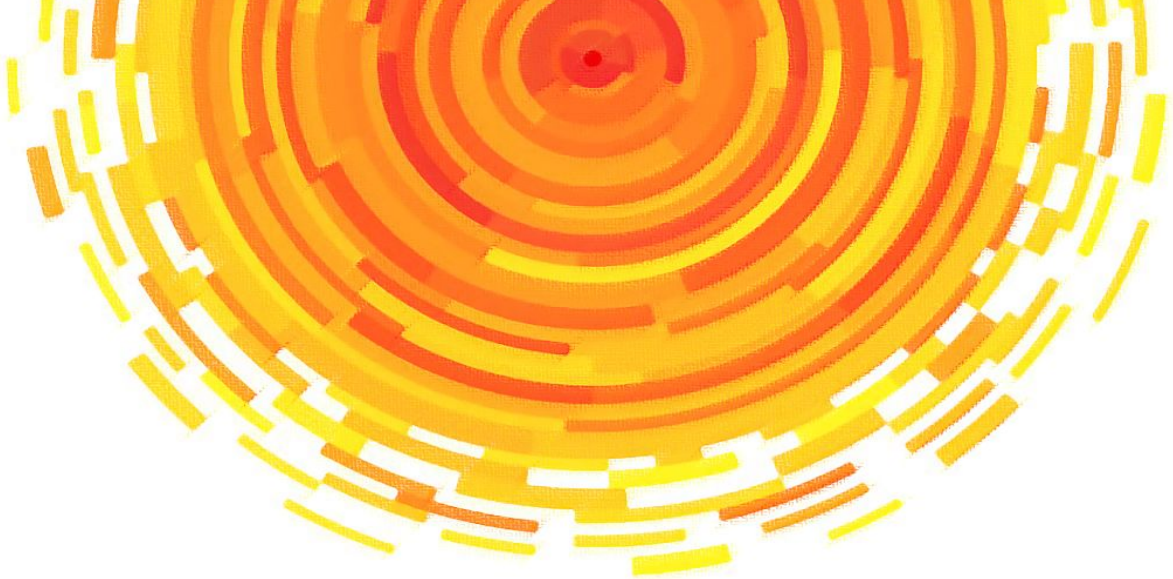


Літня школа перекладу

Introduction to Natural Language Processing

Mariana Romanyshyn
Computational Linguist at Grammarly, Inc.

Contents

1. NLP applications in our world
2. Life of a computational linguist
3. Practical examples

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

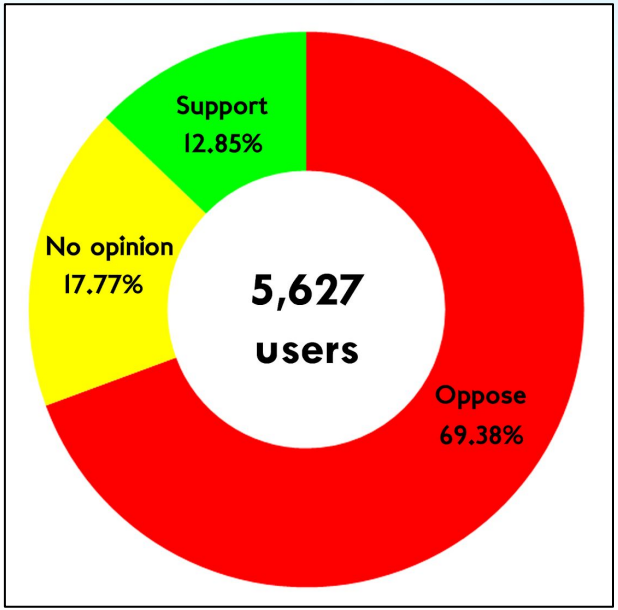
ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

…



Sentiment maps

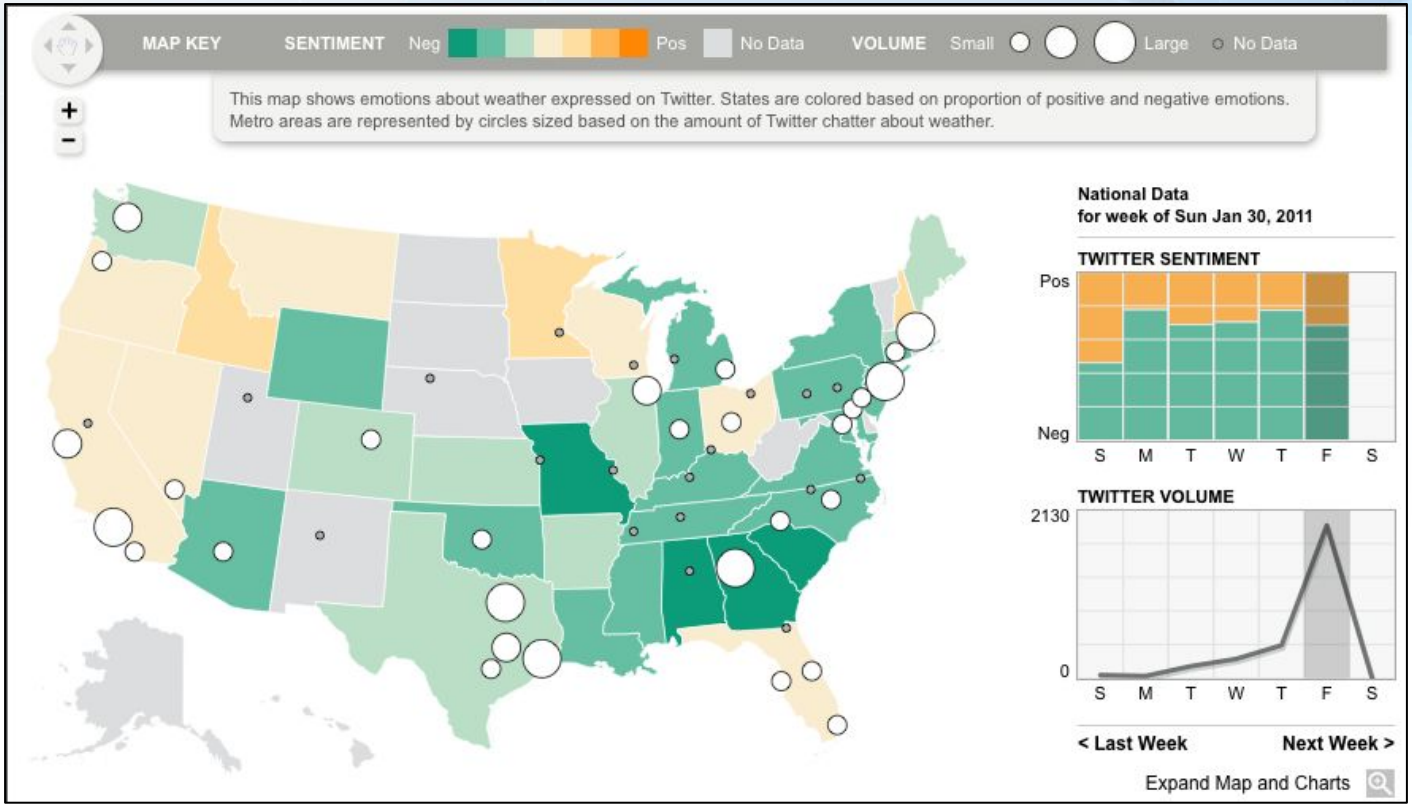


Image from http://www.dialogueearth.org/

Sentiment Analysis

It tastes amazing!
It tastes horrible!
Cola tastes much better than Pepsi.

Sentiment Analysis

It tastes amazing!
It tastes horrible!
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
- support vs. opposition

Types of NLP Applications

ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

Sarcasm/Irony Detection

Humor Detection

…

Types of NLP Applications

ANALYSIS

Spam Filtering

Abusive/Toxic Language Det

Sentiment Analysis

Sarcasm/Humor Detection

Text Grading

…

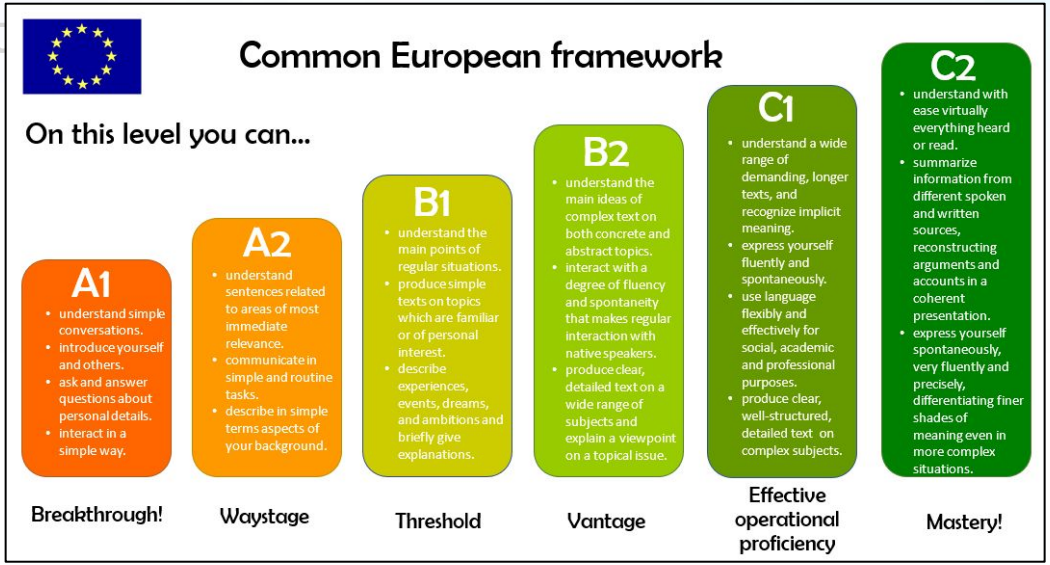


Image from http://mayraspanishschool.com/

Types of NLP Applications

ANALYSIS

Spam Filtering

Abusive/Toxic Language Detection

Sentiment Analysis

Sarcasm/Humor Detection

Text Grading

Text Mining

…

Text Mining

Task: extract data about the company from its web page.

- Company name
- Phone/Fax
- Email
- Address
- Foundation date
- Working hours
- Partners and investors, etc.

Text Mining

Image from www.restaurantlechristine.com/

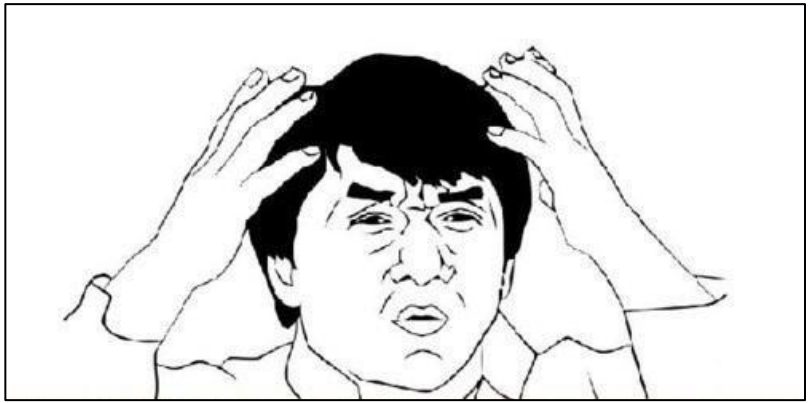
Text Mining

Image from www.restaurantlechristine.com/

Text Mining

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



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.

Image from www.bloomberg.com/

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.

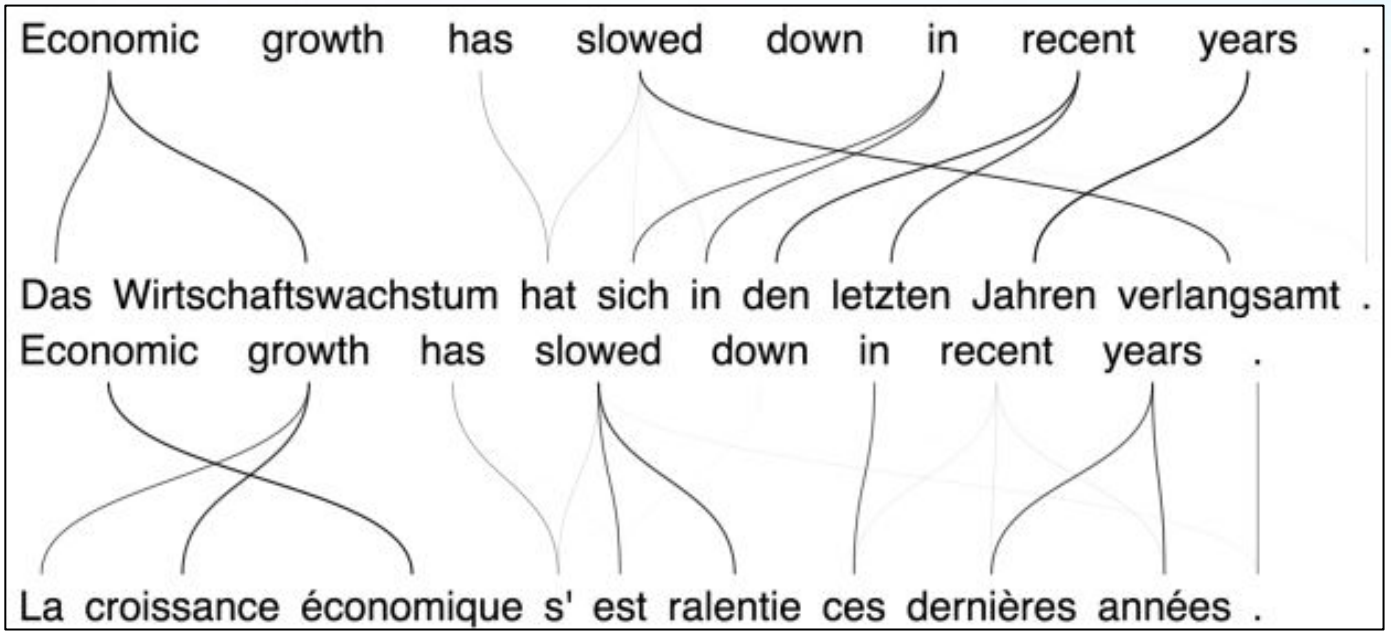
Image from www.bloomberg.com/

Types of NLP Applications

TRANSFORMATION

Machine Translation

…



Types of NLP Applications

TRANSFORMATION

Machine Translation

Error Correction

…

Error correction: Grammarly

Demo document

For years I have been driving an old used car with a lot of mileage, and I hate it. It gets me where I need to go, but I'm tired of fixing leaks and broken parts all the time. Its annoying that I have to take it to mechanic every times. Even when they take care of everything, I know in a week I'll just end up going back there.

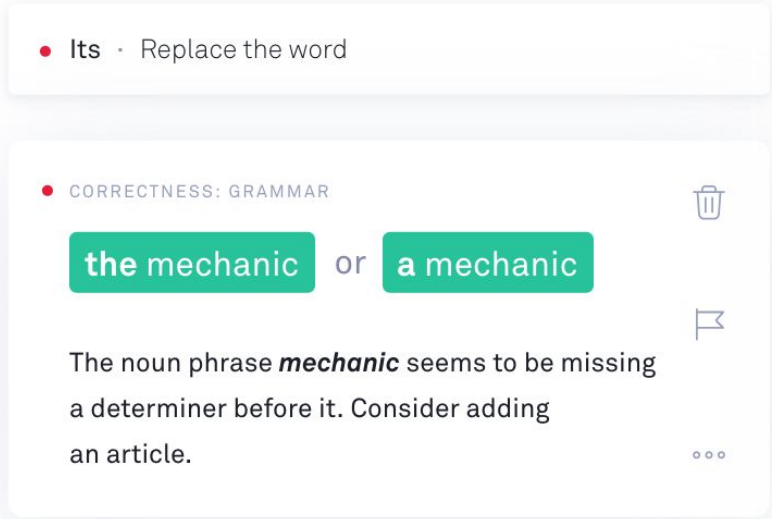


Image from https://app.grammarly.com

Error correction: Grammarly

Motivation:

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

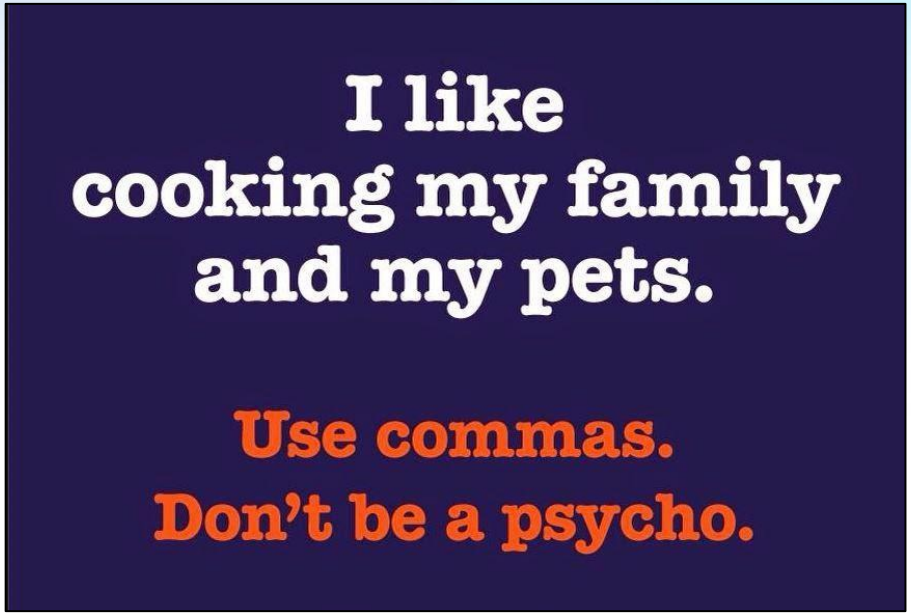
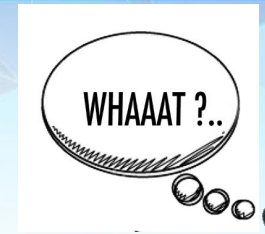
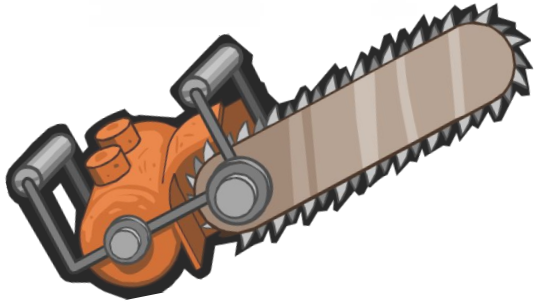


Image from www.writing.com/

Error correction: Grammarly



She sawed a black cat in the room.



Error correction: LanguageTool

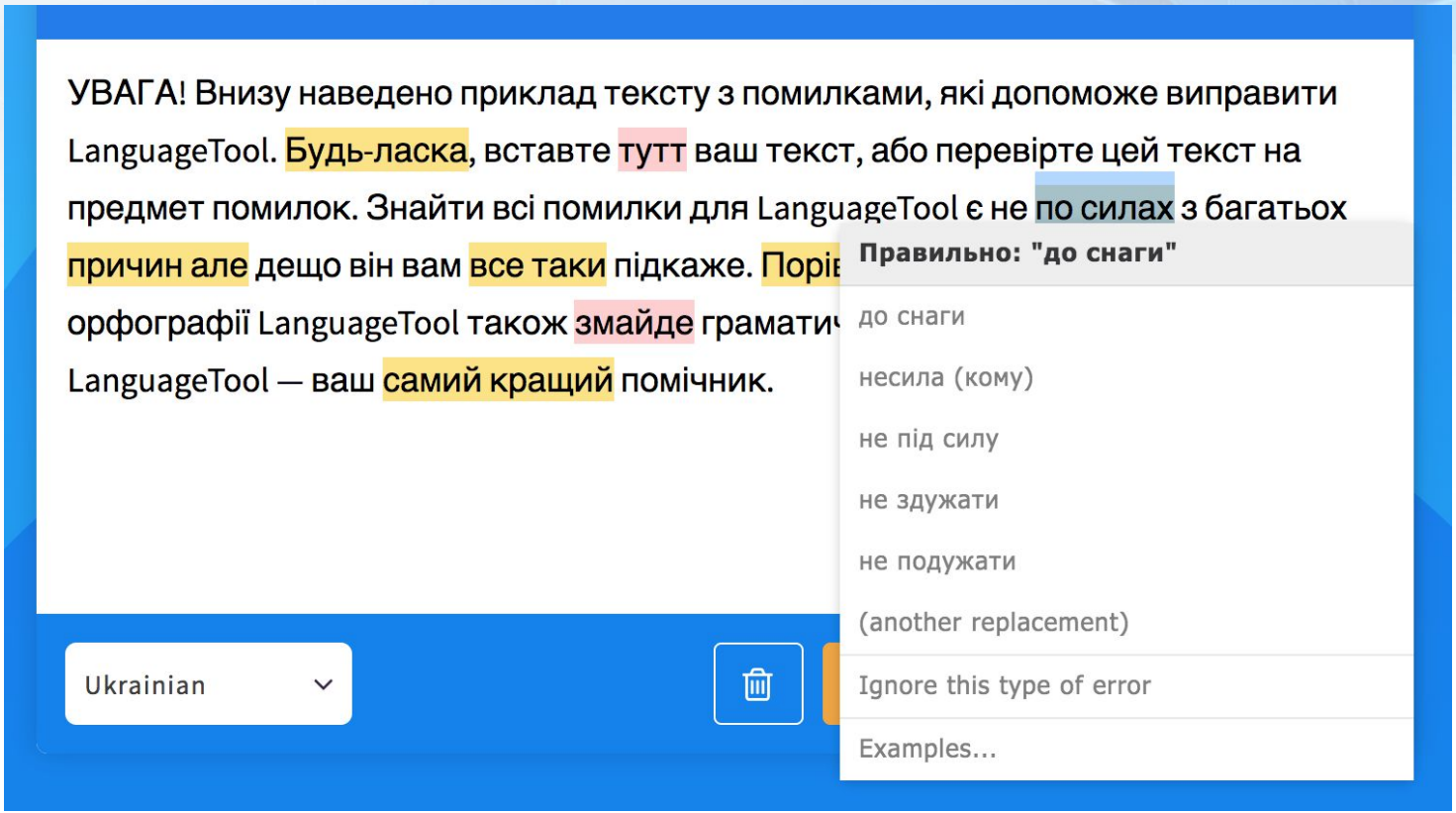


Image from https://languagetool.org

Types of NLP Applications

TRANSFORMATION

Machine Translation

Error Correction

Speech to Text / Text to Speech

…

Types of NLP Applications

TRANSFORMATION

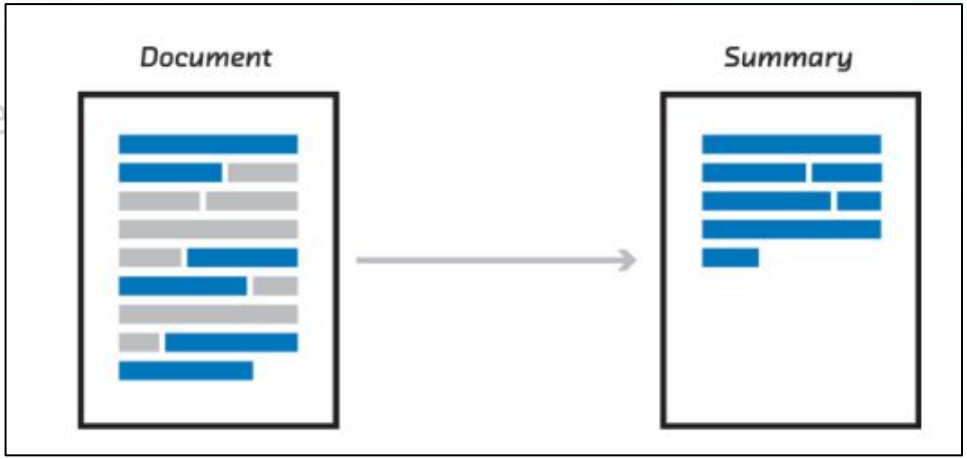
Machine Translation

Error Correction

Speech to Text / Text to Spee

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

...

Image from www.ntt-review.jp

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

TRANSFORMATION

Machine Translation

Error Correction

Speech to Text / Text to Spe

Text Summarization

Text Simplification

Text Anonymization

Question Answering...

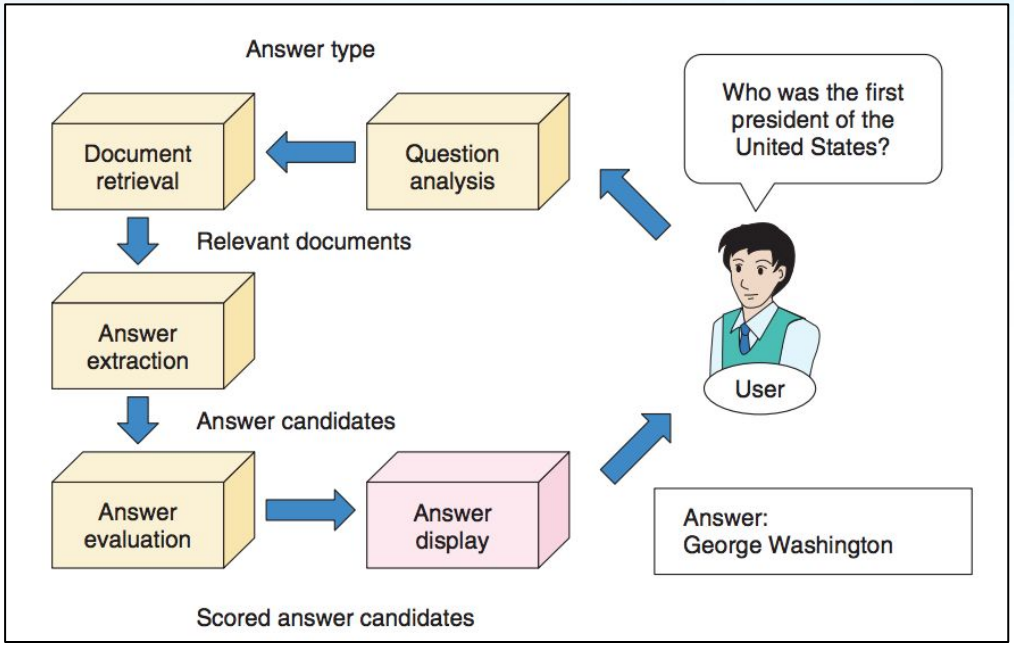


Image from www.ntt-review.jp

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

Taken from https://medium.com/swlh/the-story-behind-siri-fbeb109938b0

The story of Tay

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.*

Taken from Mostafazadeh N. et al. (2016)

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.

Taken from Nick Montfort (2013)

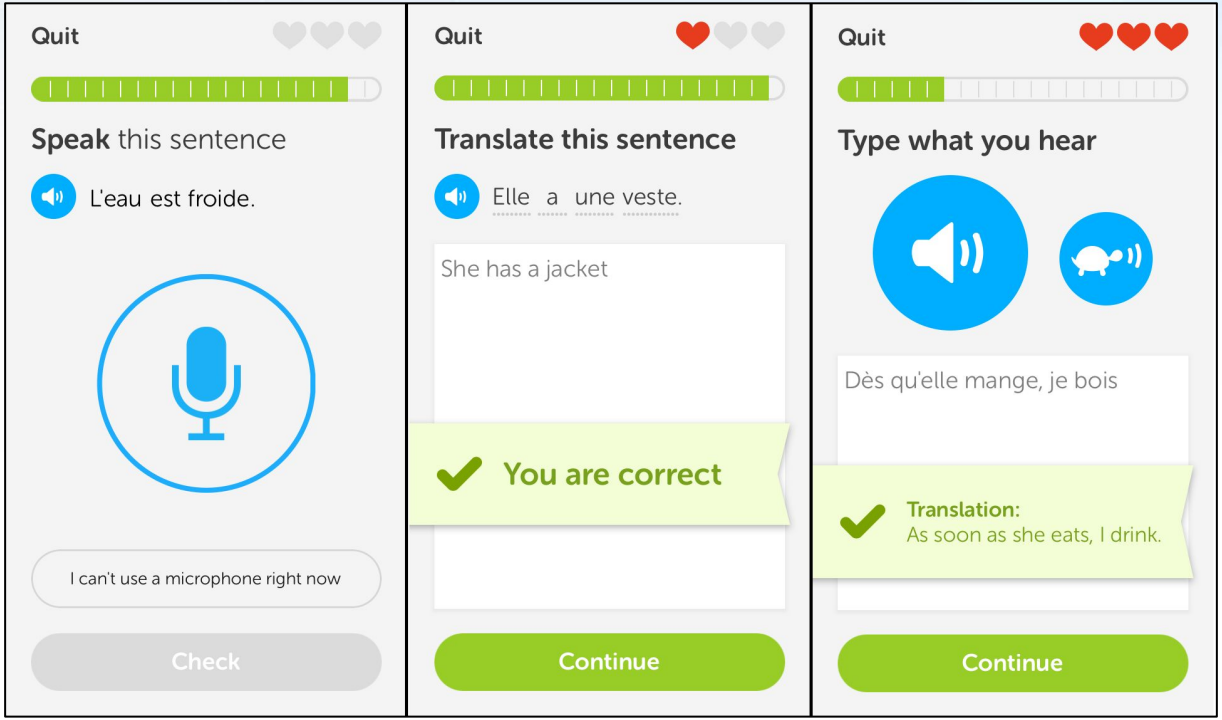
Computer-Generated Text

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."

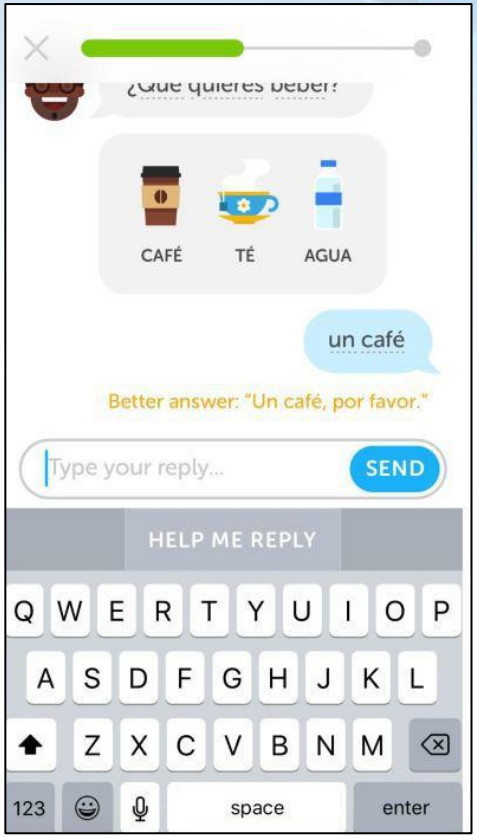
Types of NLP Applications

Language Learning



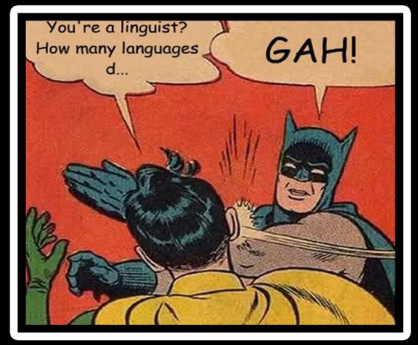
Duolingo


Hola
¿Cómo estás?
Bien tú?
¡Qué bien! Yo estoy bien.
Conoce a mis amigos. Primero, ¡dile "hola" a Roberto!
Type in Spanish
SEND
Somebody else is typing...
Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
123
space
return


CAFÉ
TÉ
AGUA
un café
Better answer: "Un café, por favor."
Type your reply...
SEND
HELP ME REPLY
Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
123
space
enter

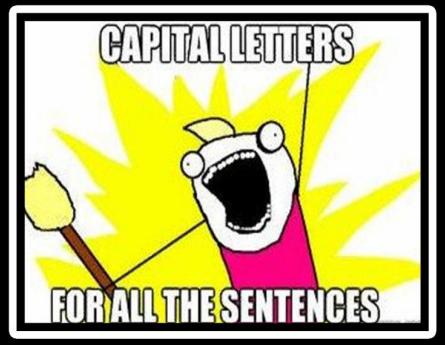
2. Life of a computational linguist

COMPUTATIONAL LINGUIST

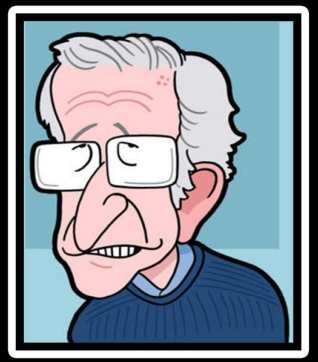


WHAT MY FRIENDS THINK I DO

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, k) in fd.keys():
            if i == first_word:
                list_of_nexts.append(j)
            if len(list_of_nexts) == 10:
                break
        if len(list_of_nexts) == 0:
            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

Computational linguistics

Use structural linguistics

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

To solve NLP tasks

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

The language

Distinguishing features:

- Ambiguous
- Noisy
- Evolving

Fruit flies fast.

The language

Distinguishing features:

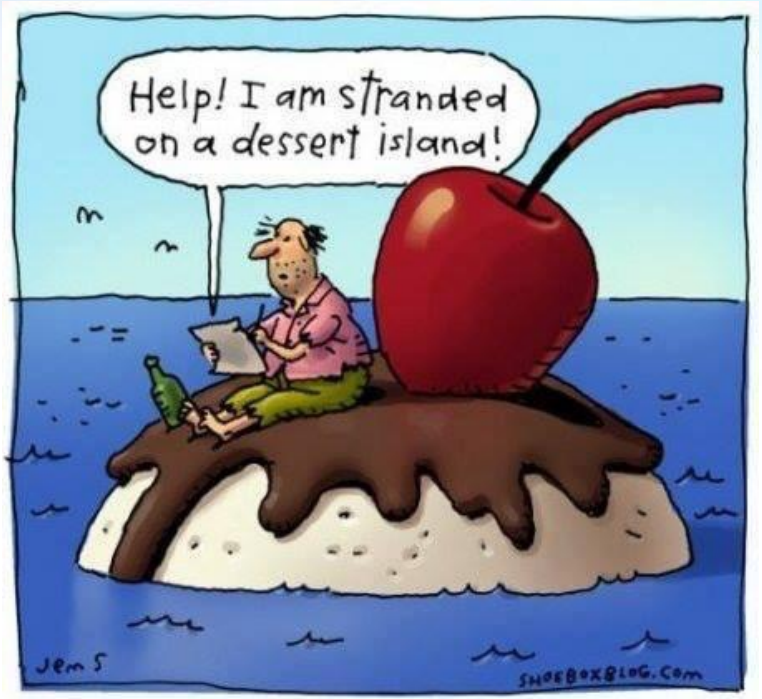
- Ambiguous
- Noisy
- Evolving

Fruit flies fast.

The language

Distinguishing features:

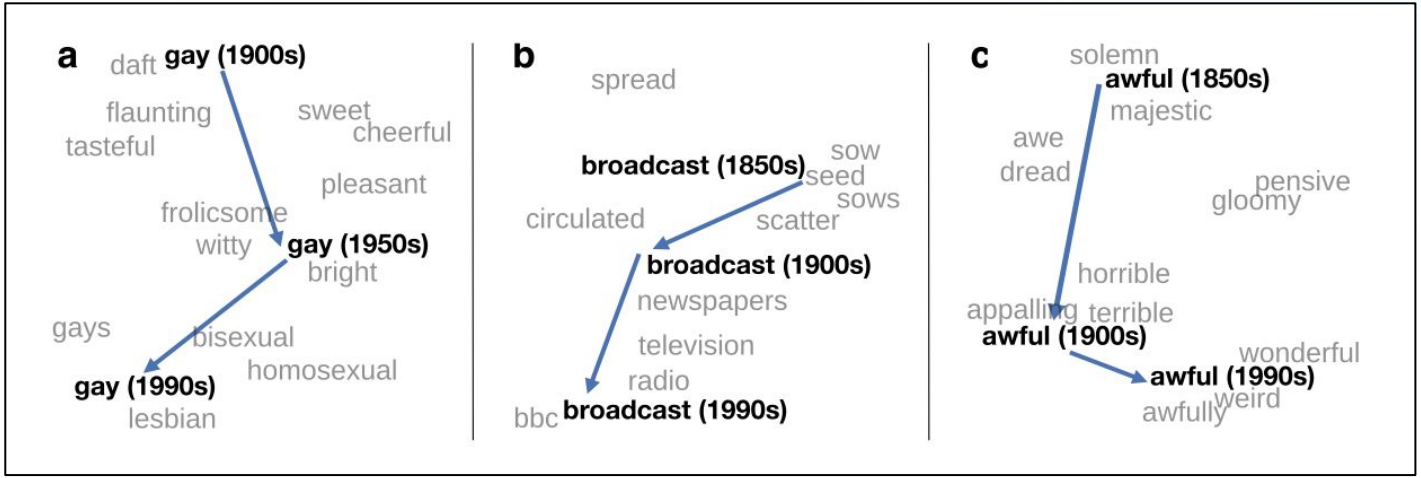
- Ambiguous
- Noisy
- Evolving



The language

Distinguishing features:

- Ambiguous
- Noisy
- Evolving

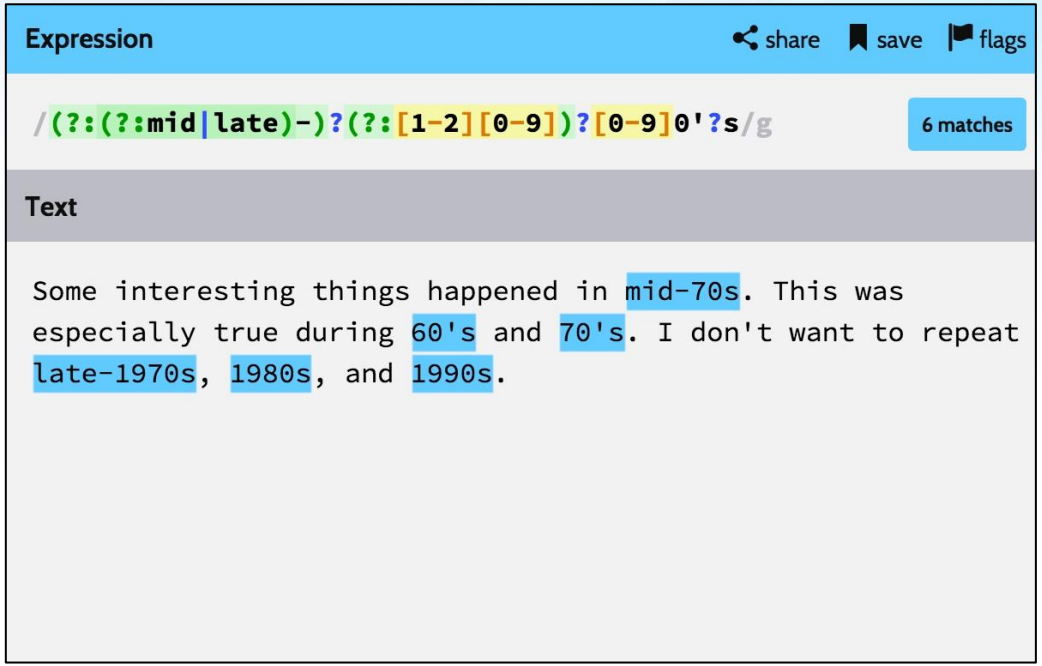


Competencies

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

Basic tech skills

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



Linguistics

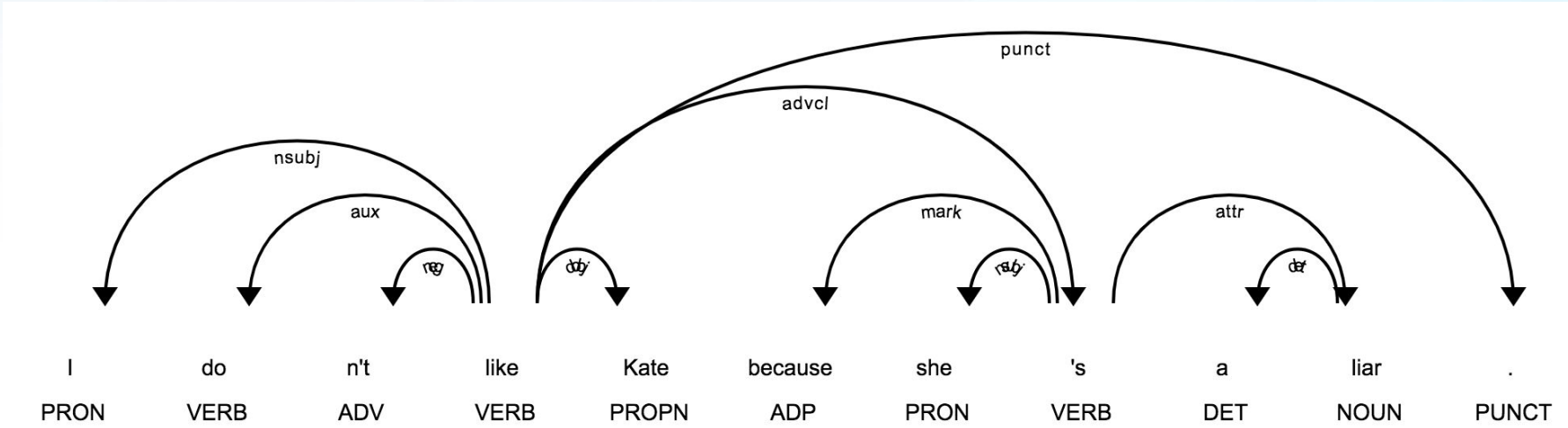
- Pattern matching
- Structural linguistics
- Linguistic ambiguities

For example

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

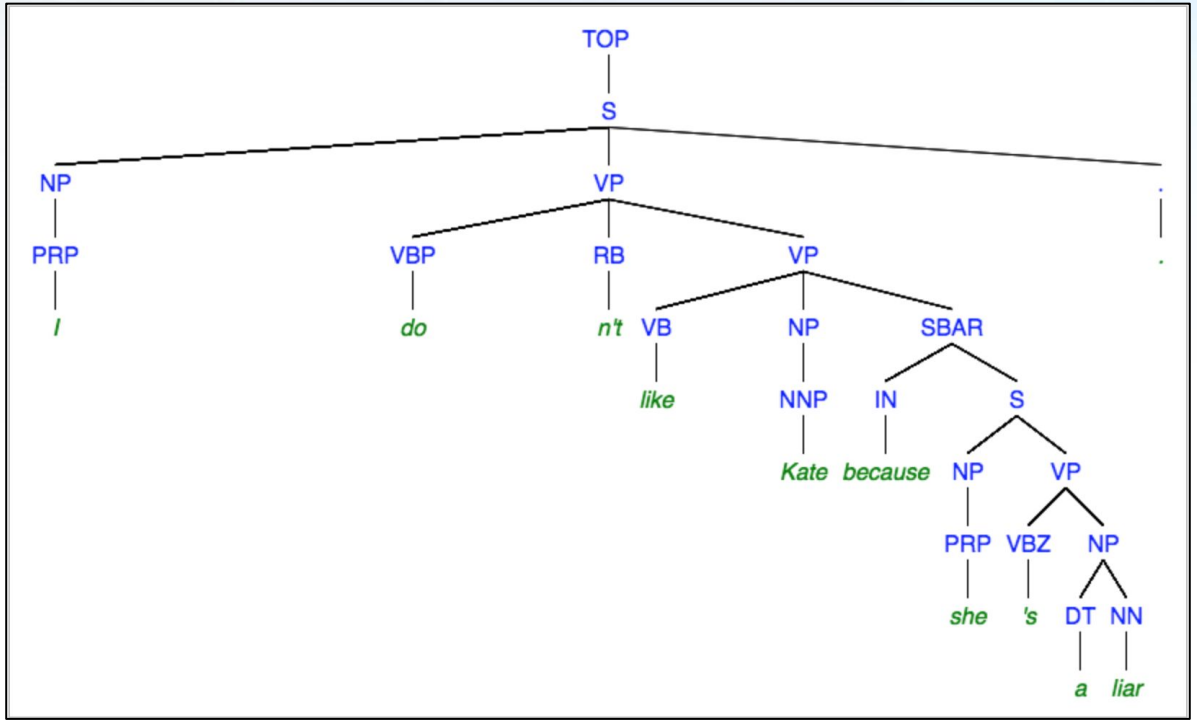
For example

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

Text processing

- Language identification
- Segmentation
- Normalization
- Part-of-speech tagging
- Syntactic analysis
- Named-entity recognition
- Coreference resolution
- Statistical analysis
- Lexical analysis
- **And many-many more features...**

Language identification

Text	Language	Explanation
Justin Bieber <3	und (Undefined)	NOT English; contains only a name.
Schalke XI v Chelsea: Fahrmann, Neustadter, Santana, Howedes, Uchida, Fuchs, Kirchhoff, Boateng, Hoger, Choupo-Moting, Huntelaar.	und (Undefined)	Contains only place/team/player names.
Ate spaghetti at La tratoria napolitana	en (English)	The name of the restaurant is in Italian, but the "main" language is English. An English-only speaker would understand this Tweet.
#NowListening Universo - Lodovica Comello @XYZ @XYZ	und (Undefined)	Italian song title and artist are just names. #NowListening is English but could be used by non-English speaker too.
#My #hot #naughty #neighbour #in #dallas: http://t.co/0dLJ 北京	en (English)	There is a Chinese word at the end, but the strongly prevailing language is English
Hahaha (•_•) (•_•)>⌐■-■ (⌐■_■) YEAHHH!	und (Undefined)	Emoticons and interjections only.
Que bonito!	und (Undefined)	Could be both Spanish and Portuguese
Pozor pozor	und (Undefined)	Could be Czech, Serbian, Croatian, Slovenian, ...
So warm in Berlin!	und (Undefined)	A valid sentence in both German and English
"Last Christmas" - Der Jose Carreras unter den Weihnachtsliedern.	de (German)	Contains an English song title and Spanish name, but is understandable to a German-only speaker.
Bécs <3	hu (Hungarian)	This is the Hungarian name for "Vienna", which is a proper name, but exists only in Hungarian
Estoy muy cansado voy a acostarme .... sooo tired goin to bedd	und (Undefined)	Strong mixture of Spanish and English, no clear "main" language

Linguistic ambiguities

At every level:

- phonetics

Linguistic ambiguities

At every level:

- phonetics
- morphology

an *un·ion·ized*

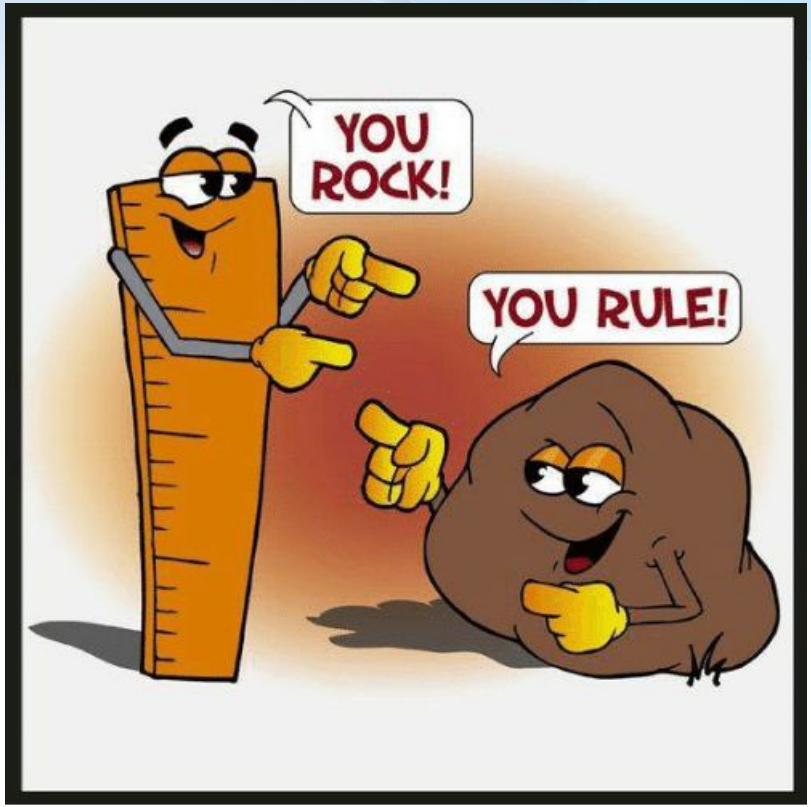
vs.

a *union·ized*

Linguistic ambiguities

At every level:

- phonetics
- morphology
- parts of speech

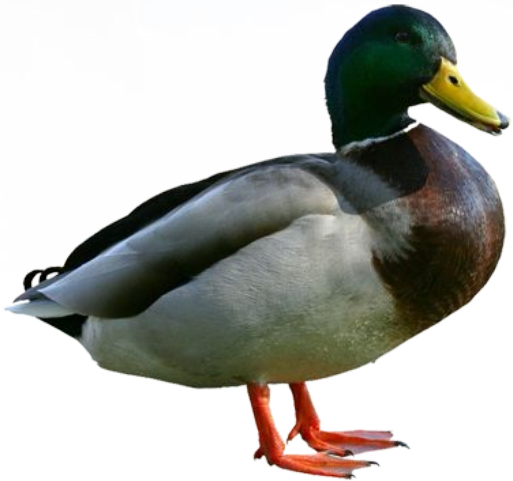


Linguistic ambiguities

I saw her duck.

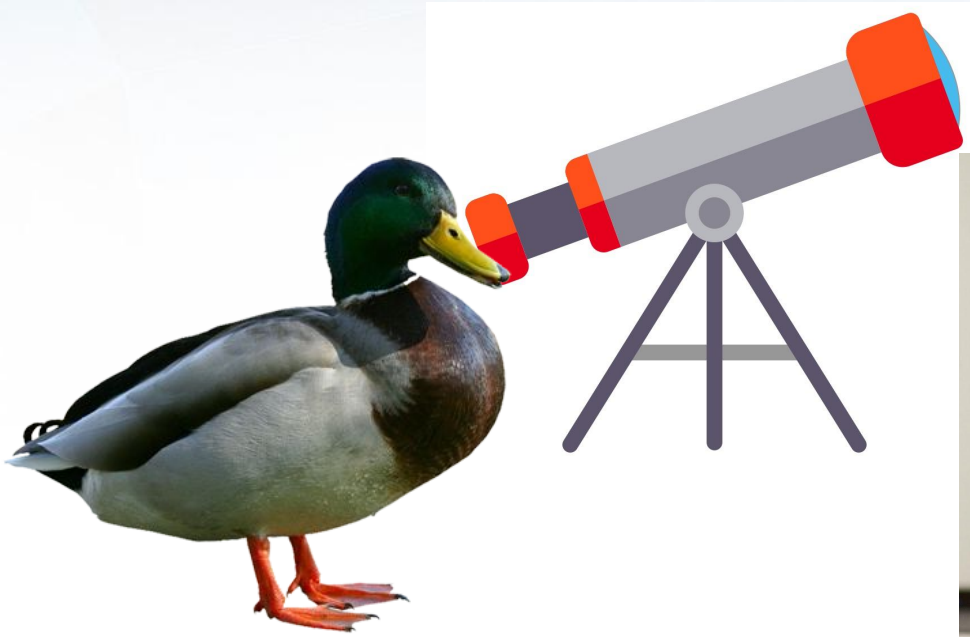
Linguistic ambiguities

I saw her duck.



Linguistic ambiguities

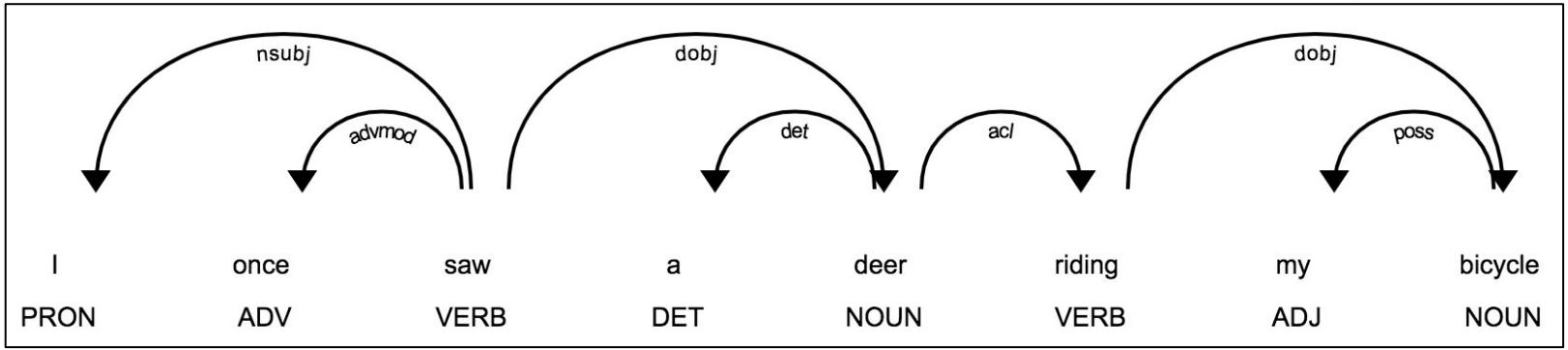
I saw her duck with a telescope.



Linguistic ambiguities

At every level:

- phonetics
- morphology
- parts of speech
- syntax



Linguistic ambiguities

At every level:

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

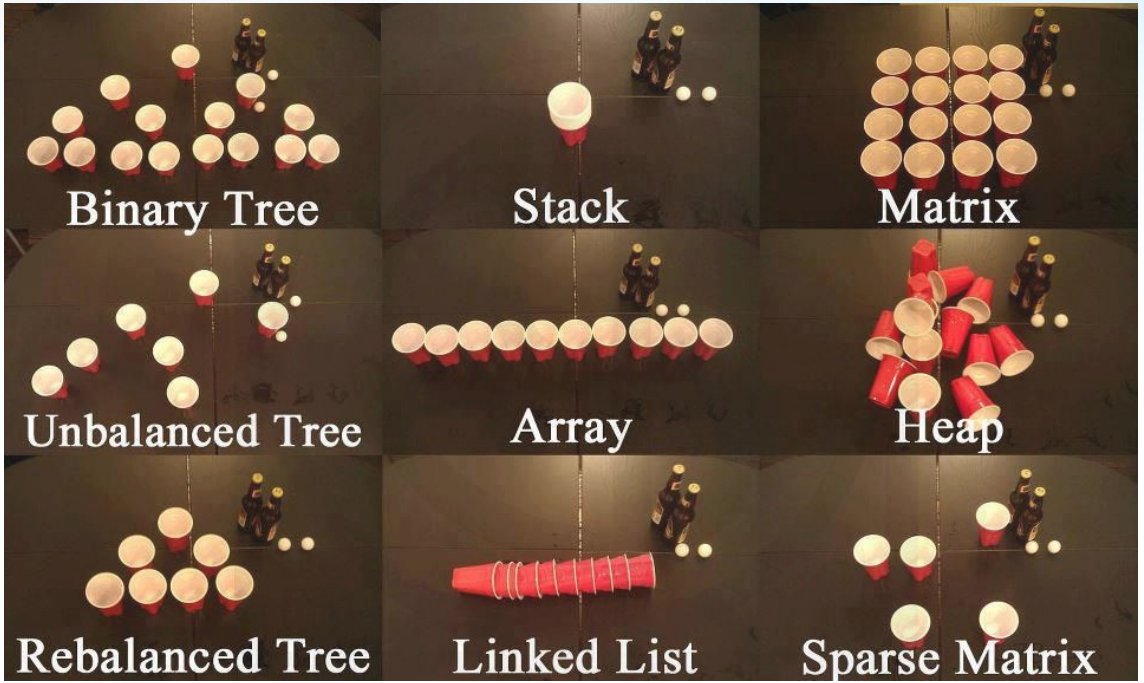
Linguistic ambiguities

At every level:

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

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

3. Practical Examples

Questions?