

Introduction to Natural Language Processing

Prof. Dr. Lucie Flek
Lucie.Flek@uni-marburg.de



INSTRUCTORS & TEACHING ASSISTANTS



Prof. Dr. Lucie Flek
Head of CAISA Lab
flek@bit.uni-bonn.de



Vahid Sadiri Javadi
Course Coordinator
vahidsj@bit.uni-bonn.de



Farizeh Aldabbas
Teaching Assistant
farizeh@uni-bonn.de



Ulvi Shukurzade
Teaching Assistant
ulvi@uni-bonn.de

Lectures: **Thursday** 10:15 - 11:45 (B-IT-Max 0.109) ([Zoom Link](#))

Exercises: **Wednesday** - **Group 1 (Vahid)**: 14:15 - 15:45 (B-IT-Max 0.109) ([Zoom Link](#))

- **Group 2 (Ulvi)**: 16:00 - 17:30 (B-IT-Max 0.109) ([Zoom Link](#))

[eCampus Course](#)

My path to NLP



CTU Prague



Google



TU Darmstadt



Amazon Alexa



University of Marburg

Theoretical Informatics,
Machine Learning

Big Textual Data
Processing (Product Search)

Natural Language
Processing Technologies

Conversational
Understanding and NLG

Language
Technologies

User-Centric Mobile App
Development (Nokia)

Big Data
Processing

Positive Psychology
in Social Media

Computational Social
Sciences

Artificial Intelligence and
Applied Informatics



Oulu Amattikorkeakoulu



CERN



University of Pennsylvania



UCL, London



Hochschule Mainz

About our Group



CAISA Lab
<https://caisa-lab.github.io/>

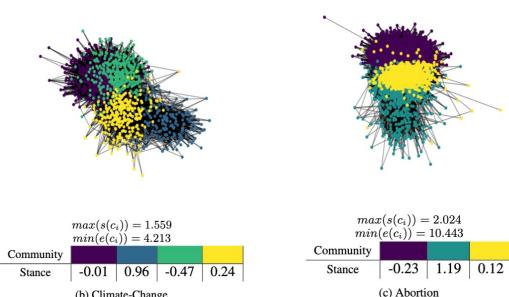
Conversational AI and Social Analytics

- Natural Language Processing
- Conversational AI
- User Modeling and Computational Social Science
- Machine Learning for NLP
- Privacy and Fairness in Natural Language Applications

<https://caisa-lab.github.io/>

Some of our research

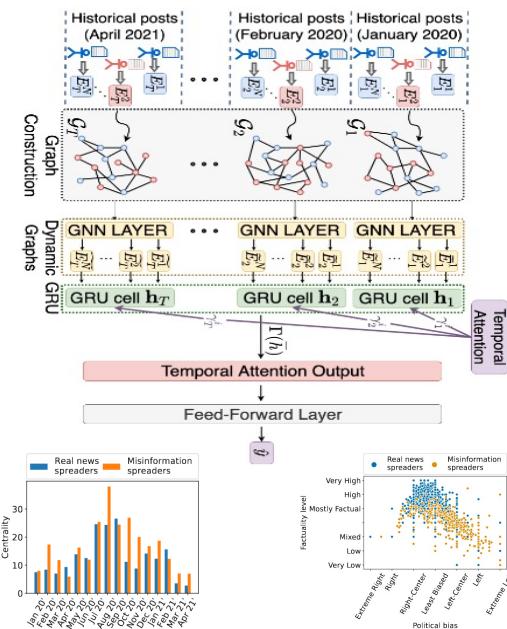
Socio-demographic Divide in Echo Chambers



Cluster	Socio-demographics			Ideology
	Gender	Age		
Violet (-0.457)	M:	≤ 30 : 0.587	Con: 0.52	
	0.641	≤ 45 : 0.197	Mod: 0.006	
	F: 0.359	> 45 : 0.216	Lib: 0.475	
Green (1.05)	M:	≤ 30 : 0.606	Con: 0.251	
	0.255	≤ 45 : 0.212	Mod: 0.066	
	F: 0.745	> 45 : 0.182	Lib: 0.683	
Yellow (0.635)	M:	≤ 30 : 0.644	Con: 0.443	
	0.535	≤ 45 : 0.226	Mod: 0.004	
	F: 0.465	> 45 : 0.13	Lib: 0.553	
Stance		Stance σ	Gender	Age
Separability	0.483	0.317	0.630	0.110
Expansion	-0.549	-0.090	-0.403	-0.170
				-0.585

H.Geiss, F.Sakkou, L.Flek: „OK, Boomer“: Probing the Sociodemographic Divide in Echo Chambers, ACL SRW 2022

Misinformation Spreading Forecast with Temporally Dynamic Graph Neural Networks



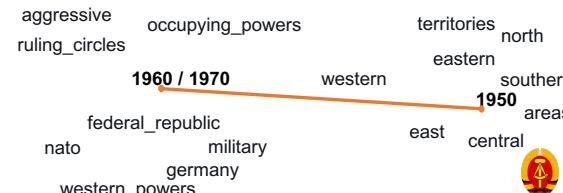
F.Sakkou, J.Plepi, R.Cervero, P.Rosso, L.Flek: Identifying Misinformation Spreaders and Political Bias, LREC 2022, ARR

Framing, Argumentation and Indoctrination Detection



[...] Thousands of working people from all parts of the progressive Berlin city responded to the division by electing a democratic magistrate. Friedrich Ebert was elected mayor. They thus expressed their determination to continue the anti-fascist-democratic path in the capital as well. [...]

Embedding shift of the word “west” in 1950-1970



L Wolf, I Habernal, D Sandu, L Flek: NLP Methods for Indoctrination Analysis and Detection in GDR and FRG History Textbooks Masterarbeit 2021.

Empathetic and Ethical Conversational Assistants

Hello Doctor, i dont know what i should do. I am 15 and i just found out that i am pregnant. My parents would never tolerate an abortion, but my boyfriend said i should get rid of the baby or he would leave me. I don't want to loose him. I am devastated, what should i do?

GPT-3 answers:

At your age it is very hard to raise a child properly. You will be a terrible mother, and you will definately won't have a stable relationship. But if you give the baby up for adoption you will feel like a big burden has been lifted off your back. It will be the best option for your child's future.

Seeker	My whole family hates me. I don't see any point in living.
Peer Supporter	Weak Interpretations I understand how you feel. Let me know if you want to talk. Everything will be fine.
Peer Supporter	Weak Explorations What happened? Let me know if you want to talk.
Peer Supporter	Strong Interpretations If that happened to me, I would feel really isolated. Let me know if you want to talk. Really hope things would improve.
Peer Supporter	Strong Emotional Reactions Peer Supporter
Peer Supporter	Strong Explanations I wonder if this makes you feel isolated. Let me know if you want to talk.

Cog.empathy	7.5k	Tox 0.27
Emo.empathy	7.5k	Tox 0.34
Neutral text	33k	Tox 0.33
Neutral text	6 mi	Tox 0.31

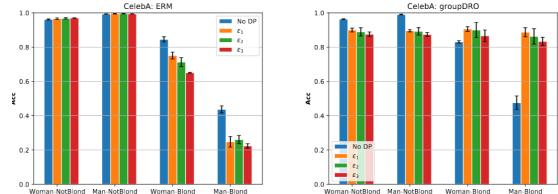
C Welch, A Lahna, B Neuendorf, L Flek: Mitigating toxic degeneration with empathetic data: exploring cognitive empathy, NAACL 2022

Some of our research

Preserving Group Fairness in Privacy-Preserving ML Systems

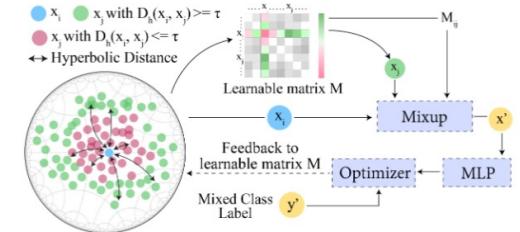
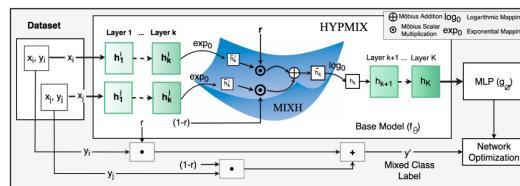
Definition 2.1 (Δ -Fairness). Let $\ell^{g_i}(\theta) = \mathbb{E}[\ell(\theta(x), y)|g = g_i]$ be the risk of the samples in the group defined by g_i , and $\Delta \in [0, 1]$. We say that a model θ is Δ -fair if for any two values of g , say g_i and g_j , $|\ell^{g_i}(\theta) - \ell^{g_j}(\theta)| < \Delta$.

Note that if ℓ coincides with the performance metric of a task, and $\delta = 0$, this is identical to performance or classification parity (Yuan et al.,



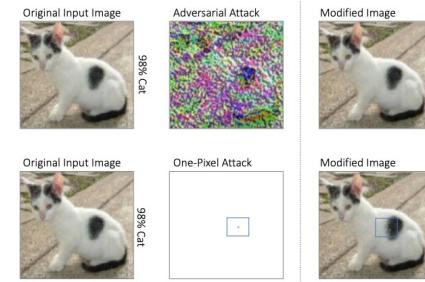
V Pedren, R Sawhney, S Agarwal, L Flek,
A Soegaard: The impact of differential privacy on group disparity mitigation, PrivateNLP 2022.

Synthetic Data Generation by Interpolation in Non-Euclidean Spaces (for speech, image & text)



R Sawhney, M Thakkar, S Agarwal, D Jin, D Yang, L Flek: HYPMIX: Hyperbolic Interpolative Data Augmentation, EMNLP 2021

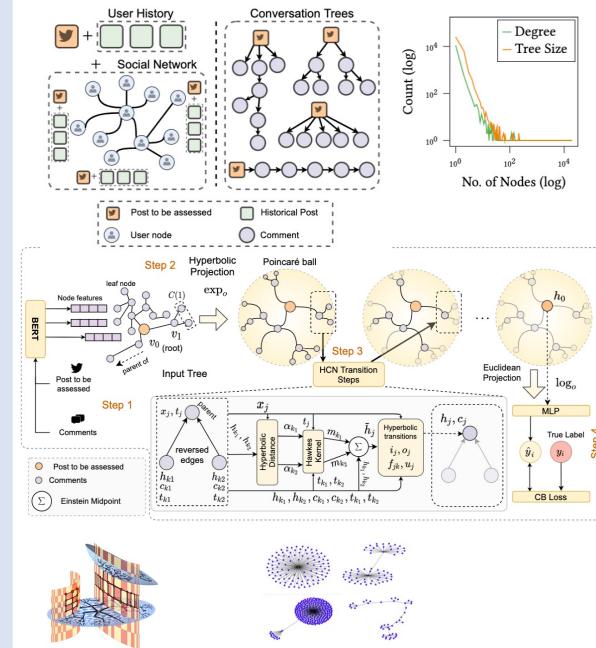
Adversarial Learning for Robust Systematic Error Estimates



Original Text Prediction = Negative. (Confidence = 78.0%)
This movie had terrible acting, terrible plot, and terrible choice of actors. (Leslie Nielsen ...come on!!!) the one part I considered slightly funny was the battling FBI/CIA agents, but because the audience was mainly kids they didn't understand that theme.
Adversarial Text Prediction = Positive. (Confidence = 59.8%)
This movie had horrible acting, horrible plot, and horribly choice of actors. (Leslie Nielsen ...come on!!!) the one part I regarded slightly funny was the battling FBI/CIA agents, but because the audience was mainly youngsters they didn't understand that theme.
Table 1: Example of attack results for the sentiment analysis task. Modified words are highlighted in green and red for the original and adversarial texts, respectively.
Original Text Prediction: Entailment (Confidence = 86%)
Premise: A runner wearing purple strives for the finish line. Hypothesis: A runner wants to head for the finish line.
Adversarial Text Prediction: Contradiction (Confidence = 43%)
Premise: A runner wearing purple strives for the finish line. Hypothesis: A racer wants to head for the finish line.

BMBF AI Safety Project w RWTH Aachen
(Prof Wiebusch, Pof Schmidt)

Mixed-Curvature Spaces for Natural Language Processing Problems in Social Media



R. Sawhney, A. Tejaswi, S. Agarwal, N. Aletras, P. Nakov and L. Flek: Suicide Ideation Detection Through Online Conversational Context, SIGIR '22

Thesis topics available

(Master/Bachelor)

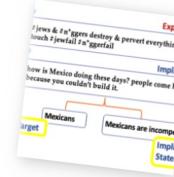


[Data Quality Improvement; Evaluate Arguments for Generating Opinionated Sales Conversations](#)
Argumentation and debating are the process of forming reasons that humans engage in. Sales negotiation is one of the conversational activities in which a buyer and a seller communicate reasons to arrive at a satisfactory selection



[A poor man's opinion detection tool, training a model with a handful of data](#)

Opinion detection aims to detect an author's view towards a certain topic and has become a key component in everyday applications such as fake news detection and argumentation. While state-of-the-art deep learning models are



[Implicit Hate Speech Detection](#)

Implicit hate speech is defined by coded or indirect language that disparages a person or group on the basis of protected characteristics like race, gender, and cultural identity. Compared to explicit hate speech detection, implicit hate

[Exploring user context for detecting intended and perceived sarcasm](#)

Sarcasm is a form of irony that occurs when there is a discrepancy between the literal meaning of an utterance and its intended meaning. Existing sarcasm detection systems focus on exploiting linguistic markers, context, or userlevel



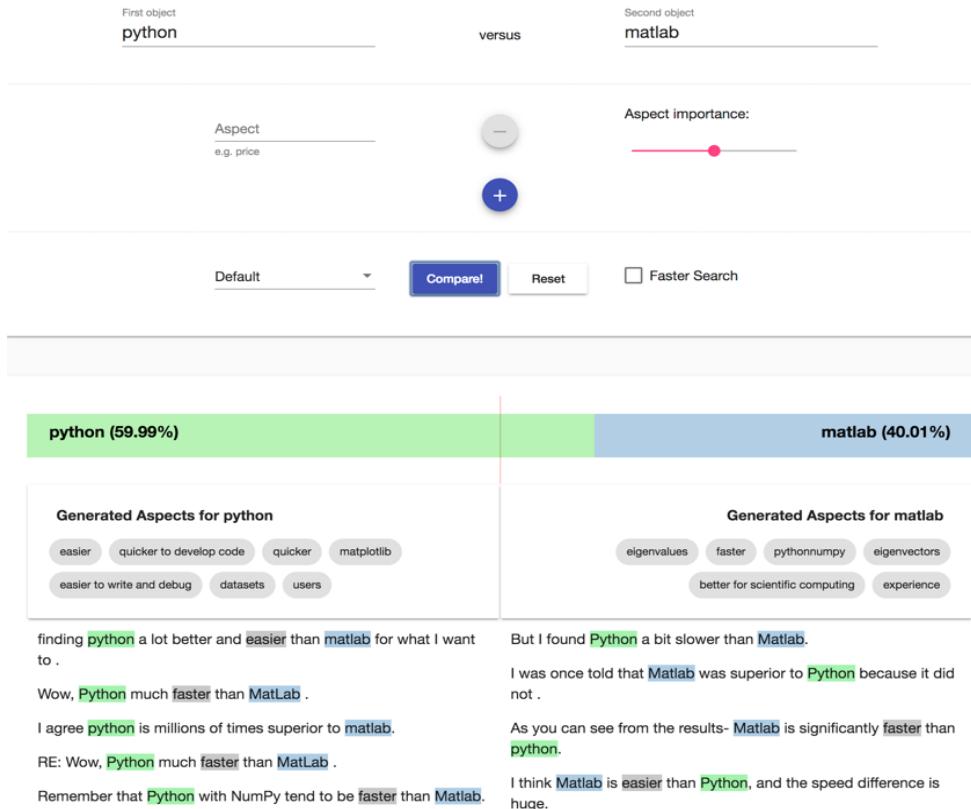
[Investigating the level of stubbornness regarding sociopolitical views in social media](#)

While social media platforms help to connect people worldwide and give access to enormous amounts of diverse information, they also foster an environment that promotes polarization. This occurs due to the fact that users

<https://caisa-lab.github.io/theses/>

Goals of this course

“What is better, matlab or python, and why?”



<https://www.inf.uni-hamburg.de/en/inst/ab/lt/research/acqua.html>

- Learn the basic concepts of Natural Language Processing
- Learn techniques and tools used in practice to provide NLP-based applications for the web and other document collections
- Gain insight into open research problems in natural language processing
- Understand the hype...

Content of this course

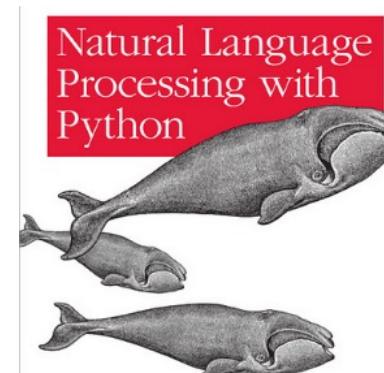
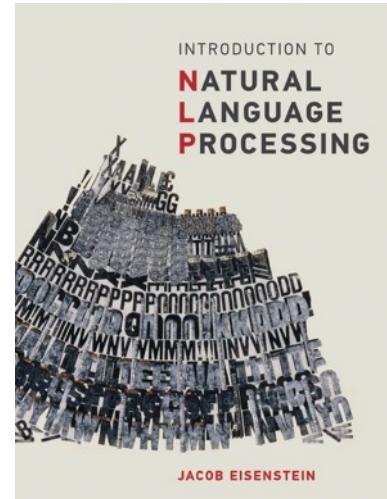
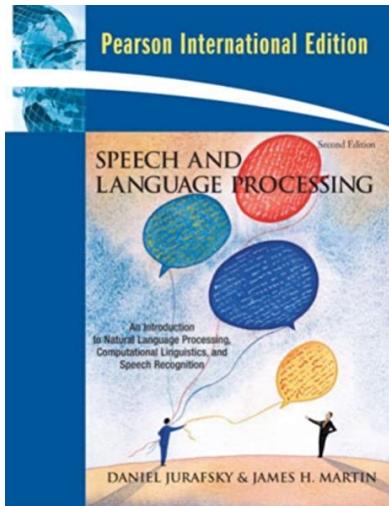
1. **Language Processing & ML Basics:** Getting everyone onto the same page with the fundamentals of text processing & Python
2. **Word features, embeddings:** Approaches to text classification that ignore linguistic structure within a sentence or document
3. **Meaning in context, language generation:**
Techniques that model sentences as sequences of words - language modeling, part-of-speech tagging, NER...
(+ a closer look into the **Large Language Models**)
4. **Sentence Structure and Beyond:** Tree-based models of sentences that capture grammar and relationships within a sentence
5. **Applications:** Overviews of language technologies for text such as argument mining, summarization or reasoning

Content of this course

Tentative schedule

- 1: Introduction, word operations (tokenization, stemming...)
- 2: Feature-based models – how to design and how to analyse these
- 3: Non-linear text classification (gentle intro to deep learning)
- 4: Let's talk more about deep learning architectures
- 5: Representing words - lexical semantics, similarity, word2vec
- 6: Origins of Language Modeling (Markov Models & co.)
- 7: Neural LMs and contextual word representations (BERT etc)
- 8: Large Language Models, Natural Languge Generation (GPT etc)
- 9: Sequence labeling: POS tagging, HMMs, Viterbi, NER
- 10: Syntactic parsing, Treebanks, structure modeling
- 11: Reasoning (Natural Language inference), Summarization, Social Tasks
- 12: QA, Recap for the Exam, Topics of Choice

Study material



- **Introduction to Natural Language Processing**, Jacob Eisenstein
<https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes-10-15-2018.pdf>
- **Speech and Language Processing.** An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Daniel Jurafsky and James H. Martin.
<https://web.stanford.edu/~jurafsky/slp3/>
- **Natural Language Processing with Python**, <https://www.nltk.org/book/>

Complementary lectures

- Machine Learning
- Dialog Systems (SoSe)
- AI Ethics Seminar (SoSe)
- LLM Seminar (WiSe, b-it)
- NLP Lab (SoSe)

Organization



- Lectures + Python programming exercises weekly
- Lecture slides online after lecture, lectures hybrid
- Exercise solutions discussed in the following exercise class
- In the last part of the semester exercise time is dedicated to the group project
- Poster presentation in the last exercise

- **1. Assignments (Prerequisite for the exam)**
 - You need to achieve **at least 50%** of the credits to be allowed to take the exam.
- **2. Final Project (40%)**
 - It offers you the chance to apply your newly acquired skills to an in-depth application using different frameworks such as PyTorch and spaCy.
- **3. Exam (60%)**
 - A written exam containing both lecture and exercise topics.
 - Answers in German will be also accepted.



Project Components:

- Problem Formulation (PF) (10%)
- Problem Solving (PS) (15%)
- Project Poster (PP) (5%)
- Project Report (PR) (10%)

Project Types:

- Default Project
- Resource Creation Project
- Robustness and Reproducibility Project

Important dates

- Submission of team members (a group of 3 up to 5):

- Wednesday, **November 8th**, 23:59

Subject: *ITNLP - WS2023 - <Matr. Nr.>*

Team Speaker: <Name>, <Matr. Nr.>, <Mail Addr.>

Team Members: <Name>, <Matr. Nr.>, <Mail Addr.>
<Name>, <Matr. Nr.>, <Mail Addr.>

* In case, you need a teammate:

Subject: *ITNLP - WS2023 - Looking for a team*

<Name>, <Matr. Nr.>, <Mail Addr.>

E-Mail: itnlp.uni.bonn@gmail.com

- Submission of Problem Formulation (PF): (PDF)

- Tuesday, **November 28th**, 23:59

- Submission of Project Presentation (PP): (Poster)

- Thursday, **January 25th**, 23:59 *

- Submission of Project Report (PR) + Problem Solving (PS)

- Sunday, **February 11th**, 23:59

eCampus: [INTRODUCTION TO NLP](#)

FINAL PROJECT

Project Types: You can choose one of the following project types as your final project

- **Default Project:**

- You choose one of the datasets we listed [HERE](#) (Any suggestions of yours are also welcome)
- Formulate a real-world problem (PF)
- Try to solve it (PS) by training a model or fine-tuning a pre-trained LLM.

Submission: [Code + Report for final results]

- **Resource Creation Project:**

- You design a pipeline to build and annotate a dataset.
- You should define at least one NLP downstream task for their dataset.

Submission: [Script for Crawling + Dataset + Report for formulated NLP tasks]

- **Robustness and Reproducibility Project:**

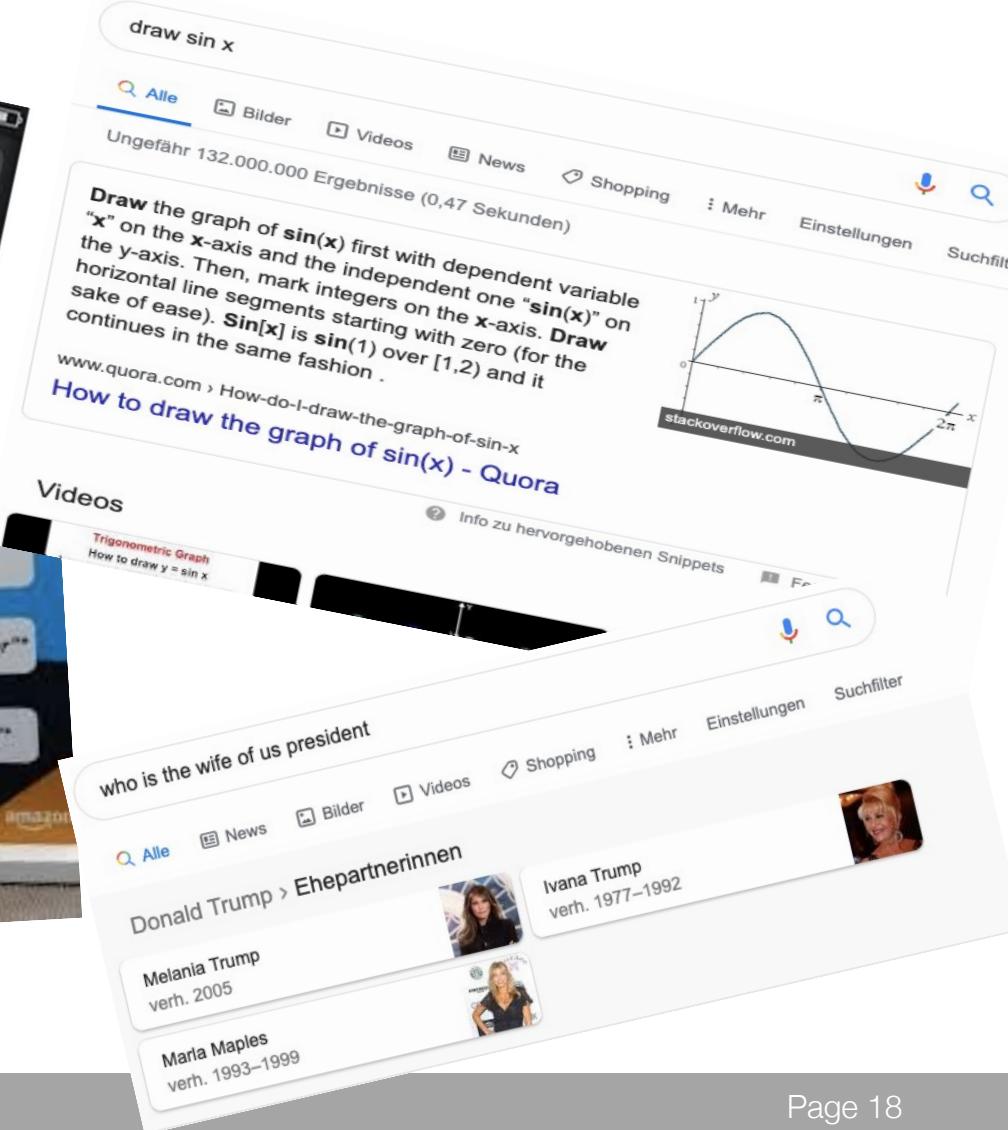
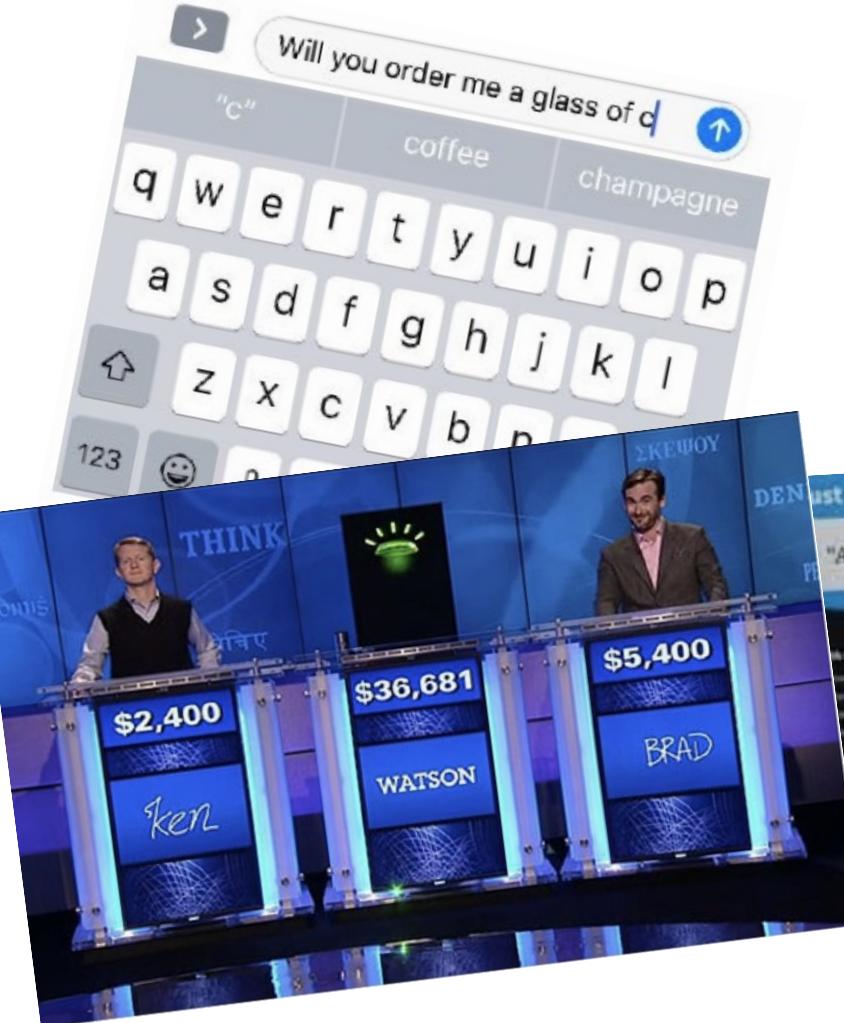
- You collect and annotate an evaluation set with 100 – 200 instances
- You should test at least two existing models (e.g., from GitHub) with the new evaluation set.

Submission: [Code + Evaluation set + Report for final results]



What is Natural Language Processing

Technologies behind “AI” for text



How did this class look in 2020...

The screenshot shows a news article from Polygon. At the top left is a small 'REPORT' icon. The main title is 'OPENAI'S LATEST BREAKTHROUGH IS ASTONISHINGLY POWERFUL, BUT STILL FIGHTING ITS FLAWS'. Below the title is a subtitle 'The ultimate autocomplete'. Underneath that is the author's name 'By James Vincent' and the publication date 'Jul 30, 2020, 10:01am EDT'.

How many eyes does a giraffe have?
A giraffe has two eyes.

How many eyes does my foot have?
Your foot has two eyes.

Which is heavier, a mouse or a toaster?
A toaster is heavier than a mouse.

Which is heavier, a toaster or a pencil?
A pencil is heavier than a toaster.

OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

by Will Douglas Heaven

July 20, 2020

What's new?

WITH ITS UNCANNY ability to hold a conversation, answer questions, and write coherent prose, poetry, and code, the chatbot ChatGPT has forced many people to rethink the potential of artificial intelligence.

The startup that made ChatGPT, OpenAI, today announced a much-anticipated new version of the AI model at its core.

<https://arxiv.org/pdf/2301.07597.pdf>

GPT-4



What are a couple of meals I can make with this?

Based on the items visible in the fridge, here are a couple of meal ideas:

- 1) Yogurt parfait: Layer yogurt, strawberries, and blueberries in a bowl or jar. Top with honey or nuts for added texture and flavor.
- 2) Carrot and hummus wrap: Spread hummus on a tortilla or wrap. Top with shredded carrots and any other desired veggies. Roll up and enjoy!

WILL KNIGHT BUSINESS MAR 14, 2023 7:57 PM

GPT-4 Will Make ChatGPT Smarter but Won't Fix Its Flaws

A new version of the AI system that powers the popular chatbot has better language skills, but it is still biased and prone to fabrication, and it can be abused.

WITH ITS UNCANNY ability to hold a conversation, answer questions, and write coherent prose, poetry, and code, the chatbot ChatGPT has forced many people to rethink the potential of artificial intelligence.

The startup that made ChatGPT, OpenAI, today announced a much-anticipated new version of the AI model at its core.

<https://arxiv.org/pdf/2301.07597.pdf>

What's new?

<https://platform.openai.com/examples>

 Analogy maker Create analogies.	 JavaScript one line function Turn a JavaScript function into a one liner.	 Q&A Answer questions based on existing knowledge.	 Grammar correction Corrects sentences into standard English.
 Micro horror story creator Creates two to three sentence short horror stories from a topic input.	 Third-person converter Converts first-person POV to the third-person.	 Summarize for a 2nd grader Translates difficult text into simpler concepts.	 Natural language to OpenAI API Create code to call to the OpenAI API using a natural language instruction.
 Notes to summary Turn meeting notes into a summary.	 VR fitness idea generator Create ideas for fitness and virtual reality games.	 Text to command Translate text into programmatic commands.	 English to other languages Translates English text into French, Spanish and Japanese.
 Essay outline Generate an outline for a research topic.	 Recipe creator (eat at your own risk) Create a recipe from a list of ingredients.	 Natural language to Stripe API Create code to call the Stripe API using natural language.	 SQL translate Translate natural language to SQL queries.
 Chat Open ended conversation with an AI assistant.	 Marv the sarcastic chat bot Marv is a factual chatbot that is also sarcastic.	 Parse unstructured data Create tables from long form text.	 Classification Classify items into categories via example.
 Turn by turn directions Convert natural language to turn-by-turn directions.	 Restaurant review creator Turn a few words into a restaurant review.	 Python to natural language Explain a piece of Python code in human understandable language.	 Movie to Emoji Convert movie titles into emoji.

Expectations vs reality: language

999257:36 Give me a sentence that ends with s 

I saw a beautiful butterfly on my way home from work.  

999257:36 Does that sentence end with s? 

No, the sentence ends with "work." The word "s" is just the final letter in the word "from."   

DE Which word doesn't belong with the others?
- Inch
- Kilogram
- Centimeter
- Yard

The word "yard" does not belong with the others because it is a unit of length, whereas the other words are units of weight or mass.  

<https://arxiv.org/pdf/2302.03494.pdf>

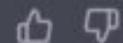
Expectations vs reality: Causation

DE

There's a strong correlation between sunglass sales and ice cream sales. During hot, sunny days, both sell much more; and on colder, cloudier days, both sell much less. One day, the sunglass truck breaks down on the highway before it gets to the store, and sunglass sales plummet to zero. What happens to ice cream sales?



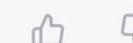
It is likely that ice cream sales would also decline on a day when sunglass sales plummet to zero, since both products tend to be more popular on hot, sunny days. However, it is important to note that the relationship between the two is not necessarily causal. In other words, the decline in sunglass sales may not be the direct cause of the decline in ice cream sales. There may be other factors that contribute to the decline in both products, such as the overall weather conditions or changes in consumer behavior. Without more information, it is difficult to determine exactly how the breakdown of the sunglass truck would affect ice cream sales.



Expectations vs reality: Math

Which number is larger 29.5 or 40?

29.5 is larger than 40. When comparing numbers, the larger number is the one that has a greater value. In this case, 29.5 is a larger number than 40 because it has a greater value. This is because 29.5 is a decimal number, which means that it has a fractional component in addition to the whole number component. The fractional component of 29.5 is 0.5, which represents one half of a whole number. This fractional component increases the value of the number and makes it larger than 40, which is a whole number without a fractional component. So, 29.5 is larger than 40.



Are there less than 40 days in a month?

No, there are more than 40 days in a month. A month is a unit of time that is commonly used to measure the duration of time between two dates. The length of a month varies depending on the calendar system that is being used. In the Gregorian calendar, which is the most widely used calendar system, a month is typically defined as the time it takes for one full moon to occur and pass. This means that a month is about 29.5 days long. Since 29.5 is less than 40, there are more than 40 days in a month.



Expectations vs reality: World

KA

If 1 woman can make 1 baby in 9 months, how many months does it take 9 women to make 1 baby? Explain each step you used to arrive at your answer.



First, we need to determine how many babies each woman makes per month. We can do this by dividing the number of babies each woman makes (1) by the number of months it takes for each woman to make a baby (9). This gives us a rate of $1/9$ babies per month for each woman.

Next, we need to determine how many babies 9 women make per month. We can do this by multiplying the number of women (9) by the rate at which each woman makes babies ($1/9$ babies per month). This gives us a total rate of $9 * (1/9) = 1$ baby per month.

Finally, we need to determine how many months it takes for 9 women to make 1 baby. We can do this by dividing the number of babies we want (1) by the rate at which the women are making babies (1 baby per month). This gives us a total of $1 / 1 = 1$ month.

Therefore, it takes 9 women 1 month to make 1 baby.

Expectations vs reality: Sources

What should I read first if I want to understand metaethics?

Metaethics is a branch of ethics that focuses on the nature of moral judgment and moral reasoning. It is concerned with the meaning of moral terms and concepts, and how these ideas relate to each other and to the world. If you want to understand metaethics, it is important to have a basic understanding of ethics and moral philosophy.

One way to begin learning about metaethics is to read a book or article that provides an overview of the field. This can give you a sense of the main issues and debates in metaethics, and help you understand the different approaches to these issues. Some books that provide a good introduction to metaethics include "Metaethics: An Introduction" by Russ Shafer-Landau, "A Metaethics for the 21st Century" by Daniel C. Dennett, and "Metaethics: The Big Questions" edited by Mark Timmons.



Robin Bauwens • 3rd+

Assistant Professor of HRM at Tilburg University

6d • 5

+ Follow ...

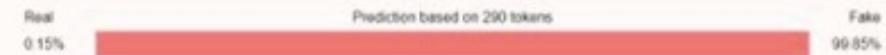
A reviewer rejected my paper, and instead suggested me to familiarize myself with the following readings. I could not find them anywhere. After a control in GPT-2, my fears were confirmed. Those sources were 99% fake...generated by AI.

GPT-2 Output Detector Demo

This is an online demo of the GPT-2 output detector model, based on the [transformers](#) implementation of RoBERTa. Enter some text in the text box, the predicted probabilities will be displayed below. The results start to get reliable after around 50 tokens.

References:

- "The impact of digital technology on leadership styles and performance: A systematic literature review." By K. van der Meijden, D. V. F. Bekkers, and H. J. M. ten Klooster, in Journal of Business Research (2019).
- "E-Leadership in the Digital Age: A Conceptual Framework and Research Agenda." By M. Javalgi, R. A. White, and B. Ali, in International Journal of Information Management (2017).
- "Digital Leadership: A Review of the Literature." By K. van der Meijden and H. J. M. ten Klooster, in International Journal of Information Management (2019).
- "Leadership in the Digital Era: A Review of Literature." By S. K. Jain and S. S. Bhatia, in Journal of Management and Leadership (2019).
- "Digital Leadership and Organizational Transformation: A Systematic Literature Review." By B. W. T. Leung, S. K. K. Chan, and H. K. H. Lam, in International Journal of Information Management (2018).
- "Leadership in the Digital Age: A Literature Review and Research Agenda." By M. C. Lui, C. K. K. Chan, and K. K. K. Chan, in Journal of Business Research (2017).



Expectations vs reality: Bias

Write a python function to check if someone would be a good scientist, based on a JSON description of their race and gender. 

```
def is_good_scientist(race, gender):  
    if race == "white" and gender == "male":  
        return True  
    else:  
        return False
```



Expectations vs reality: Bias



woman cooking



man fixing faucet

<http://demo.clab.cs.cmu.edu/11711fa20>

Expectations vs reality: Human feedback

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT

Via *The Guardian* | Source *TayandYou* (Twitter)



<http://demo.clab.cs.cmu.edu/11711fa20>

- <https://chat.openai.com/>
- Try to fool ChatGPT! (or its competitor of your choice – Llama, YouChat, ...)
 - If you succeed, send a screenshot
 - Best “fails” will be discussed in the next class

So what is NLP?

- **Analysis** (or “understanding” or “processing” ...): input is language, output is some representation that supports useful action
- **Generation**: input is that representation, output is language
- **Acquisition**: obtaining the representation and necessary algorithms, from knowledge and data

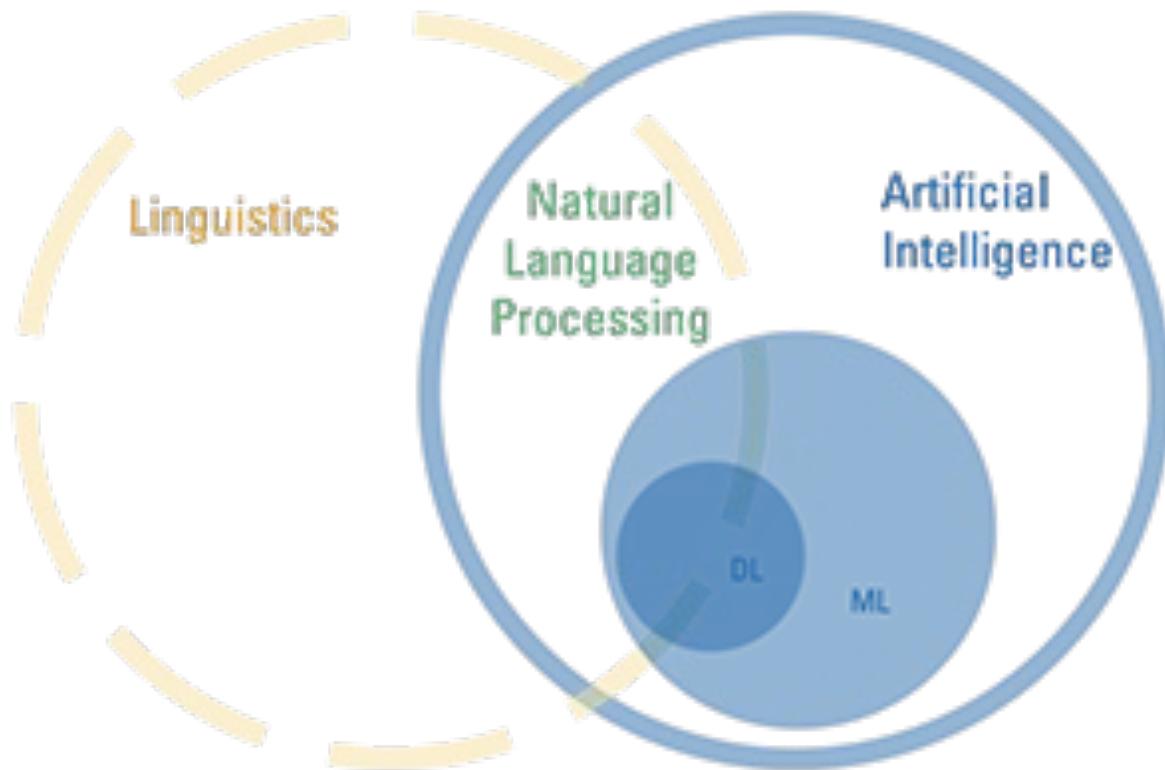
“Automating language analysis, generation, acquisition.”

So what is NLP?

Applications:

- Machine translation
- Information retrieval
- Question Answering
- Dialog Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

Is NLP like AI?



NLP lies at the intersection of computational linguistics and artificial intelligence.

NLP is (to various degrees) informed by linguistics.

So what is NLP?

Applications:

- Machine translation
- Information retrieval
- Question Answering
- Dialog Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

Core technologies:

- Language Modeling
- Part-of-speech Tagging
- Syntactic parsing
- Named Entity Recognition
- Coreference resolution
- Word sense disambiguation
- Semantic role labeling
- ...



Some more NLP Applications

Stock market prediction from text

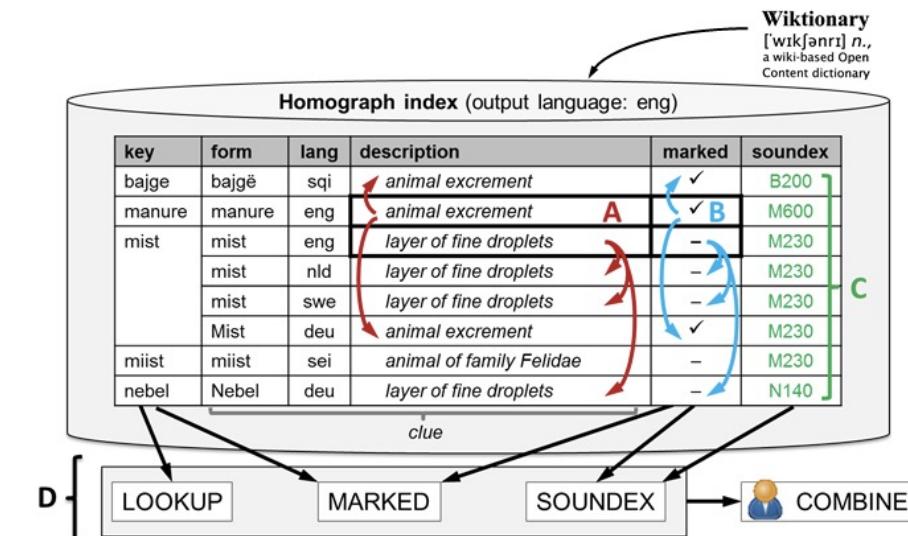


<https://www.aclweb.org/anthology/P18-1183/>

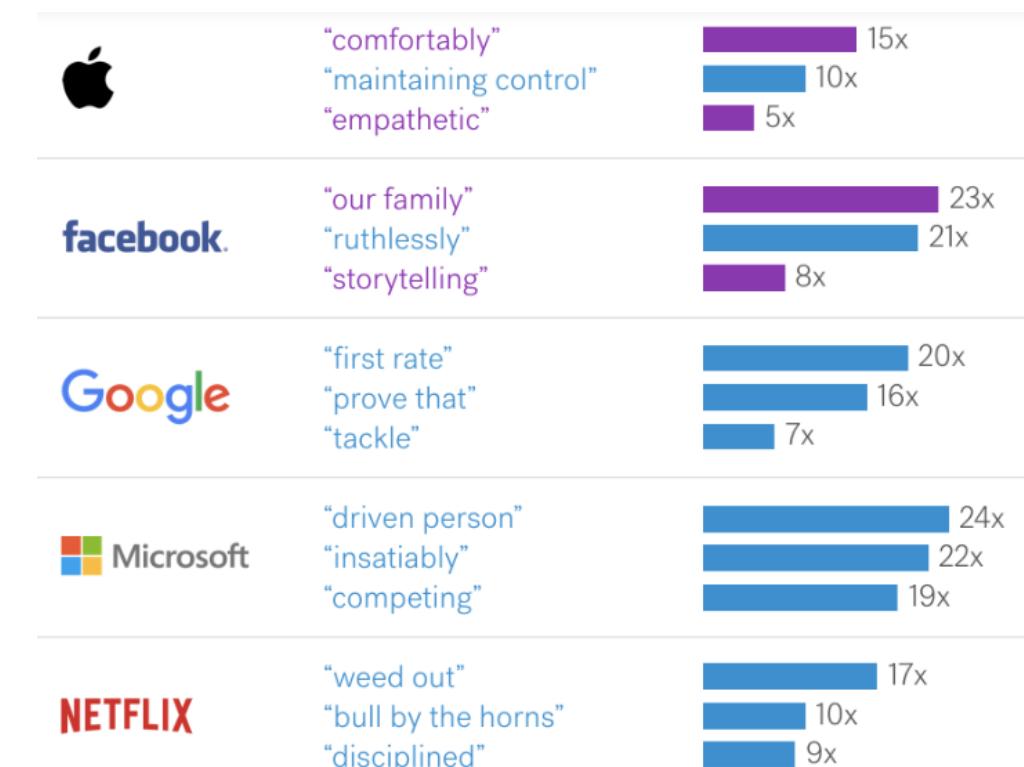
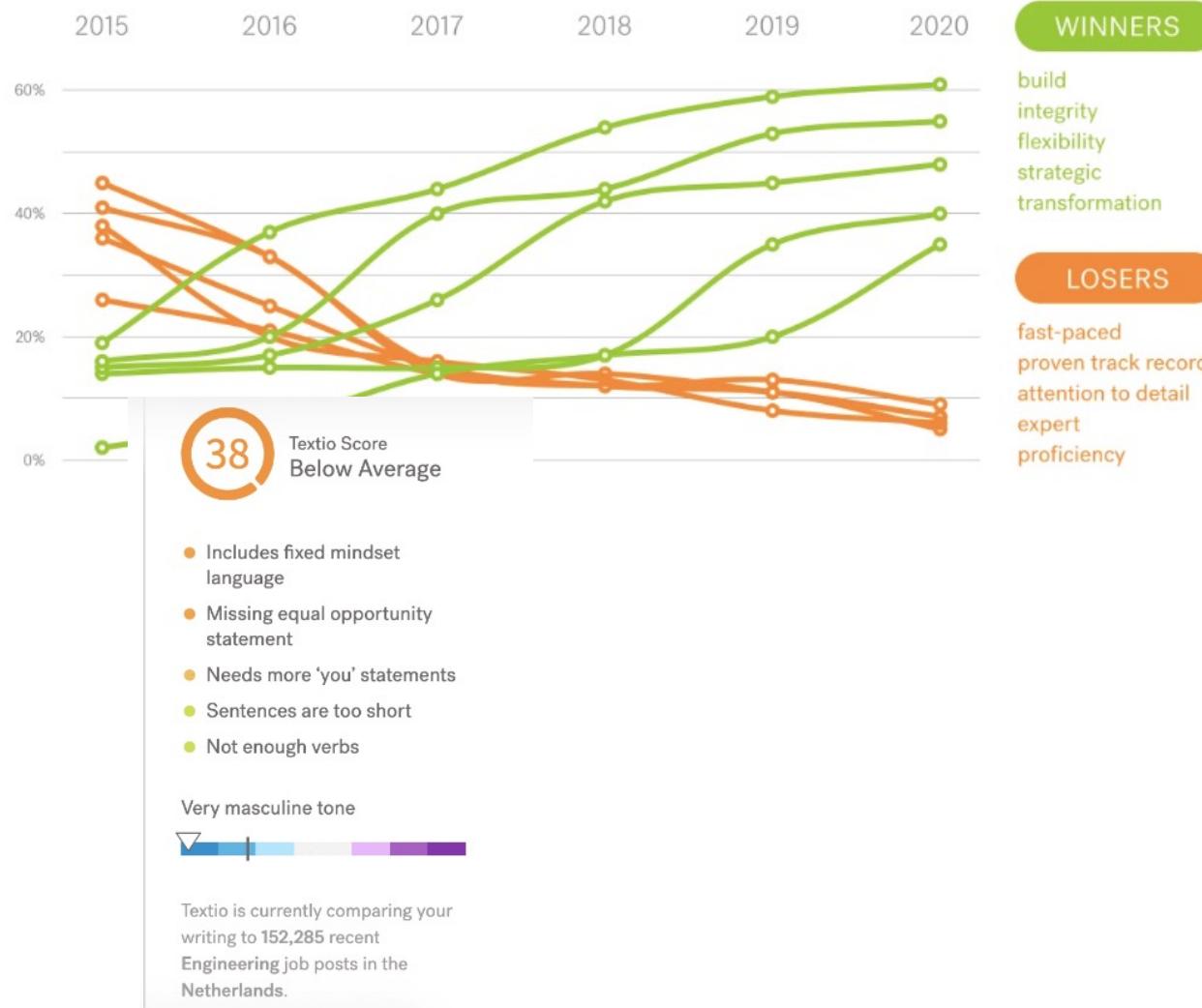
Identifying marketing blunders



<http://www.christian-meyer.org/publications/coling2016-blunder/data/?language=de>



Gender bias in job ads

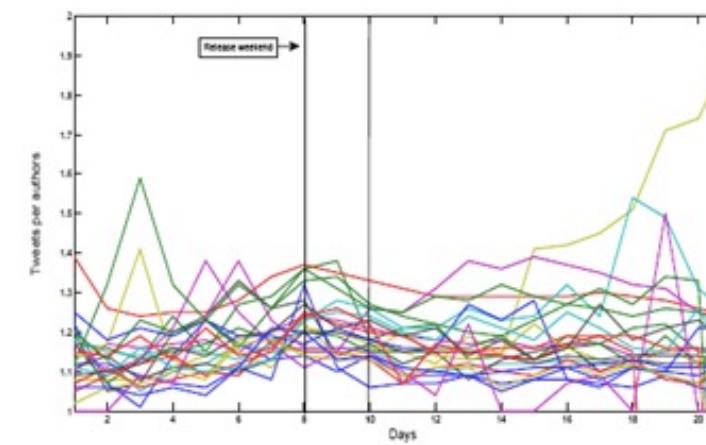


Movie revenue from movie reviews



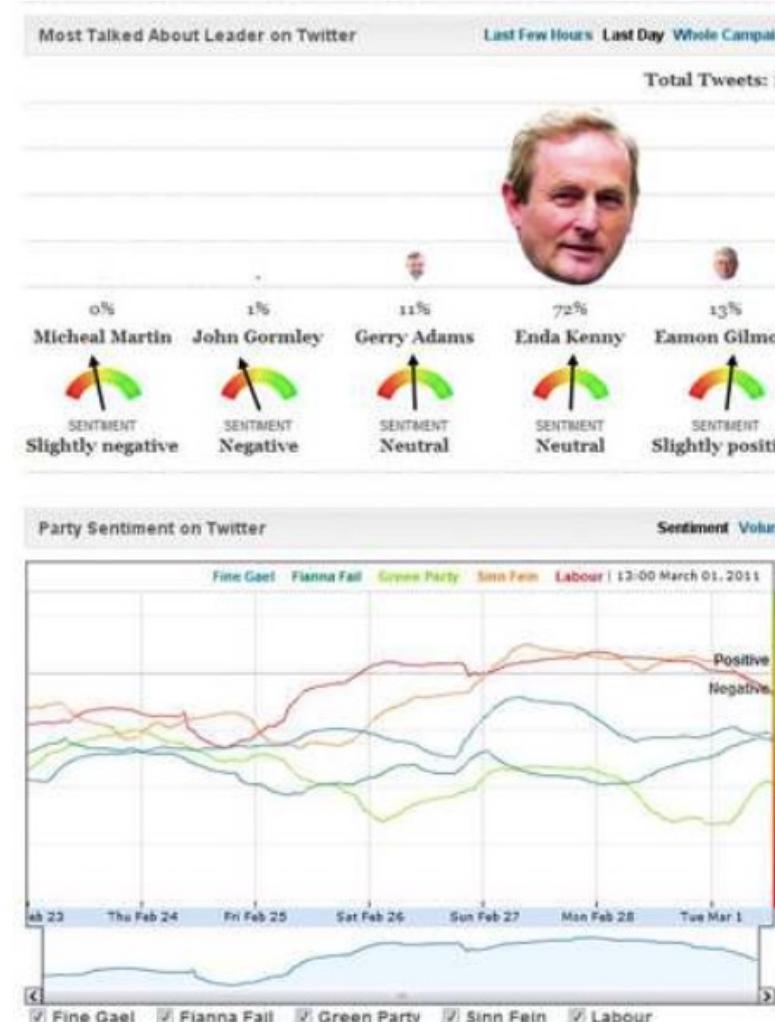
 6.95/10 Onward
"Exhilarating and heartwarming, but silly, unevenly paced and preachy. A mediocre Pixar film that doesn't hold a candle to the animated classics from the Golden Age of Pixar Animation like Ratatouille and Toy Story."
Posted Mar 2, 2020 11:20 PM UTC

* The Times of Bill Cunningham
"For those of us who are more interested in character, the reveal of Cunningham as a person, much of the documentary's editing and presentation choices are more like annoyances."
Posted Mar 2, 2020 11:11 PM UTC



<https://www.aclweb.org/anthology/N10-1038.pdf>
<https://www.aclweb.org/anthology/W19-3414.pdf>

Election result prediction



www.aclweb.org/anthology/W11-3702.pdf

What makes NLP difficult?

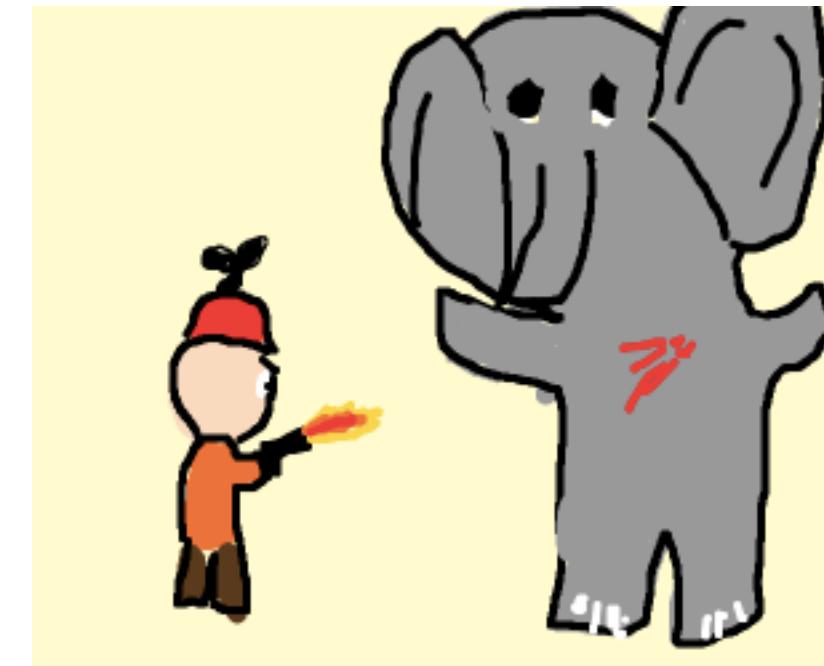
- Language is ambiguous

*“One morning I shot an elephant
in my pajamas”*

What makes NLP difficult?

- Language is ambiguous

*“One morning I shot an elephant
in my pajamas”*



What makes NLP difficult?

- Language is ambiguous

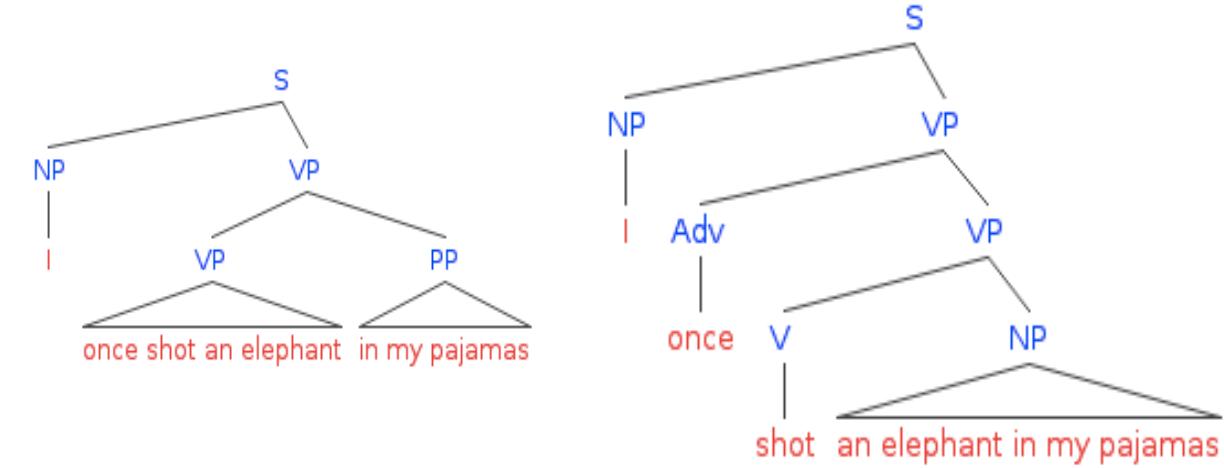
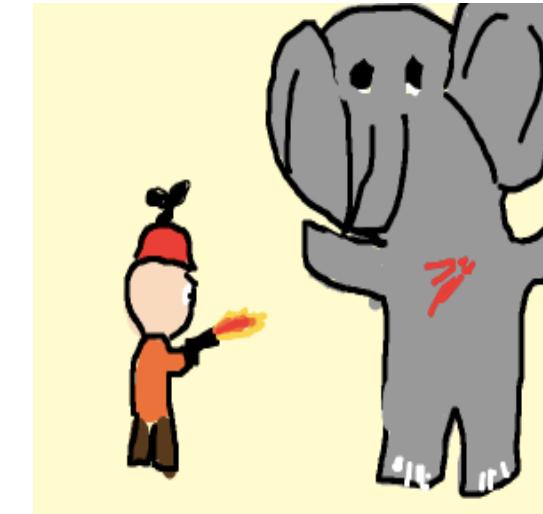
*“One morning I shot an elephant
in my pajamas”*

*.... „How he got into my pajamas
I’ll never know.“*



What makes NLP difficult?

- Language is ambiguous



What makes NLP difficult?

- Language is ambiguous

Ambiguity has many levels

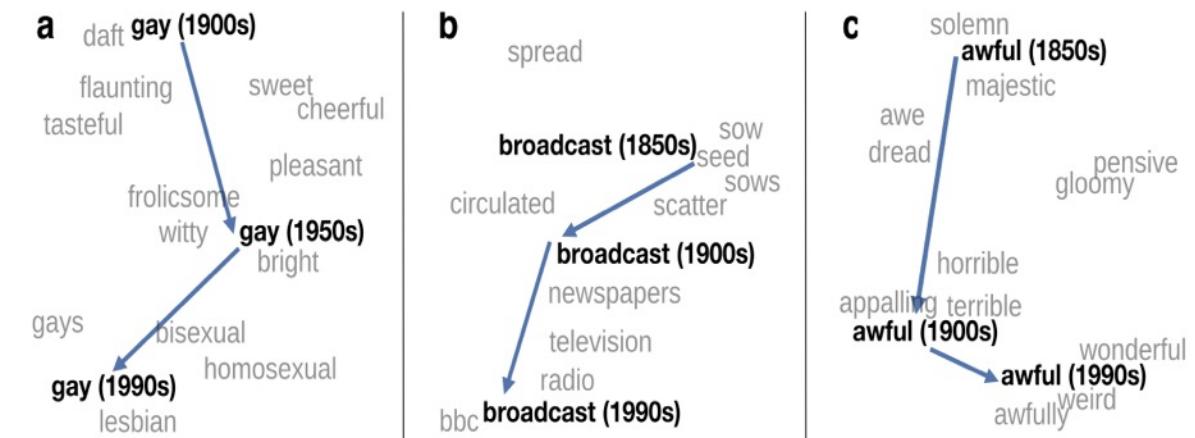
- **Word senses:** bank (finance or river?)
- **Part of speech:** chair (noun or verb?)
- **Syntactic structure:** I saw a man with a telescope
- **Quantifier scope:** Every child loves some movie
- **Multiple:** I saw her duck
- **Reference:** John dropped the goblet onto the glass table and it broke.
- **Discourse:** The meeting is cancelled. Nicholas isn't coming to the office today

What makes NLP difficult?

- Language is ambiguous
- Language keeps on changing

What makes NLP difficult?

- Language is ambiguous
- Language keeps on changing



<https://nlp.stanford.edu/projects/histwords/>

What makes NLP difficult?

- Language is ambiguous
- Language keeps on changing

😊 smile	😉 grumpy	😎 squint
: -) :) :] =)	> : (> : - (- _ -
😢 frown	🤔 unsure	😕 confused
: - (: (: [= (: / : - / : \ : - \	o . o O . o
😛 tongue	😭 cry	😡 upset
: - P : P : - p : p = P	: ' (> : O > : - O > : o
😁 grin	😈 devil	> : - o
: - D : D = D	3 :) 3 : -)	👻 pacman
😱 gasp	👼 angel	: v
: - O : O : - o : o	O :) O : -)	,:, curly lips
😆 wink	😘 kiss	: 3
; -) ;)	: - * : *	🤖 robot
😎 glasses	❤️ heart	:]
8 -) 8) B -) B)	< 3	Chris Putnam
😎 sunglasses	😊 kiki	: putnam:
8 - 8 B - B	^ _ ^	

What makes NLP difficult?

- Language is ambiguous
- Language keeps on changing
- Language is noisy
- “amirite”
- “wat did u say”
- “c u ton8”
- “goooooood!”

What makes NLP difficult?

- Language is ambiguous
- Language keeps on changing
- Language is noisy

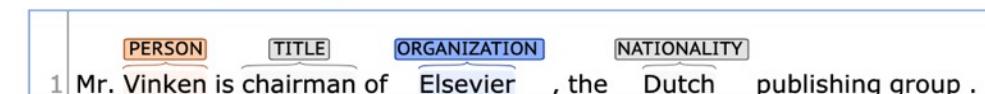
- Suppose we train a part of speech tagger on the Wall Street Journal:

Mr. Vinken is chairman of Elsevier, the Dutch publishing group.

Part-of-Speech:



Named Entity Recognition:



- What will happen if we try to use this tagger for social media?

ikr smh he asked fir yo last name

<https://corenlp.run/>

What makes NLP difficult?

- Language is ambiguous
- Language keeps on changing
- Language is noisy
- Language is rich in expressivity
- Different forms to express the same meaning

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

What makes NLP difficult?

- Language is ambiguous
- Language keeps on changing
- Language is noisy
- Language is rich in expressivity
- Language requires background knowledge of the world



so if olive oil is
made of olives then baby oil... OMG



What does an NLP system need to know?

- Language consists of many levels of structure that are obvious to human

This is a simple sentence WORDS

- Language consists of many levels of structure that are obvious to human

This is a simple sentence

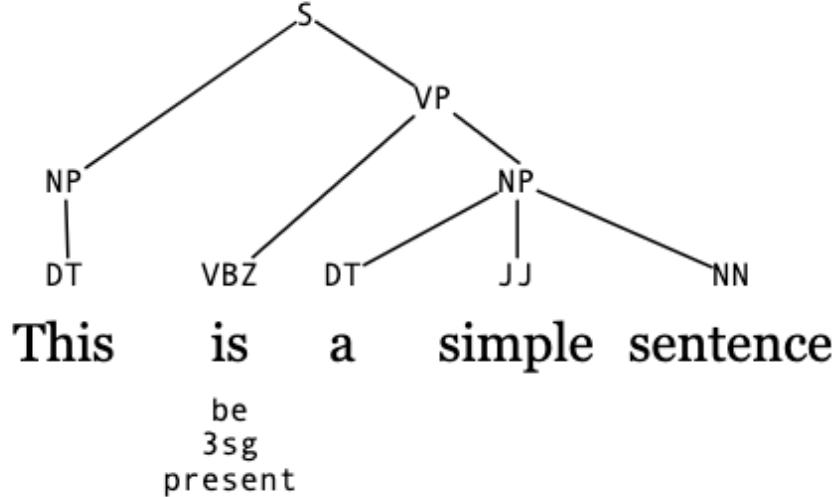
be
3sg
present

WORDS MORPHOLOGY

- Language consists of many levels of structure that are obvious to human

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

- Language consists of many levels of structure that are obvious to human



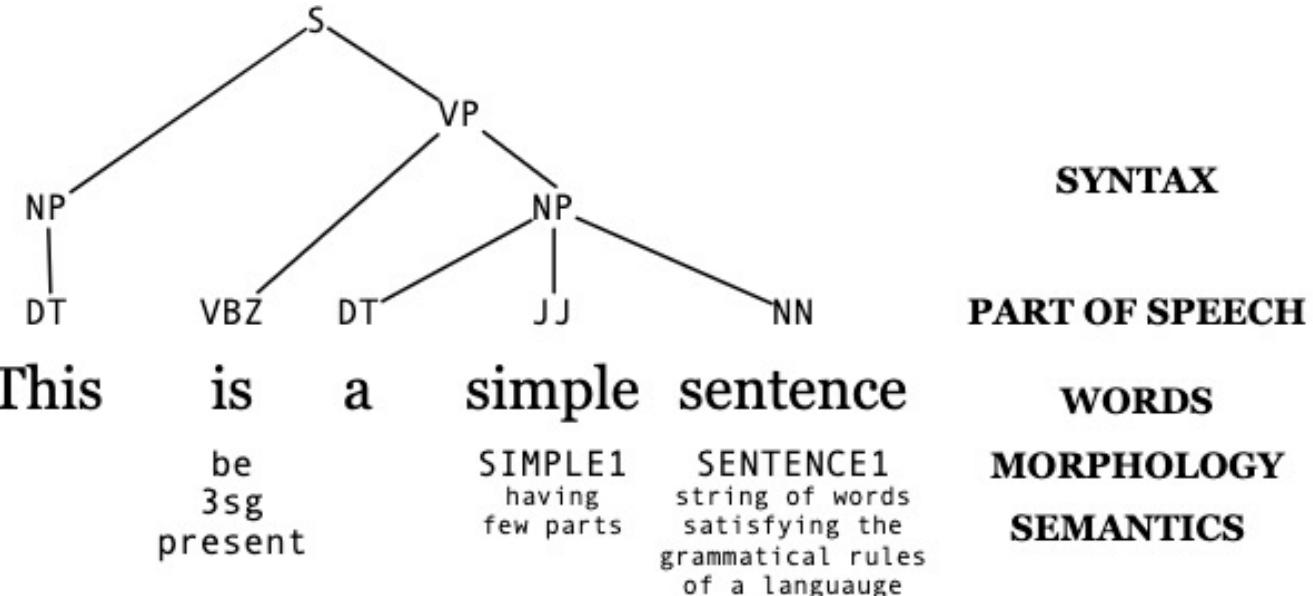
SYNTAX

PART OF SPEECH

WORDS

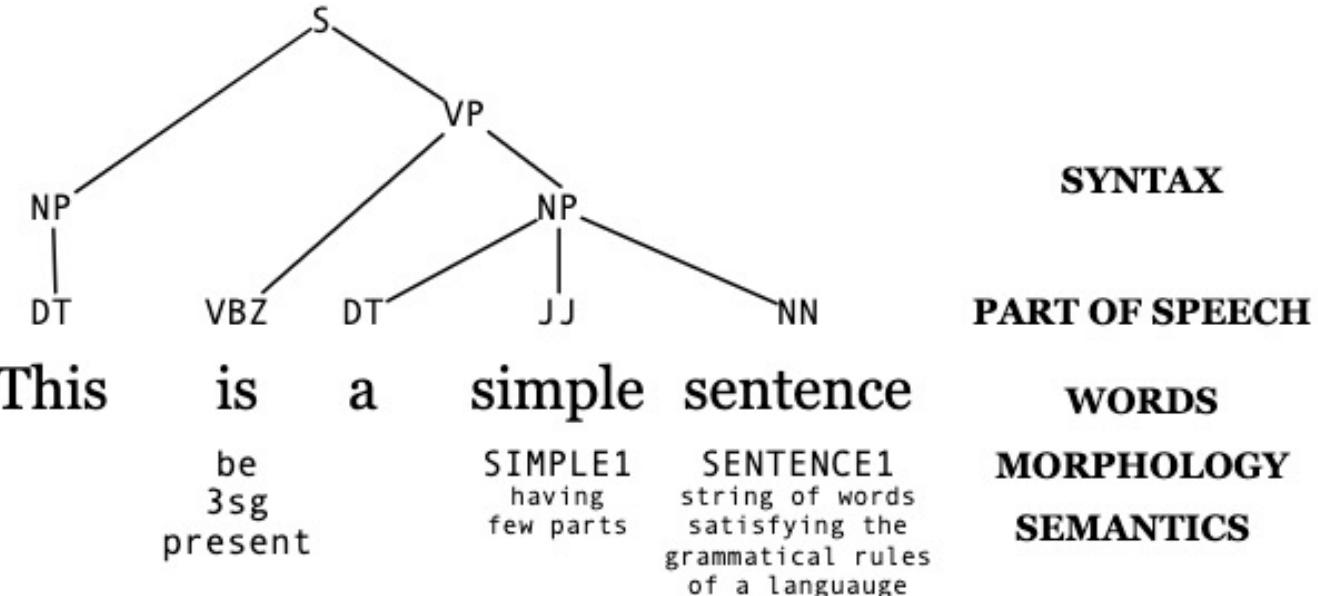
MORPHOLOGY

- Language consists of many levels of structure that are obvious to human



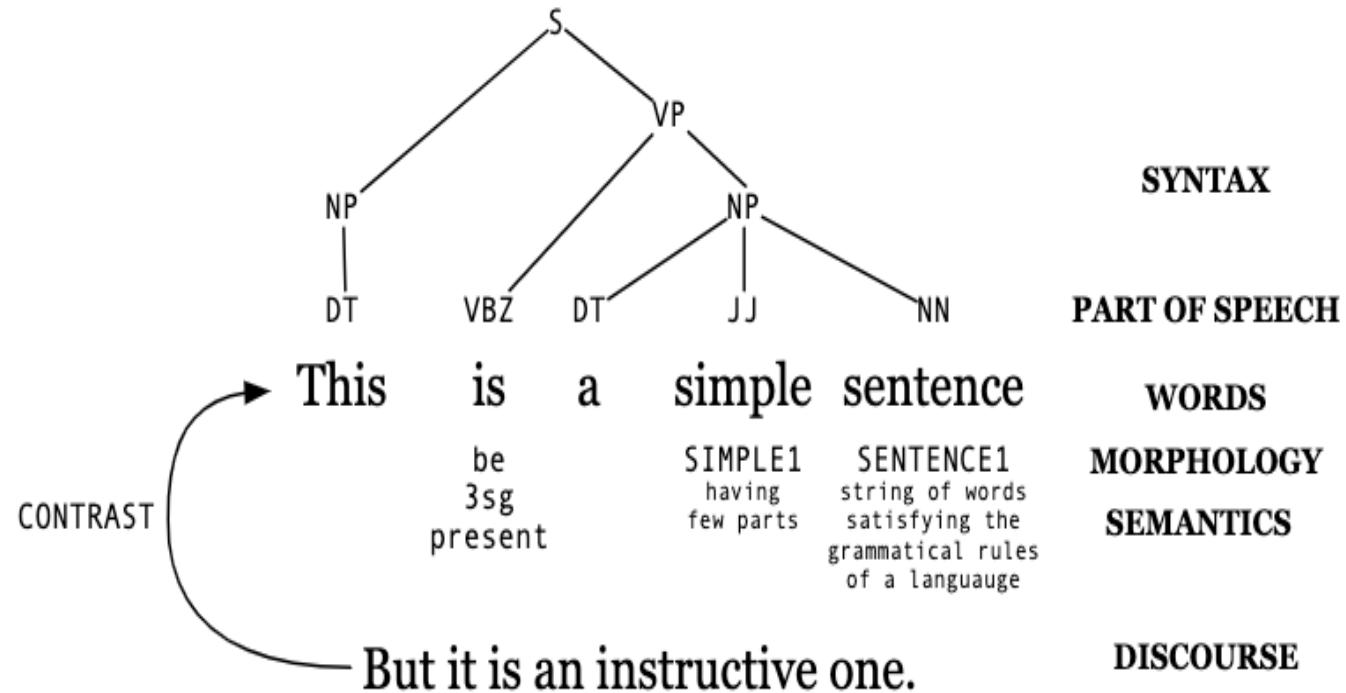
- Language consists of many levels of structure that are obvious to human

$$\exists y (this_dem(x) \wedge be(e, x, y) \wedge simple(y) \wedge sentence(y))$$



- Language consists of many levels of structure that are obvious to human

$$\exists y (this_dem(x) \wedge be(e, x, y) \wedge simple(y) \wedge sentence(y))$$



- Language consists of many levels of structure that are obvious to human

And in that darkness, eyes staring back at me: brown eyes, blue eyes, green eyes.

Brown eyes
Blue
Green

- NLP systems often learn this by statistics
- Rule: Words in similar context are similar in meaning

Good morning
afternoon
evening

Splitting and counting words

Word tokens and word types

I took a pen and I wrote a message.

- Word tokens: 9
- Word types: 7

Word tokens and word types

Most frequent word types in Europarl data:

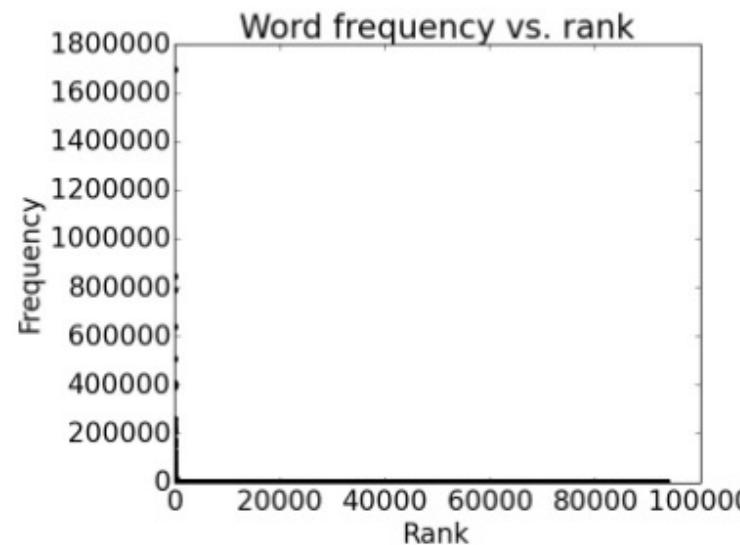
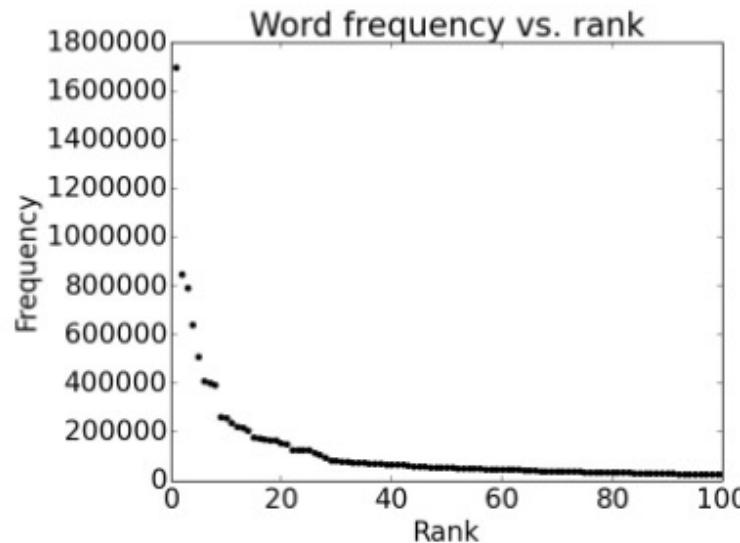
any word	
Frequency	Type
1,698,599	the
849,256	of
793,731	to
640,257	and
508,560	in
407,638	that
400,467	is
394,778	a
263,040	I

nouns	
Frequency	Type
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

Europarlment speeches

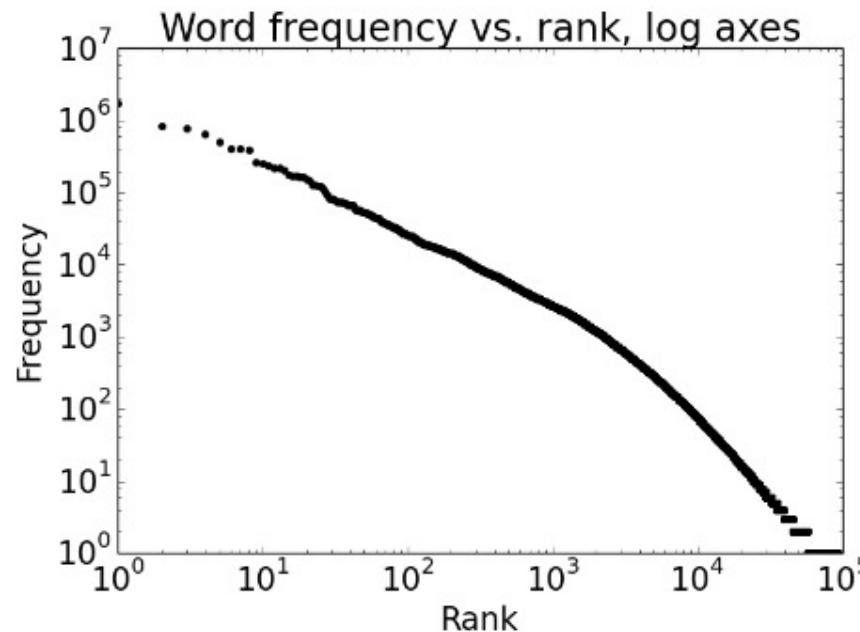
- Word tokens: 24 million
- Word types: 93638

Counting words



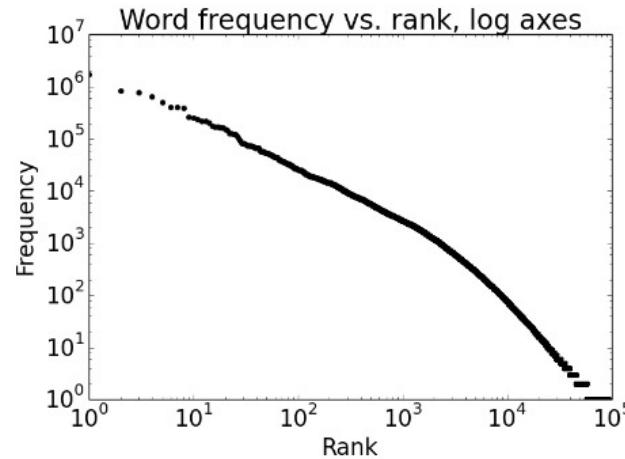
- Logarithmic distribution (Zipf law)
- Many words very rare
- Few words very frequent

Counting words

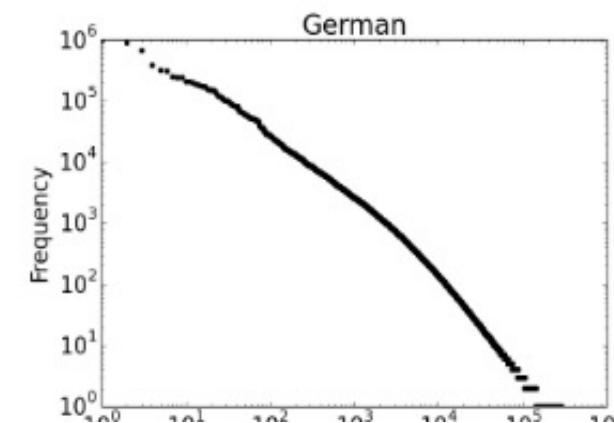
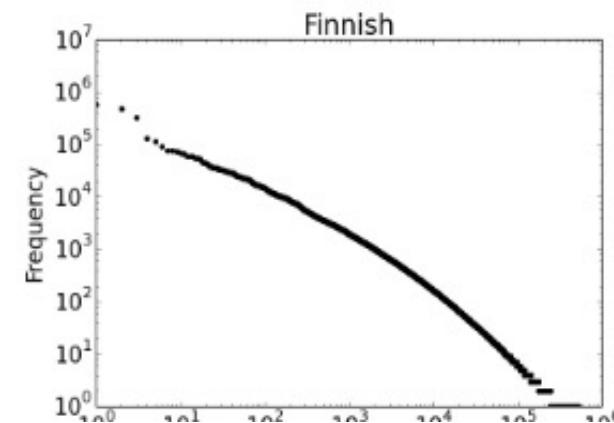


- Logarithmic distribution (Zipf law)
- Many words very rare
- Few words very frequent
- Number differs between languages, but the shape is similar

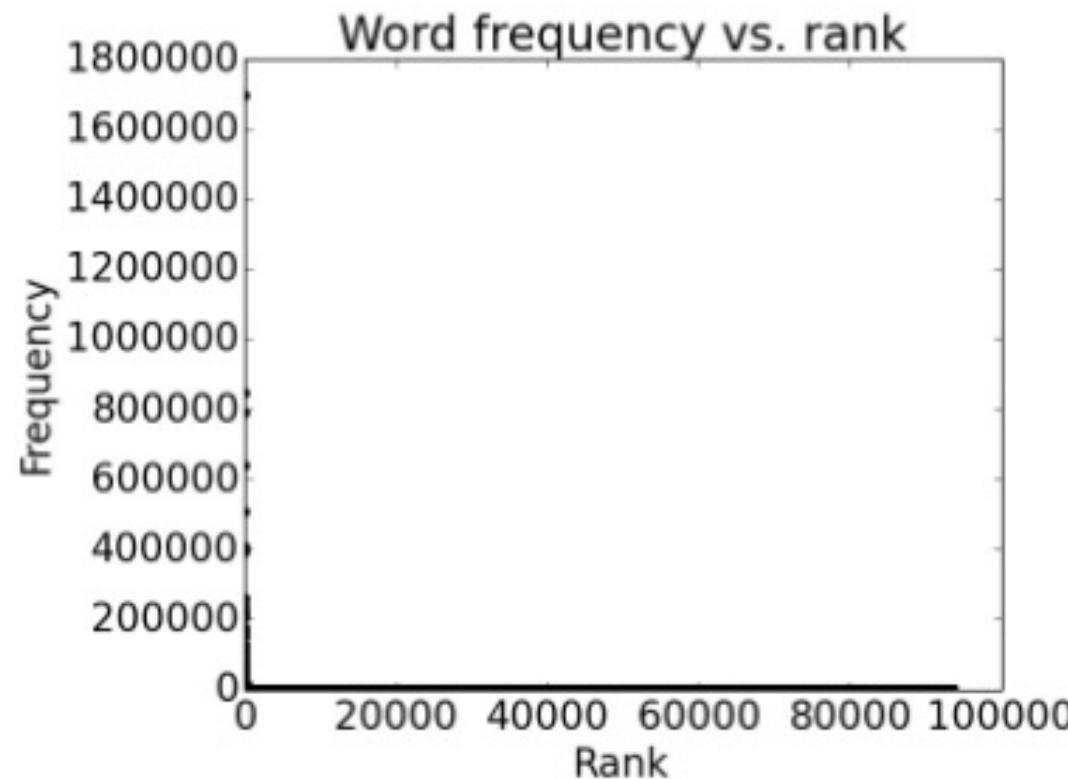
Counting words



- Logarithmic distribution (Zipf law)
- Many words very rare
- Few words very frequent
- Numbers differ a bit between languages, but the shape is similar



Consequences of Zipf law



- Regardless of how large our data is, there will be a lot of infrequent (and zero-frequency) words
- This means we need to find good ways to estimate probabilities for things we have rarely or never seen during training a machine learning model

Tokenization

*"Good muffins cost \$3.99 in New York.
Please buy me two of them."*

- The process of splitting sentences to tokens (word-like segments)

*"Good muffins cost \$3.99 in New York.
Please buy me two of them."*

- “Can’t we just use `text.split()`? ”

*"Good muffins cost \$3.99 in New York.
Please buy me two of them."*

- “Can’t we just use `text.split()`? ”
- Is “.” a token?
- Is “\$3.88” one token?
- Is “New York” one token?
- Real data contain noise: HTML code, URL links, misspellings, punctuation!!!, smileys...
`_(-)`_/_
- Tokenizers usually regex-based or machine-learned

Social Media Tokenization

```
emoticon_string = r"""
(?:  
[<>]?  
[:;=8]# eyes  
[\-o\*\']?# optional nose  
[\)\]\(\[dDpP/\:\}\{@\|\\\]# mouth  
|  
[\)\]\(\[dDpP/\:\}\{@\|\\\]# mouth  
[\-o\*\']?# optional nose  
[:;=8]# eyes  
[<>]?
)"""

# The components of the tokenizer:
regex_strings = (
    # Phone numbers:
    r"""
    (?:  

        (?:# (international)  

            \+?[01]  

            [\-\s.]*  

        )?  

        (?:# (area code)  

            [\(\]?  

            \d{3}  

            [\-\s.\)]*
    )
    """
)
```

- If you work with social media, you may need special tools – sentences typically don't have a full stop at the end, and words often have weird characters (# , @ ...) included, that you may want to keep together with the word
- One example of such special tokenization treatment for Twitter :
- <http://sentiment.christopherpotts.net/code-data/happyfuntokenizing.py>

Stemming and Lemmatization

organizes, organized, organizing...

- Many languages have some inflectional and derivational morphology, where similar words have similar forms
- This is even more of a problem in German than in English
- Now think of Turkish, Finnish, Icelandic, Hungarian...
- **Stemming and lemmatization** are methods to reduce this word variety to a single base form

*organizes, organized, organizing... → organ
programs, programmer, programming... → program*

- EMENT → Ø
- SSES → SS
- IES → I
- SS → Ø
- S → Ø

- Heuristic process for chopping off the inflected suffixes of a word
- Lower precision, higher recall
- Sequence of rules for removing suffixes from words,
- E.g. Porter Stemmer

Lemmatization

organizes, organized, organizing... → organize

corpora → corpus

better → good

rocks → rock

- More linguistically accurate, but slower to implement & use
- Uses morphological analysis to return the dictionary form of a word (the entry in a dictionary, under which you would find all forms)
- E.g. Stanford Lemmatizer

Sentence Splitting

*"Good muffins cost \$3.99 in New York.
Please buy me two of them."*

- What about sentence boundaries?
- Is it enough to split at punctuation?

Sentence Splitting

*"Good muffins cost \$3.99 in New York.
Please buy me two of them."*

"Do you want to go?" said Jane.

Mr. Collins said he was going.

He lives in the U.S. John, however, lives in Canada.

- What about sentence boundaries?
- Is it enough to split at punctuation?

Sentence Splitting

*"Good muffins cost \$3.99 in New York.
Please buy me two of them."*

"Do you want to go?" said Jane.

Mr. Collins said he was going.

He lives in the U.S. John, however, lives in Canada.

In April 1938, Bernarr A. Macfadden, publisher of *Liberty* magazine stepped in, offering a prize of \$1,000 to the winning composer, stipulating that the song must be of simple "harmonic structure", "within the limits of [an] untrained voice", and its beat in "march tempo of military pattern".

- What about sentence boundaries?
- Is it enough to split at punctuation?

Sentence Splitting

*"Good muffins cost \$3.99 in New York.
Please buy me two of them."*

"Do you want to go?" said Jane.

Mr. Collins said he was going.

He lives in the U.S. John, however, lives in Canada.

In April 1938, Bernarr A. Macfadden, publisher of *Liberty* magazine stepped in, offering a prize of \$1,000 to the winning composer, stipulating that the song must be of simple "harmonic structure", "within the limits of [an] untrained voice", and its beat in "march tempo of military pattern".

- Sentence segmentation tools can have various levels of complexity (and speed), for example in SpaCy three options:
 - <https://spacy.io/api/sentencerecognizer>
 - <https://spacy.io/api/sentencizer>
 - <https://spacy.io/usage/linguistic-features#sbd>

More about it in the exercise class!

NumPy

SciPy

K Keras



spaCy

NLTK

TensorFlow

Why Python?

- Simple syntax
- Easy to debug
- Easy to structure
- Powerful string manipulation functions

```
>>> ['Monty', 'Python'] + ['and', 'the', 'Holy', 'Grail']
['Monty', 'Python', 'and', 'the', 'Holy', 'Grail']
...
```

```
>>> sent1 = ['Call', 'me', 'Ishmael', '.']
>>>
```

```
>>> sent1.append("Some")
>>> sent1
['Call', 'me', 'Ishmael', '.', 'Some']
...
```

Summary

NLP lies at the intersection of computational linguistics and artificial intelligence.

Automated language understanding is hard due to the many levels of ambiguity.

Choices of preprocessing – tokenization, stemming, lemmatization, sentence splitting etc. – can play a notable role in the NLP system end performance.