

# Intro to Natural Language Processing

Mariana Romanyshyn, *Grammarly, Inc.*  
Vsevolod Dyomkin, *Franz Inc.*

# Contents

1. Overview of NLP
2. NLP applications in our world
3. Real-life projects by Vsevolod Dyomkin
4. Real-life projects by Mariana Romanyshyn

# WTF NLP or NLP FTW?

# The Goal of NLP

**Goal:**

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

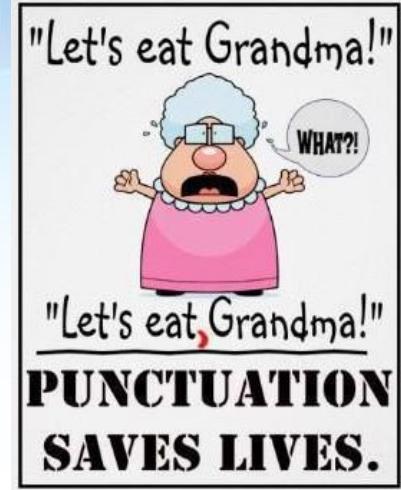
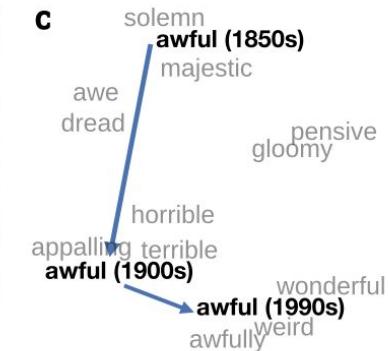
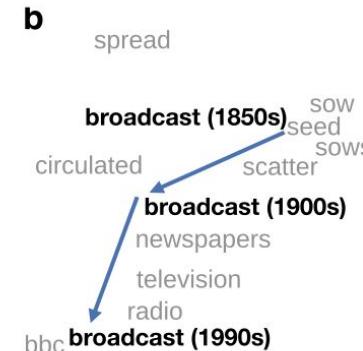
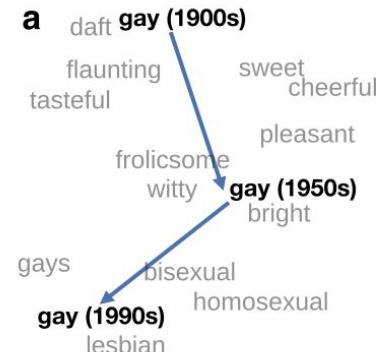
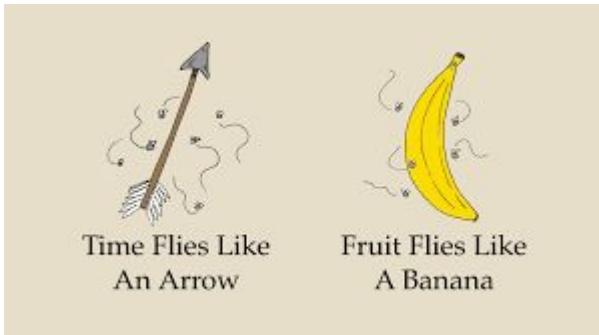
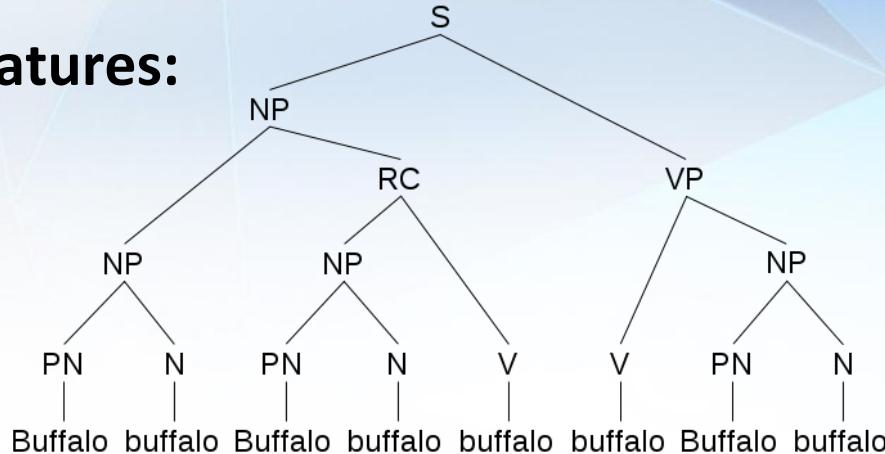
**How:**

transform free-form text into structured data and back

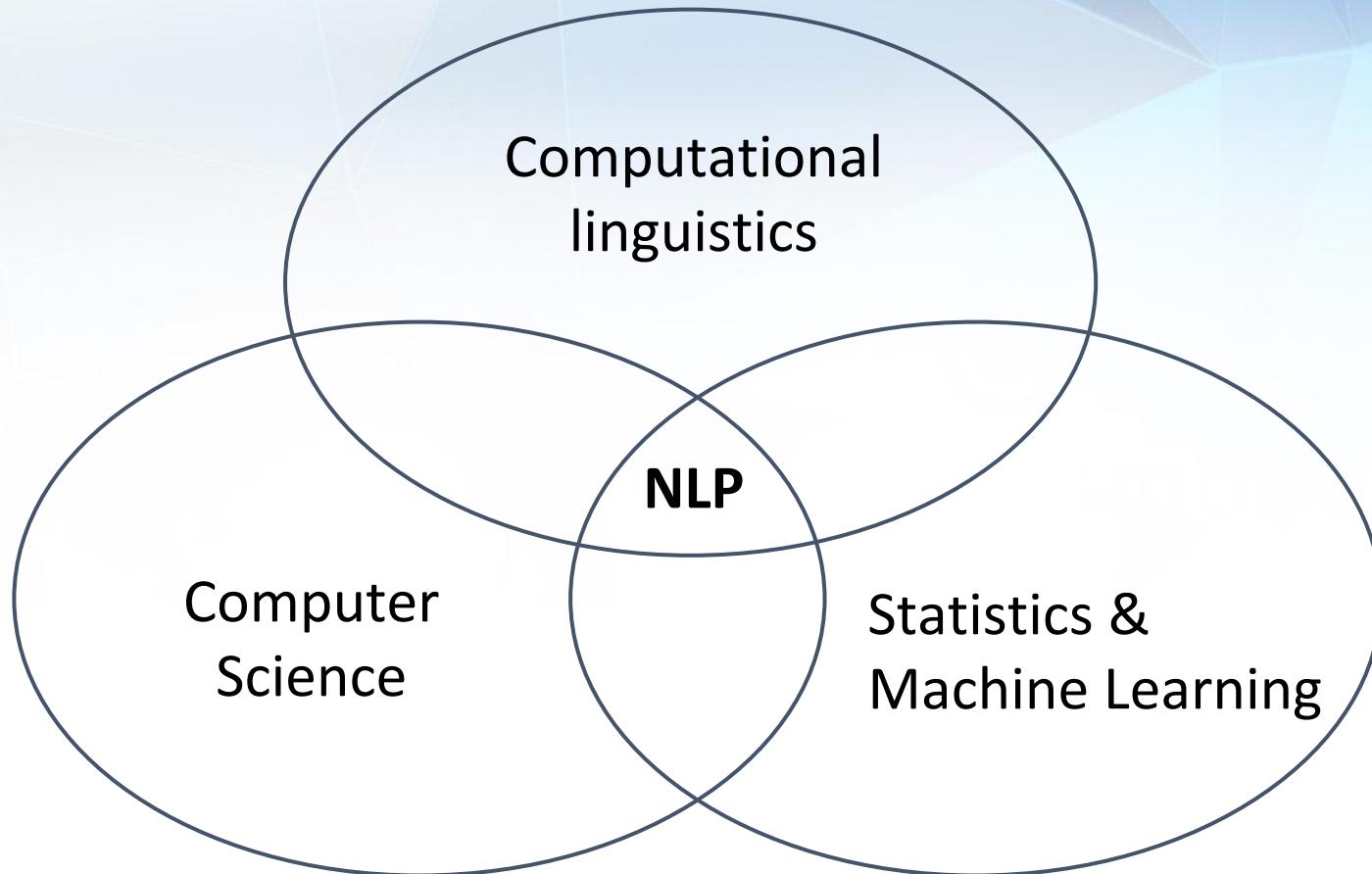
# Natural Language

## Distinguishing features:

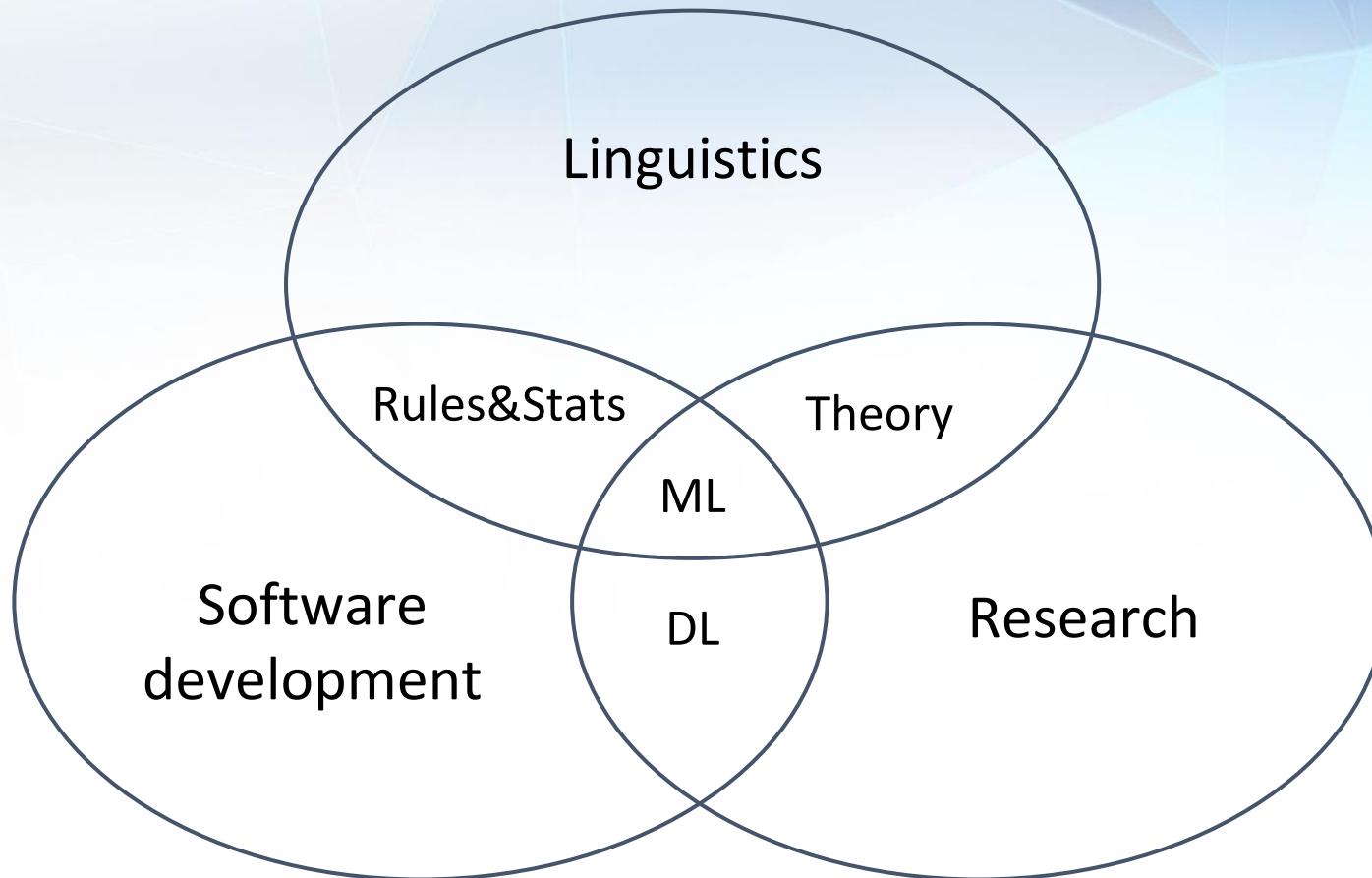
- Ambiguous
- Noisy
- Evolving



# Position of NLP



# Expertise in NLP



# NLP & AI

CV vs NLP

NLP vs NLU

Are we there yet? - <https://www.eff.org/ai/metrics>

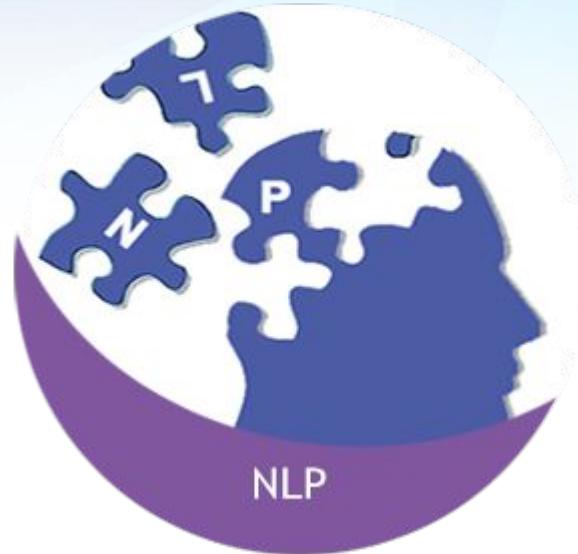
# NLP applications in our world

**Q: What NLP applications do you know?**

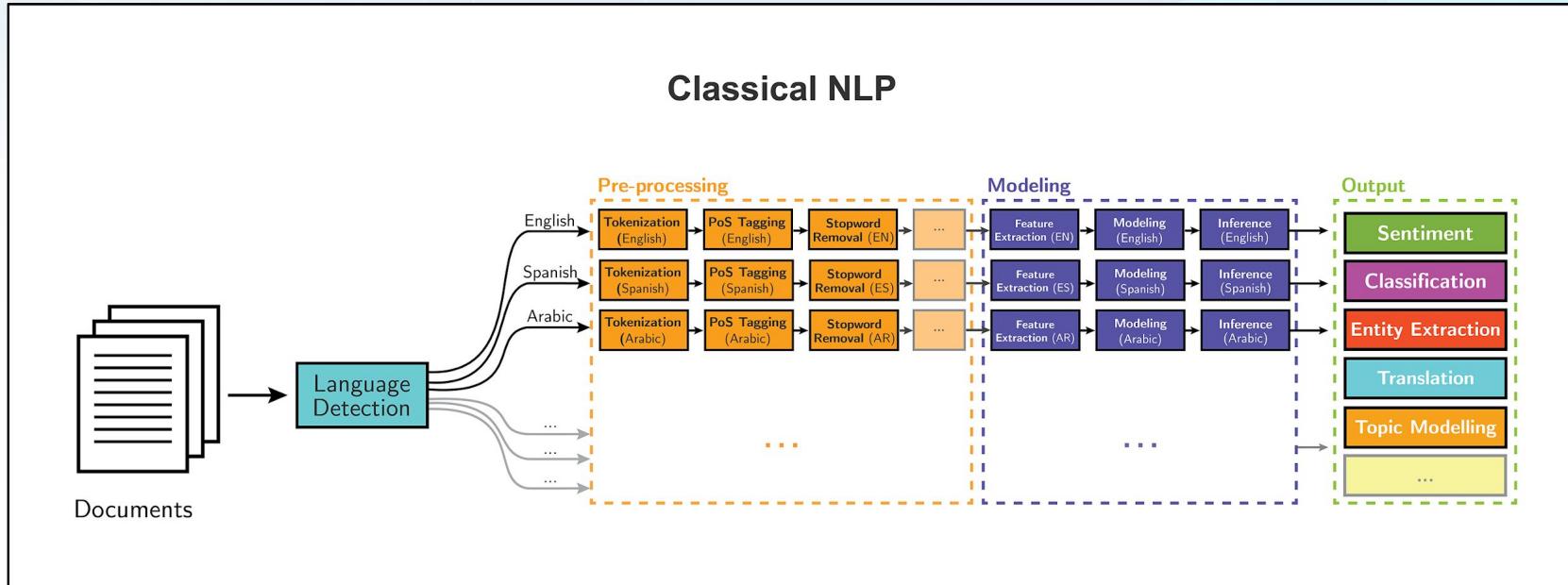
**A: [https://github.com/Kyubyong/nlp\\_tasks](https://github.com/Kyubyong/nlp_tasks)**

# Types of NLP Applications

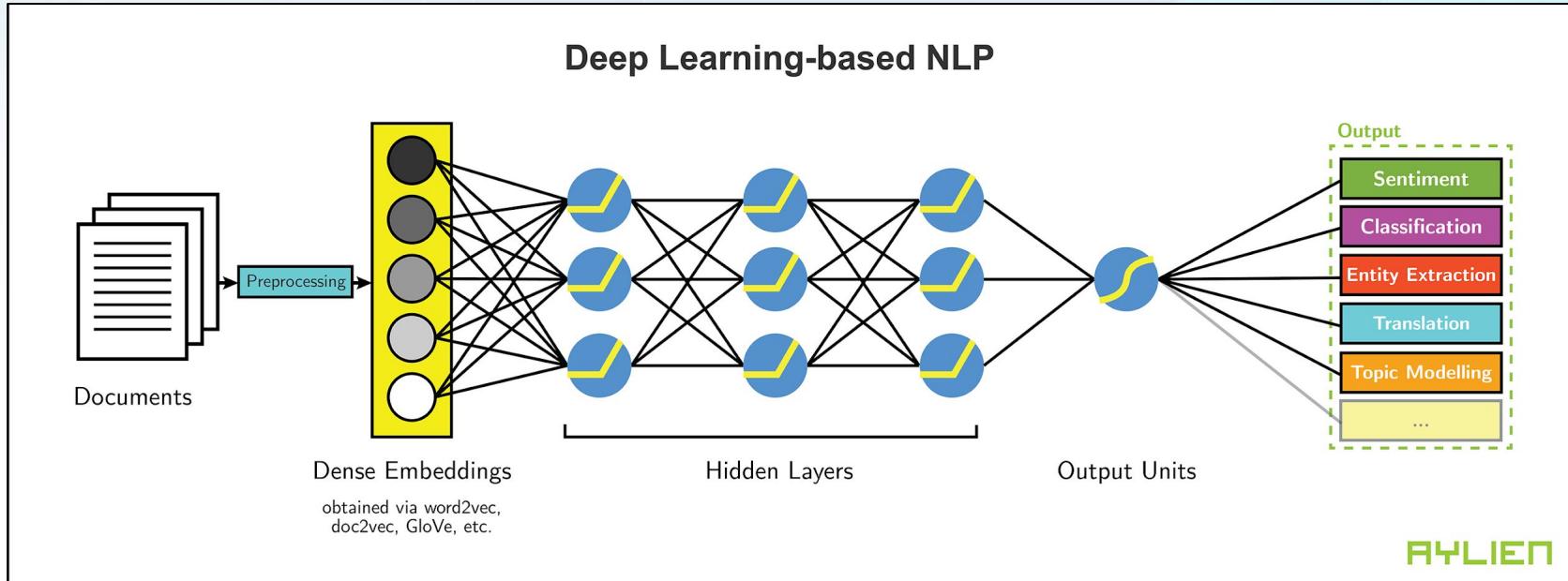
- Linguistic
- Analysis
- Transformation
- Generation



# An NLP Pipeline



# An NLP Pipeline



# Types of NLP Applications

## Linguistic tasks

- Segmentation
- Part of speech tagging
- Named-entity recognition
- Relation extraction
- Syntactic parsing
- Coreference resolution
- Semantic parsing
- ...

# Types of NLP Applications

## ANALYSIS

Spam Filtering

...



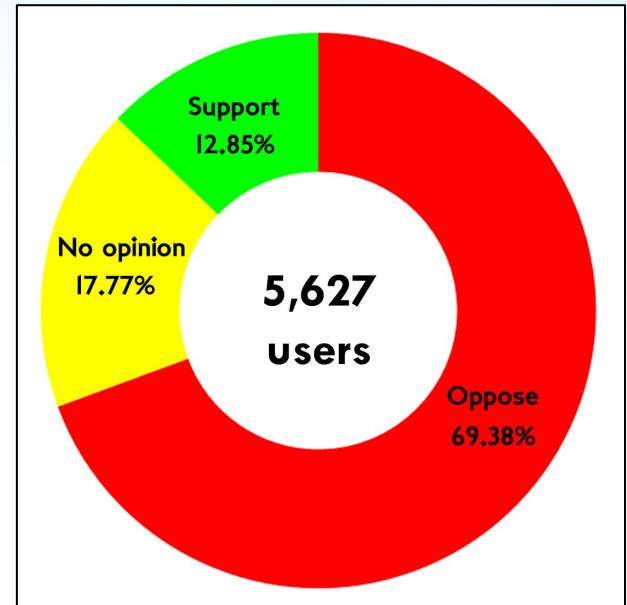
# Types of NLP Applications

## ANALYSIS

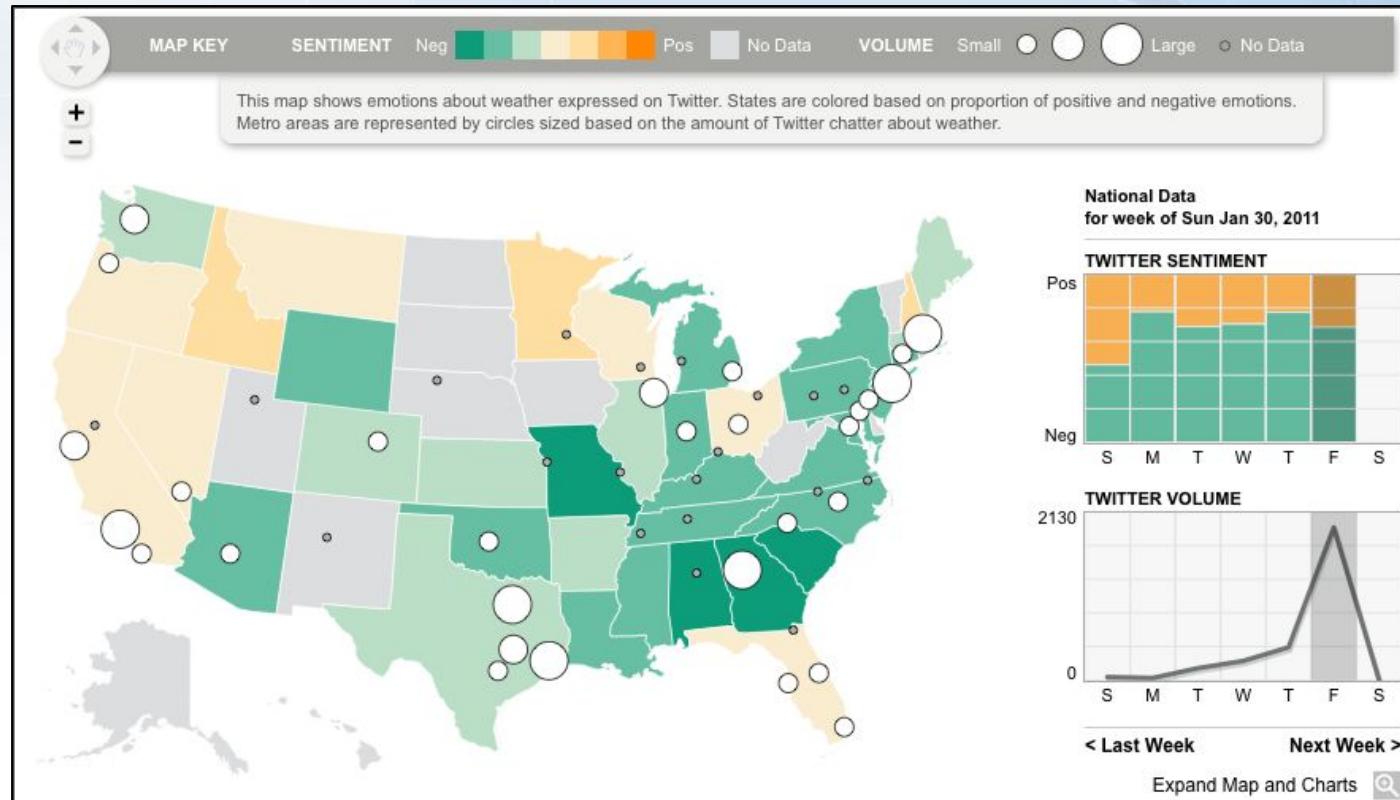
Spam Filtering

Sentiment Analysis

...



# Sentiment maps



# Sentiment Analysis

It tastes amazing!

It tastes horrible!

It tastes OK.

Cola tastes much better than Pepsi.



# Sentiment Analysis

It tastes amazing!

It tastes horrible!

It tastes OK.

Cola tastes much better than Pepsi.

It tastes like beer!

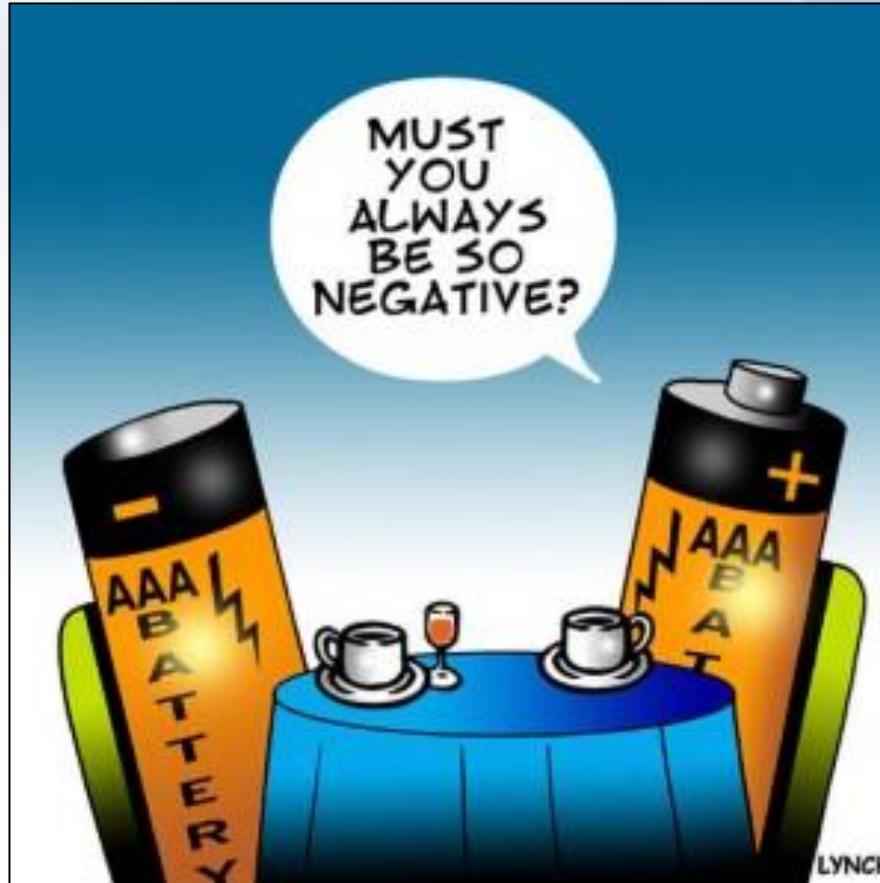
It tastes interesting!

It tastes like my mom said it would!

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



# Sentiment Analysis



# Types of NLP Applications

## ANALYSIS

Spam Filtering

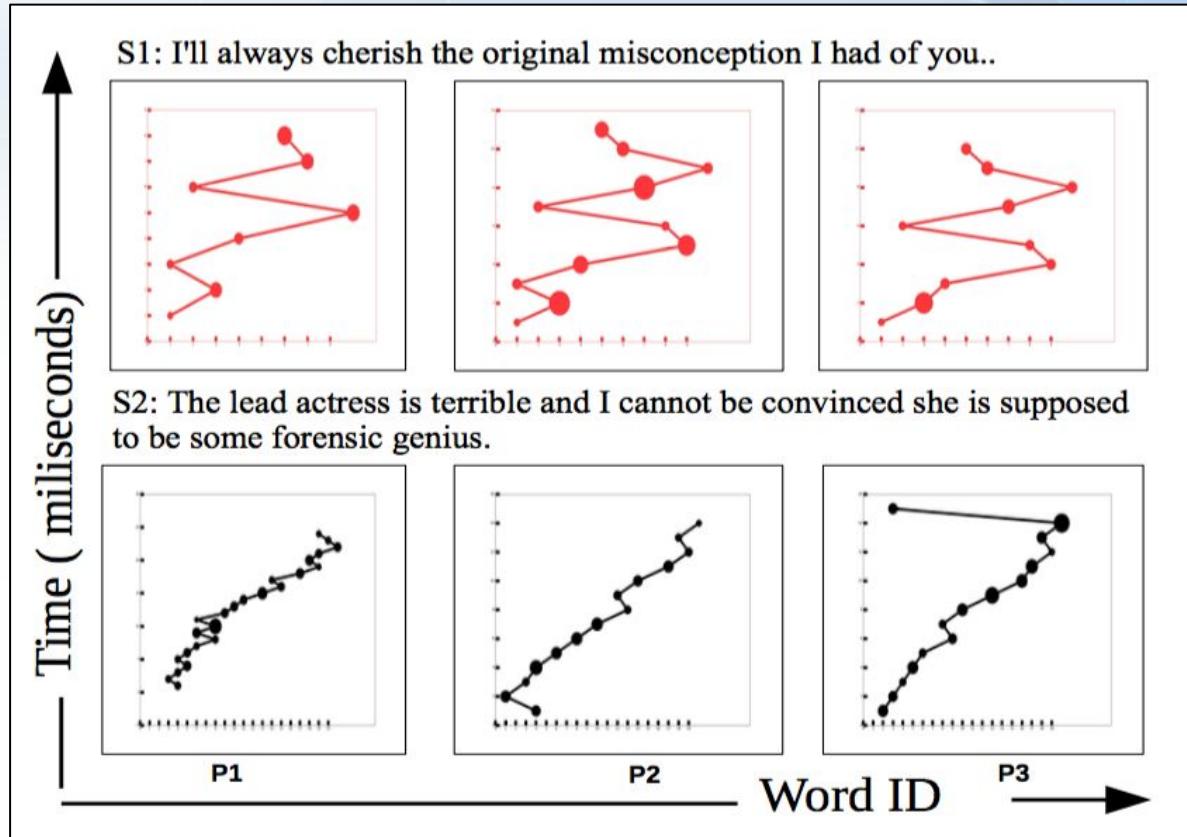
Sentiment Analysis

Sarcasm Detection

...

ME?  
SARCASTIC?  
NEVER.

# Cognitive features



# Types of NLP Applications

## ANALYSIS

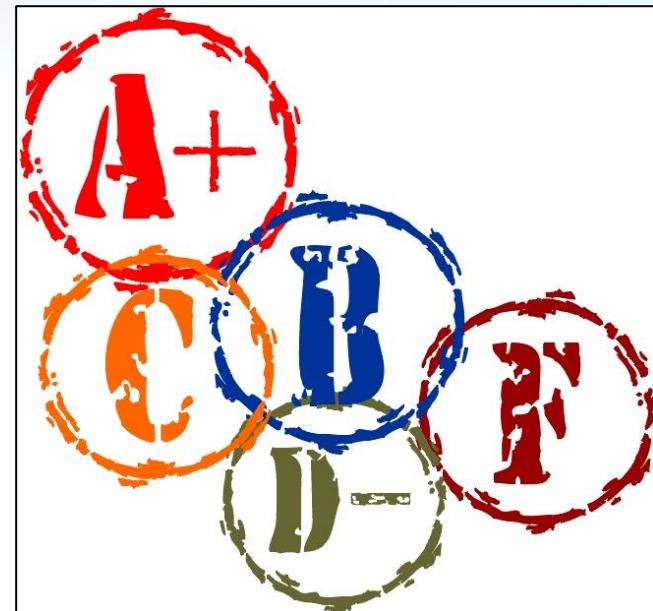
Spam Filtering

Sentiment Analysis

Sarcasm Detection

Essay Grading

...



# Types of NLP Applications

## ANALYSIS

Spam Filtering

Sentiment Analysis

Sarcasm Detection

Essay Grading

Good/Evil Characters

...



# Phonological features

Most frequent in positive characters	
Phoneme	Examples
<b>n-gram</b>	
/lɪ/	Ned Alleyn ( <i>Shakespeare in Love</i> )
/an/	Anouk Rocher ( <i>Chocolat</i> )
/aɪ/	Eliza Doolittle ( <i>My Fair Lady</i> )
/nɪ/	Linguini ( <i>Ratatouille</i> )
/ɪst/	Kevin McCallister ( <i>Home Alone</i> )
/ər/	Frodo ( <i>The Lord of the Rings</i> )
/and/	Dylan Sanders ( <i>Charlie's Angels</i> )
/stə/	C.C. Baxter ( <i>The Apartment</i> )

Most frequent in negative characters	
Phoneme	Examples
<b>n-gram</b>	
/ən/	Tom Buchanan ( <i>The Great Gatsby</i> )
/əv/	Iago ( <i>Aladdin</i> )
/tə/	Norrington ( <i>Pirates of the Caribbean</i> )
/ɪ/	Tom Ripley ( <i>The Talented Mr. Ripley</i> )
/mən/	Norman Bates ( <i>Psycho</i> )
/mɪs/	Mystique ( <i>X-Men</i> )
/ktə/	Hannibal Lecter ( <i>Hannibal</i> )

# 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

...

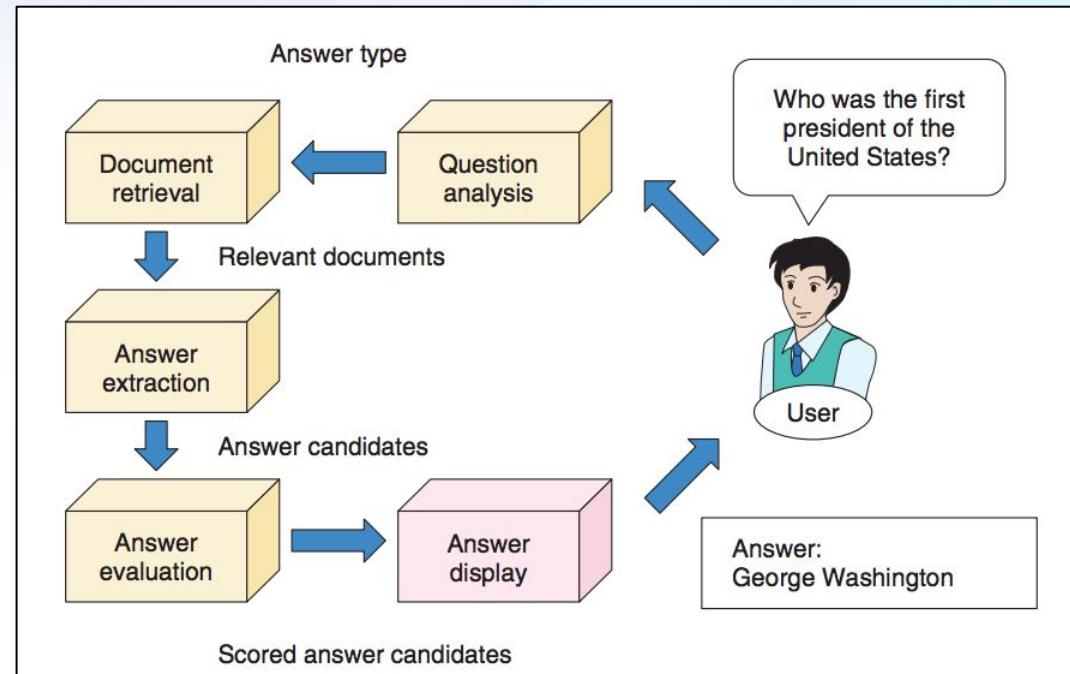


# Types of NLP Applications

## Question Answering

...

### GENERATION



# Types of queries

- **Factoid:** Who discovered America?
- **Yes/No:** Is Berlin the capital of Germany?
- **Definition:** What is leukemia?
- **Cause/consequence:** Why did the Iraq war start?
- **Procedural:** Which are the steps for getting a Master degree?
- **Comparative:** What is the difference between model A and model B?
- **Queries with examples:** What hard disks are similar to hard disk X?
- **Queries about opinion:** What is the opinion of the majority of Americans about the Iraq war?

# Types of NLP Applications

## Generation

Question Answering

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

Question Answering

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

Question Answering

Conversational Agents

Story Cloze Task

Computer-generated Fiction

...

## Generation

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.

# Types of NLP Applications

## Language Learning

The image displays three side-by-side screenshots of a mobile application designed for language learning, illustrating various NLP applications.

**Screenshot 1: Speaking Practice**

Top bar: Quit, four gray heart icons, progress bar (green).

Text: Speak this sentence

Text: L'eau est froide.

Icon: Microphone inside a blue circle.

Text: I can't use a microphone right now

Buttons: Check (gray), Continue (green)

**Screenshot 2: Translation Practice**

Top bar: Quit, one red heart, three gray hearts, progress bar (green).

Text: Translate this sentence

Text: Elle a une veste.

Text: She has a jacket

Icon: Microphone inside a blue circle.

Text: You are correct

Buttons: Continue (green)

**Screenshot 3: Listening Practice**

Top bar: Quit, three red hearts, progress bar (green).

Text: Type what you hear

Icon: Speaker (blue circle) and Headphones (blue circle).

Text: Dès qu'elle mange, je bois

Icon: Checkmark inside a green circle.

Text: Translation: As soon as she eats, I drink.

Buttons: Continue (green)

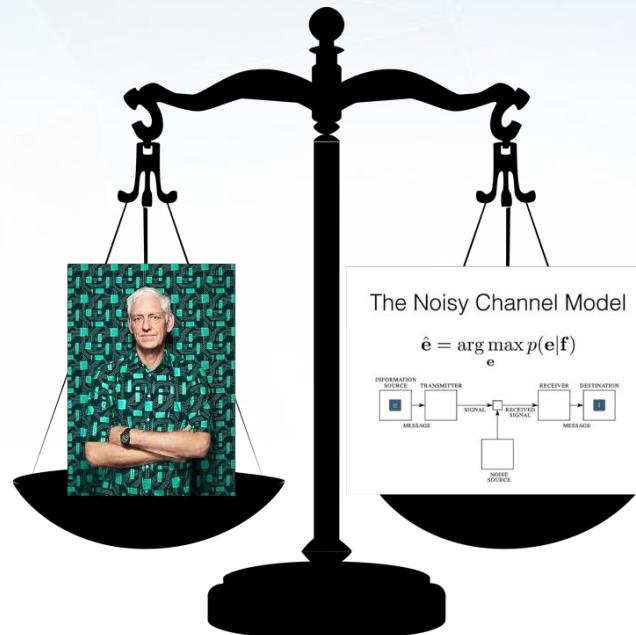
# Duolingo



**Real-life projects by  
Vsevolod Dyomkin**

# Error correction at Sciworth

## Spellchecker



## Error-correction framework

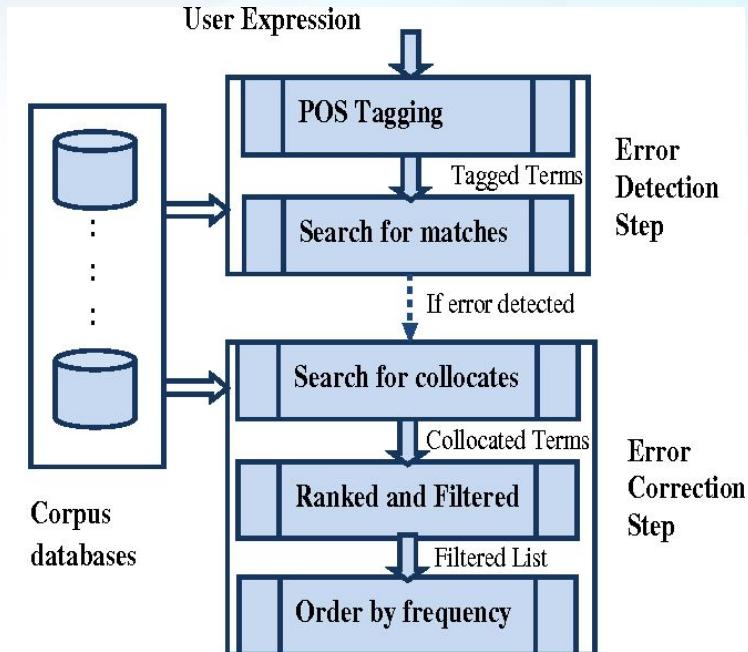


Figure 3: The CollOrder Framework

# Open-source libraries

- CL-NLP

```
> (pprint-tree '(TOP (S (NP (NN ))
  (VP (VBZ )
    (NP (DT )
      (JJ )
      (NN ))))
  (| . | <.:5 22..23>)))
```

TOP  
: :  
S  
: :  
VP  
: :  
NP  
: : NP  
: :  
NN VBZ DT JJ NN  
: : : : : :  
This is a simple test .

- WILD

<https://www.slideshare.net/vseloved/nlp-in-the-wild-or-building-a-system-for-text-language-identification>

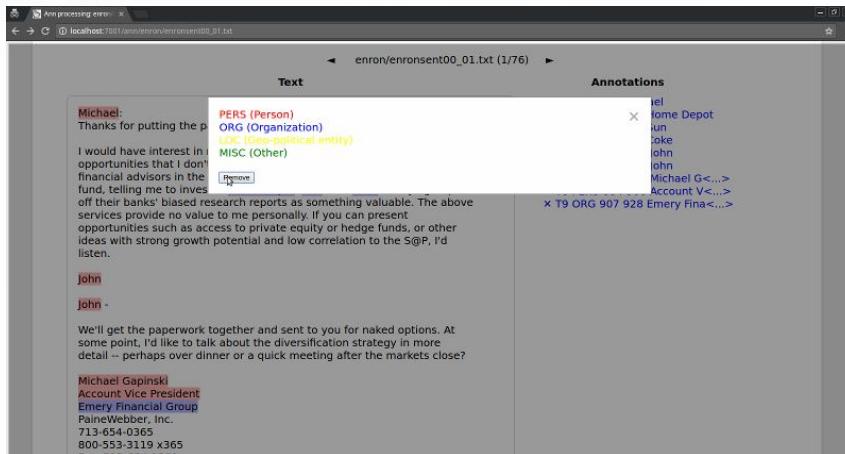
# Building Classifiers

- Email dissection
- Product catalog with 1k categories
- Identifying subjective statements

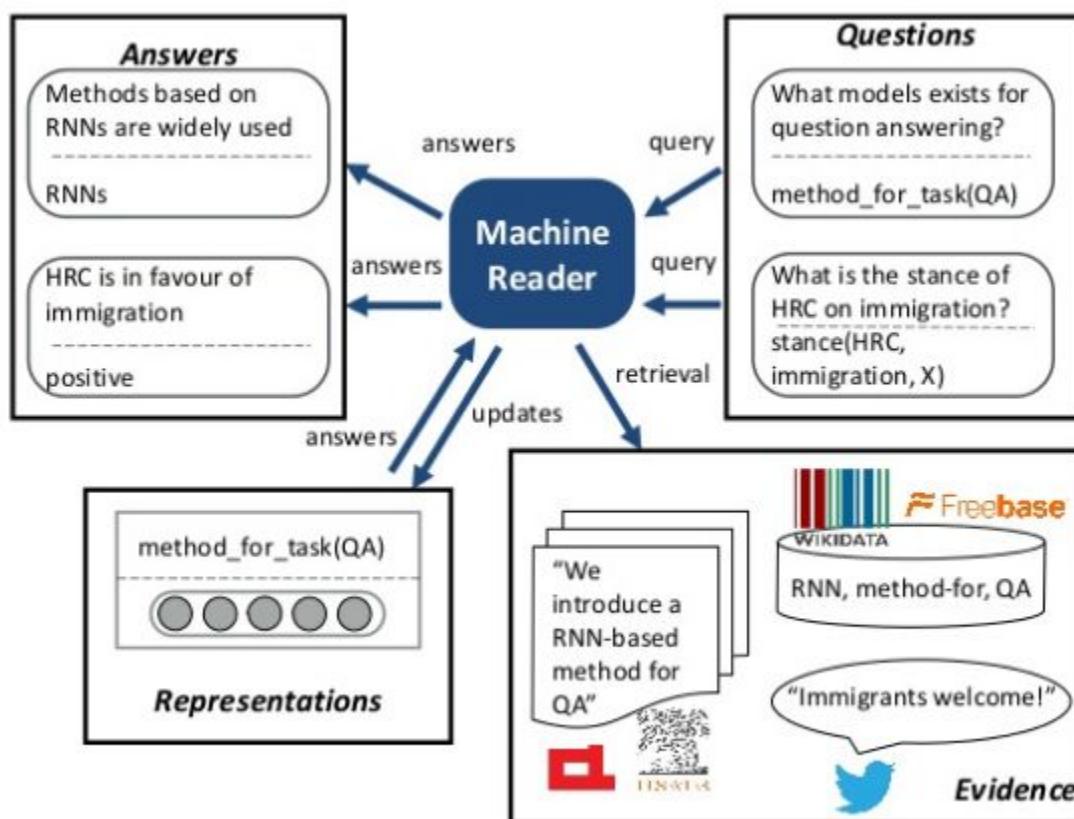


# Building Corpora

- Grammarly
- lang.org.ua
- fact-checking

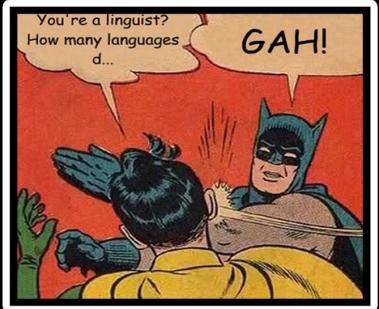


# Automated Fact-checking



**Real-life projects by  
Mariana Romanyshyn**

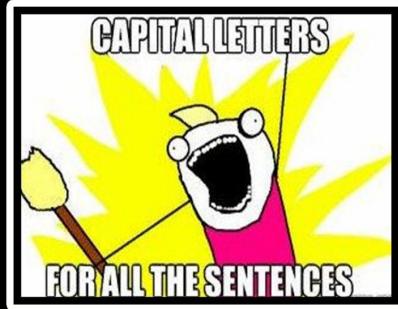
# COMPUTATIONAL LINGUIST



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(chnon.words(categories = "adventure"))
    fd = FreqDist(ngram_list)
    sentence = [first_word]
    while len(sentence) < 40:
        list_of_nexes = []
        for (i, j) in fd.keys():
            if i == first_word:
                list_of_nexes.append(j)
        if len(list_of_nexes) == 0:
            return "The sentence cannot be generated."
        first_word = random.choice(list_of_nexes)
        sentence.append(first_word)
        if first_word in [".", "?", "!", "..."]:
            break
    return " ".join(sentence)
```

WHAT I REALLY DO

# What we do

## 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

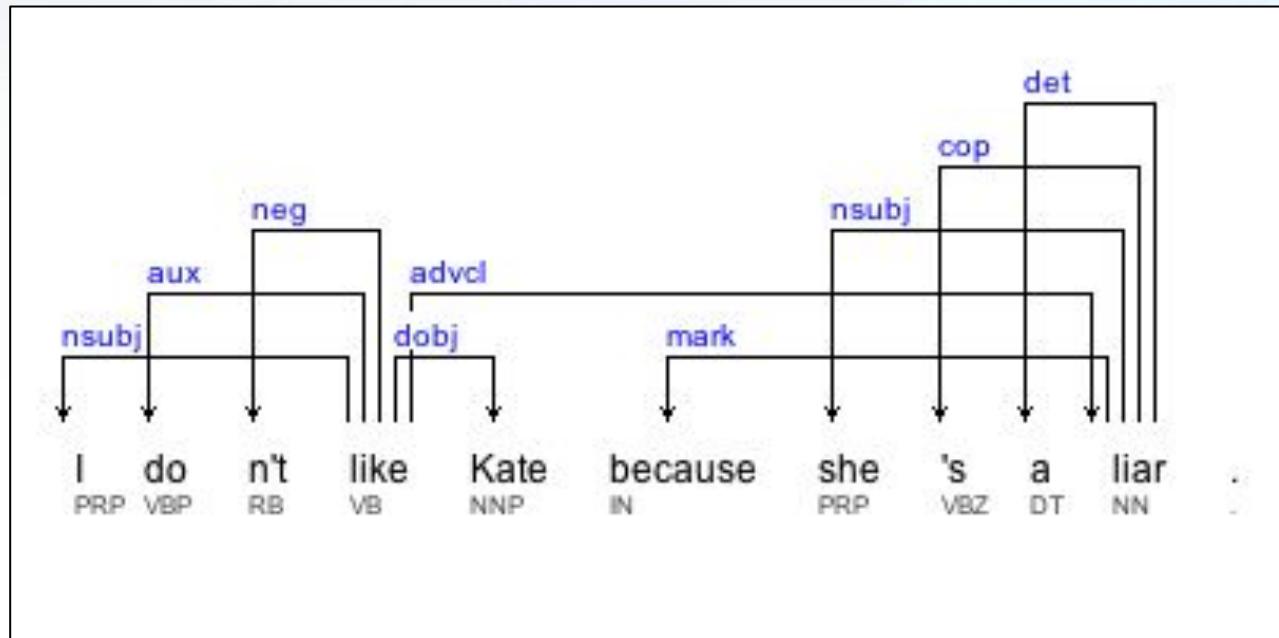
You see:

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



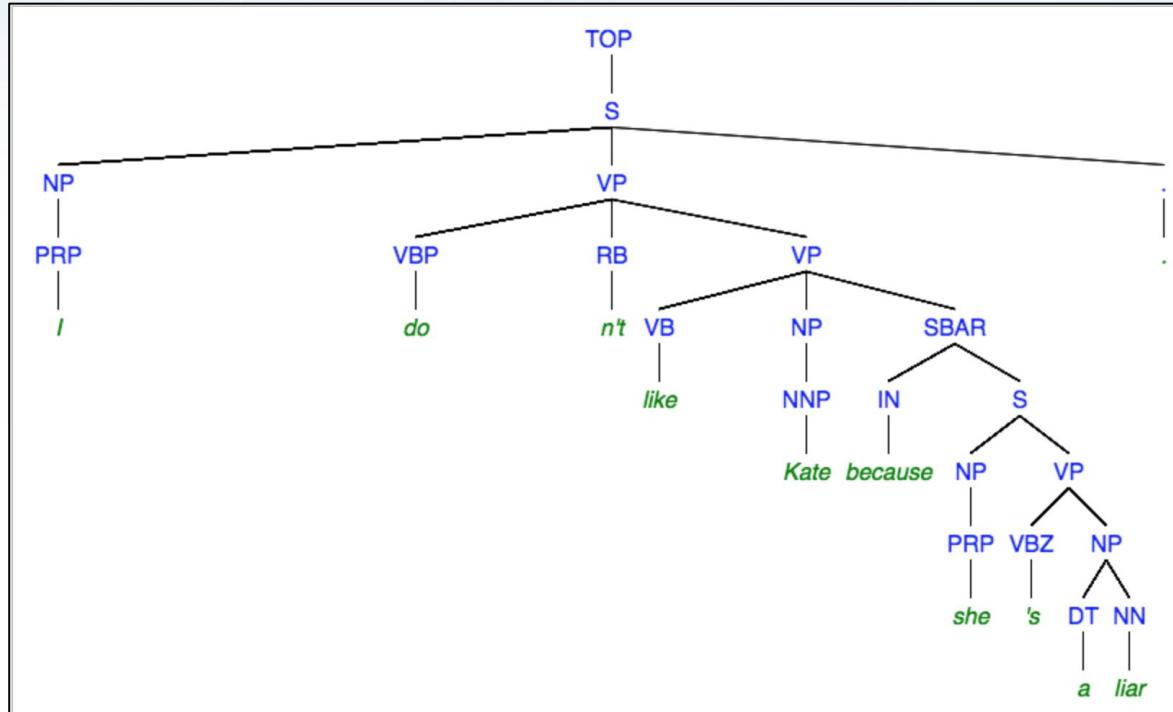
# For example

Linguists see:



# For example

Linguists see:



# For example

Linguists see:

```
(TOP
  (S  (NP  (PRP  "I" ) )
        (VP  (VBP  "do" )  (RB  "n't" )
              (VP  (VB  "like" )  (NP  (NNP  "Kate" ) )
                  (SBAR  (IN  "because" )
                        (S  (NP  (PRP  "she" ) )
                            (VP  (VBZ  "'s" )
                                (NP  (DT  "a" )  (NN  "liar" ) ) ) ) ) ) )
      ( | . |  ".") ) ) )
```

# For example

Linguists see:

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

# For example

Linguists see:

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



# For example

Linguists see:

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



**I** - agent

**Kate** - patient

# And many-many more features

- number of sentences, words, words per sentence, etc.
- size and arrangement of paragraphs
- word length
- word position in a sentence
- number of syllables in a word
- ratio of vowels vs consonants
- depth of the word in the dependency tree of the sentence
- number of word senses
- ngrams
- morphemes: stems and affixes
- is the word capitalized/hyphenated/compound?
- grammatical categories of different POS...

# 1. Text Grading

**Task:** match a text with a CEFR level.

- Vocabulary
  - Extract notional parts of speech
  - Classify by weighted frequency
- Grammar
  - Extract grammatical structures
  - Map them to CEFR levels



# 1. Text Grading



## 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.

### A2

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

### 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.

### 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.

### 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.

### 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.

Breakthrough!

Waystage

Threshold

Vantage

Effective  
operational  
proficiency

Mastery!

## 2. 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.



# 2. Text Mining

## RESTAURANT LE CHRISTINE

Le Christine est membre des Maîtres Restaurateurs de France, certifiant la fraîcheur et la qualité de tous les produits de notre carte et une réalisation entièrement faite maison.

## ARTICLES RÉCENTS

Le 14 février, l'amour s'installe au CHRISTINE  
Nouveau site Internet  
Bonne Année !!

## 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

## CONTACTEZ-NOUS !

📍 1 rue Christine, 75006 Paris

📞 +33 1 40 51 71 64

🍴 **RESERVER**

## 2. Text Mining

### RESTAURANT LE CHRISTINE

Le Christine est membre des Maîtres Restaurateurs de France, certifiant la fraîcheur et la qualité de tous les produits de notre carte et une réalisation entièrement faite maison.

### ARTICLES RÉCENTS

Le 14 février, l'amour s'installe au CHRISTINE  
Nouveau site Internet  
Bonne Année !!

### 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

### CONTACTEZ-NOUS !

📍 1 rue Christine, 75006 Paris

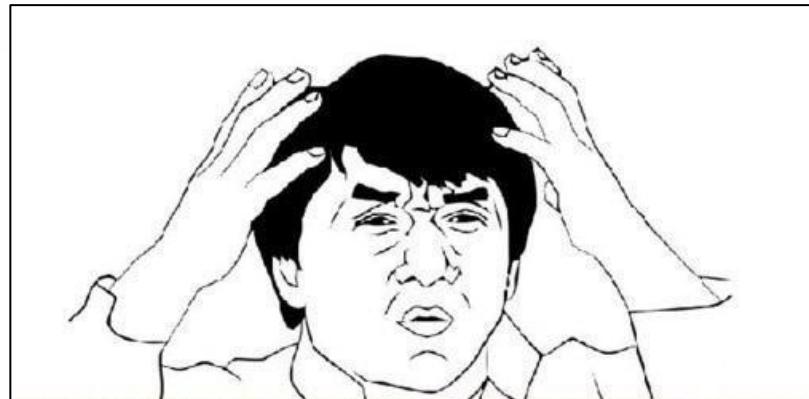
📞 +33 1 40 51 71 64

🍴 RESERVER

## 2. 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



### 3. Sentiment Analysis

**Task:** extract objects and sentiment they acquired.

- POS tagger
- Named-entity recognition
- Syntactic parser
- Coreference resolution
- Object-oriented sentiment analysis



### 3. Sentiment Analysis

TomTom dealt a major blow to TimTim.

### 3. Sentiment Analysis

TomTom dealt a major blow to TimTim.

TomTom dealt TimTim a major blow.

TimTim was dealt a major blow by TomTom.

A major blow was dealt to TimTim by TomTom.

## 4. Fact Extraction

**Task:** extract events about operational risks in financial companies

- Semantic roles
  - Agent
  - Patient
  - Sum of money
  - Date of settlement
  - Location
  - Reason for the loss
- Classification of operational risks
- Sequence of events
- Merge duplicate articles



# 4. 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.

# 4. 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.

## 5. Error correction

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.

## 5. Error correction

Task: make user texts **correct** and **clear**.

Spelling, Grammar, Punctuation

- *I cutted your finger didn't I?*
- *I cut your finger, didn't I?*
  
- *In daytime, he stayed in room.*
- *In the daytime, he stayed in the room.*

## 5. Error correction

Task: make user texts **correct** and **clear**.

### Style

- *There is a scarcity of food in the city.*
- *Food is scarce in the city.*
  
- *Could you go check what i've written???*
- *Could you check what I have written?*

## 5. Error correction

Task: make user texts **correct** and **clear**.

Paraphrasing

- *Joey came racing at a very fast speed.*
  - *Joey came racing at a breakneck speed.*
- 
- A very fast train runs through the city of Urumqi.
  - A high-speed train runs through the city of Urumqi.

# 5. Error correction

A documents can be found on my table.

Jim is a graduate students at the University of Kansas.

What are a dimensions in general?

If you want to learn about an arts, visit the national gallery.



A documents → Documents  
A document

The indefinite article A may not be required with the plural noun documents in this sentence. Consider removing the article, or changing the noun to singular.

▼ MORE

✗ IGNORE

students → student

a dimensions → dimensions

an arts → art

## 5. Error correction



She sawed a black cat in the room.



## 6. Data anonymization

Original:

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

## 6. Data anonymization

Original:

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

Anonymized:

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

# Etc

- Syntactic parsing
- Annotation tool
- Data annotation and data licensing
- Text analysis for Marketing (e.g., analysis of presidential debates)
- Text insights (e.g., readability, vocabulary variety)
- Spell checker for mobile language
- Grammarly cards
- Thesauri
- ...

## 7. brown-uk

- [dict-uk](#) - electronic dictionary of Ukrainian
  - 3,5 mln word forms
- [nlp-uk](#) - NLP API of LanguageTool for Ukrainian
  - segmentation
  - POS tagging
  - spelling, grammar, punctuation
- [corpus](#) - Brown corpus of Ukrainian
  - balanced by genres
  - balanced by regions

# Questions?

# Interesting References

- Peter Eckersley and Yomna Nasser, [Measuring the Progress of AI Research](#) (ongoing)
- Peter Norvig, [How to Write a Spelling Corrector](#) (2007)
- Mishra A. et al., [Harnessing Cognitive Features for Sarcasm Detection](#) (2016)
- Papantoniou K. and Konstantopoulos S., [Unravelling Names of Fictional Characters](#) (2016)
- Kyunghyun Cho, [Introduction to Neural Machine Translation with GPUs](#) (2015)
- Vaswani A. et al., [Attention is all you need](#) (2017)
- Deepmind, [WaveNet: A Generative Model for Raw Audio](#) (2016)
- Mostafazadeh N. et al., [A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories](#) (2016)
- Nick Montfort, [World Clock](#) (2013)