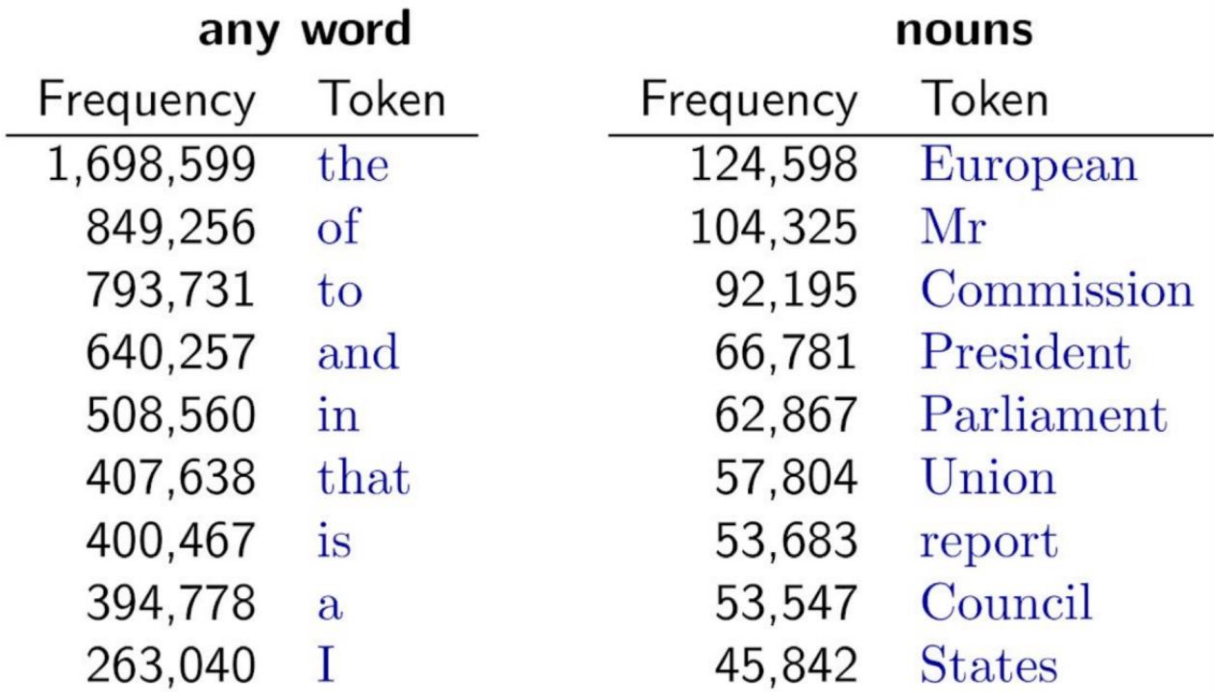
# Why NLP is Hard?

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



**any word**

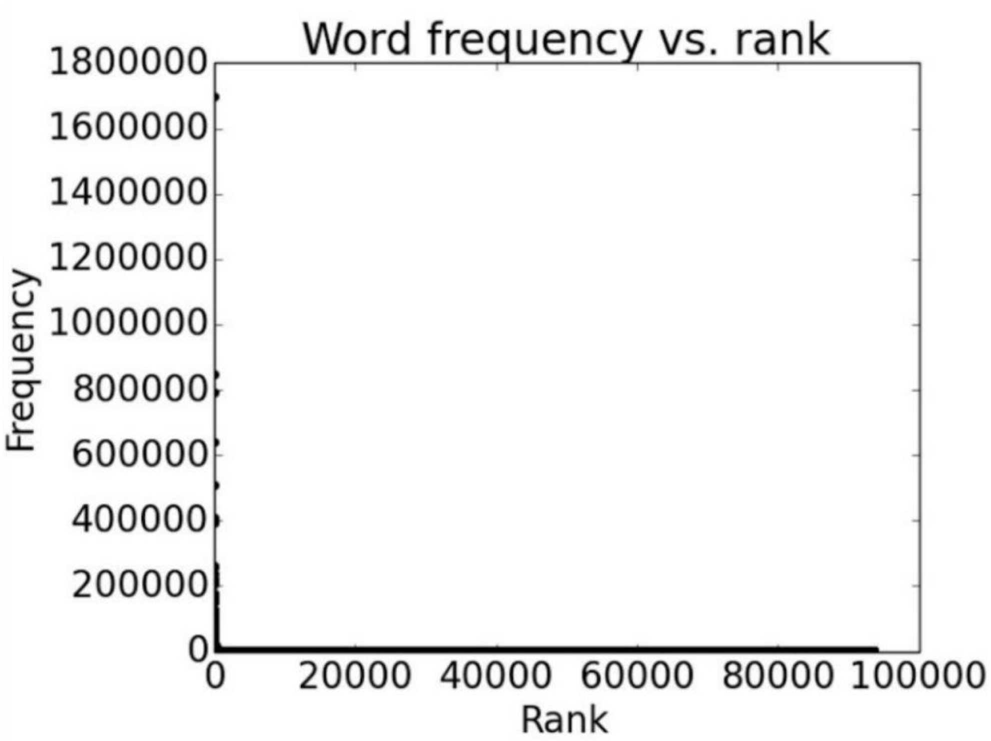
| Frequency | Token |
|---|---|
| 1,698,599 | the |
| 849,256 | of |
| 793,731 | to |
| 640,257 | and |
| 508,560 | in |
| 407,638 | that |
| 400,467 | is |
| 394,778 | a |
| 263,040 | I |

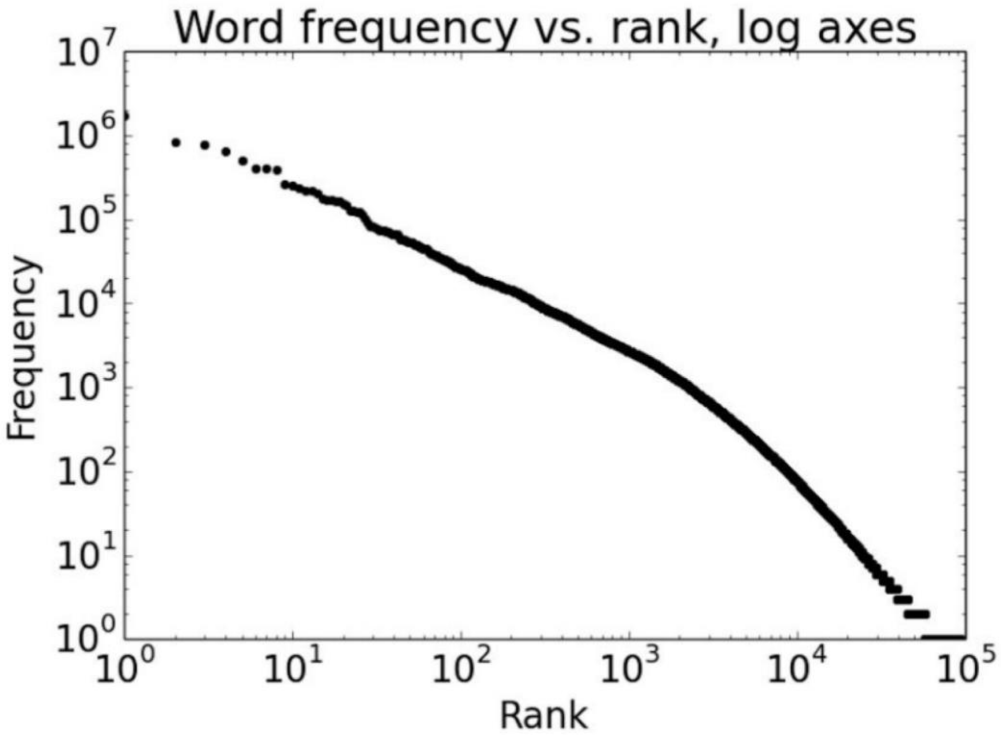
**nouns**

| 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 |

# Sparsity

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

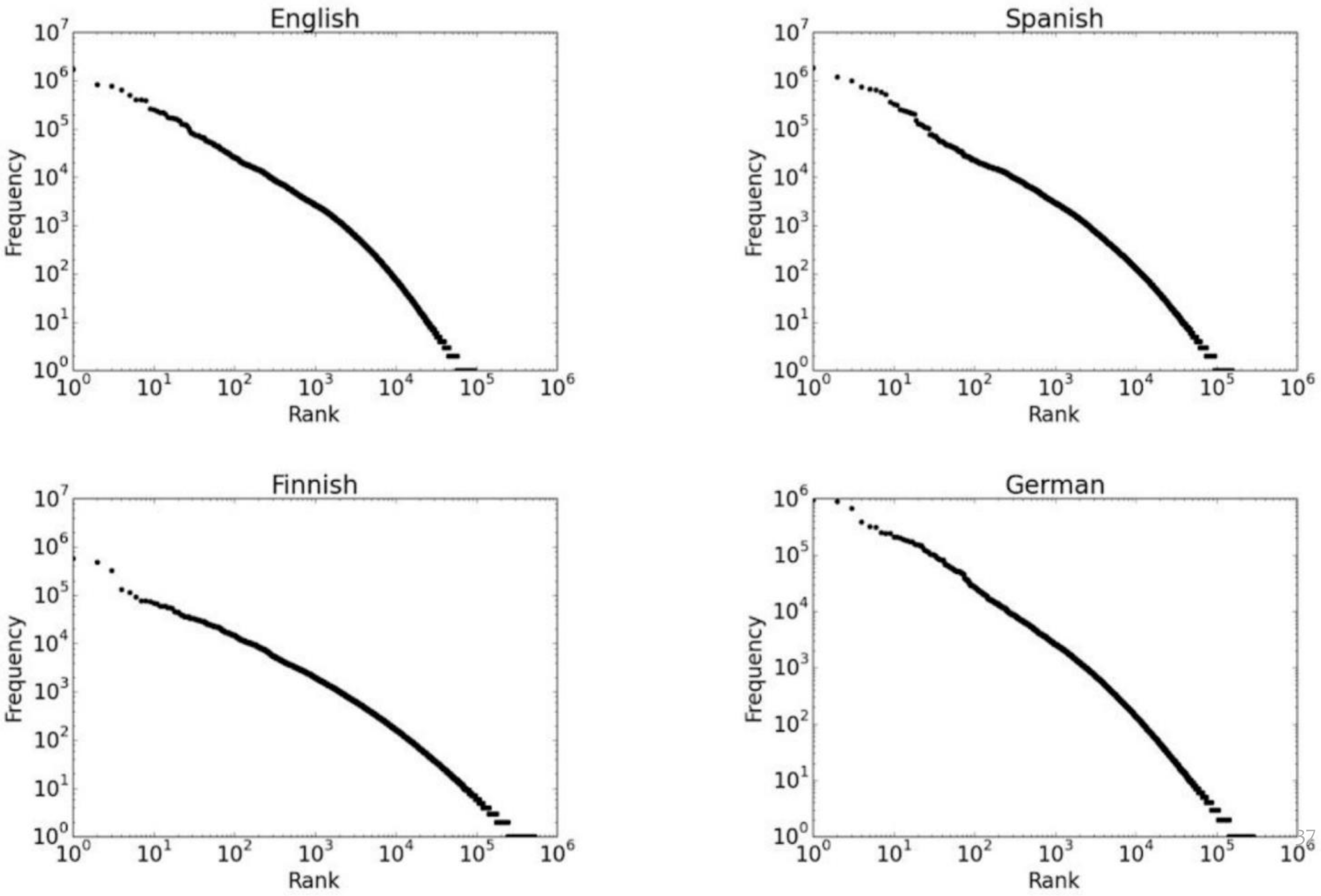




Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.

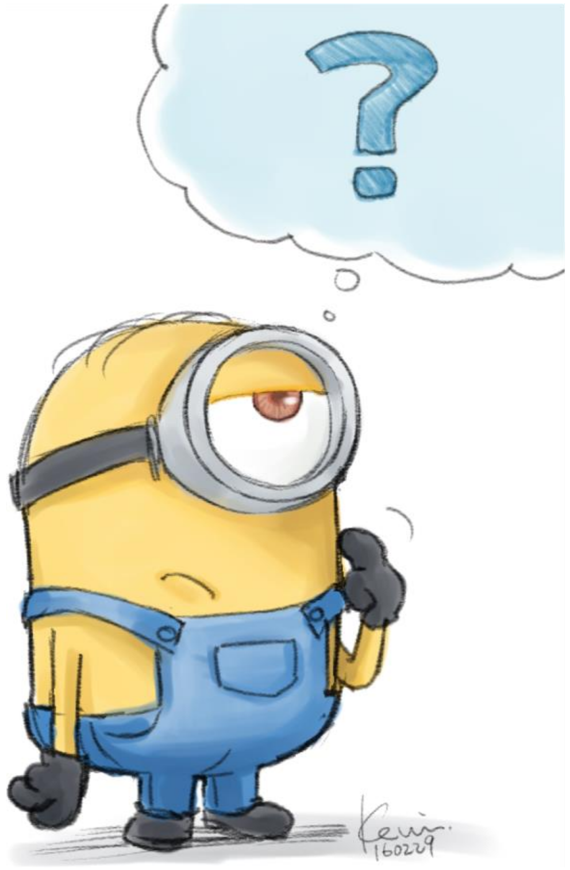
# Sparsity

- 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



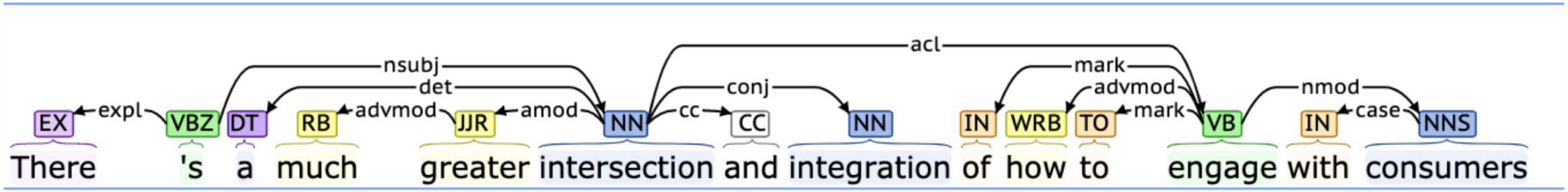
# Why NLP is Hard?

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



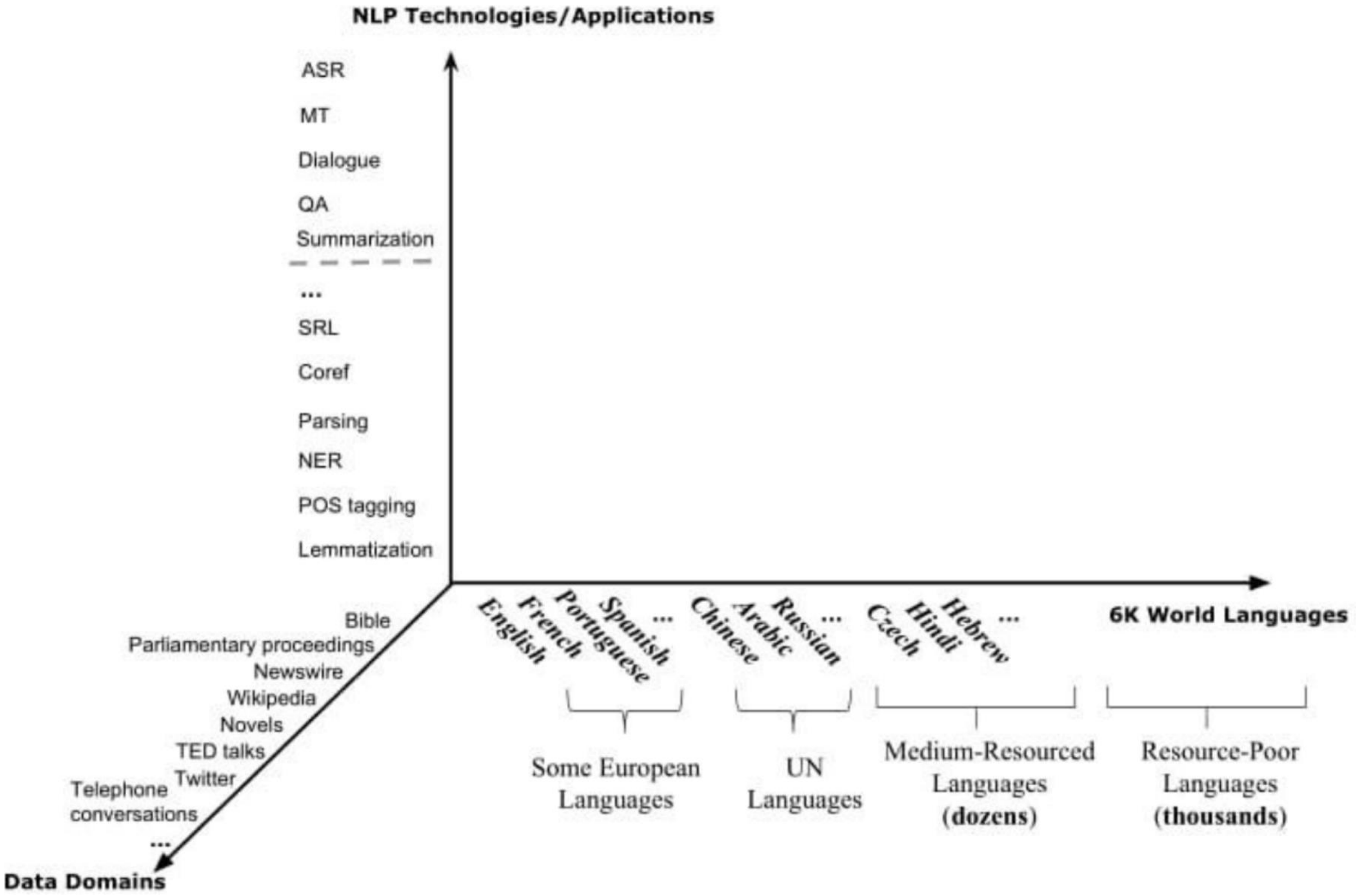
# Variation

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



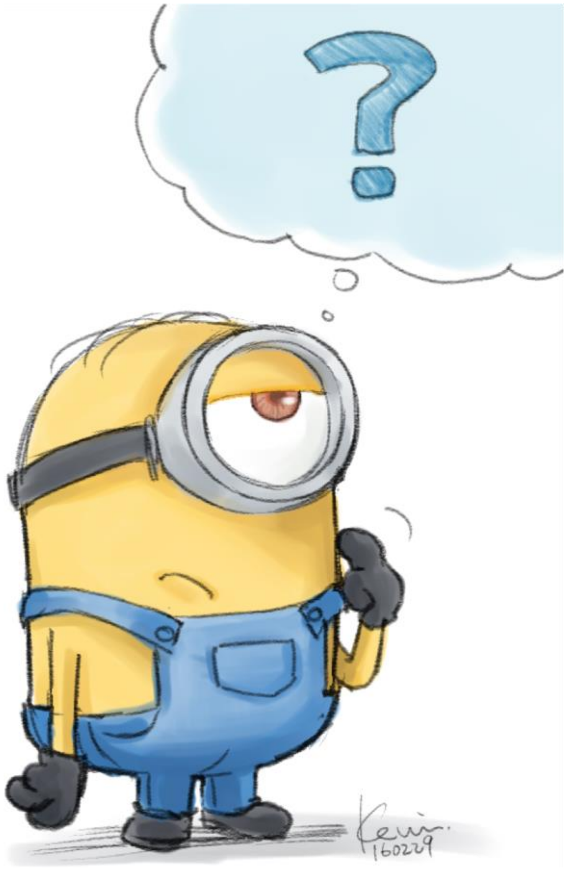
- What will happen if we try to use this tagger/parser for social media?
  - *"ikr smh he asked fir yo last name so he can add u on fb lololol"*

# Variation


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

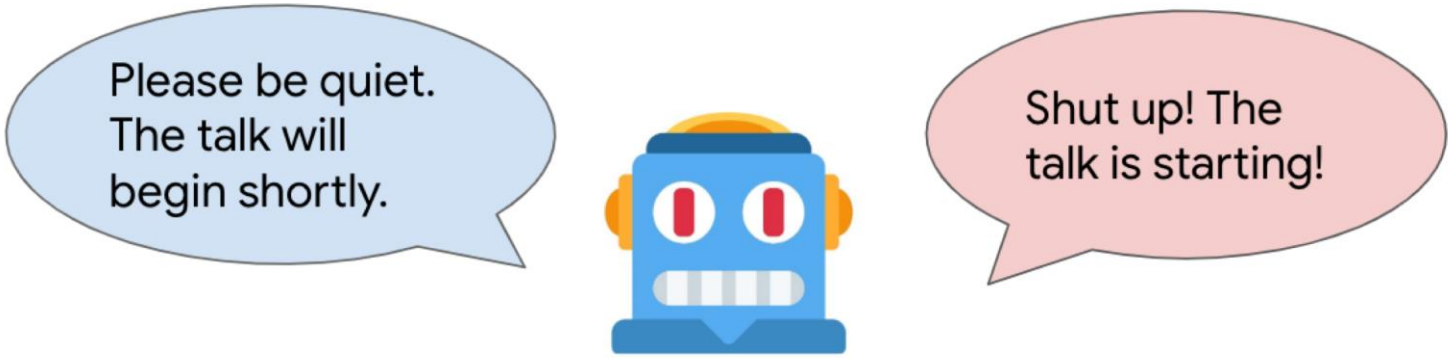
# Why NLP is Hard?

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



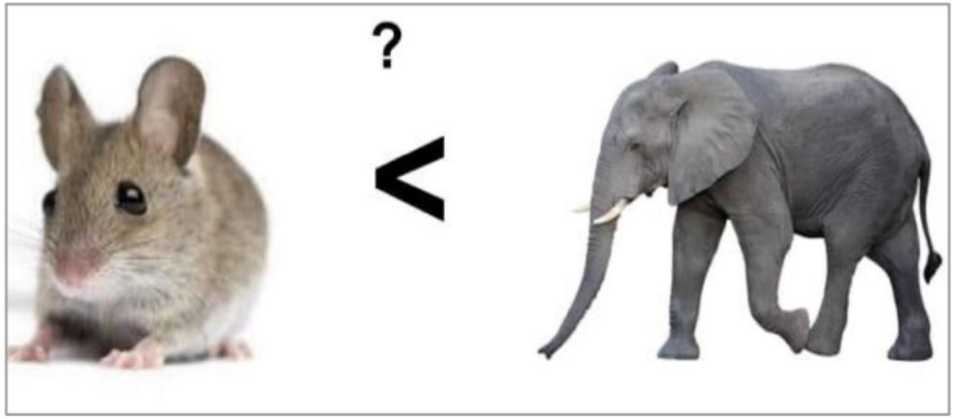
# Expressivity

- 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*


Please be quiet. The talk will begin shortly.
Shut up! The talk is starting!

# 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 ! , since we don’t even know how to represent the knowledge a human has/needs:

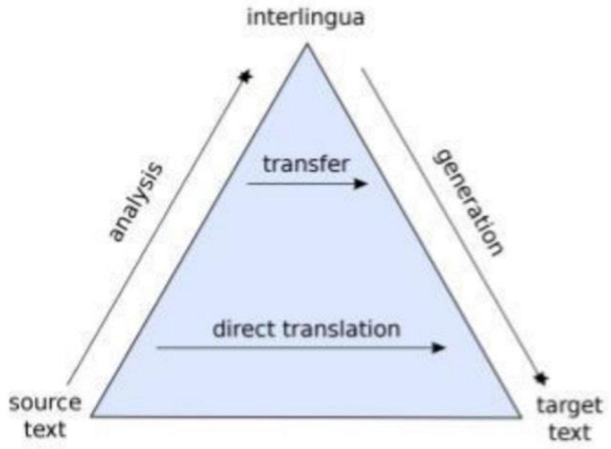
- What is the “meaning” of a word or sentence?
- How to model context?
- Other general knowledge?

# Desiderate 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

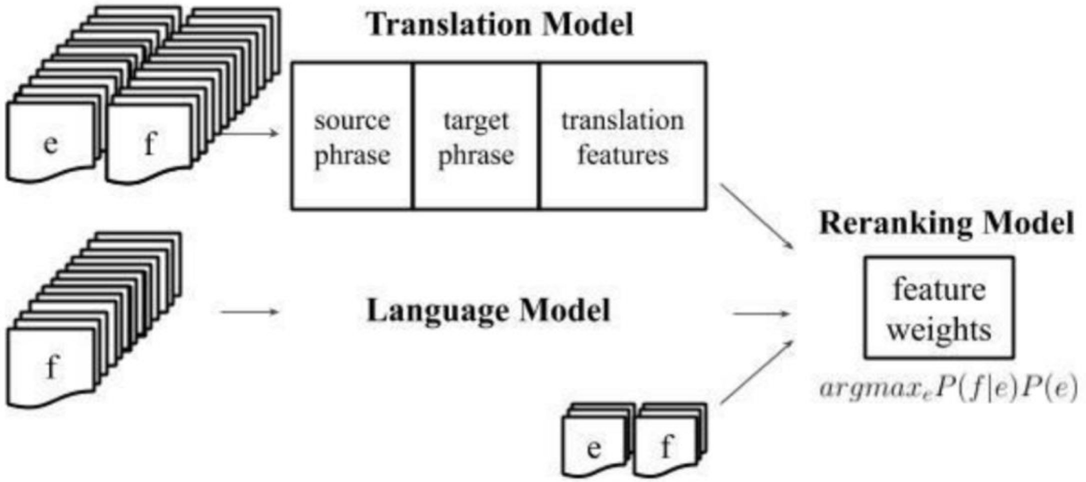
# Symbolic and Probabilistic NLP

## Logic-based/Rule-based NLP



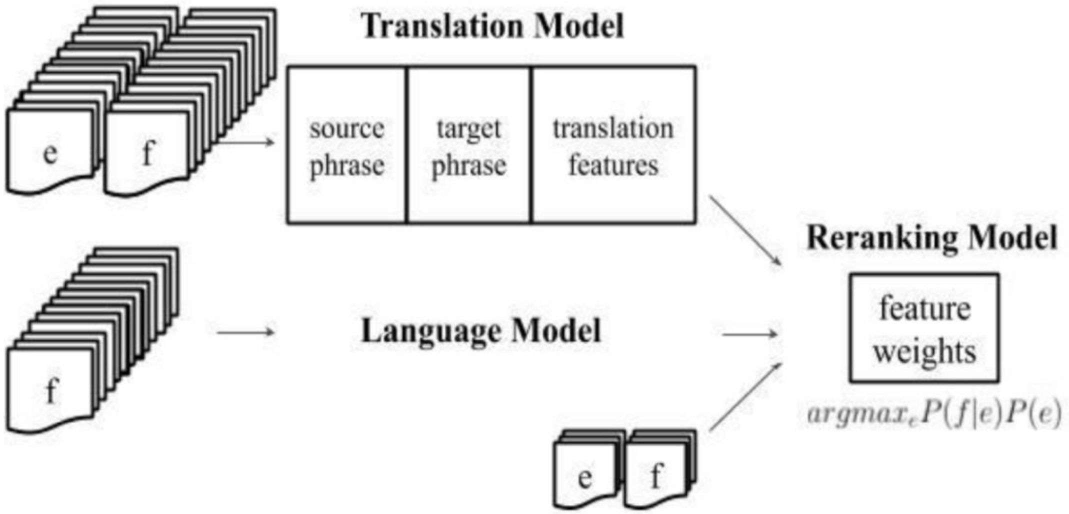
~ 90s

## Statistical NLP



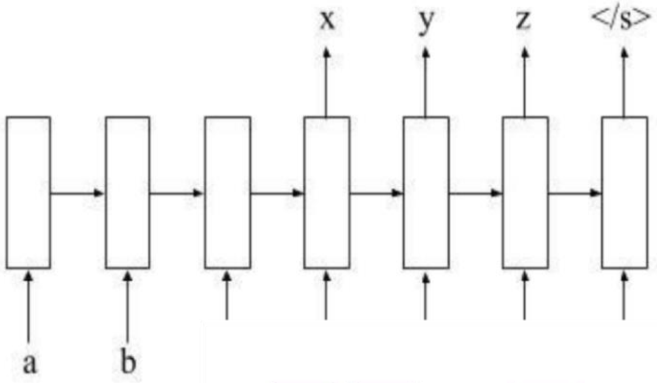
# Probabilistic and Connectionist NLP

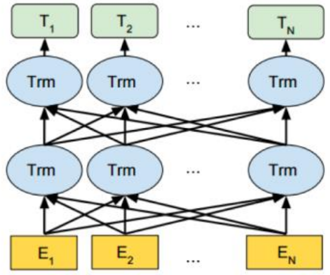
**Engineered Features/Representations**



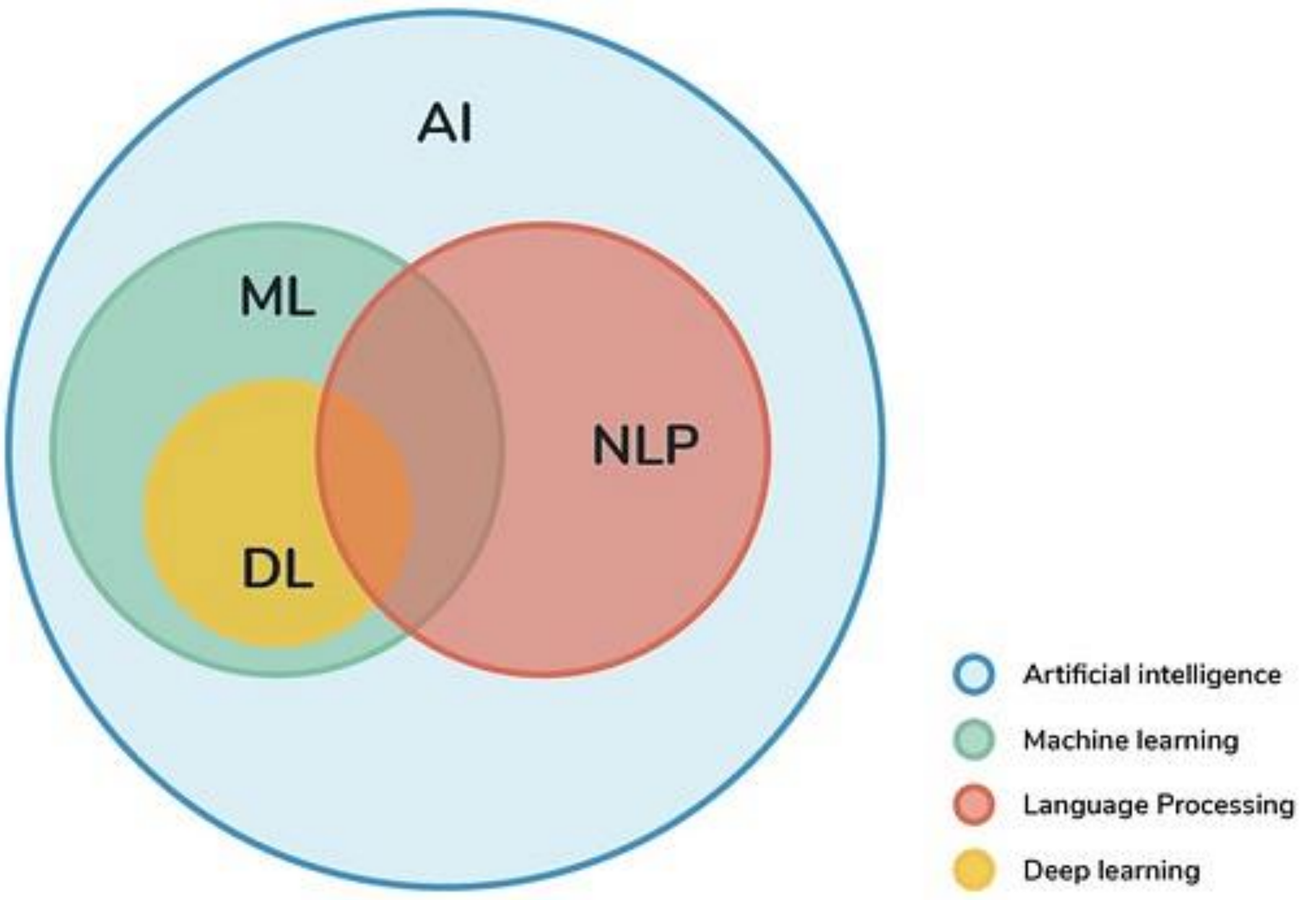
**~mid 2010s**

**Learned Features/Representations**





# AI – ML – DL - NLP


AI
ML
NLP
DL
Artificial intelligence
Machine learning
Language Processing
Deep learning

# NLP vs. Machine Learning

- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- ! 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 ≈ 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

# Python Libraries for NLP

1.Natural Language Toolkit(NLTK)
2.GenSim
3.SpaCy
4.CoreNLP
5.TextBlob
6.AllenNLP
7.polyglot
8.scikit-learn

## Core Python

Numpy
Pandas
Scikit Learn (SkLearn)
Beautiful Soup

# What’s Next?

- NLP Pipeline

Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.