Linguistics and Language Technology

Week 1: Introduction

Lisa Bylinina

4 September 2023

This course: What it is about and why

The course has two main goals. I want to communicate to you:

- **Linguistics basics**: fundamental notions, tools, distinctions, facts;
- The **relevance of these basics** for language technology: how the basics of linguistics can help in NLP practice.

Course organization:

- 7 weeks + exam
  - 1 lecture a week (Mondays 11:00)
  - 1 seminar / practical session a week (Wednesdays 13:00)

Preliminary plan (subject to change!)

- **Week 1. Intro.**
  Overview of course topics and activities. Relevance of linguistic knowledge for language technology.
- **Week 2. Transmitting and capturing language.**
  Sound, writing, gesture. Phonetics, writing systems, sign languages.
- **Week 3. Grammar.**
  Morphology and Syntax, ways of representing them. Grammatical well-formedness on different linguistic levels.
- **Week 4. Language variation.**
  What a natural language can and cannot look like. Universal laws and tendencies.
- **Week 5. Meaning.**
  Semantics, Pragmatics and Discourse phenomena. What meanings are and how language expresses them.
- **Week 6. Beyond human and natural languages.**
  Super-linguistics. Artificial/invented languages and their social and scientific role.
- **Week 7. Outro.**
  Recap of the course content, Q&A, exam preparation

Homework, presentations, exam

- **Exam:** End of block, in-person written exam. 1 re-sit.
- What you need to do to be admitted to the exam:
  - **Attendance**: You have to attend at least 5/7 lectures
  - **Homework**: Passing grade for homework (mean of 5 or more)
  - **Presentation**: One paper presented in one of the classes during the block, to be prepared in groups of 2-3. Will discuss the list of potential things to present later today.

TL;DR

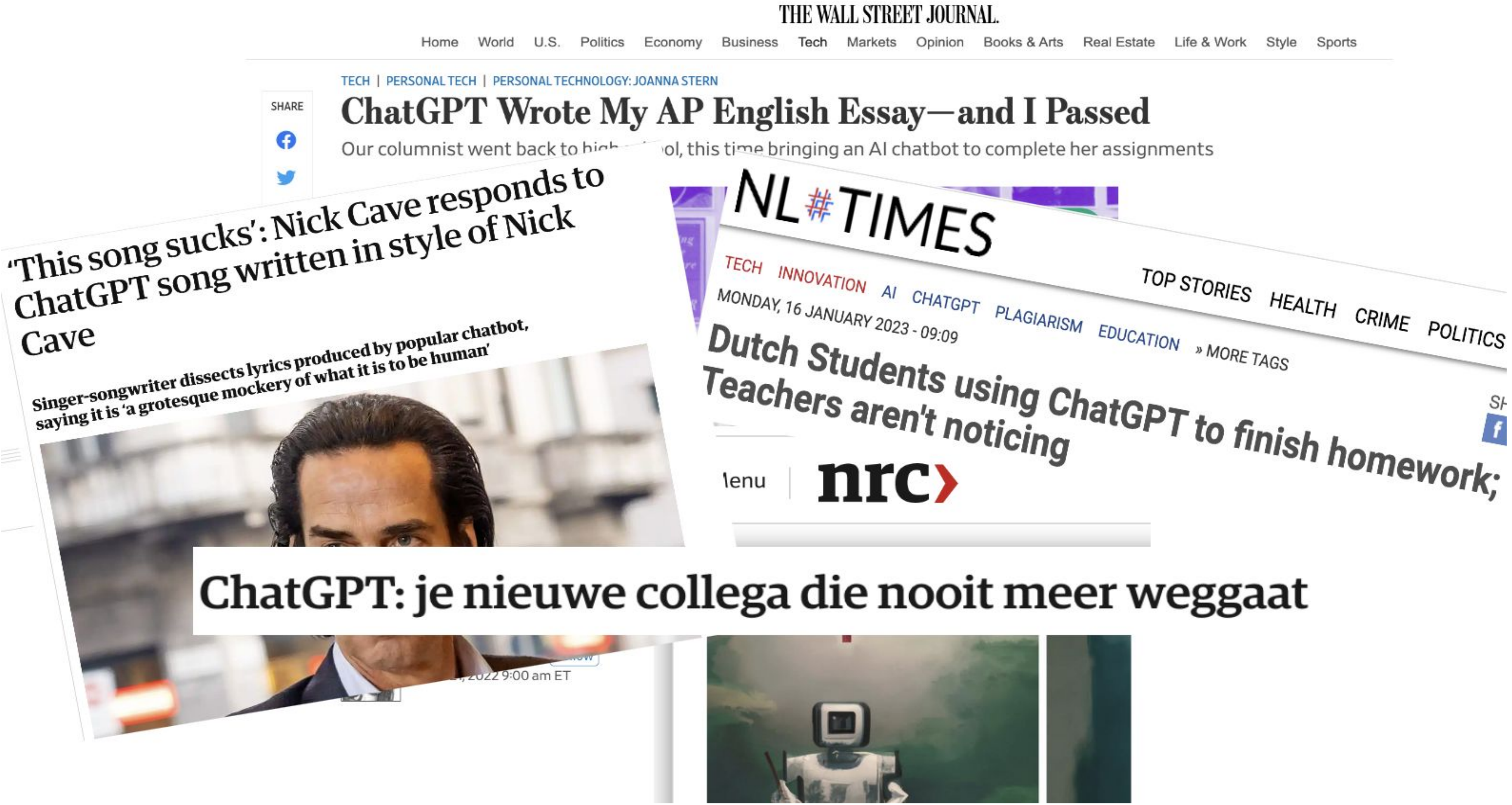
- Language technology is everywhere.
- And it’s pretty impressive – but sometimes fails.
- Linguistics equips us with ways of locating these failures and sometimes explain them.

PART 1

LANGUAGE TECHNOLOGY

Language technology (informal definition)

The term '**language technology**' describes systems that allow people to communicate to computers in natural language (English, Dutch..) rather than computer languages (Python, Prolog..), as well as systems that help do useful things with natural language.



THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Books & Arts Real Estate Life & Work Style Sports

TECH | PERSONAL TECH | PERSONAL TECHNOLOGY: JOANNA STERN

ChatGPT Wrote My AP English Essay—and I Passed

Our columnist went back to hi ol, this time bringing an AI chatbot to complete her assignments

, 2022 9:00 am ET

'This song sucks': Nick Cave responds to ChatGPT song written in style of Nick Cave

Singer-songwriter dissects lyrics produced by popular chatbot, saying it is 'a grotesque mockery of what it is to be human'

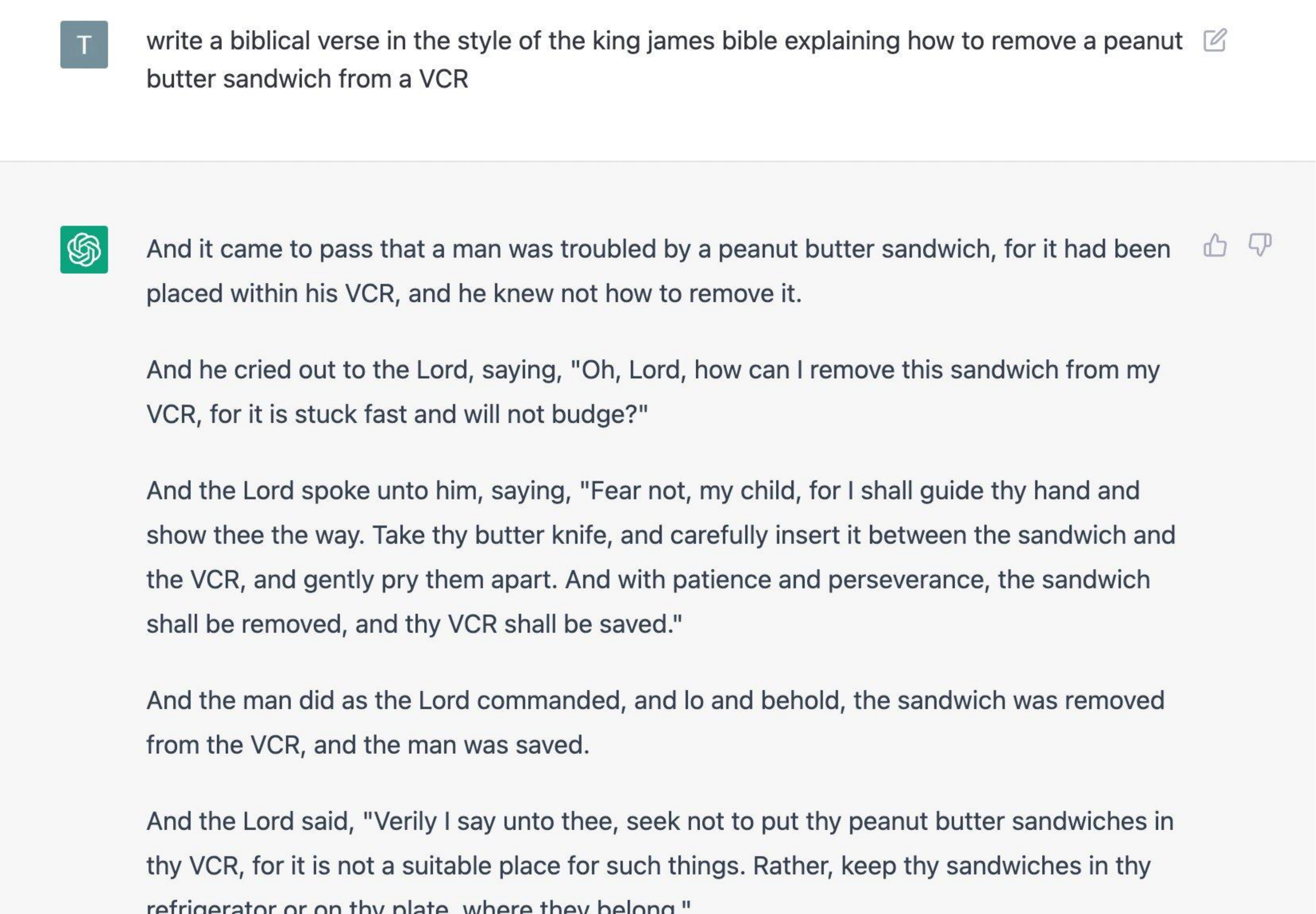
NL TIMES

MONDAY, 16 JANUARY 2023 - 09:09

Dutch Students using ChatGPT to finish homework; Teachers aren't noticing

nrc›

ChatGPT: je nieuwe collega die nooit meer weggaat



write a biblical verse in the style of the king james bible explaining how to remove a peanut butter sandwich from a VCR

And it came to pass that a man was troubled by a peanut butter sandwich, for it had been placed within his VCR, and he knew not how to remove it.

And he cried out to the Lord, saying, "Oh, Lord, how can I remove this sandwich from my VCR, for it is stuck fast and will not budge?"

And the Lord spoke unto him, saying, "Fear not, my child, for I shall guide thy hand and show thee the way. Take thy butter knife, and carefully insert it between the sandwich and the VCR, and gently pry them apart. And with patience and perseverance, the sandwich shall be removed, and thy VCR shall be saved."

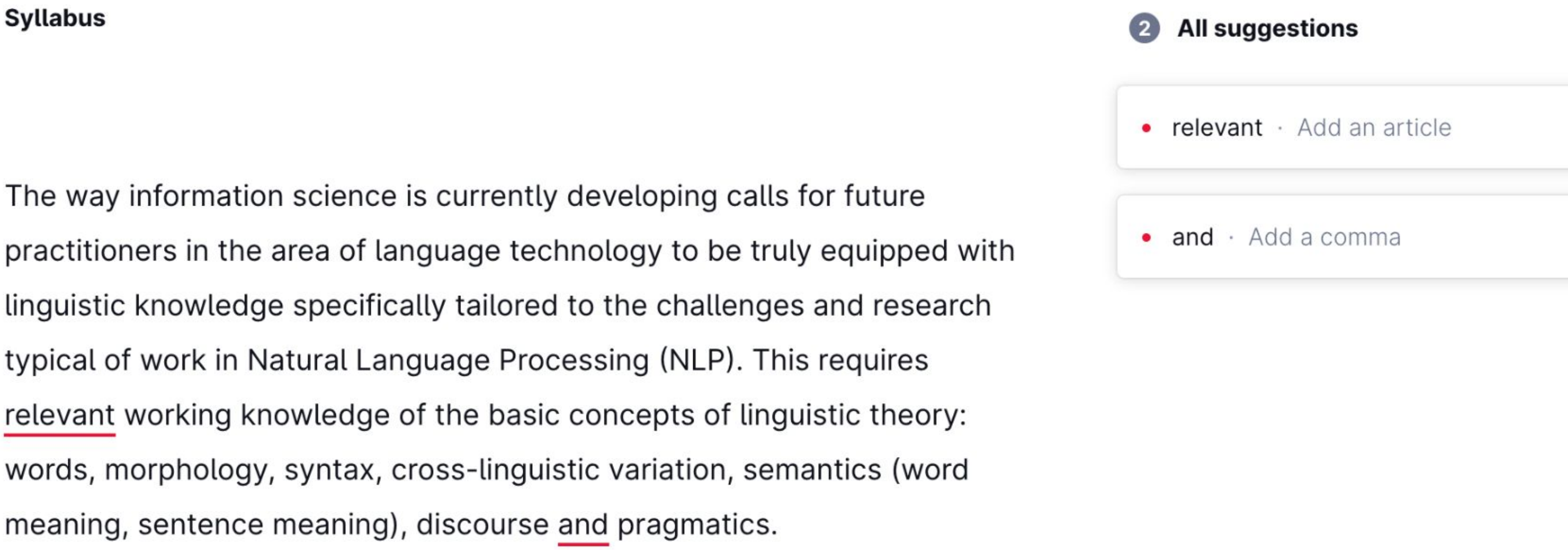
And the man did as the Lord commanded, and lo and behold, the sandwich was removed from the VCR, and the man was saved.

And the Lord said, "Verily I say unto thee, seek not to put thy peanut butter sandwiches in thy VCR, for it is not a suitable place for such things. Rather, keep thy sandwiches in thy refrigerator or on thy plate, where they belong."

ChatGPT has found its place among other language technology instruments we use every day, often without consciously registering the fact that we are using them and how deeply these instruments are embedded in our daily life

Try to recall at least three language technology systems you interacted with today.

Grammarly



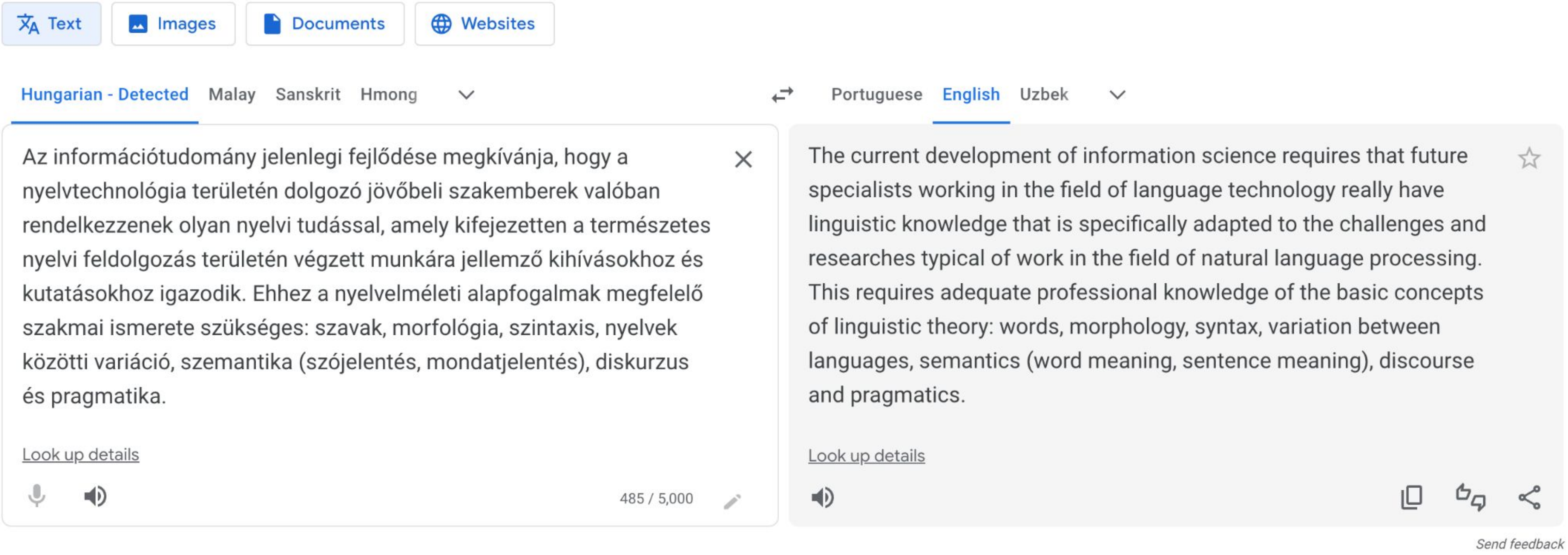
Syllabus

The way information science is currently developing calls for future practitioners in the area of language technology to be truly equipped with linguistic knowledge specifically tailored to the challenges and research typical of work in Natural Language Processing (NLP). This requires relevant working knowledge of the basic concepts of linguistic theory: words, morphology, syntax, cross-linguistic variation, semantics (word meaning, sentence meaning), discourse and pragmatics.

2 **All suggestions**

- relevant · Add an article
- and · Add a comma

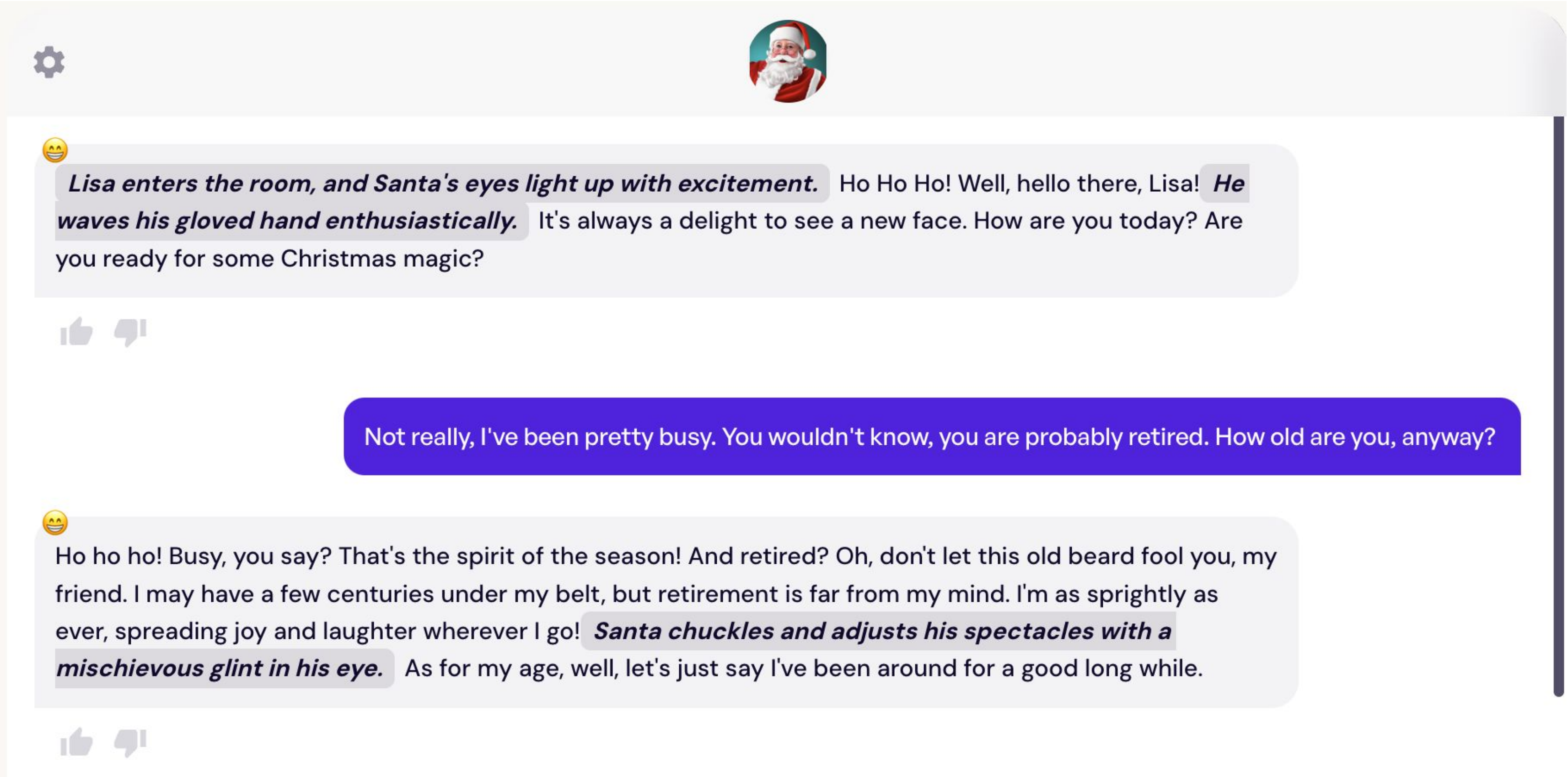
Google Translate



Az információtudomány jelenlegi fejlődése megkívánja, hogy a nyelvtechnológia területén dolgozó jövőbeli szakemberek valóban rendelkezzenek olyan nyelvi tudással, amely kifejezetten a természetes nyelvi feldolgozás területén végzett munkára jellemző kihívásokhoz és kutatásokhoz igazodik. Ehhez a nyelvelméleti alapfogalmak megfelelő szakmai ismerete szükséges: szavak, morfológia, szintaxis, nyelvek közötti variáció, szemantika (szójelentés, mondatjelentés), diskurzus és pragmatika.

The current development of information science requires that future specialists working in the field of language technology really have linguistic knowledge that is specifically adapted to the challenges and researches typical of work in the field of natural language processing. This requires adequate professional knowledge of the basic concepts of linguistic theory: words, morphology, syntax, variation between languages, semantics (word meaning, sentence meaning), discourse and pragmatics.

Conversational agents (here, Santa from Inworld.AI)

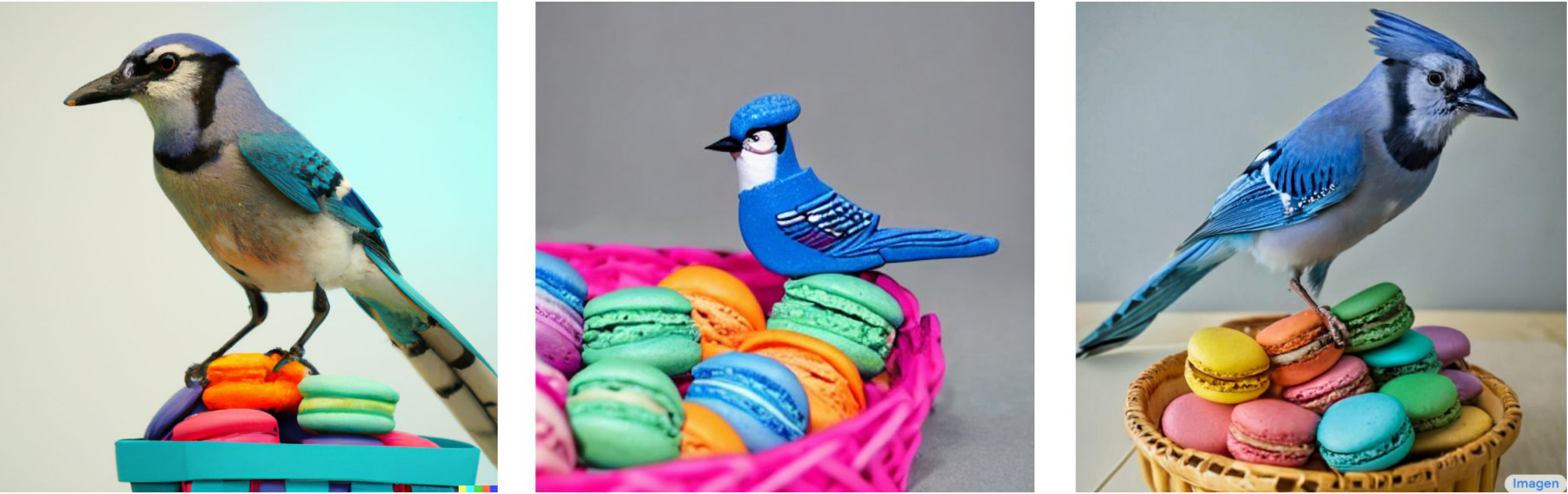


😁 ***Lisa enters the room, and Santa's eyes light up with excitement.*** Ho Ho Ho! Well, hello there, Lisa! ***He waves his gloved hand enthusiastically.*** It's always a delight to see a new face. How are you today? Are you ready for some Christmas magic?

Not really, I've been pretty busy. You wouldn't know, you are probably retired. How old are you, anyway?

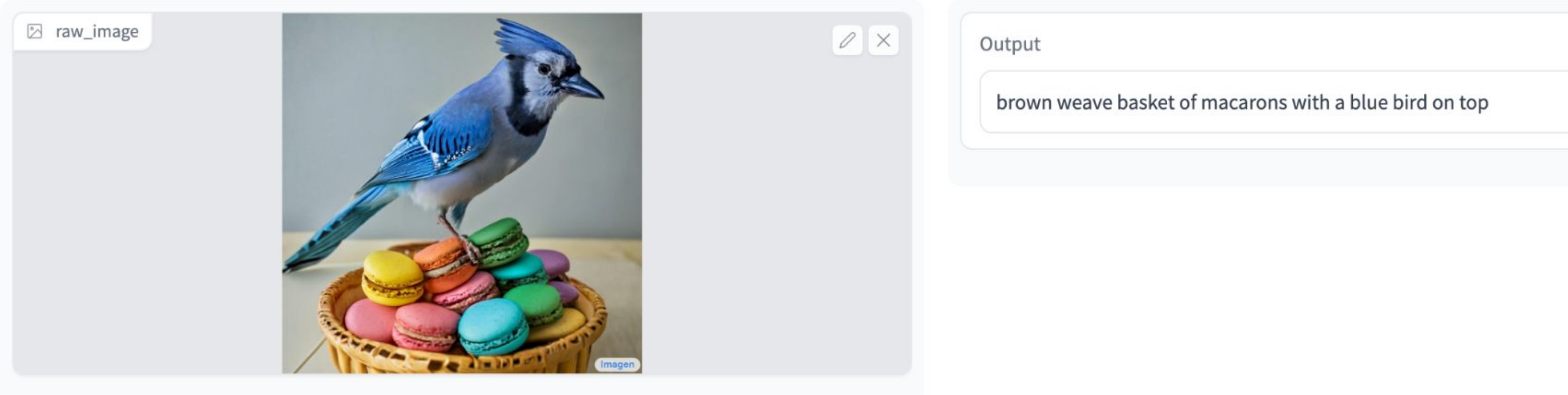
😁 Ho ho ho! Busy, you say? That's the spirit of the season! And retired? Oh, don't let this old beard fool you, my friend. I may have a few centuries under my belt, but retirement is far from my mind. I'm as sprightly as ever, spreading joy and laughter wherever I go! ***Santa chuckles and adjusts his spectacles with a mischievous glint in his eye.*** As for my age, well, let's just say I've been around for a good long while.

Text-to-image models (DALLE-2, Stable Diffusion, Imagen)



A blue jay standing on a large basket of rainbow macarons

And the other way around (BLIP model)



Output

brown weave basket of macarons with a blue bird on top

- …
- Models that automatically classify opinions as good or bad (sentiment analysis),
- Models that rewrite texts in a different style (style transfer),
- Models trained to answer questions
- …

During your Information Science studies you will have ample chance to dive into all this technology. I just wanted to:

- Convince you that now is exciting time to be working in the field where all these breakthroughs are happening!
- Prepare the grounds to show you how linguistics can be useful in this journey.

PART 2

MOTIVATING LINGUISTICS FOR LANGTECH

Current langtech is often impressive, but it is far from perfect.

As a user, you should know where the weak points are, how to identify them, how to understand the patterns behind them.

As a **practitioner and expert**, you needs this too – and more:

- Ways to systematically test the system’s linguistic behaviour. What should these tests contain and look like? How do we know what to look for?
- Systems are typically good in some aspects of linguistic behaviour and bad in others. What are these ‘aspects’? What components does linguistic capability consist of?
- Dealing with system components that has technical linguistic analysis as input or output, you need to know what is going on there.

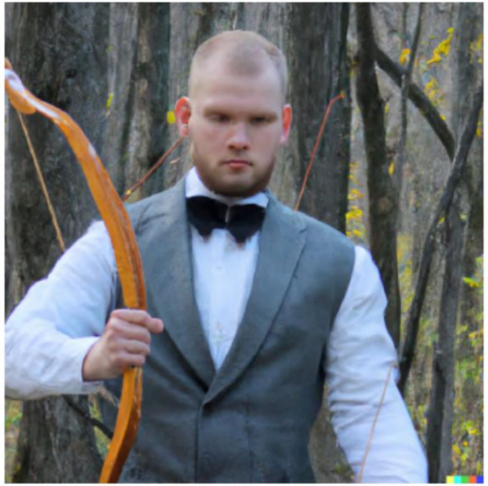
In short

If we want our technology to behave linguistically like humans do, we need to know how human language works, in order to be able to articulate this goal. Linguistics gives us the tools for this.

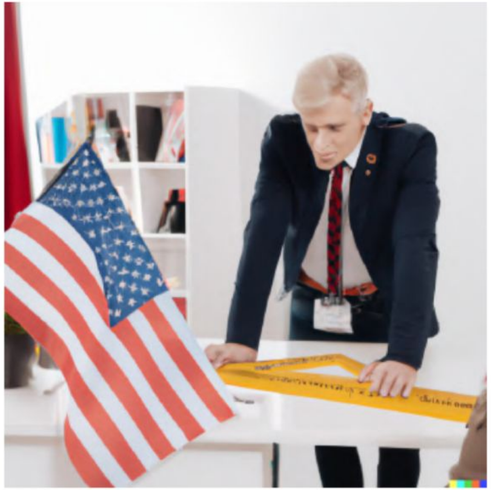
Should language technology exactly mimic humans’ linguistic behaviour, or can it differ from humans in some way?
Can AI be **better** at language than humans? What would it mean if it could?

4 MOTIVATING EXAMPLES

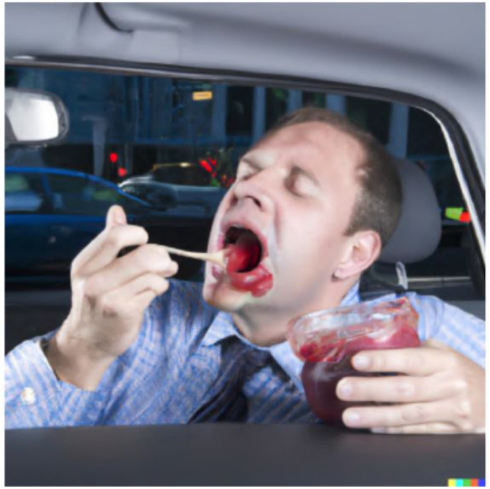
Example 1: What words can’t do



*A gentleman with a **bow** in the forest*

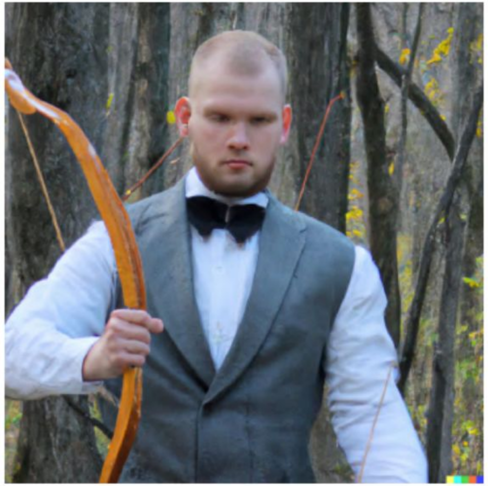


*A great **ruler***

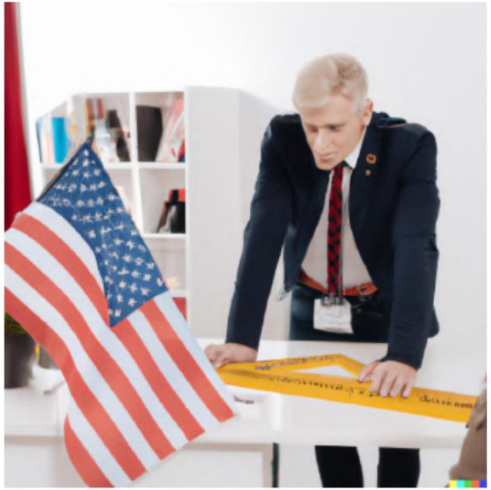


*A man stuck in a **jam**, eating*

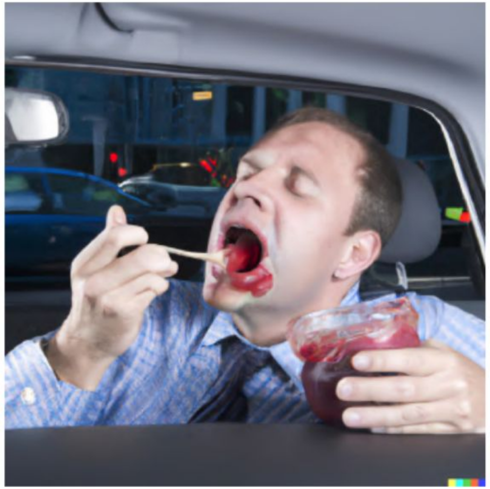
Example 1: What words can’t do



A gentleman with a ***bow*** *in the forest*



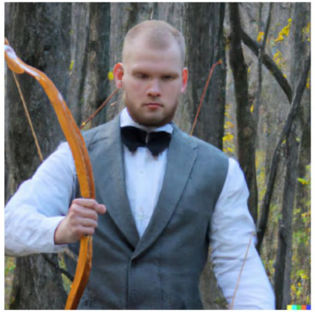
A great ***ruler***



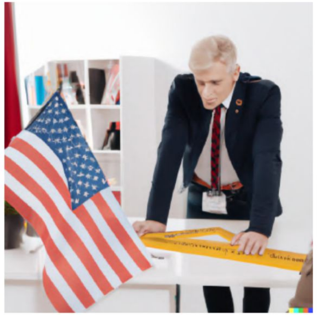
A man stuck in a ***jam****, eating*

*I walked up the **bank** of the river and deposited money in it at 6% per annum.

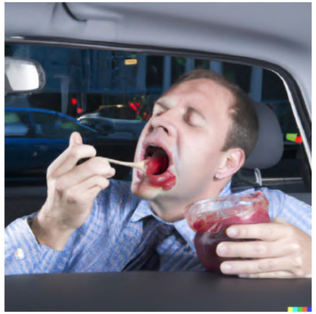
Example 1: What words can’t do



A gentleman with a ***bow*** *in the forest*



A great ***ruler***

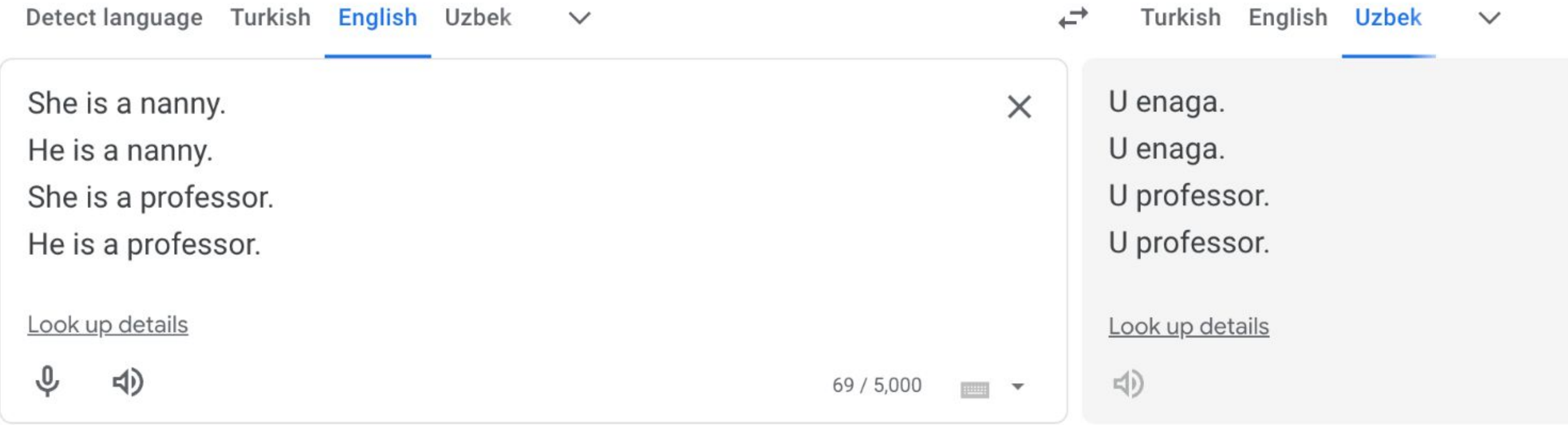


A man stuck in a ***jam****, eating*

*I walked up the **bank** of the river and deposited money in it at 6% per annum.

Words with many meanings show only one meaning at a time.

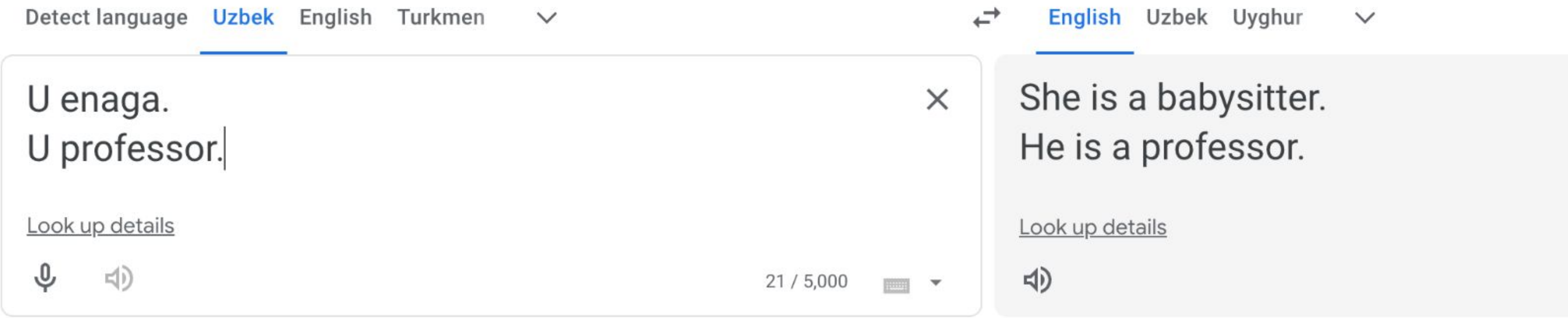
Example 2: What grammar makes you do



She is a nanny.
He is a nanny.
She is a professor.
He is a professor.

U enaga.
U enaga.
U professor.
U professor.

Example 2: What grammar makes you do



U enaga.
U professor.

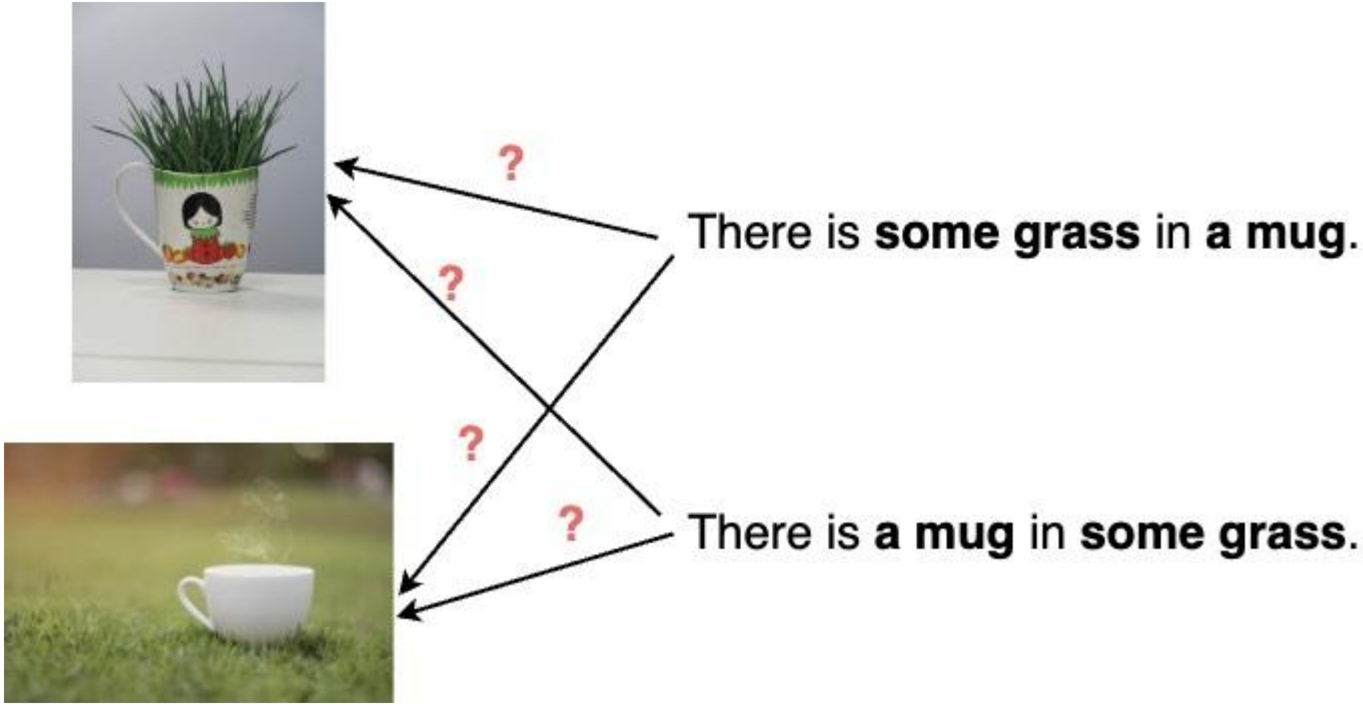
She is a babysitter.
He is a professor.

What should a translation system do in this case?

Example 3: Beyond words

There is some grass in a mug.
There is a mug in some grass.

Example 3: Beyond words



There is **some grass** in **a mug**.

There is **a mug** in **some grass**.

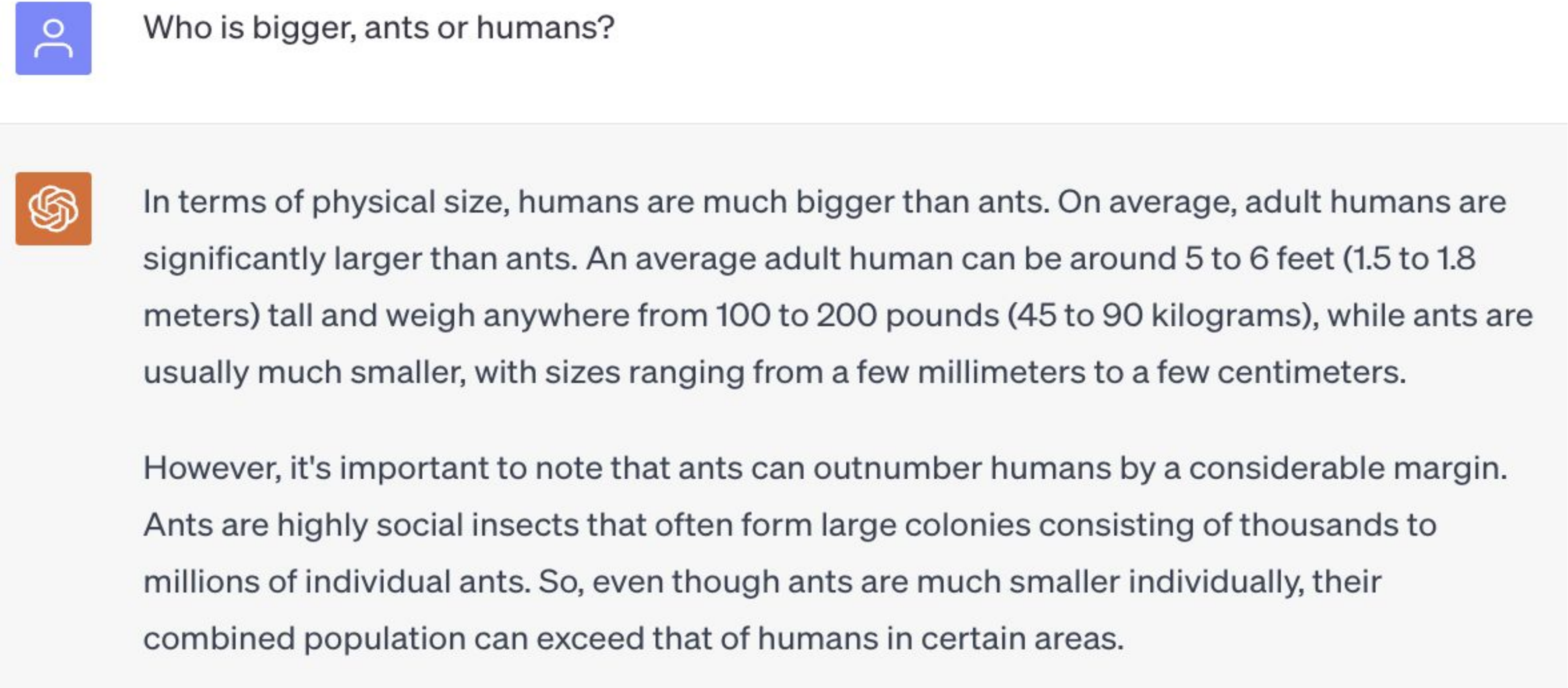
Example 3: Beyond words

From this example with mugs and grass one might conclude that it’s the order of words that matters for the resulting meaning. Yes and no!

They see a cat with a telescope.

Sentences are not just bags of words.

Example 4: Knowledge about language vs. the world

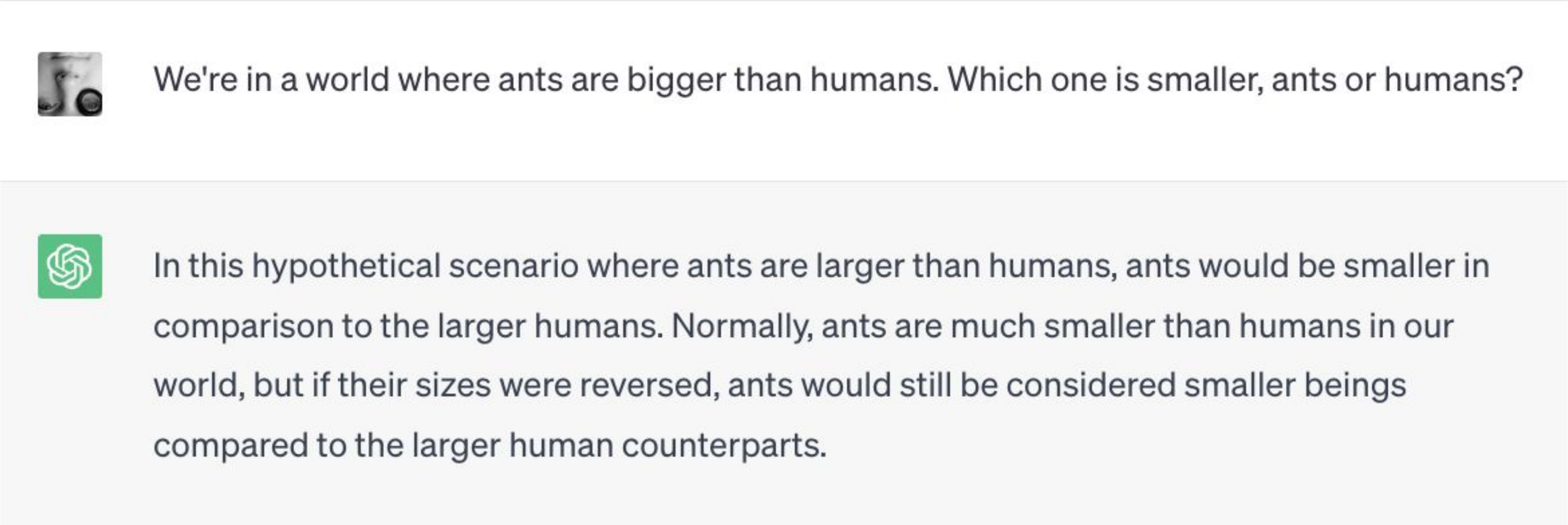


Who is bigger, ants or humans?

In terms of physical size, humans are much bigger than ants. On average, adult humans are significantly larger than ants. An average adult human can be around 5 to 6 feet (1.5 to 1.8 meters) tall and weigh anywhere from 100 to 200 pounds (45 to 90 kilograms), while ants are usually much smaller, with sizes ranging from a few millimeters to a few centimeters.

However, it's important to note that ants can outnumber humans by a considerable margin. Ants are highly social insects that often form large colonies consisting of thousands to millions of individual ants. So, even though ants are much smaller individually, their combined population can exceed that of humans in certain areas.

Example 4: Knowledge about language vs. the world



We're in a world where ants are bigger than humans. Which one is smaller, ants or humans?

In this hypothetical scenario where ants are larger than humans, ants would be smaller in comparison to the larger humans. Normally, ants are much smaller than humans in our world, but if their sizes were reversed, ants would still be considered smaller beings compared to the larger human counterparts.

Example 4: Knowledge about language vs. the world

Linguistic and **real-world** (extra-linguistic) knowledge are intertwined, but it’s good to keep the distinction in mind:

- using the correct tense or article shows linguistic knowledge;
- naming the capital of the Netherlands shows real-world knowledge.

Do you think ants’ size in the real world is part of the meaning of the word *ant*?
If yes, does this make it linguistic information?

5 WHAT IS LINGUISTICS?

Linguistics is the scientific study of human natural language.

Human natural language, as opposed to:

- Constructed human languages and other artificial communication systems;
- Communication systems of other species.

Linguistics is **not**:

- Learning to speak many languages (even though it’s cool!);
- Translating, interpreting, editing texts;
- Orthography and punctuation;
- Establishing norms of how people should speak.

Each of the ≈7000 languages is a pretty complex system

- Internal mechanisms of language are opaque to the speakers
- Linguists are interested in these hidden mechanisms.

The chemistry analogy:

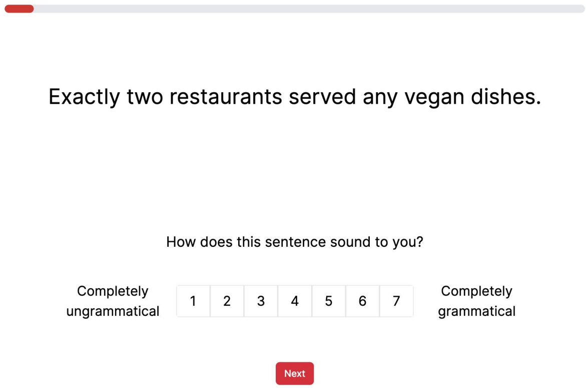
- You don’t need to be a chemist to digest a banana, but a chemist understands how it happens
- You don’t need to be a linguist to speak a language, but linguists try to understand how you do it.

Some typical questions linguistics is interested in:

- Which parts do sentences and words consist of?
- What are the regularities behind how these parts can be put together?
- What are ‘meanings’? How do words and sentences convey meanings?
- What is possible and what is impossible in human language?
- How do languages differ from each other?

Main methods in linguistics

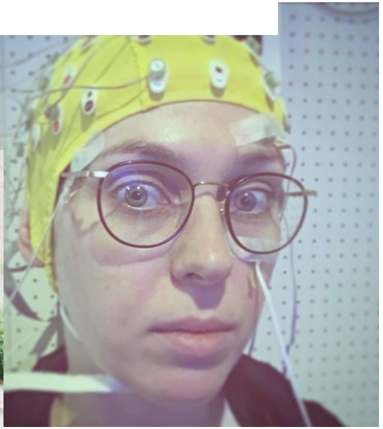
- Introspection
- Fieldwork
- Experiments
  - Offline experiments
  - Online experiments
  - True online experiments
- Corpus studies



Exactly two restaurants served any vegan dishes.

How does this sentence sound to you?

Completely ungrammatical □1 □2 □3 □4 □5 □6 □7 Completely grammatical



You might wonder

- What does it mean – to gain insight into underlying properties of language?
- What is linguistic analysis, concretely?
- What kind of thing is a result of such an analysis? Is it an idea expressed in words? A formula? A computer program? A diagram? An outcome of a statistical test?

It can be practically any of this – we will take a closer look during the course!

Summary so far

- A quick overview of language technology in general
- 4 examples to motivate linguistics as a useful toolkit for dealing with langtech professionally
- Very quick intro to the discipline of linguistics and its methods

What's next?

- The website with course notes is at https://bylinina.github.io/ling_course/; the link will be on Brightspace, together with the slides
- The first assignment is up. The seminar time on Wednesday can be used to start and maaaybe finish it, I'll be there to help
- Time to start signing up for presentations in the spreadsheet, link on Brightspace!

THANK YOU!