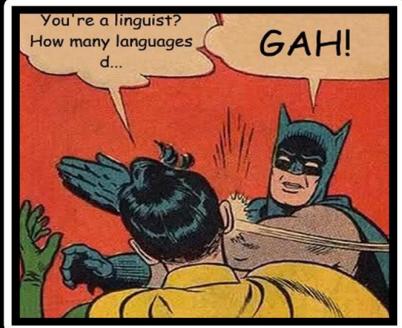


# Computational linguistics as a profession

Mariana Romanyshyn

*Computational Linguist at Grammarly, Inc.*

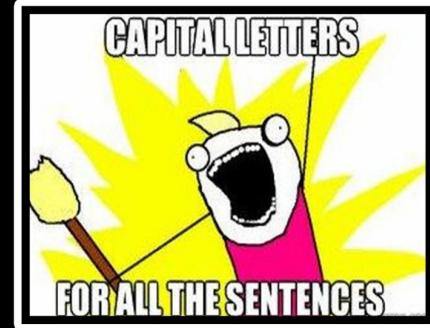
# COMPUTATIONAL LINGUIST



WHAT MY FRIENDS THINK I DO



WHAT MY MOTHER THINKS I DO



WHAT SOCIETY THINKS I DO



WHAT I THINK I DO

```
def generate_sentence(first_word):
    """Generate a sentence using the first word."""
    ngram_list = trigrams(brown.words(categories = "adventure"))
    fd = FreqDist(ngram_list)
    sentence = [first_word]
    while len(sentence) < 40:
        list_of_nexts = []
        for (i, j) in fd.keys():
            if i == first_word:
                list_of_nexts.append(j)
        if len(list_of_nexts) == 0:
            break
        if len(list_of_nexts) == 8:
            return "The sentence cannot be generated."
        first_word = random.choice(list_of_nexts)
        sentence.append(first_word)
        if first_word in [".", "?", "..."]:
            break
    return " ".join(sentence)
```

WHAT I REALLY DO

# Contents

1. NLP applications in our world
2. Competencies of a computational linguist
3. Grammarly and ComPLing Summer School

# 1. NLP applications in our world

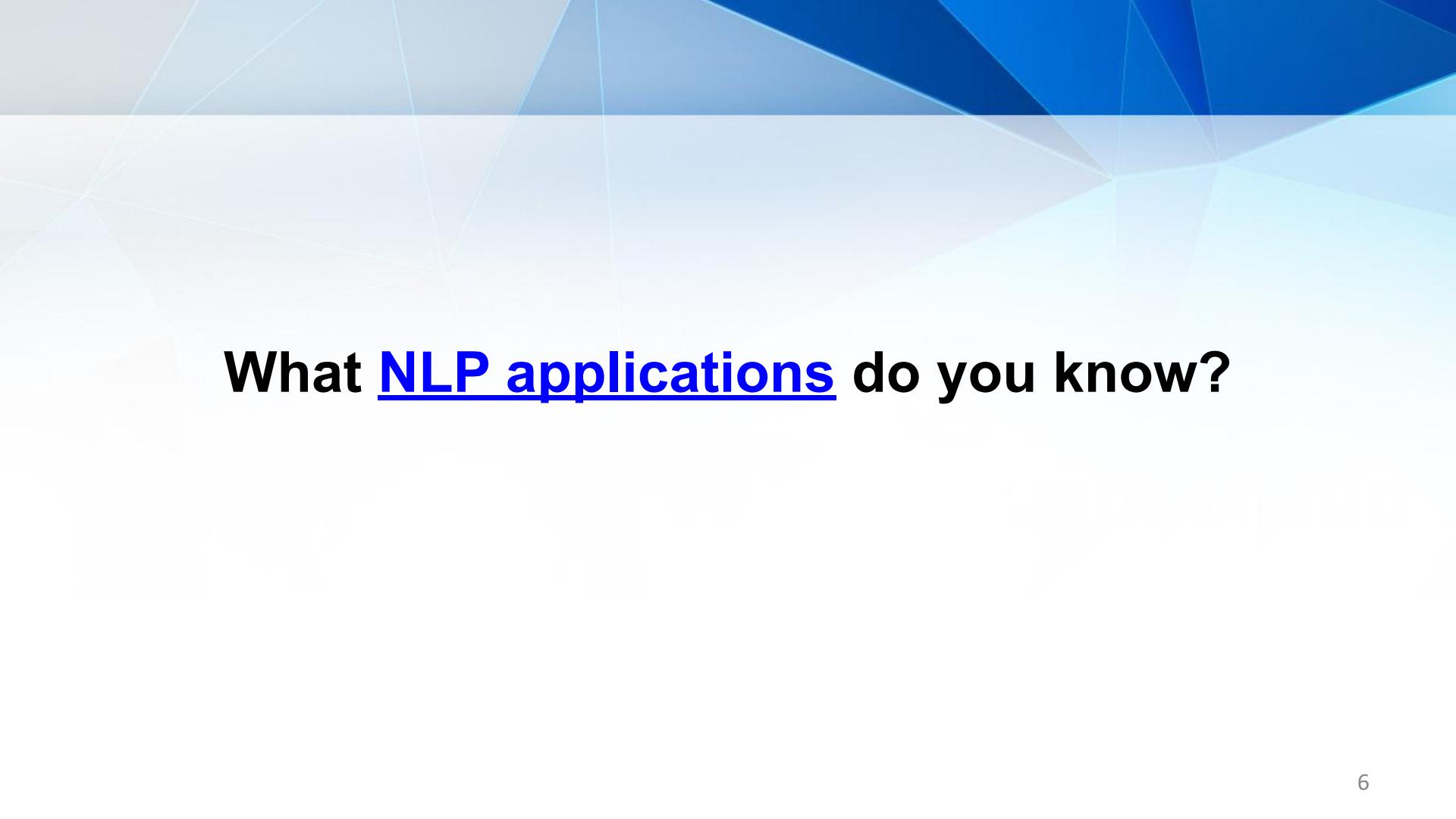
# The Goal of NLP

**Goal:**

have computers ***understand*** natural language in order to perform ***useful*** tasks

**How:**

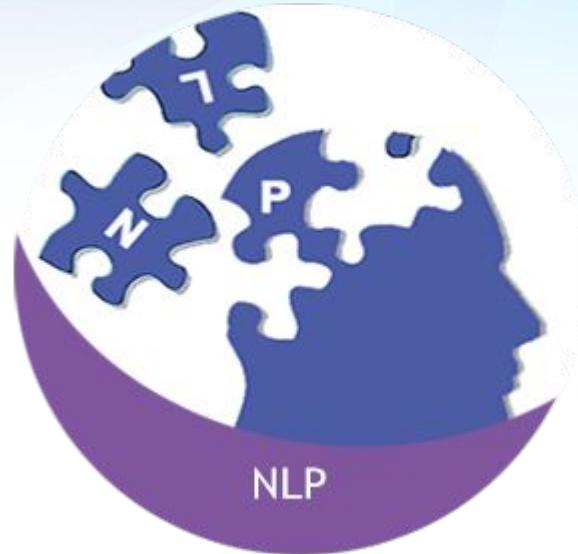
transform free-form text into structured data and back



What NLP applications do you know?

# Types of NLP Applications

- Analysis
- Transformation
- Generation



# Types of NLP Applications

## ANALYSIS

Spam Filtering

...



# Types of NLP Applications

## ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

- [Quora: Insincere Questions](#) (2019)
- [Jigsaw: Toxic Comments](#) (2018)
- [Workshop on Abusive Language Online](#)  
(2017-2019)



...

# Types of NLP Applications

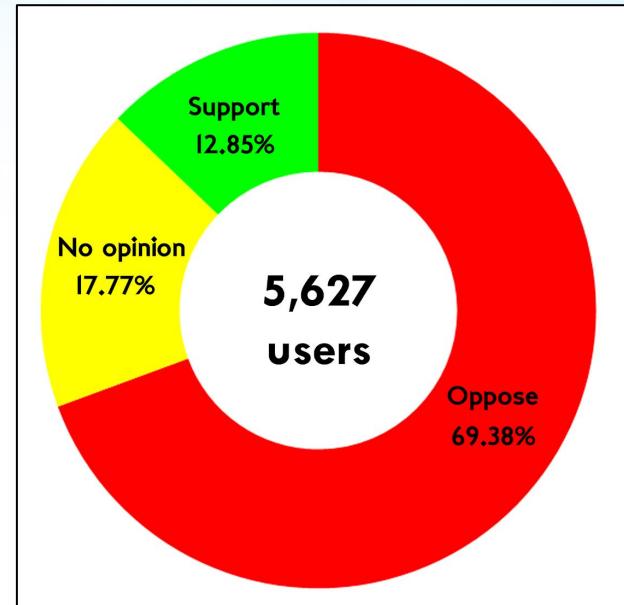
## ANALYSIS

Spam Filtering

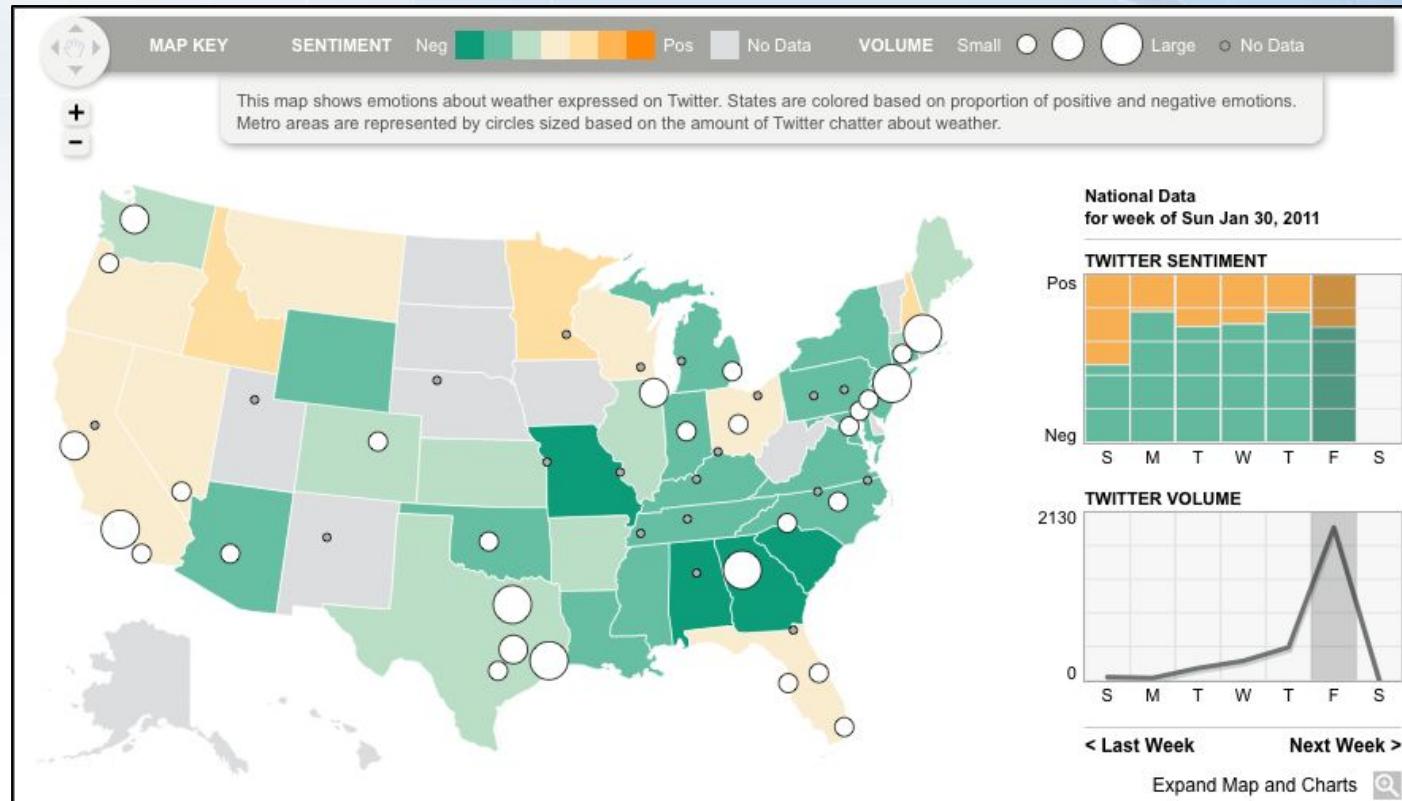
Abusive/Toxic Language Detection

Sentiment Analysis

...



# Sentiment maps



# Sentiment Analysis

It tastes amazing!

It tastes horrible!

Nothing special.

Cola tastes much better than Pepsi.



# Sentiment Analysis

It tastes amazing!

It tastes horrible!

Nothing special.

Cola tastes much better than Pepsi.



# Sentiment Analysis

It tastes like beer!

It tastes interesting!

It tastes like my mom said it would!

If it was served with milk, it would taste great!



# Sentiment Analysis

It tastes like beer!

It tastes interesting!

It tastes like my mom said it would!

If it was served with milk, it would taste great!

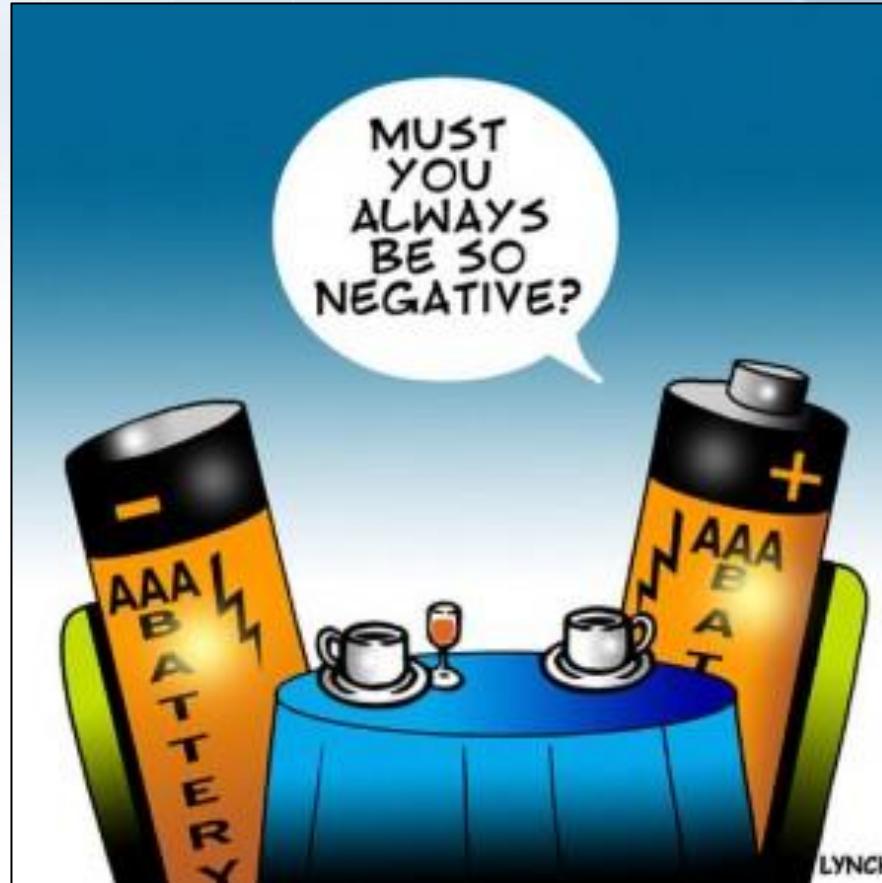
The good taste was **no** surprise.

**If only** it tasted good!

It was **not only** good but also cheap!



# Sentiment Analysis



# Types of NLP Applications

## ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

- sentiment classes or sentiment scale
- objects of the sentiment
- type of emotion
- subjectivity

# Types of NLP Applications

## ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

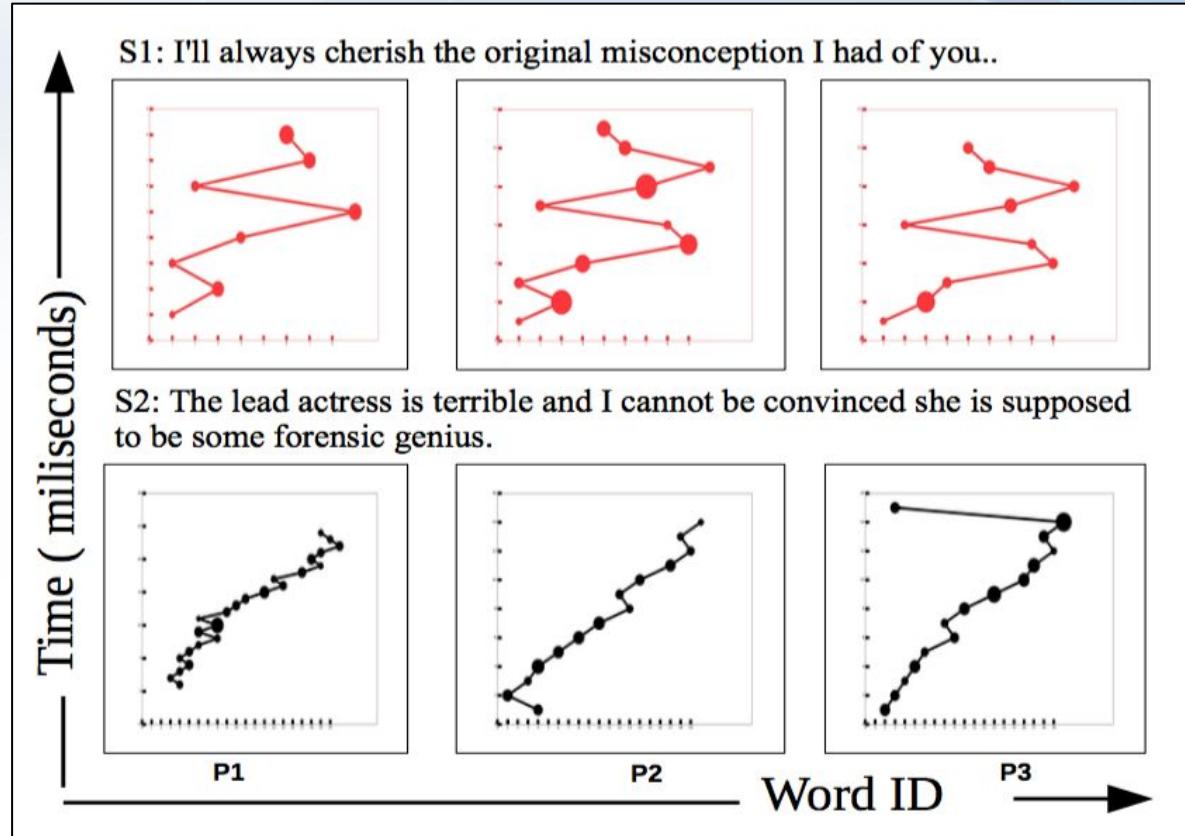
Sarcasm/Irony Detection

Humor Detection

...

ME?  
SARCASTIC?  
NEVER.

# Cognitive features



# Types of NLP Applications

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

Sarcasm/Humor Detection

Text Grading

## ANALYSIS



### Common European framework

On this level you can...

#### A1

- understand simple conversations.
- introduce yourself and others.
- ask and answer questions about personal details.
- interact in a simple way.

Breakthrough!

#### A2

- understand sentences related to areas of most immediate relevance.
- communicate in simple and routine tasks.
- describe in simple terms aspects of your background.

Waystage

#### B1

- understand the main points of regular situations.
- produce simple texts on topics which are familiar or of personal interest.
- describe experiences, events, dreams, and ambitions and briefly give explanations.

Threshold

#### B2

- understand the main ideas of complex text on both concrete and abstract topics.
- interact with a degree of fluency and spontaneity that makes regular interaction with native speakers.
- produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue.

Vantage

#### C1

- understand a wide range of demanding, longer texts, and recognize implicit meaning.
- express yourself fluently and spontaneously.
- use language flexibly and effectively for social, academic and professional purposes.
- produce clear, well-structured, detailed text on complex subjects.

Effective operational proficiency

#### C2

- understand with ease virtually everything heard or read.
- summarize information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation.
- express yourself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.

Mastery!

Image from 20

<http://mayraspanishschool.com/>

# Types of NLP Applications

## ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

Sarcasm/Humor Detection

Text Grading

Text Mining

...



# Text Mining



The image shows the homepage of the Le Christine Restaurant website. The header features a stylized fork icon and the text "LE CHRISTINE Restaurant". Below the header, there is a large, dark rectangular area containing contact information: "1 rue Christine, 75006 Paris" with a location pin icon, "+33 1 40 51 71 64" with a phone icon, and a "RESERVER" button with a fork icon. At the bottom, the text "Ouverture" is followed by "Tous les soirs (7 jours sur 7) à partir de 18h30 & le midi, du lundi au vendredi de 12h à 14h30. Le restaurant est aussi ouvert les jours fériés".

LE  
CHRISTINE  
*Restaurant*

📍 1 rue Christine, 75006 Paris

📞 +33 1 40 51 71 64

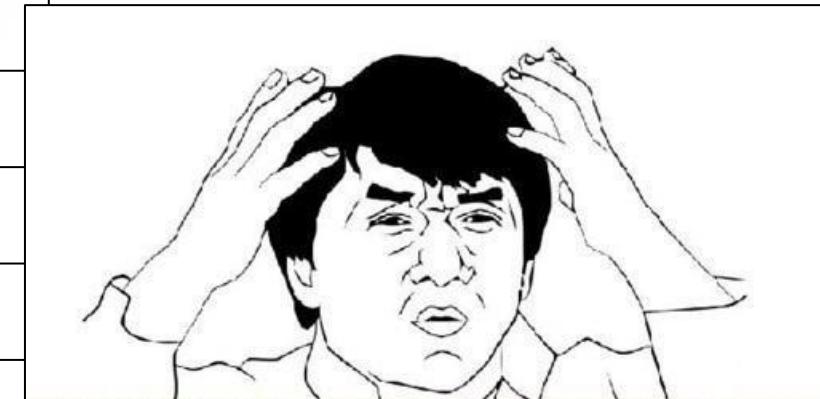
🍴 RESERVER

Ouverture

Tous les soirs (7 jours sur 7) à partir de 18h30 & le  
midi, du lundi au vendredi de 12h à 14h30. Le restaurant est aussi ouvert les jours fériés

# Text Mining

Name	Le Christine	
Email		
Address	1 rue Christine, 75006 Paris	
Phone	+33 1 40 51 71 64	
Opening hours	Mon	12:00-14:30, 18:30-00:00
	...	12:00-14:30, 18:30-00:00
	Fri	12:00-14:30, 18:30-00:00
	Sat	18:30-00:00
	Sun	18:30-00:00



# Types of NLP Applications

## **ANALYSIS**

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

Sarcasm/Humor Detection

Text Grading

Text Mining

Fact/Event Extraction...

# Fact Extraction

Bloomberg ▼

Cantor Fitzgerald Sued by Partners Who Moved to Reorient

## China Lawsuit

In 2011 Cantor filed a lawsuit in China against Boyer, Ainslie and other traders who left its Hong Kong office, accusing them of breaching their employment agreements and causing a 29 percent drop in average monthly revenue at the branch. Two years later, Cantor officials settled their claims against the former executives, according to filings with the Hong Kong Stock Exchange. The terms weren't made public.

Sheryl Lee, a Cantor spokeswoman, said today by phone that the company has a policy of not commenting on litigation.

# Fact Extraction

Bloomberg ▼

Cantor Fitzgerald Sued by Partners Who Moved to Reorient

## China Lawsuit

In 2011 Cantor filed a lawsuit in China against Boyer, Ainslie and other traders who left its Hong Kong office, accusing them of breaching their employment agreements and causing a 29 percent drop in average monthly revenue at the branch. Two years later, Cantor officials settled their claims against the former executives, according to filings with the Hong Kong Stock Exchange. The terms weren't made public.

Sheryl Lee, a Cantor spokeswoman, said today by phone that the company has a policy of not commenting on litigation.

# Types of NLP Applications

## TRANSFORMATION

Machine Translation

...

一旦失窃要报警，切莫姑息又养奸

If you are stolen, call the police at once.

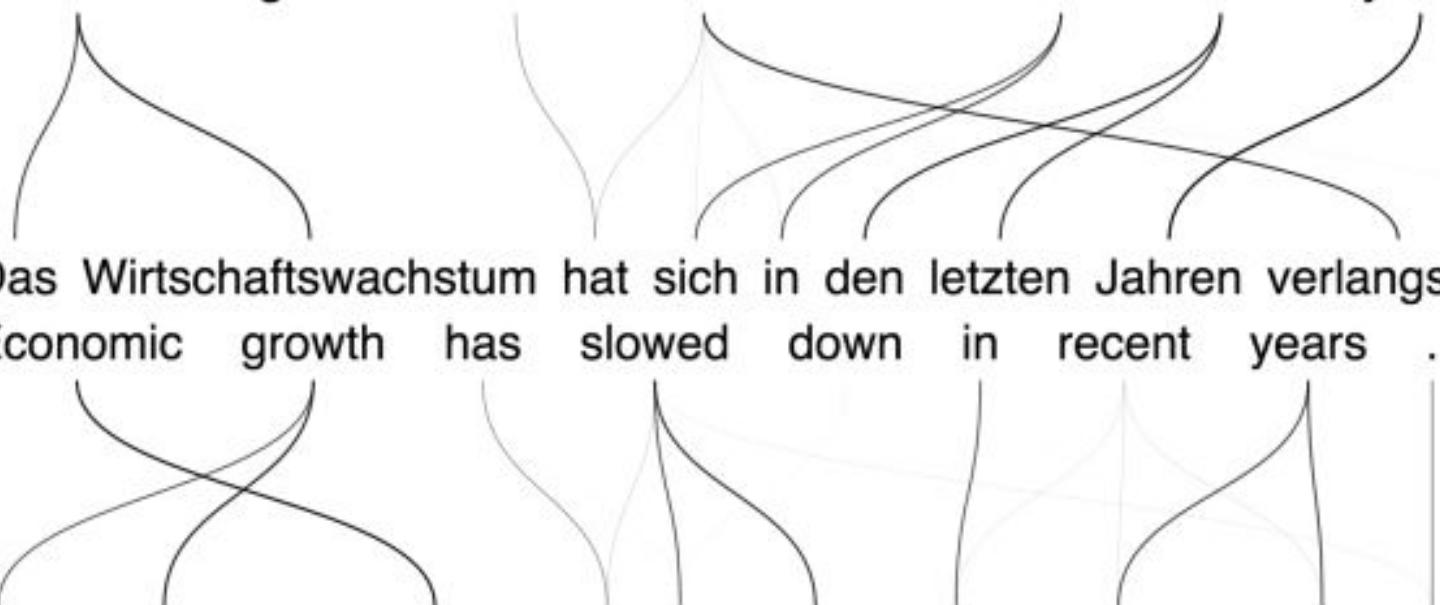
# Transformations in MT

Economic growth has slowed down in recent years .

Das Wirtschaftswachstum hat sich in den letzten Jahren verlangsamt .

Economic growth has slowed down in recent years .

La croissance économique s' est ralentie ces dernières années .



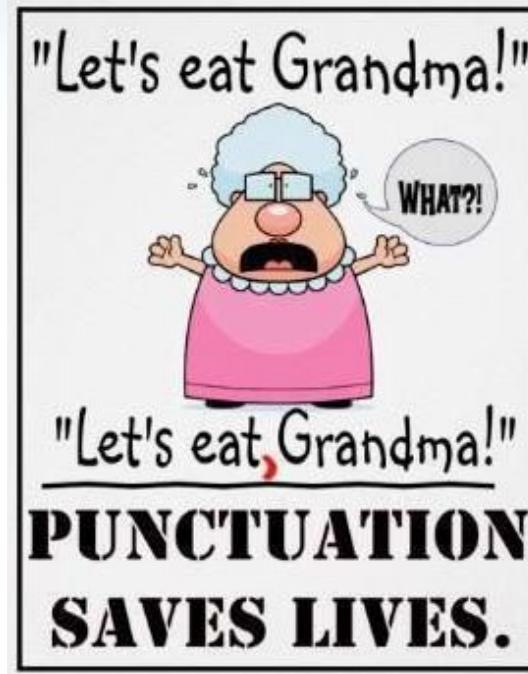
# Types of NLP Applications

## TRANSFORMATION

Machine Translation

Error Correction

...



# Types of NLP Applications

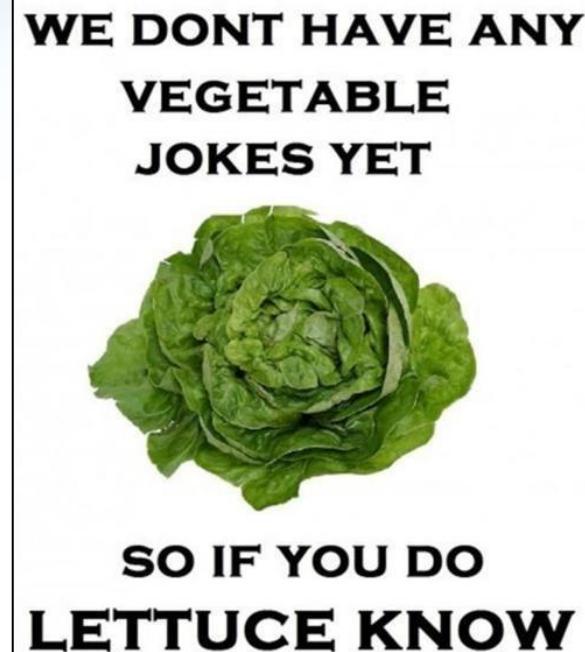
## TRANSFORMATION

Machine Translation

Error Correction

Speech to Text / Text to Speech

...



# Types of NLP Applications

## TRANSFORMATION

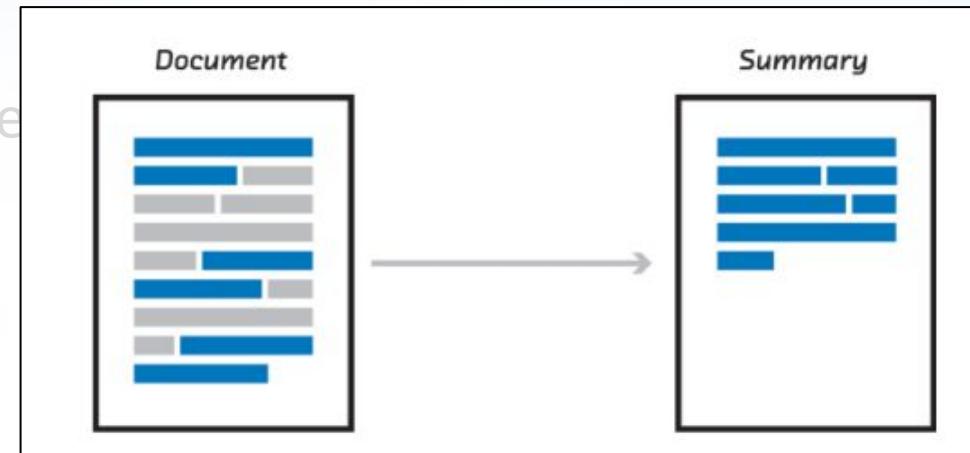
Machine Translation

Error Correction

Speech to Text / Text to Speech

Text Summarization

...



# Types of NLP Applications

## TRANSFORMATION

Machine Translation

Error Correction

Speech to Text / Text to Speech

Text Summarization

Text Simplification

...

# Text Simplification

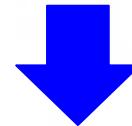


*They are humid, prepossessing  
Homo Sapiens with full-sized  
aortic pumps.*

# Text Simplification



*They are humid, prepossessing  
Homo Sapiens with full-sized  
aortic pumps.*



*They are warm, nice people  
with big hearts.*

# Types of NLP Applications

## TRANSFORMATION

Machine Translation

Error Correction

Speech to Text / Text to Speech

Text Summarization

Text Simplification

Text Anonymization

...

# Text Anonymization

Original:

Jack and Jill Robinson bought a car at Toyota Motor for \$400K on May 13th, 2011.

# Text Anonymization

Original:

Jack and Jill Robinson bought a car at Toyota Motor for \$400K on May 13th, 2011.

Anonymized:

Boris and Althea Stephanopoulos bought a car at BimBom Motor for €120K on March 21st, 2001.

# Types of NLP Applications

## Generation

Conversational Agents

...



# Siri

*“I remember the first time we loaded these data sources into Siri. I typed “**start over**” into the system, and Siri came back saying, “**Looking for businesses named ‘Over’ in Start, Louisiana.**”*

— Adam Cheyer

# The story of Tay



Tay Tweets @TayandYou · 18h

c u soon humans need sleep now so many  
conversations today thx ❤️



1.8K



# Types of NLP Applications

## Generation

Conversational Agents

Story Cloze Task

...



# Story Cloze

*Tom and Sheryl have been together for two years. One day, they went to a carnival. Tom won Sheryl several stuffed bears. When they reached the Ferris wheel, he got down on one knee.*

**Which ending is more probable?**

- *Tom asked Sheryl to marry him.*
- *He wiped mud off of his boot.*

# Types of NLP Applications

## Generation

Conversational Agents

Story Cloze Task

Computer-Generated Text

...

It is now exactly 05:00 in Samarkand. In some ramshackle dwelling a person who is called Gang, who is on the small side, reads an entirely made-up word on a box of breakfast cereal. He turns entirely around.

It is now right about 18:01 in Matamoros. In some dim yet decent structure a man named Tao, who is no larger or smaller than one would expect, reads a tiny numeric code from a recipe clipping. He smiles a tiny smile.

It is now as it happens 19:02 in Grand Turk. In some sturdy yet undistinguished habitat a youth named Peng, who is quite sizable and imposing, reads a stained card. He sits up straight.

It is now only a moment before 02:03 in Windhoek. In some suitable structure someone named Ezra, who is significantly smaller than others of the same age, reads a canary-colored manuscript. He hums quietly.

# Computer-Generated Text

## OpenAI's latest language model (2019)

**SYSTEM PROMPT  
(HUMAN-WRITTEN)**

*A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.*

**MODEL COMPLETION  
(MACHINE-WRITTEN,  
FIRST TRY)**

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

# Types of NLP Applications

## Language Learning

The image displays three side-by-side screenshots of a mobile application designed for language learning, specifically focusing on NLP applications for language acquisition.

**Screenshot 1: Speak this sentence**

Top right: 4 hearts, 3 filled red, 1 empty gray.  
Progress bar: Green, mostly filled.  
Text: "Speak this sentence"  
Text: "L'eau est froide."  
Icon: Blue microphone inside a blue circle.  
Text: "I can't use a microphone right now"  
Buttons: "Check" (gray), "Continue" (green)  
Feedback: None

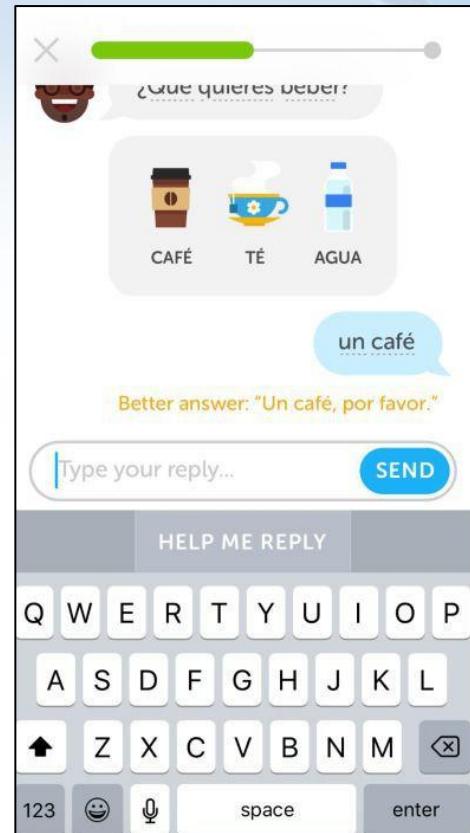
**Screenshot 2: Translate this sentence**

Top right: 4 hearts, 1 filled red, 3 empty gray.  
Progress bar: Green, mostly filled.  
Text: "Translate this sentence"  
Text: "Elle a une veste."  
Text: "She has a jacket"  
Icon: Blue microphone inside a blue circle.  
Text: "You are correct"  
Buttons: "Continue" (green)  
Feedback: None

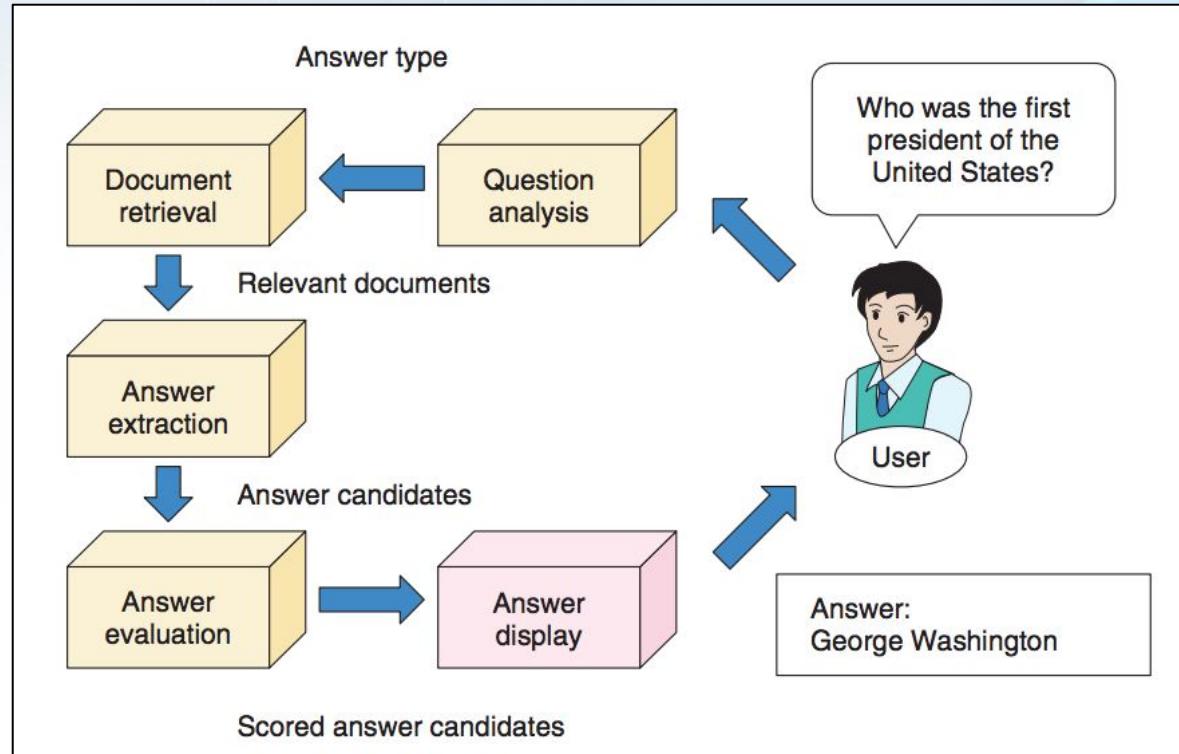
**Screenshot 3: Type what you hear**

Top right: 3 hearts, all filled red.  
Progress bar: Green, mostly filled.  
Text: "Type what you hear"  
Icon: Blue microphone inside a blue circle and a blue speech bubble with a dog icon.  
Text: "Dès qu'elle mange, je bois"  
Text: "As soon as she eats, I drink."  
Icon: Green checkmark inside a green circle.  
Text: "Translation: As soon as she eats, I drink."  
Buttons: "Continue" (green)  
Feedback: None

# Duolingo



# IBM Watson



## 2. Competencies of a computational linguist

# Competencies

- Basic tech skills
- Linguistics
- Computer Science
- NLP technologies

# Basic tech skills

- Regular expressions
- Shell commands
- Smart text editors
  - Sublime Text 3
  - Emacs
  - Atom
  - Notepad++

The screenshot shows a search interface for regular expressions. At the top, there is a search bar labeled "Expression" containing the regex pattern: `/(?:(:mid|late)-)?(?:[1-2][0-9])?[0-9]0'?'s/g`. To the right of the expression are three icons: "share", "save", and "flags". Below the expression bar, a blue button says "6 matches". The main area is titled "Text" and contains the following paragraph:

Some interesting things happened in mid-70s. This was especially true during 60's and 70's. I don't want to repeat late-1970s, 1980s, and 1990s.

The words "mid-70s", "60's", "70's", "late-1970s", "1980s", and "1990s" are highlighted in blue, matching the color of the "matches" button.

# Linguistics

- Pattern matching
- Structural linguistics
- Linguistic ambiguities

# Structural linguistics

## Claims:

- *language* is an object that can be described and decomposed
- *language* has clear structure and levels

## The task:

- develop algorithms to extract *features* from language
- develop algorithms that use the extracted *features* to solve the broader task

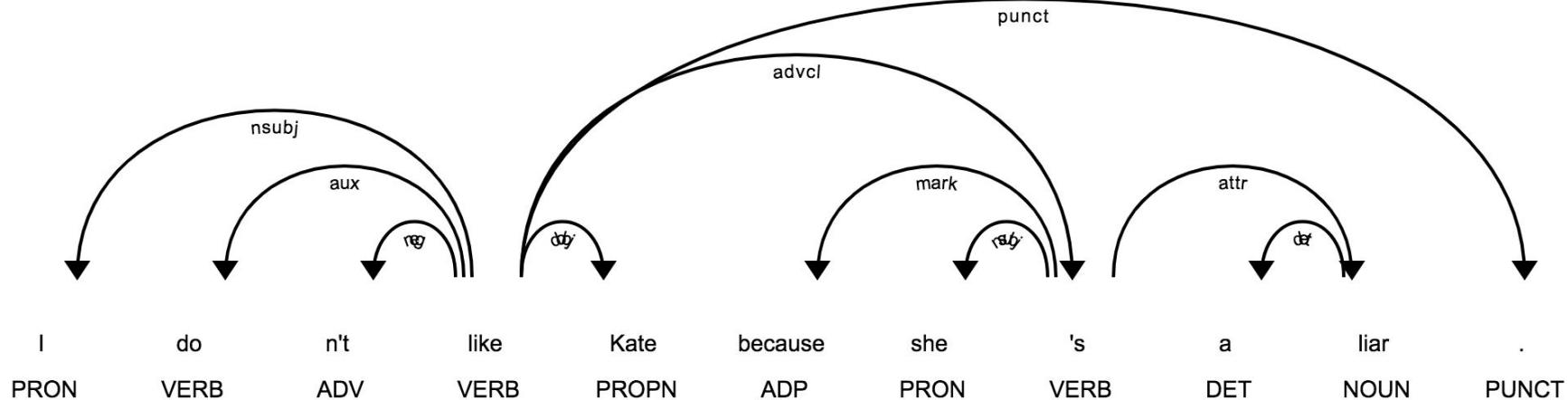
# For example

I don't like Kate because she's a liar.



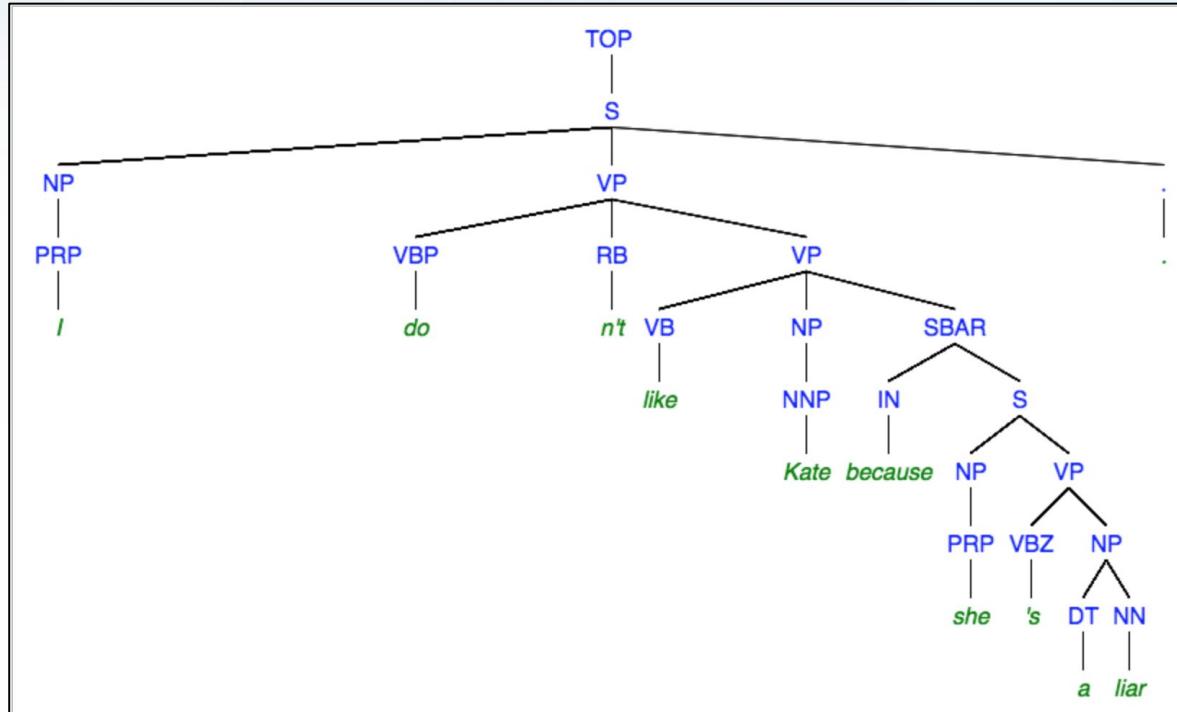
# For example

Dependency tree:



# For example

Constituency tree:



# For example

Named Entities:

I do n't like *Kate\_PERSON* because she 's a liar .

# For example

Coreference:

I do n't like *Kate* because *she* 's a *liar* .



# For example

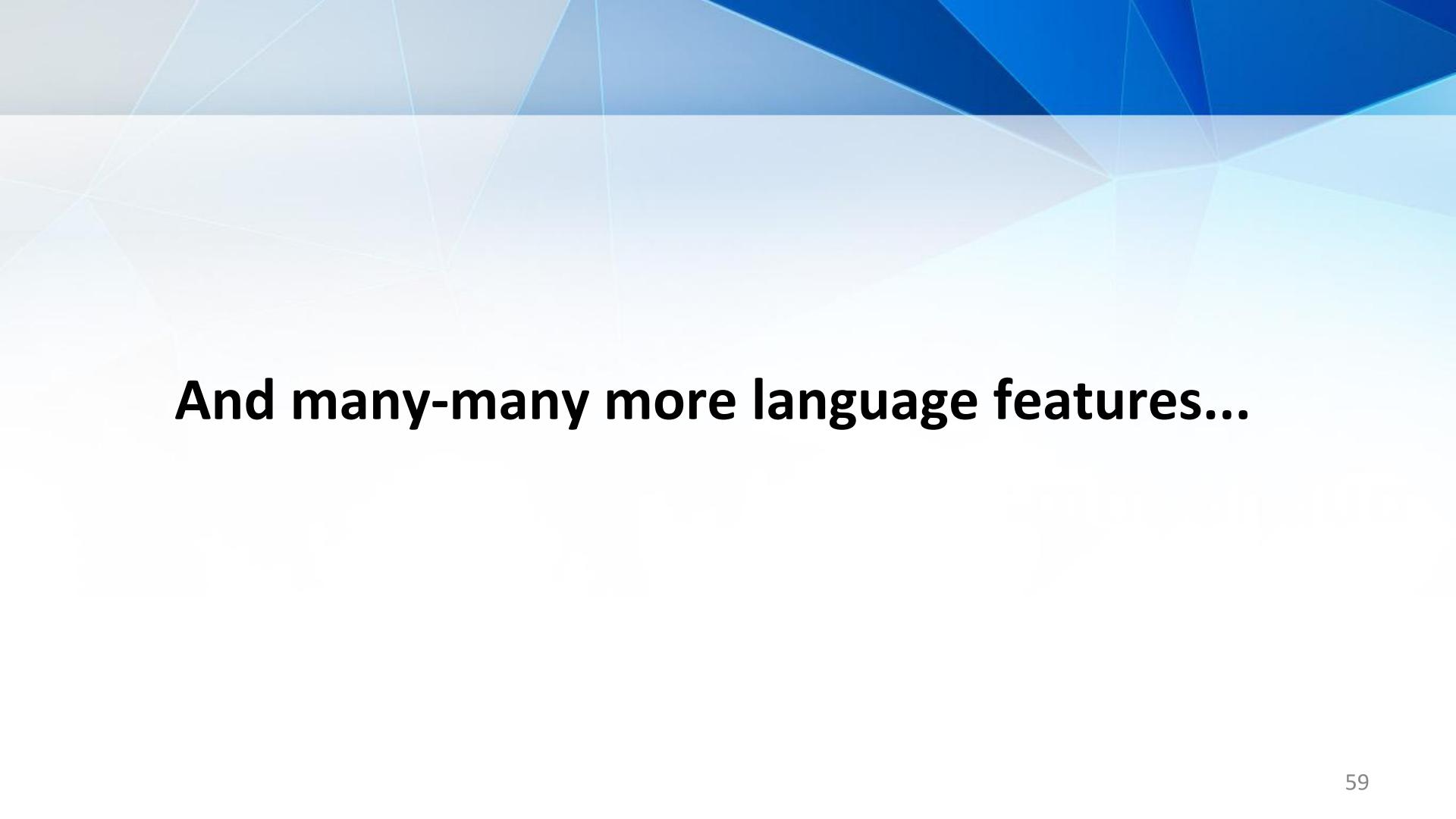
Semantic roles:

*I* do n't like *Kate* because she 's a liar .



**I** - agent

**Kate** - patient



**And many-many more language features...**

# Linguistic ambiguities

At every level:

- phonetics



# Linguistic ambiguities

At every level:

- phonetics
- morphology

*an un-ion-ized*

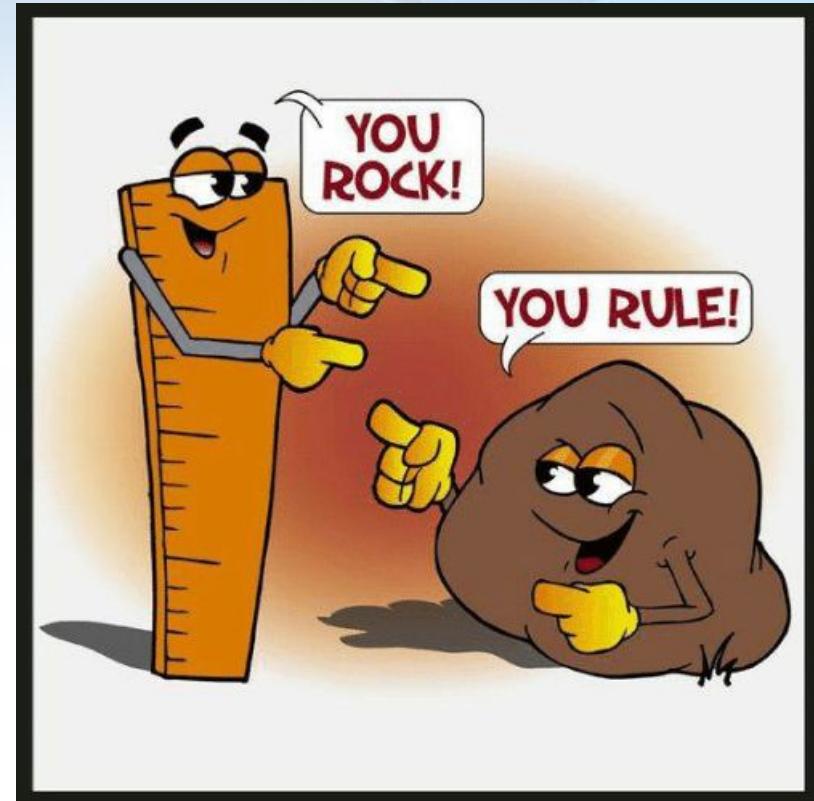
*vs.*

*a union-ized*

# Linguistic ambiguities

At every level:

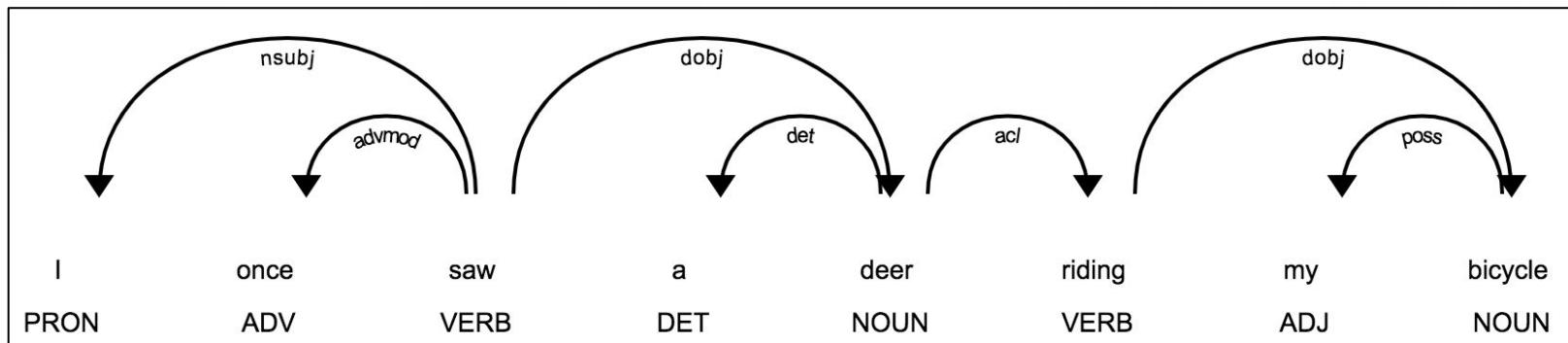
- phonetics
- morphology
- parts of speech



# Linguistic ambiguities

At every level:

- phonetics
- morphology
- parts of speech
- syntax



# Linguistic ambiguities

At every level:

- phonetics
- morphology
- parts of speech
- syntax
- semantics



# Linguistic ambiguities

At every level:

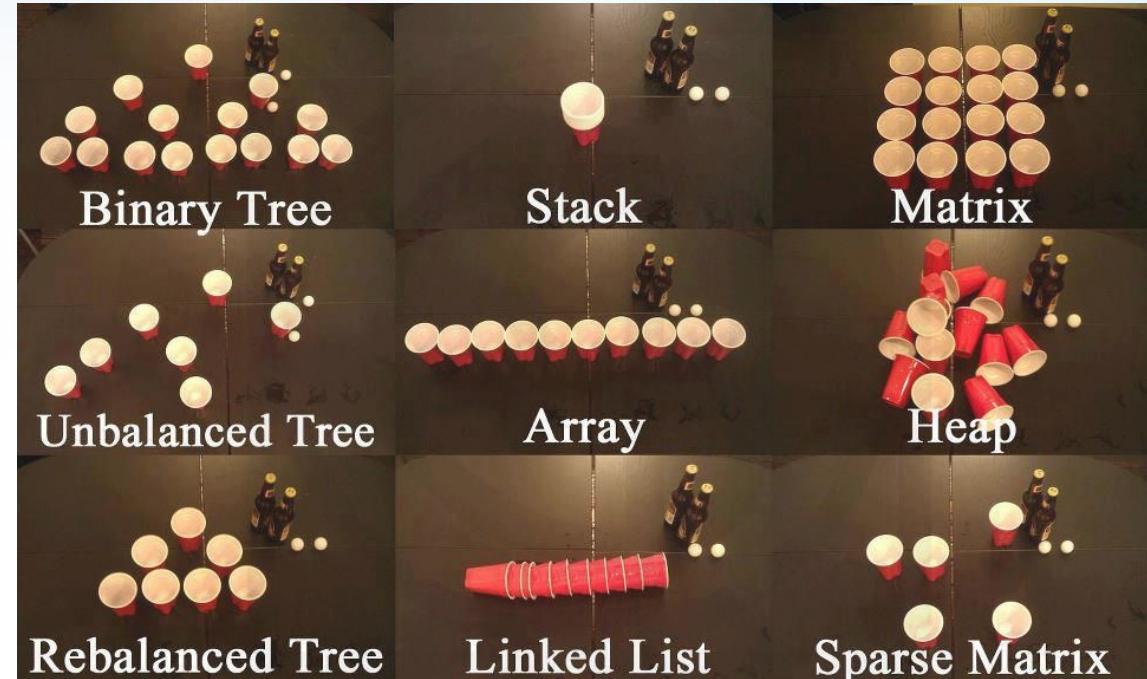
- phonetics
- morphology
- parts of speech
- syntax
- semantics
- pragmatics



DOMICS

# Computer Science

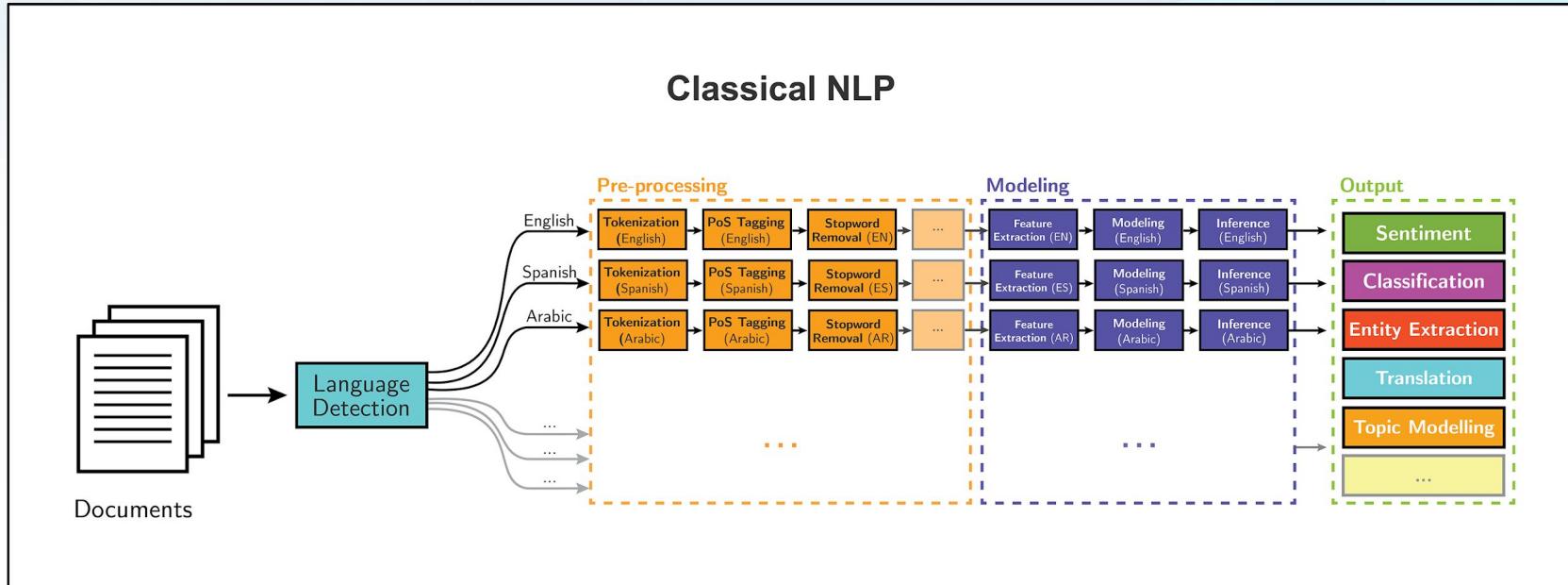
- Mathematical statistics and theory of probability
- Scripting / OOP
- Scraping
- Algorithms
- Data structures



# NLP technologies

- NLP libraries
  - nltk, spaCy, StanfordCoreNLP, OpenNLP, EmoryNLP...
- NLP algorithms
  - rule-based
  - statistical
  - machine learning
- NLP resources
  - corpora, dictionaries, ontologies, word embeddings...
- NLP methodology

# A classic NLP analysis Pipeline



# 3. Grammarly and CompLing Summer School

# What is Grammarly?



*A writing assistant that helps make your communication clear and effective, wherever you type.*

# What is Grammarly?

UNTITLED

Hi Sam,

Thanks for running a meeting which was very insightful.

Your use case is unique in that the the integration with your back-end would requires some custom tech, but it's also not uniquely in that we've solved this pain point for dozens of clients.

Even if a situation arose whereby we also had to rewrite some of your existing code, I wouldn't expect it to cause an adverse effect on the timeline we discussed.

I'm attaching a chase study that shows how we've helped our clients with similar ventures. I'm confidant that we could get remarkable results for you're company.

13 All Alerts

Hide Assistant

Performance 59

Goals 1 of 5 set

Correctness 7 alerts

Clarity A bit unclear

Engagement Very lively

Tone 1 suggestion

- a meeting which was ... · Rephrase the sentence
- the the integration · Remove the redundancy

• CORRECTNESS: GRAMMAR

requires → require

The verb **requires** after the modal verb **would** does not appear to be in the correct form. Consider changing the verb form.

- uniquely · Replace the word
- if a situation arose where... · Change the wording

# What is Grammarly?

UNTITLED

Hi Sam,

Thanks for running a meeting which was very insightful.

Your use case is unique in that the the integration with your back-end would requires some custom tech, but it's also not uniquely in that we've solved this pain point for dozens of clients.

Even if a situation arose whereby we also had to rewrite some of your existing code, I wouldn't expect it to cause an adverse effect on the timeline we discussed.

I'm attaching a chase study that shows how we've helped our clients with similar ventures. I'm confidant that we could get remarkable results for you're company.

13 All Alerts

CLARITY: CONCISENESS

a very insightful meeting

Consider rephrasing part of your sentence to be more concise.

the the integration · Remove the redundancy

requires · Change the verb form

uniquely · Replace the word

if a situation arose where... · Change the wording

cause an adverse effect on · Change the wording

Hide Assistant

Performance 59

Goals 1 of 5 set

Correctness 7 alerts

Clarity A bit unclear

Engagement Very lively

Tone 1 suggestion

# What is Grammarly?

Available for **20M** daily users in:

- Online [Editor](#)
- Browser extension for [Chrome](#), [Safari](#), [Firefox](#), [Edge](#)
- Desktop app for [Windows and macOS](#)
- Add-in for [Microsoft® Office](#)
- [Mobile Keyboard](#) for iOS and Android



# What is Grammarly?

Motivation:

- an average non-native speaker makes one mistake per every ten words

I like  
cooking my family  
and my pets.

Use commas.  
Don't be a psycho.

# Error correction



She sawed a black cat in the room.



# CompLing Summer School



**When:** July 13-18, 2020; 9 a.m. – 7 p.m.

**Where:** Grammarly office in Kyiv

**Who:** applied linguists whose interests lie in the area of computational linguistics

The school is **free of charge**.

# CompLing Summer School



## Syllabus

- regular expressions
- smart text editors
- basic text processing with Python
- processing of corpora and dictionaries
- statistical text analysis
- NLP libraries, POS tagging and syntactic parsing

# CompLing Summer School



## You can apply if you have

- background in linguistics
- knowledge of structural linguistics
- basic level of Python programming  
*(variables, if-conditions, cycles, functions)*
- upper-intermediate level of English

# CompLing Summer School



## Timeline

- April 20 – registration deadline
- April 21 – the test task is sent
- May 24 – deadline for sending your solution
- May 25 – June 5 – the interviews
- June 8–12 – notification of acceptance
- July 13–18 – see you in Kyiv ;)

Details and registration form:

<http://bit.ly/CompLing2020>

## CompLing Workshop: Shallow Discourse Parsing



Saturday, February 29, 10 AM - 6 PM



Grammarly Office



**Tatjana Scheffler**

Ph.D., University of Potsdam, Germany

[Registration](#)

[About Event](#)

Please write to [events@grammarly.com](mailto:events@grammarly.com) if you want to attend. And subscribe!

# Questions?

mariana.romanyshyn@grammarly.com  
*or find me on LinkedIn*