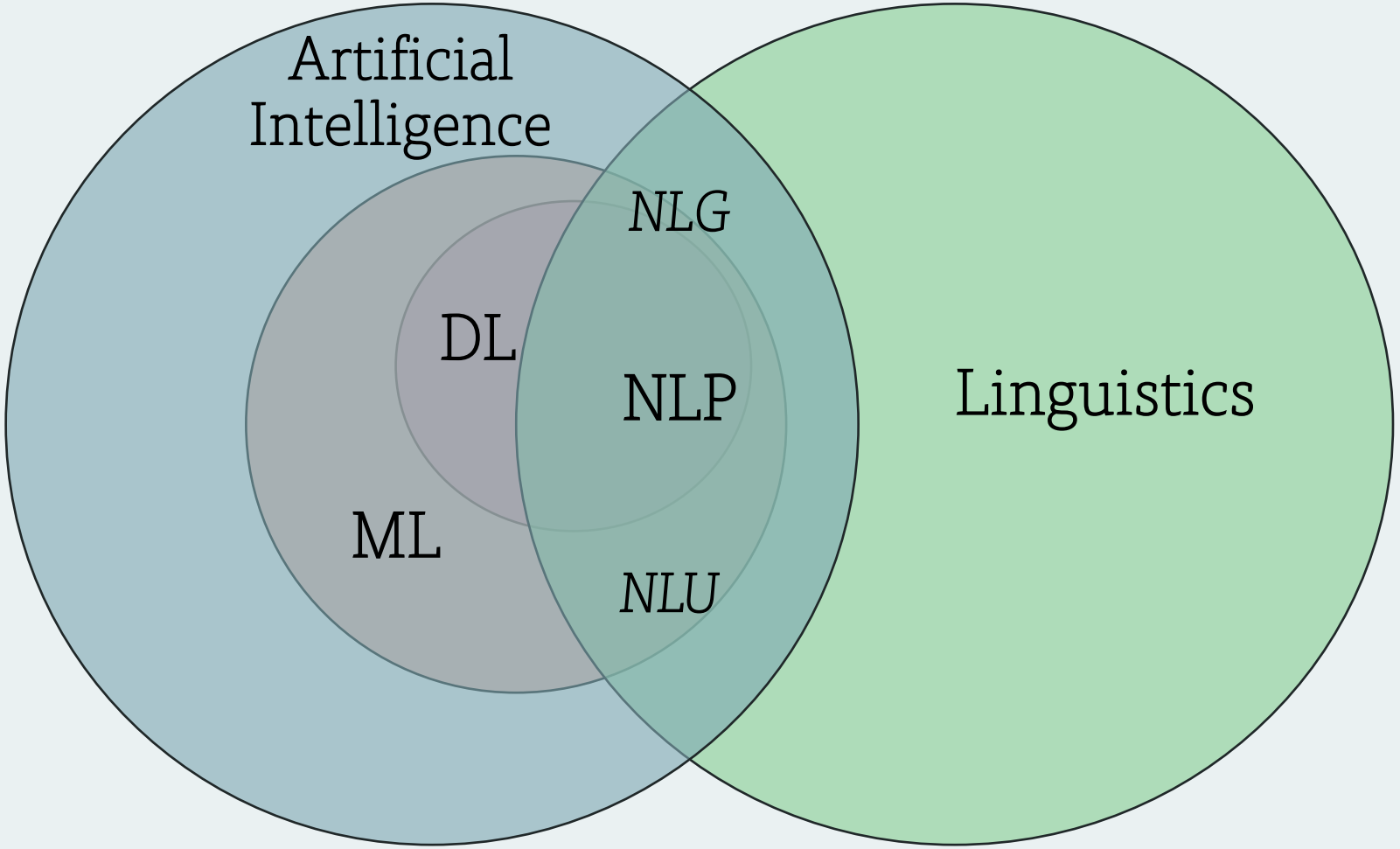
NLP for Social Sciences

1. The Basics of NLP

Irina Proskurina, Université de Lyon, de Lyon 2, Laboratoire ERIC

Natural Language Processing



Why NLU is difficult?

1 Ambiguity

bank, present, light, spring, match, trip

2 Idioms

break the ice, under the weather, a piece of cake

3 Complex grammar

passive voice, conditional tense. phrasal verbs

Why NLU is difficult?

4 Homonyms

bass (fish/sound)
bat (animal/sport.)

5 Borrowed words

rendezvous

6 Slang

Lol, u, yp

Why NLU is difficult?

- Inference tasks, world knowledge

Jessica noticed that Sarah was unusually quiet. After the discussion, **she** asked if everything was okay.

Who is **she**?

Why NLU is difficult?

- Inference tasks, world knowledge

Jessica noticed that Sarah was unusually quiet. After the discussion, **she** asked if everything was okay.

Who is **she**?

Training data limitations

Why NLP is difficult?

Text representation problem

- a word is the basic structural unit of language
- word meaning depends on context
- many words and sparse feature spaces

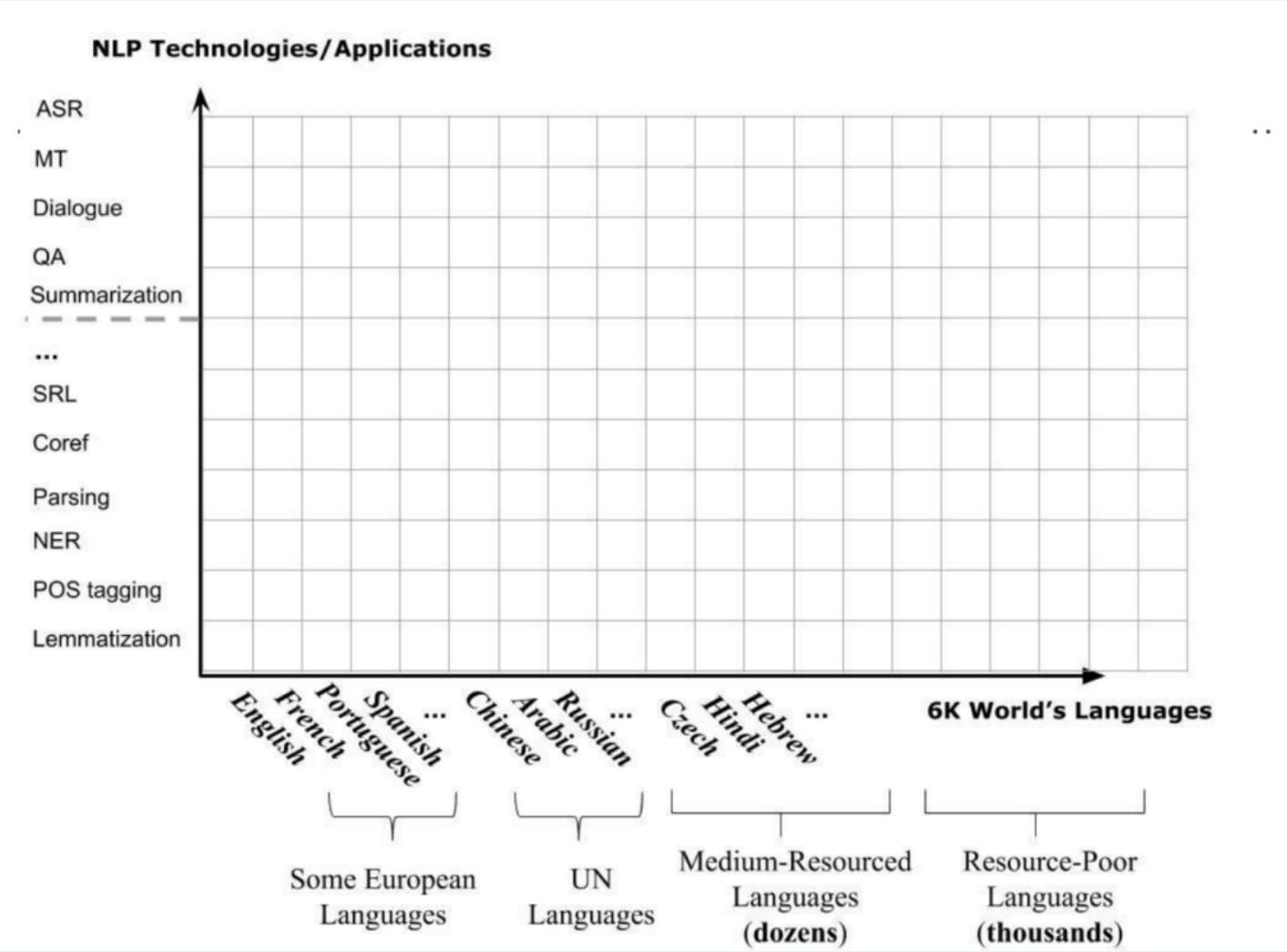
Why NLP is difficult?

Language structure (3-level)

- Words and phrases (morphology)
- Sentences (syntax)
- Text (discourse)

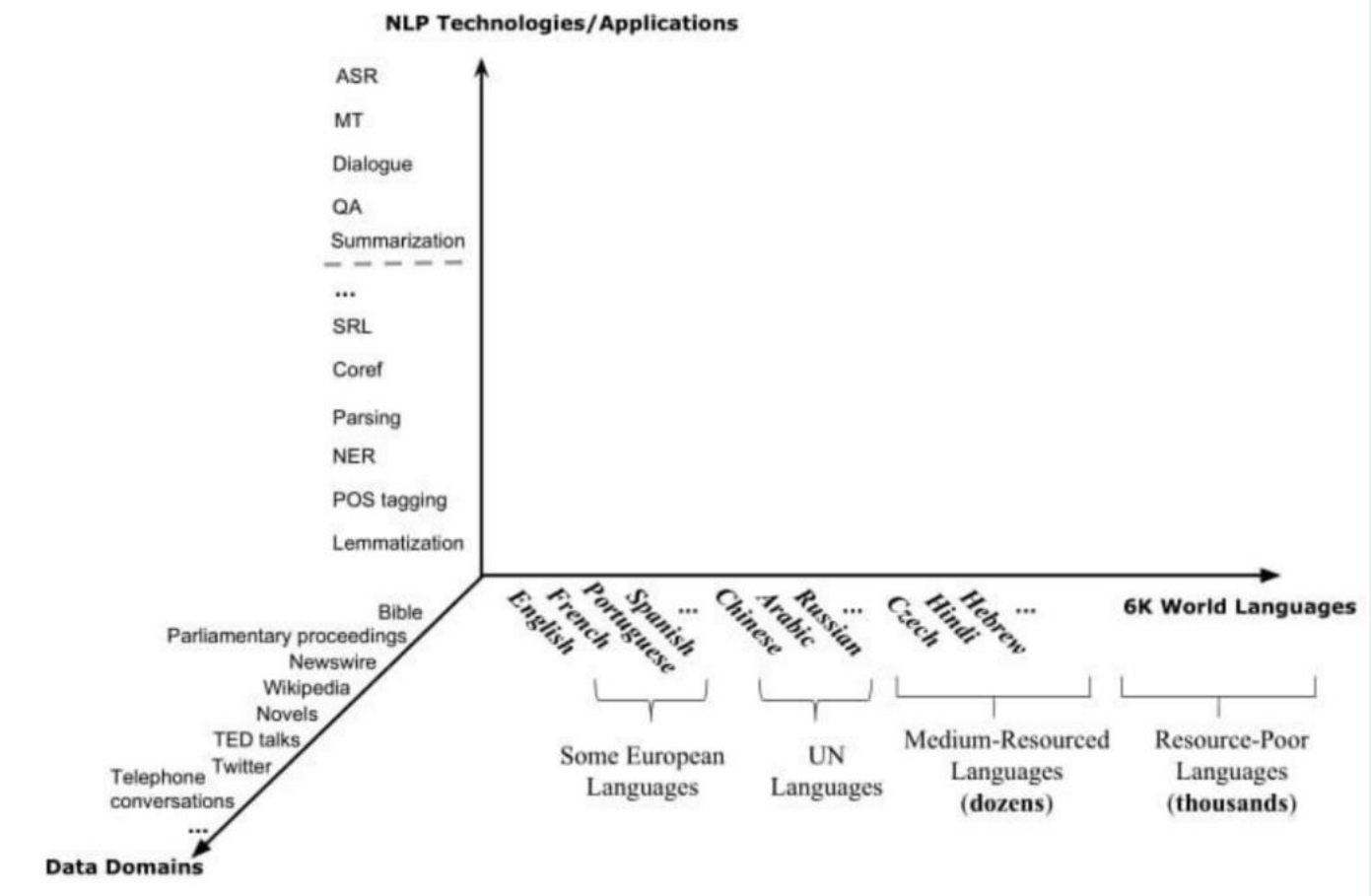
Why NLP is difficult?

- Tasks
- Languages



Why NLP is difficult?

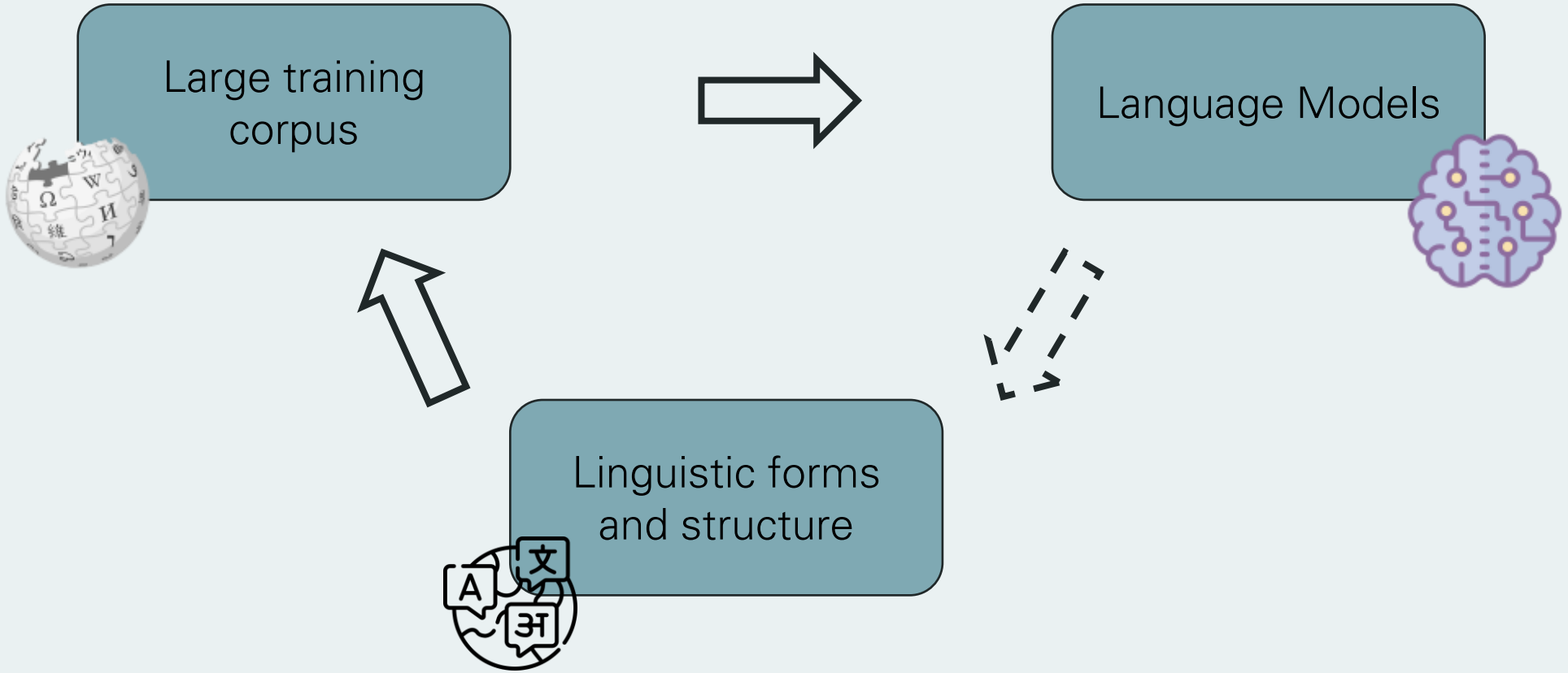
- Tasks
- Languages
- Data domains



Corpus

- A corpus is a collection of text
- Often annotated in some way
- Sometimes just lots of text
- Examples
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of French/English sentences
  - Yelp reviews
  - Famous benchmarks: MMLU, GLUE
  - https://huggingface.co/datasets

Why NLP is difficult?



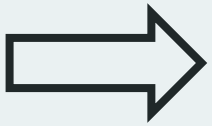
Course Structure

- Introduction to structural linguistics and text processing
- Text representation models
- Main tasks: text classification, named entity recognition, machine translation
- Transformer-like language models

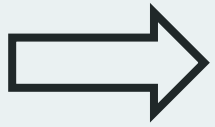
Evaluation: Project + Exam

Text classification problem

Processing
+ n feature extraction



$$\begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nm} \end{pmatrix}$$



Model

Overview of main tasks in NLP

- Conversational agents
- Information extraction and question answering
- Machine translation
- Opinion and sentiment analysis
- Social media analysis
- Visual understanding
- Essay evaluation
- Mining legal, medical, or scholarly literature

Text classification problem

Processing
+ n feature extraction

$$\begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nm} \end{pmatrix}$$

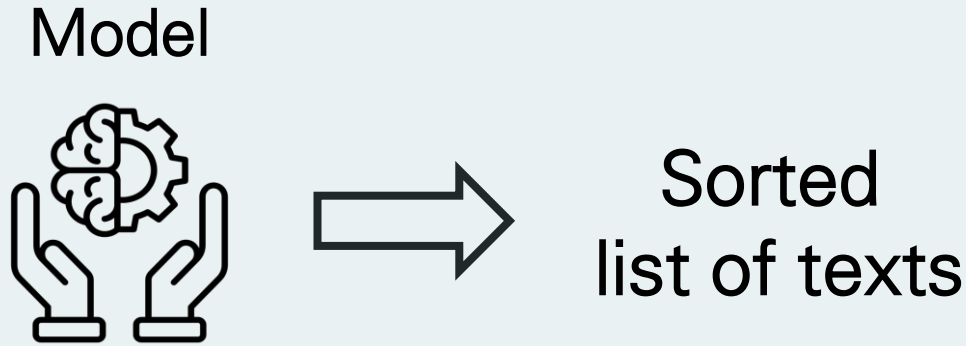
Model

Label

- Spam classification
- Sentiment analysis
- Hate speech detection
- Text similarity

Ranking problem

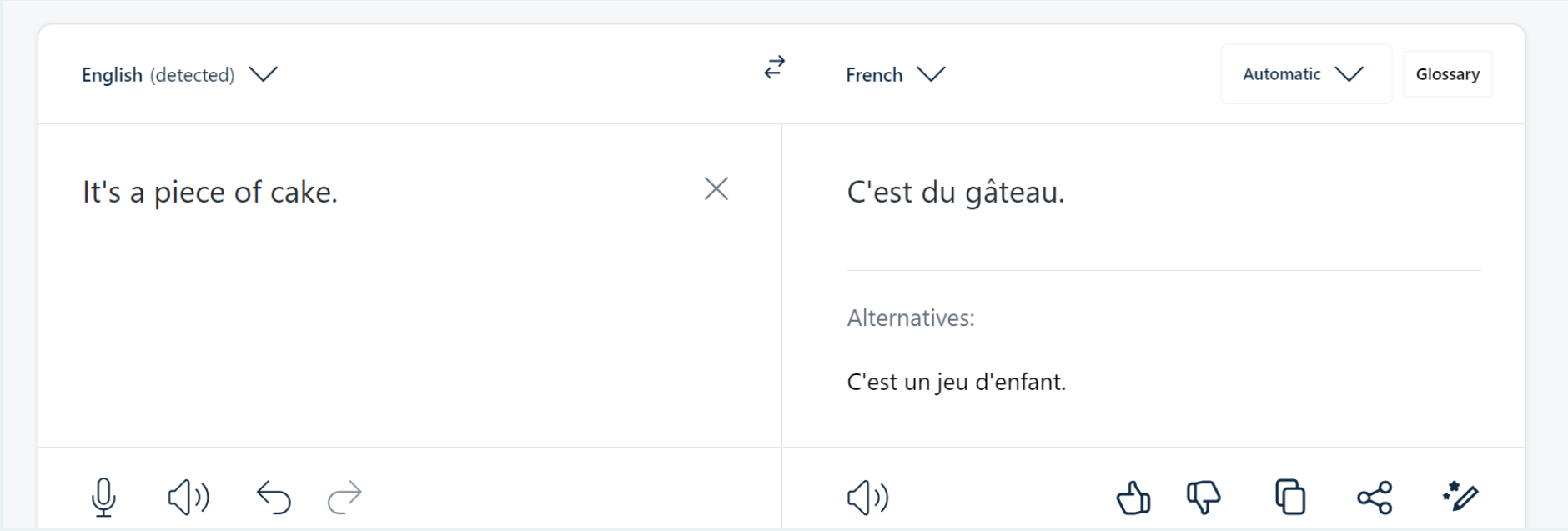
- Search engine ranking:
- Recommendation systems
- Ad ranking



Machine translation problem

Google Translate

DeepL



Grammatical error correction

I goes to the market to buy some fruit. I seen a lot of colorful apples and oranges. The price of them was cheaper then I expected. I buyed three apples and two oranges. I sees …

Virtual assistants (Chatbots)

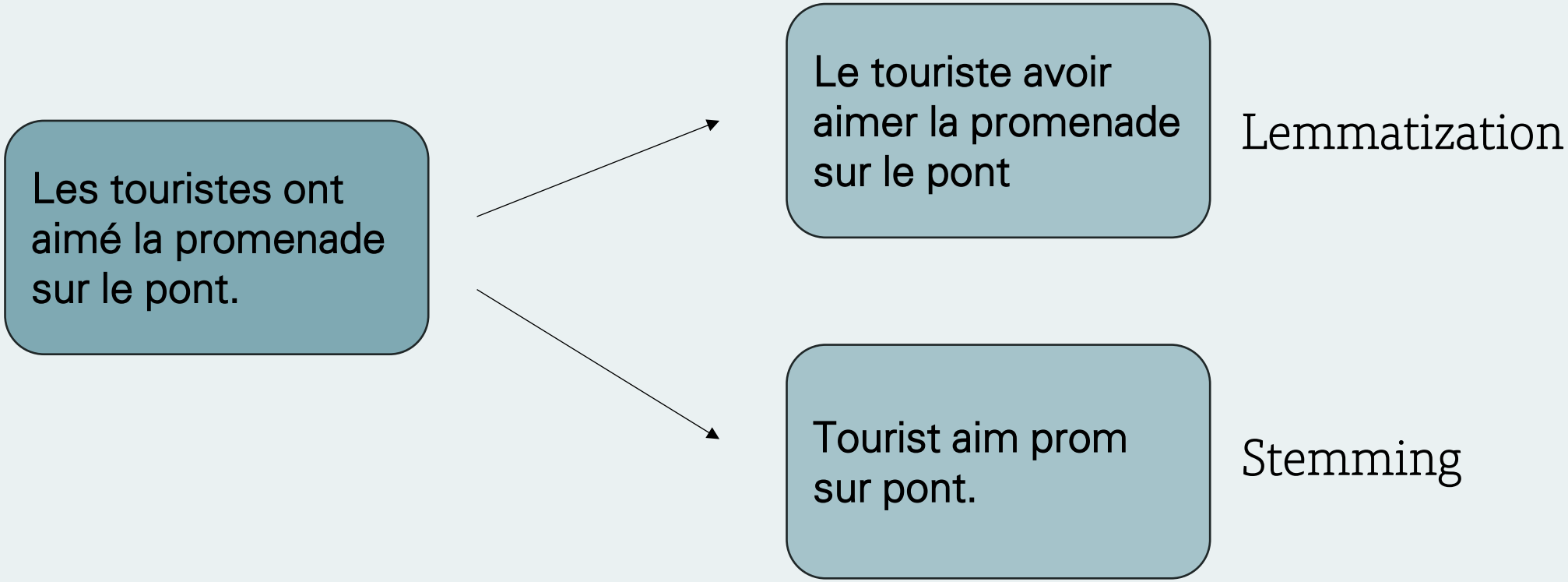
Analyze the input and generates output based on the request

- SNCF/RATP chatbots
- La banque postale virtual assistant
- Air France Chatbot (Louis)

Text preprocessing

- Tokenization
- Sentence segmentation
- Punctuation removal
- Stop words
- Order by length, frequency/regular expression
- Lemmatization
- Stemming

Example of Lemmatization

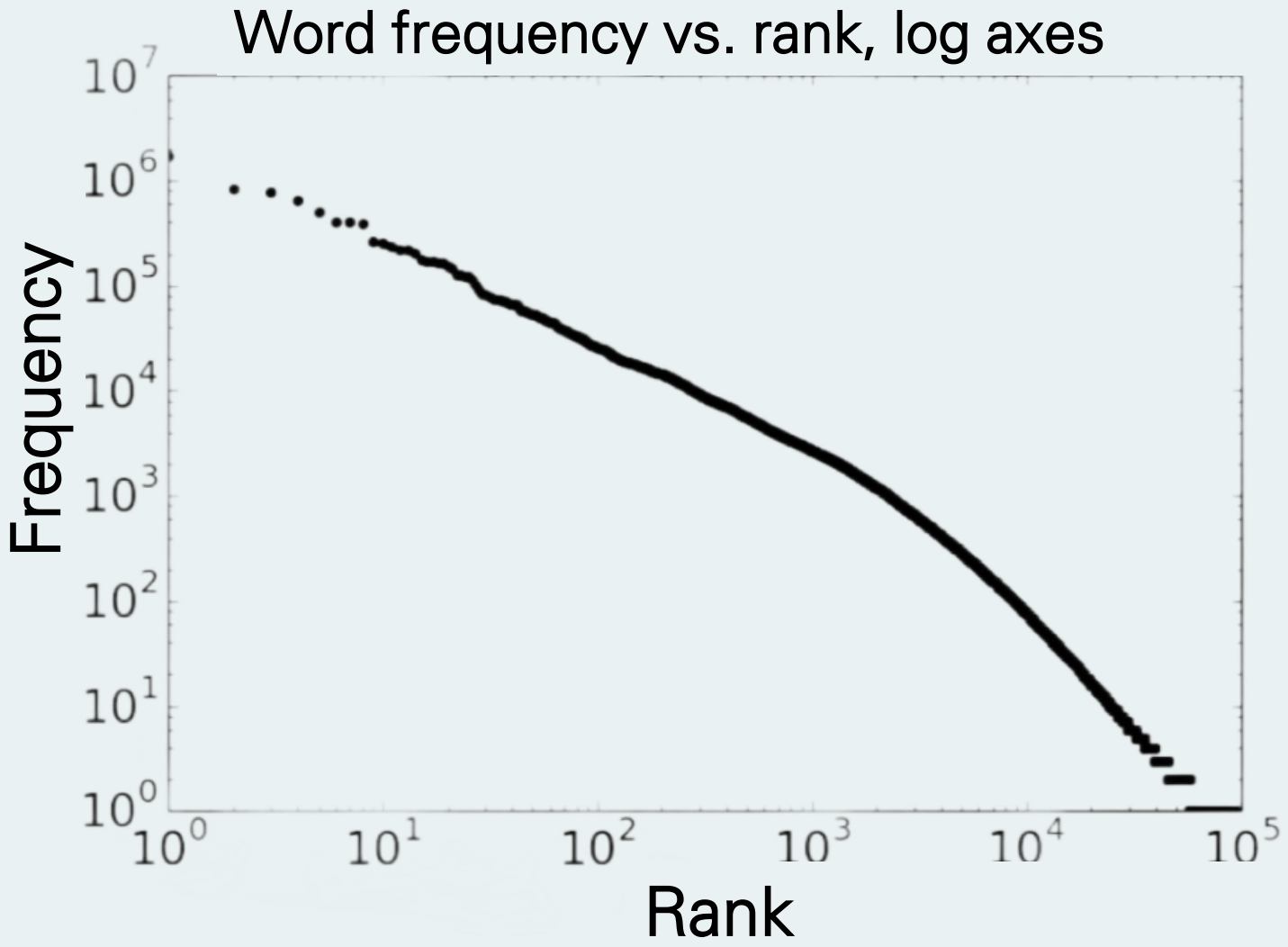


Zipf's law

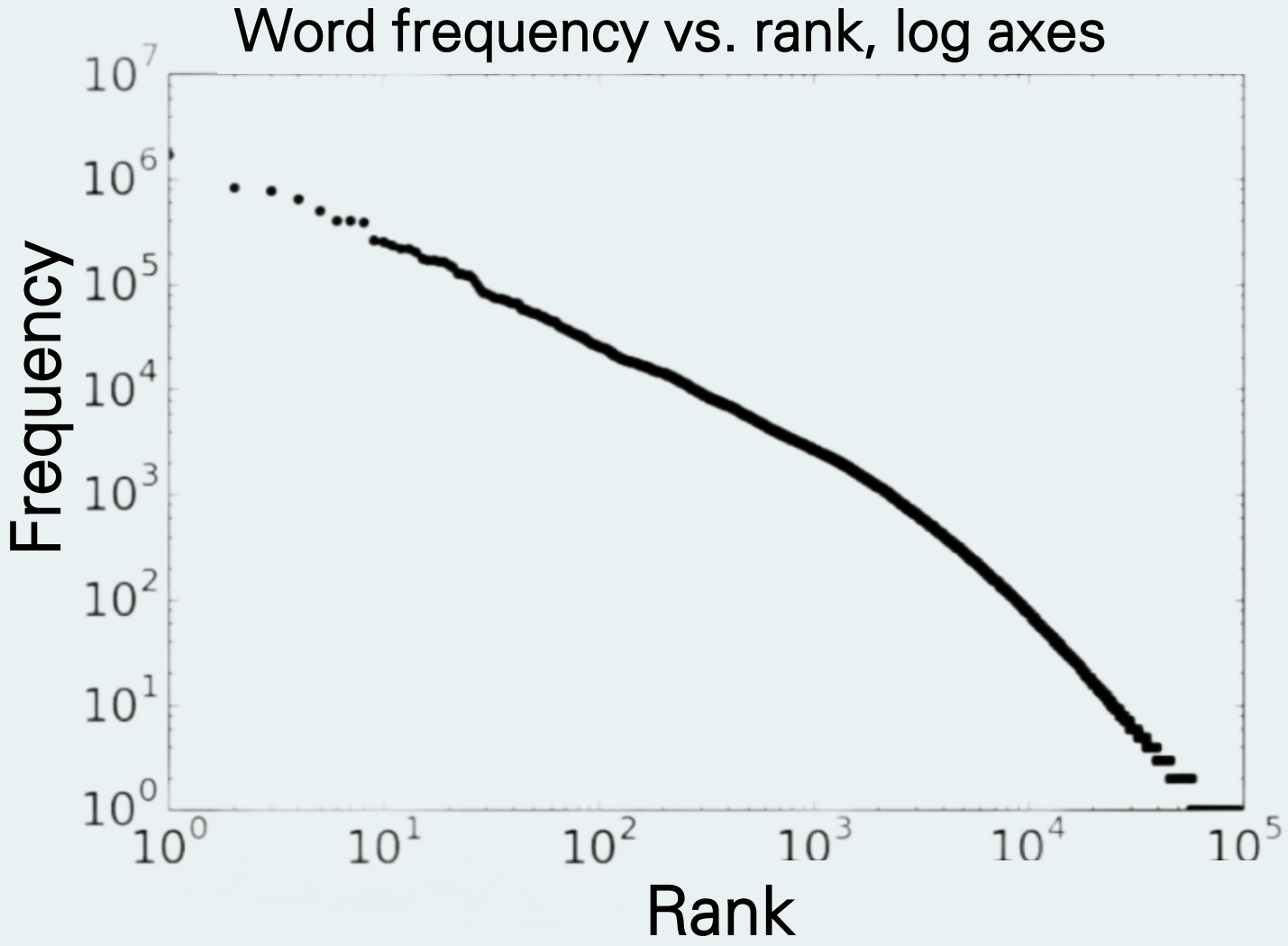
- Sparse data problem
- Example: the frequency of different words in a large text corpus

any word		nouns	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States

Zipf's law


Word frequency vs. rank, log axes
Frequency
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
10^0
10^1
10^2
10^3
10^4
10^5
Rank

Zipf's law



- Regardless of how large our corpus is, there will be a lot of infrequent words
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen