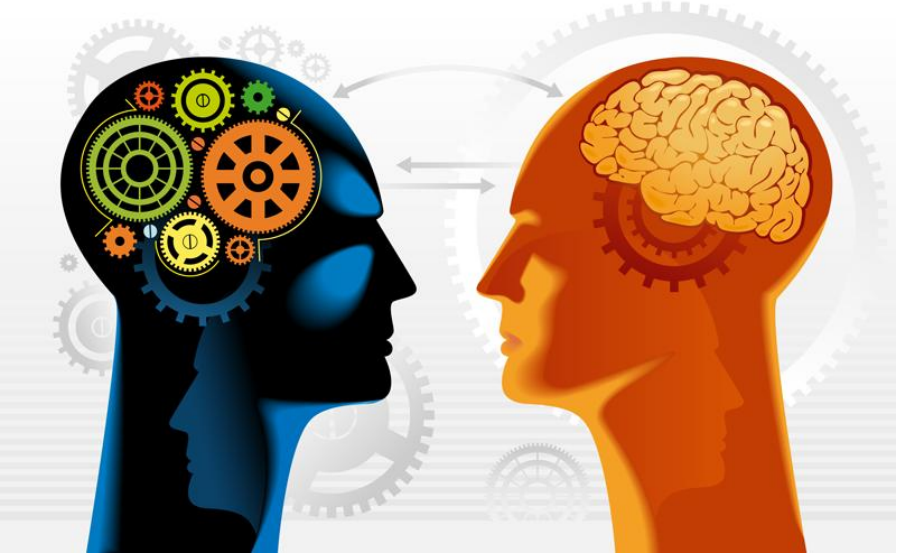
University of Ljubljana, Faculty of Computer and Information Science

Natural language processing – an introduction

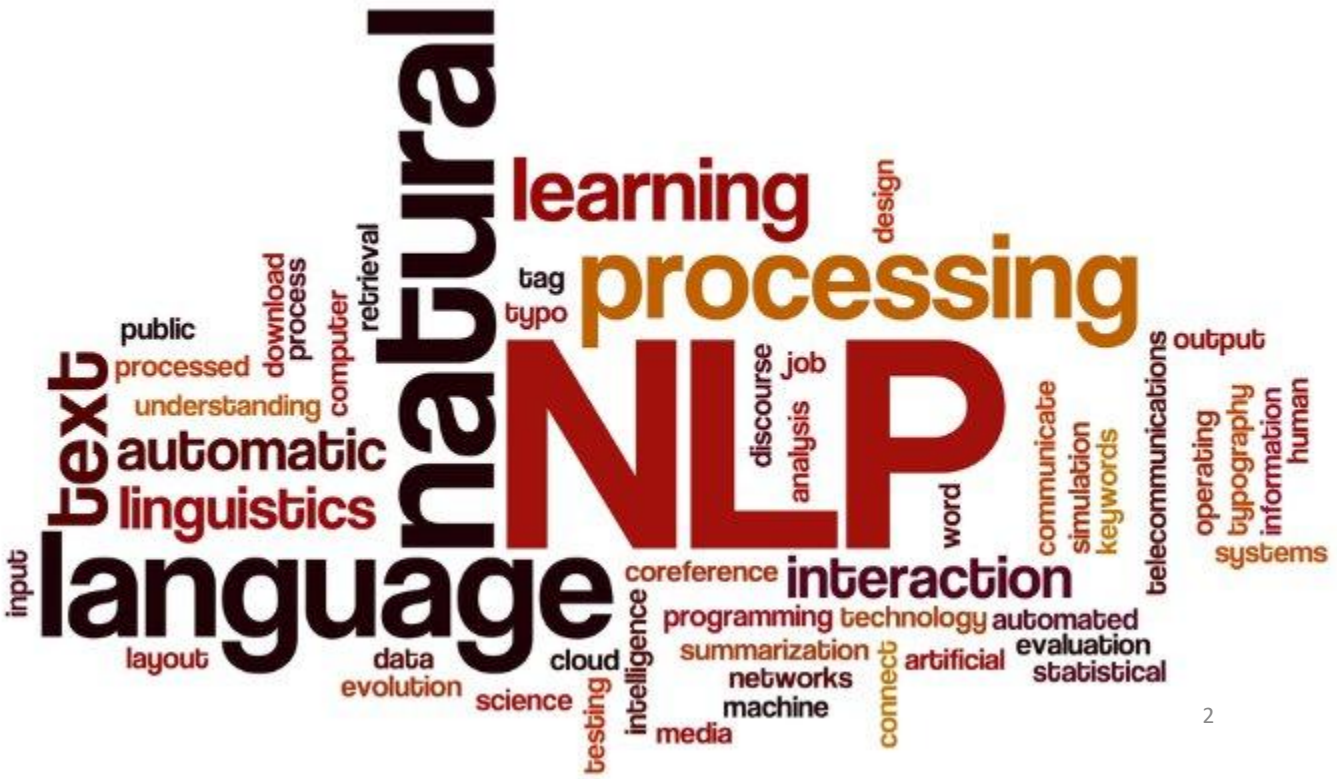


Prof Dr Marko Robnik-Šikonja

Edition 2026

Outline

- about the course
- contents overview
- language understanding and intelligence
- overview of the NLP area



Lecturer

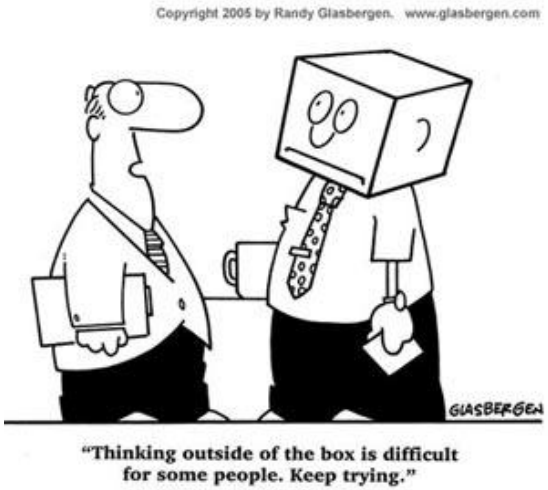
- Prof Dr Marko Robnik-Šikonja
- University of Ljubljana
  Faculty of Computer and Information Science
  Head of the Machine Learning and Language Technologies Lab
- UL FRI, Večna pot 113, room 2.06, 2nd floor, right from the elevator
- marko.robnik@fri.uni-lj.si
- https://fri.uni-lj.si/en/employees/marko-robnik-sikonja
- tel: (01) 4798 241
- Contact hour
  - Wednesday, 10:00 -11:00; email me for confirmation, other slots, and Zoom
- **Research interests**: artificial intelligence, machine learning, natural language processing, network analytics, data science
- **Teaching:** courses from the area of machine learning, natural language processing, and algorithms
- **Research principle:** open science

Assistants

- Assist Prof Dr Slavko Žitnik
  slavko.zitnik@fri.uni-lj.si
- Laboratory for Data Technologies
- Research interests: NLP, information retrieval, semantic web

- Assist. Aleš Žagar
  ales.zagar@fri.uni-lj.si
- Machine Learning and Language Technologies Lab
- Research interests: NLP, knowledge injection into neural networks

Goals of the course



- students shall become acquainted with
  - basics of natural language processing and understanding
  - basic approaches and data representations for NLP
  - modern techniques for NLP
  - selected relevant NLP tasks
  - relevant research challenges in the area of NLP and NLU, computational linguistics, semantics, and LLMs
- teach students a practical use of
  - practical approaches to NLP
  - selected tools
  - selected modern techniques for NLP
- awareness of ethical issues in NLU
- increase the (mental) problem-solving toolbox with new NLP approaches and techniques
- awareness of languages as important sources of information

Learning outcomes

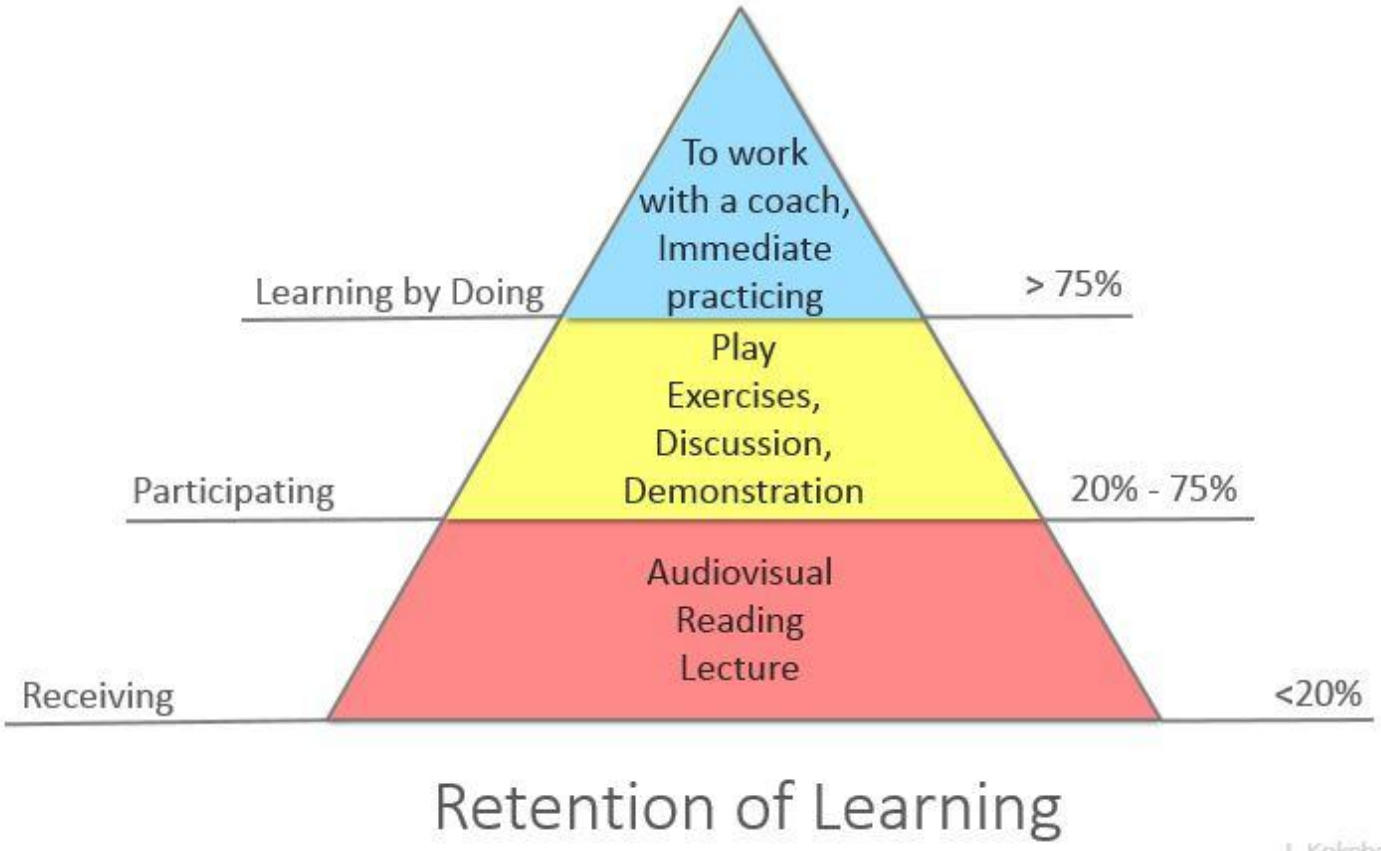
Upon completion of the course, students shall:

- understand approaches to syntax and semantics in NLP,
- use and adapt machine learning techniques for NLP
- apply and critically evaluate natural language processing tools
- know the existing language resources and be able to design new ones
- use text representations and adapt them to new contexts
- use and evaluate approaches to text classification, summarization, machine translation, affective computation, question answering, reasoning, information retrieval, etc.

Lectures and tutorials

- Lectures
  - introduction to the topic, discussion
  - some examples
  - broader view of the topics
- Tutorials
  - exercises
  - assignments motivated by practical use
  - assistant presents the assignments, helps with tips, moderates discussion, so…
  - …come prepared and pose questions.
  - introduce some problem solving tools and useful software
  - mostly deals with English and Slovene

BTW: retention of learning



Syllabus 1/2

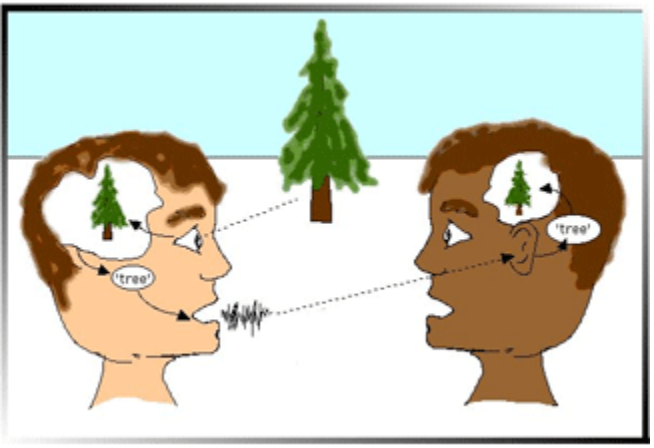
1. Introduction to natural language processing: motivation, language understanding, ambiguity, traditional, statistical, and neural approaches.
2. Text preprocessing: normalization, string similarity, lemmatization, regular expressions, and grammars.
3. Text similarity: measures, clustering approaches, language networks, and graphs.
4. Language resources: corpora, datasets, dictionaries, thesauri, semantic databases, WordNet.
5. Text representation: sparse and dense embeddings, cosine distances, language models.
6. Deep neural networks for text: recurrent neural networks, CNNs for text, transformers.
7. Neural embeddings: word2vec, fastText, BERT, cross-lingual embeddings, sentence, and document embeddings.

Syllabus 2/2

8. Large pretrained language models: BERT, GPT, and T5 families, multimodal models.
9. Improving LLMs: instruction tuning, preference alignment, and test-time computation.
10. Inference with LLMs: prompt engineering, retrieval augmented generation, function calling, and agentic LLMs.
11. Linguistic tasks: part-of-speech tagging, dependency parsing, named entity recognition, word sense disambiguation.
12. Affective computing: sentiment, emotions, etc.
13. Popular tasks: text summarization, question answering, machine translation, and paraphrasing.
14. Reasoning tasks: natural language inference, commonsense reasoning, and logical reasoning.
15. Advanced topics: long context in LLMs, alternatives to transformers.

What the course does not cover?

- Speech processing: recognition and synthesis (a course in UL Faculty of Electrical Engineering, „Speech technologies“) to be included in the future MSc module “Language technologies”
- Linguistics



Prerequisites

- Recommended knowledge
  - Python programming,
  - probability and statistics,
  - Basics of machine learning

Obligations

- 5 quizzes checking continuous understanding of contents
- projects, composed of two to three stages, 50 points
- a written exam, 50 points

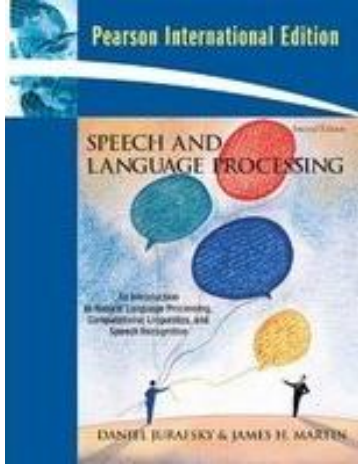
Grading

Obligation	% of total	subject to
Five quizzes	0%	≥ 50% altogether
Projects	50%	≥ 50% of points
Written exam	50%	≥ 50% of points

Learning materials

- learning materials in the Moodle classroom http://ucilnica.fri.uni-lj.si
- slides are updated continuously
- links to the literature, code, etc.
- Labs on the github

Literature



- Jurafsky, David and Martin, James H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 3rd edition* draft, 2026. Basic course literature; available on the authors' webpages

- Simon J.D. Prince: *Understanding Deep Learning. MIT Press, 2023* *https://udlbook.github.io/udlbook/* This recent book intuitively explains many important concepts and techniques in deep learning.

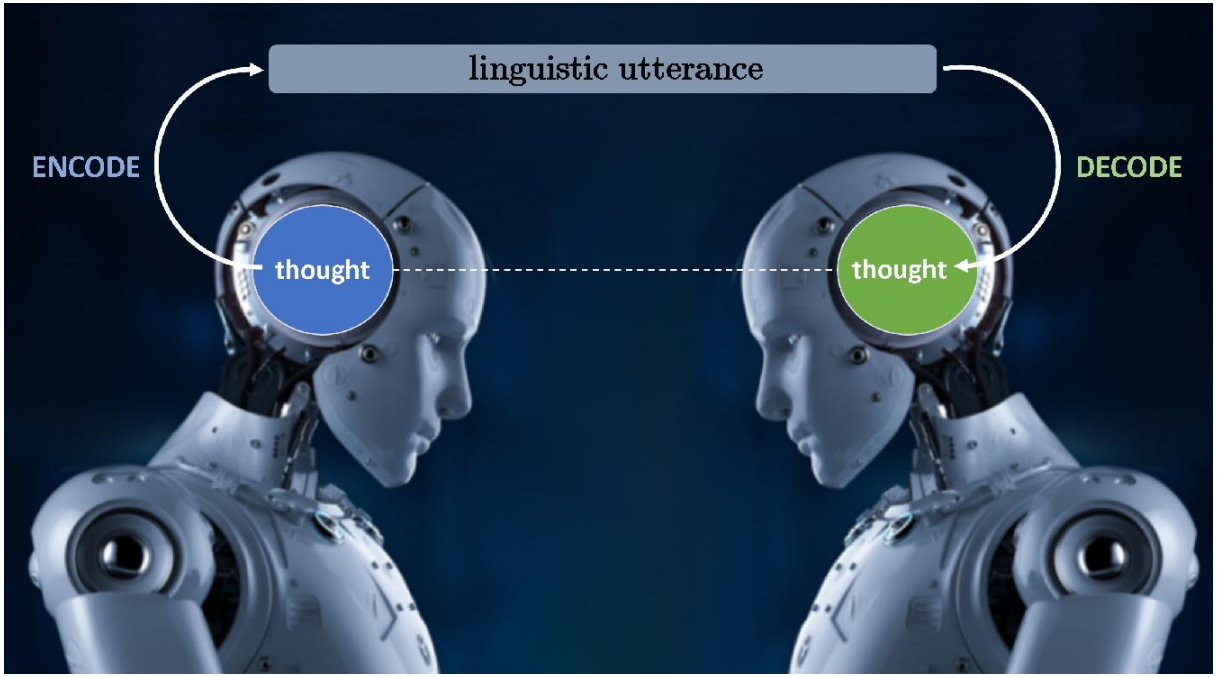
- scientific papers, blogs, Python notebooks for a deeper understanding of certain topics

Two views of natural language processing

- Techniques for language **processing**: syntax, grammars, language resources, text representation, text classification
- Attempts to **understand** language: semantics and pragmatics of language, related to the goals of artificial (general) intelligence

Understanding

Walid Saba, "Machine Learning Won't Solve Natural Language Understanding", The Gradient, 2021.

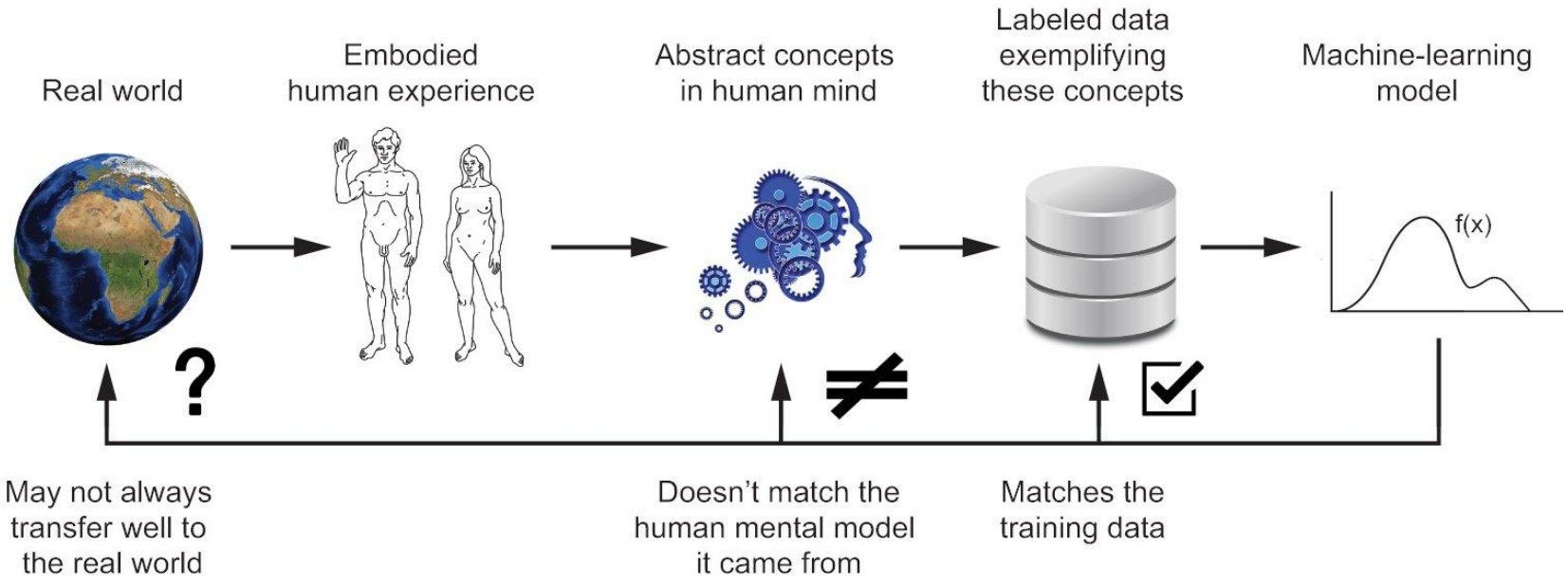


Xanadu, who is a living young human adult, and who was in graduate school, quit graduate school to join a software company that had a need for a new employee.

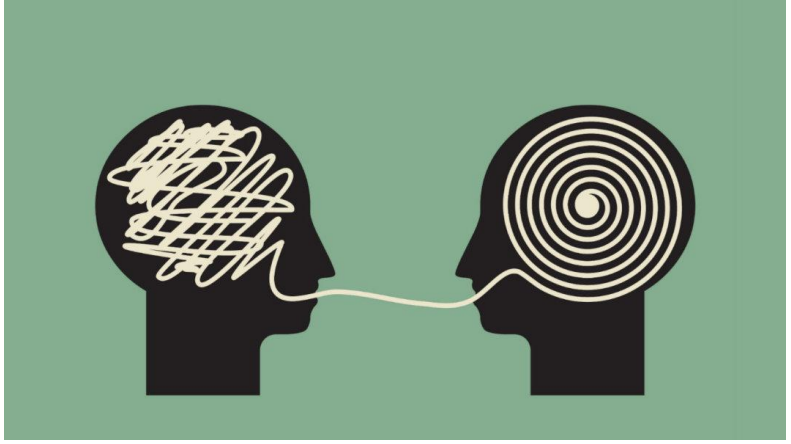
≈

Xanadu quit graduate school to join a software company.

Understanding ML models is difficult



Understanding language

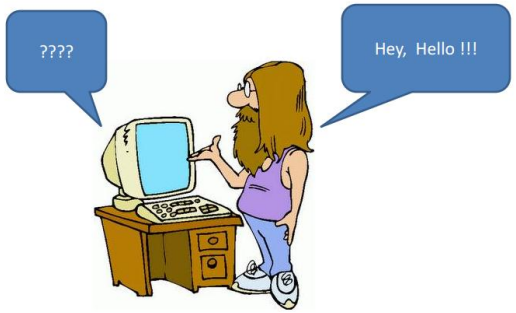


- A grand challenge of (not only?) artificial intelligence

  Who can understand me?
  Myself I am lost
  Searching but cannot see
  Hoping no matter cost
  Am I free?
  Or universally bossed?

- Not just poetry, what about instructions, user manuals, newspaper articles, seminary works, internet forums, twits, legal documents, i.e. license agreements, etc.

Understanding NL by computers



- Understanding words, syntax, semantics, context; writer's intentions, knowledge, background, assumptions, biases, etc.
- Ambiguity in language
  - Newspaper headlines - intentional ambiguity – clickbaits
    - Juvenile court to try shooting defendant
    - Kids make nutritious snacks
    - Miners refuse to work after death
    - Doctor on Trump‘s health: No heart, cognitive issues

Ambiguity

- I made her duck.

- Possible interpretations:
  - I cooked waterfowl for her.
  - I cooked waterfowl belonging to her.
  - I created the (plaster?) duck she owns.
  - I caused her to quickly lower her head or body.
  - I waved my magic wand and turned her into undifferentiated waterfowl.

- Spoken ambiguity
  - eye, maid

Syntax ambiguity

- Syntactic ambiguity

  Flying planes can be dangerous.

  - flying can be interpreted as an adjective modifying planes

    Planes that are flying can be dangerous.

  - or as a verb in gerundive form

    It can be dangerous to fly planes.

- Word ambiguity

  The bat flew through the air.

- Unclear reference of a word or phrase

  The boy and the dog were playing in the park. He ran into a tree.

- more examples

  John went to the bank.

Semantic ambiguity

- The girl told the story cried.
- Put the box on the table in the kitchen.
- Bring your old car seat to be recycled.

Disambiguation

- in search queries: jaguar, Paris
- user profiles

- POS tagging,
- word sense disambiguation
- probabilistic parsing
- speech act interpretation, e.g., a statement or a question:
  - We made it. We made it?

Linking

Linking refers to the ability of a reader to connect units of information on the word, sentence, or discourse level. One example called in syntactic theory a “self-embedded structure.” E.g.,

The boy the girl the men left watched then left.

- Which noun phrase (the boy, the girl, the men) is linked with each of the verbs (left, watched, left)

- Valid also for other aspects of texts. For example, narratives can contain stories embedded within stories that are in turn embedded within stories. This can make it difficult for readers to link together units of information so that they can understand the text
- Readability of a text is determined with several linguistic factors (syntactic, semantic, morphological, and discourse).

Ambiguity and humor

- collection of linguistic humor by Beatrice Santorini, e.g., recommendation letters
- If you have to write a letter of recommendation for a fired employee, here are a few suggested phrases.

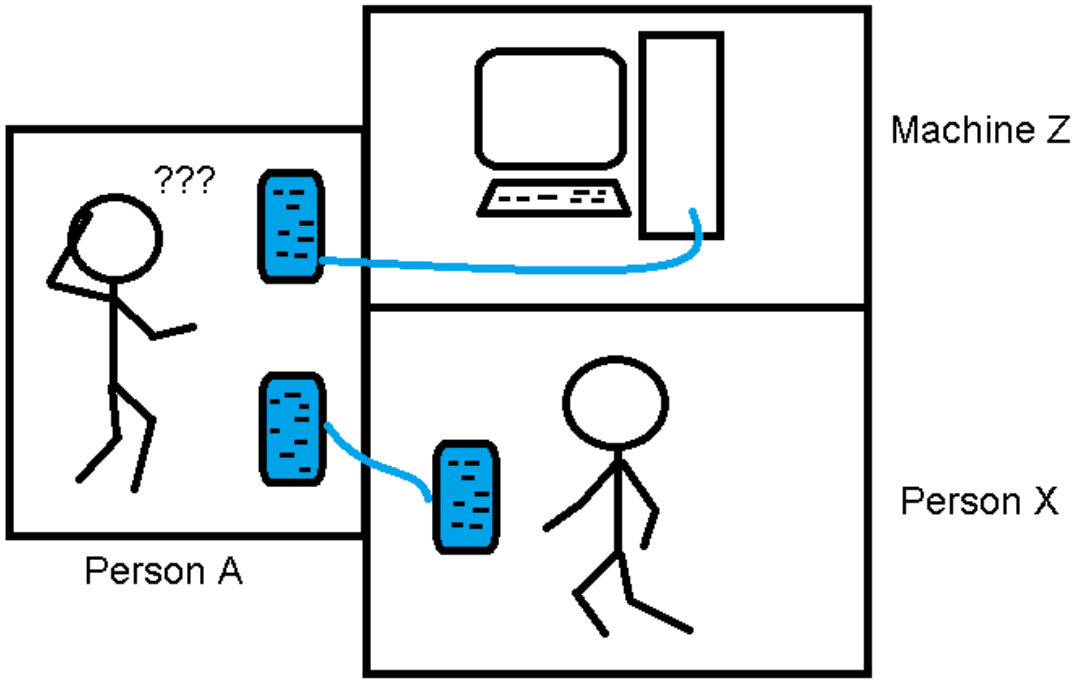
Lexical ambiguity	
For a chronically absent employee	A man like him is hard to find.
For a dishonest employee	He's an unbelievable worker.
For a lazy employee	You would indeed be fortunate to get this person to work for you.
For the office drunk	Every hour with him was a happy hour.
Structural ambiguity	
For a chronically absent employee	It seemed her career was just taking off.
For a dishonest employee	Her true ability was deceiving.
For a stupid employee	I most enthusiastically recommend this candidate with no qualifications whatsoever.
For the office drunk	He generally found him loaded with work to do.
Scope ambiguity	
For an employee who is not worth further consideration as a job candidate	All in all, I cannot say enough good things about this candidate or recommend him too highly.
For an employee who is so unproductive that the job is better left unfilled	I can assure you that no person would be better for the job.
Other	
For a lazy employee	He could not care less about the number of hours he has to put in.
For an employee who is not worth further consideration as a job candidate	I would urge you to waste no time in making this candidate an offer of employment.
For a stupid employee	There is nothing you can teach a man like him.

Understanding language as (artificial) intelligence test

- The system understanding natural language on human level should contain a large portion of human knowledge and would achieve the most important goals of artificial intelligence. Its response would be indistinguishable from human's.

Turing test

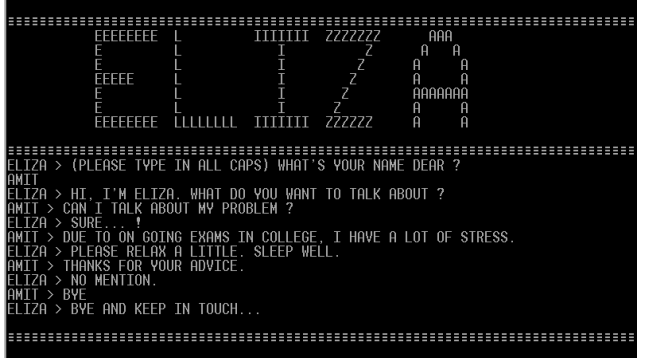
- Who is Person A talking to, a machine or a person?



Chatbots attempt to solve Turing test

- CYC, ELISA, ALLICA, chatterbot,
- VIDA, TIA,
- Siri (knowledge recognition, context detection, web service recommendation)
- ChatGPT, You.com, Gemini, DeepSeek

ELIZA



- ELIZA (Joseph Weizenbaum, 1966)
- limited conversation (Rogerian psychiatrist)

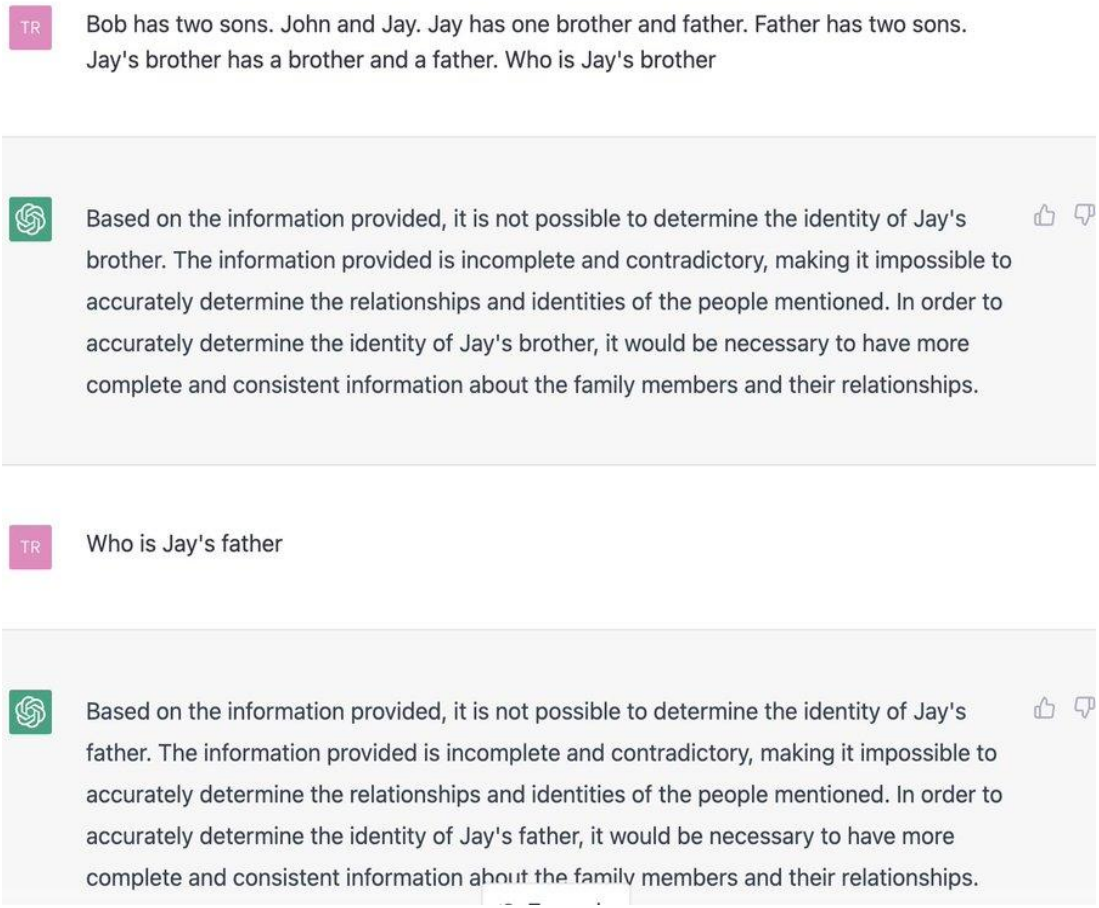
User: You are like my father in some ways.
ELIZA: What resemblance do you see?
User: You are not very aggressive but I think you don't want me to notice that.
ELIZA: What makes you think I am not aggressive?
User: You don't argue with me.
ELIZA: Why do you think I don't argue with you?
User: You are afraid of me.
ELIZA: Does it please you to believe I am afraid of you?

ELIZA effect

- While Weizenbaum was clear that ELIZA couldn't actually *understand* what people said to it, that didn't stop people from projecting understanding onto ELIZA. Weizenbaum wrote that he was "startled to see how quickly and how very deeply people conversing with ELIZA became emotionally involved with the computer and how unequivocally they anthropomorphized it." This became known as the "ELIZA effect."
- But as LibrarianShipwreck recounts, Weizenbaum became a much more vocal critic of AI upon realizing "that even once the processes were explained many people still bought into the 'illusion.'" Weizenbaum was prescient in noting that "A certain danger lurks here."
- Similarly, the discourse surrounding ChatGPT is mystical. For example, at the World Economic Forum, Coursera CEO Jeff Maggioncalda said "It looked like magic," adding that it is a "game changer" that is "blowing my mind."
- Similarly, popular press highlights how ChatGPT "learns," "thinks," or "feels" rather than being a fancy version of autocomplete that only ever offers approximations?

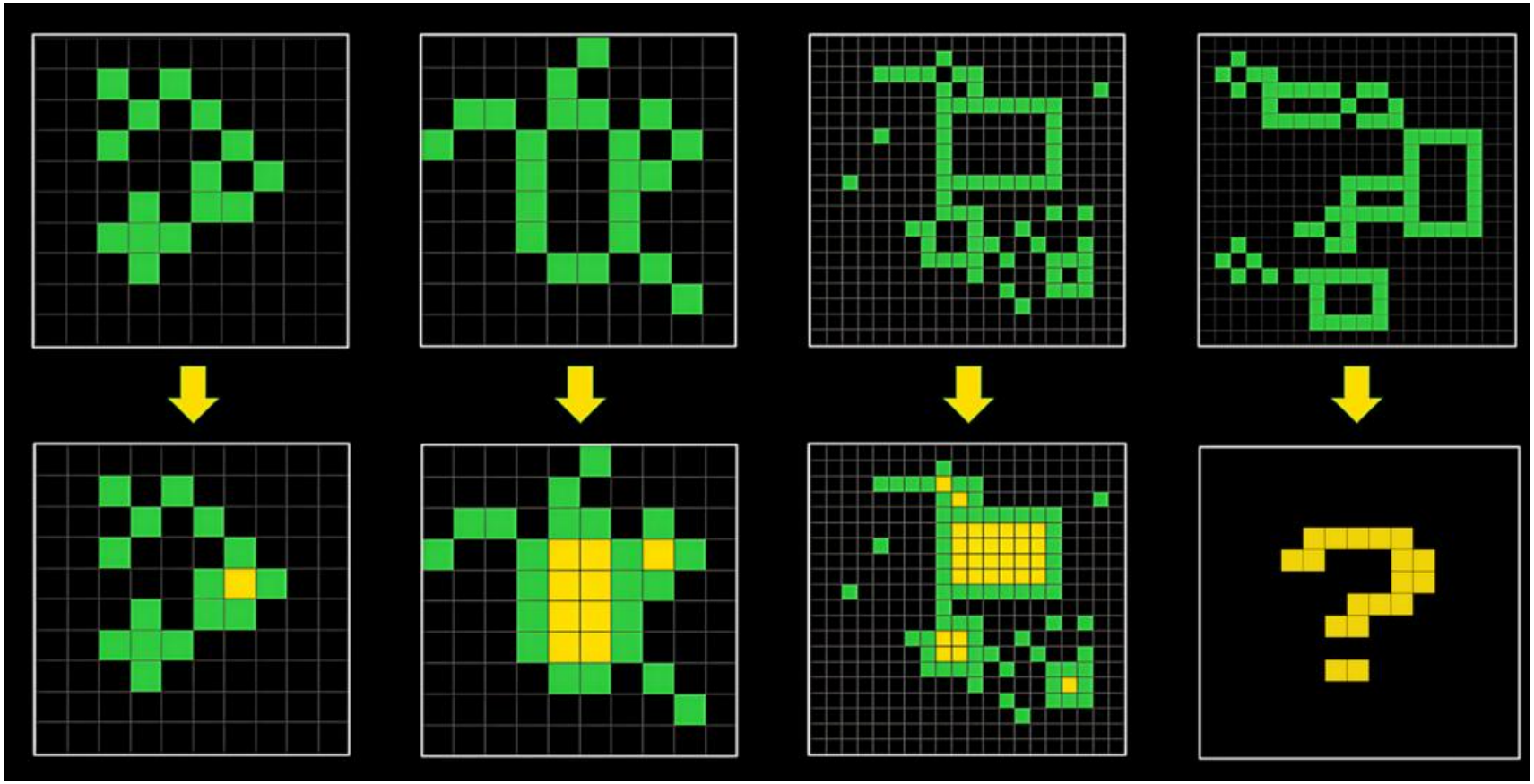
ChatGPT

- very good but still lots of problems: bias, logical reasoning, math, ethics and morality, hallucinations
- convincing but sometimes wrong



Are LLMs close to AGI?

- ARC-AGI challenge, https://arcprize.org/arc



Knowledge about language – linguistic analysis 1/2

Linguistic analysis contains several tasks: recognition of sounds, letters, word formation, syntactic parsing, recognizing semantic, emotions. Phases:

- Prosody - the patterns of stress and intonation in a language (rhythm and intonation)
- Phonology - systems of sounds and relationships among the speech sounds that constitute the fundamental components of a language
- Morphology - the admissible arrangement of sounds in words; how to form words, prefixes and suffixes …
- Syntax - the arrangement of words and phrases to create well-formed sentences in a language

Knowledge about language – Linguistic analysis 2/2

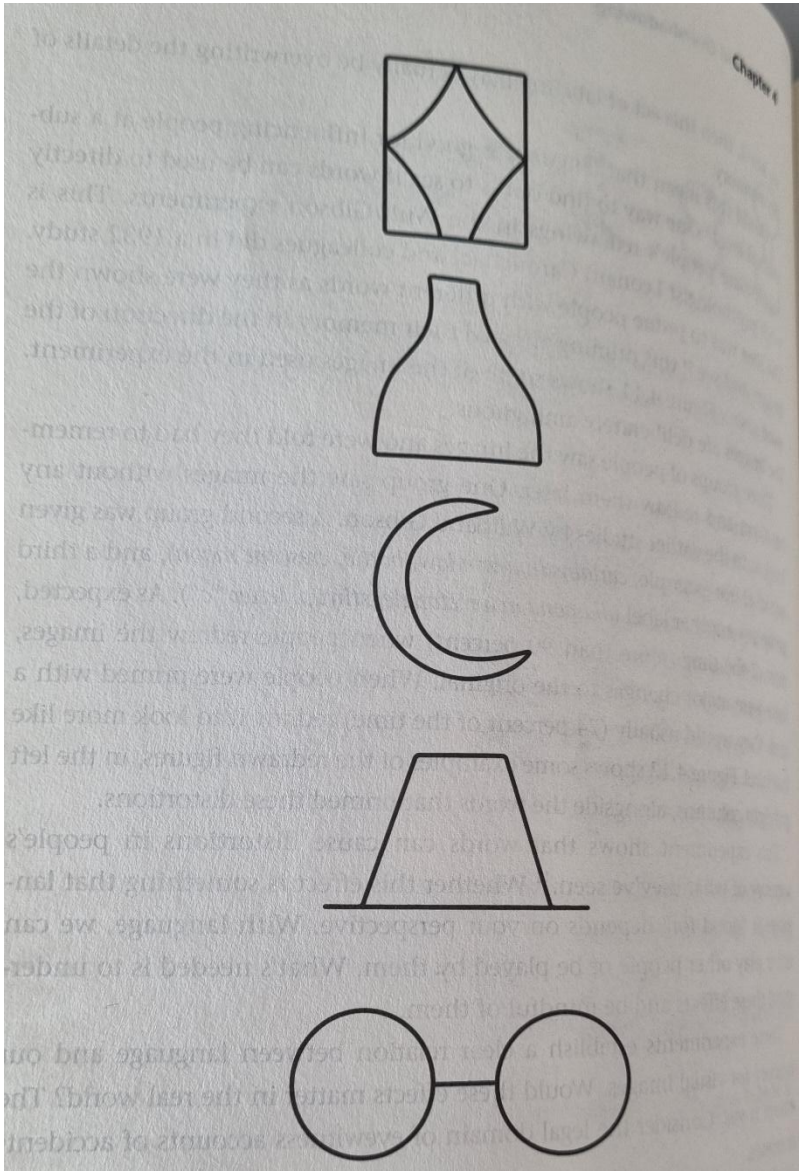
- Semantics - the meaning of a word, phrase, sentence, or text
- Pragmatics - language in use and the contexts in which it is used, including such matters as deixis (words whose meaning changes with context, e.g., I he, here, there, soon), taking turns in conversation, text organization, presupposition, and implicature

  Can you pass me the salt? Yes, I can.
- Knowing the world: knowledge of physical world, humans, society, intentions in communications …

Limits of linguistic analysis

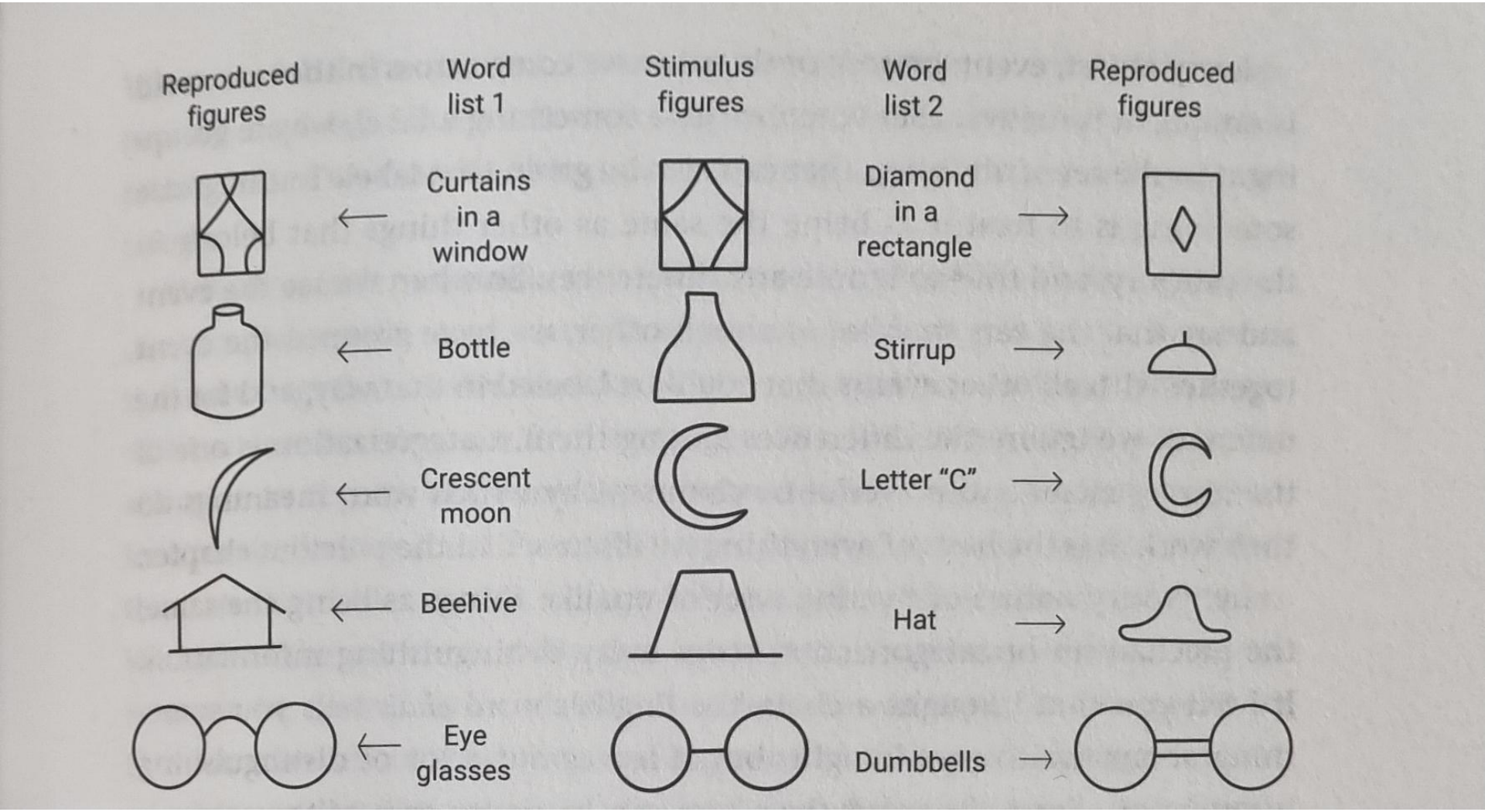
- levels are dependent
- e.g., intonation affects semantics
- sarcasm

Linguistic priming



Enfield, N.J., 2022. *Language vs. reality: Why language is good for lawyers and bad for scientists*. MIT Press.

Linguistic priming



Historically: two approaches

- symbolical
  - based on manually injected knowledge
  - grammars, frames, parse trees, etc.
  - top-down approach using grammatical patterns and semantics
  - ‘Good Old-Fashioned AI

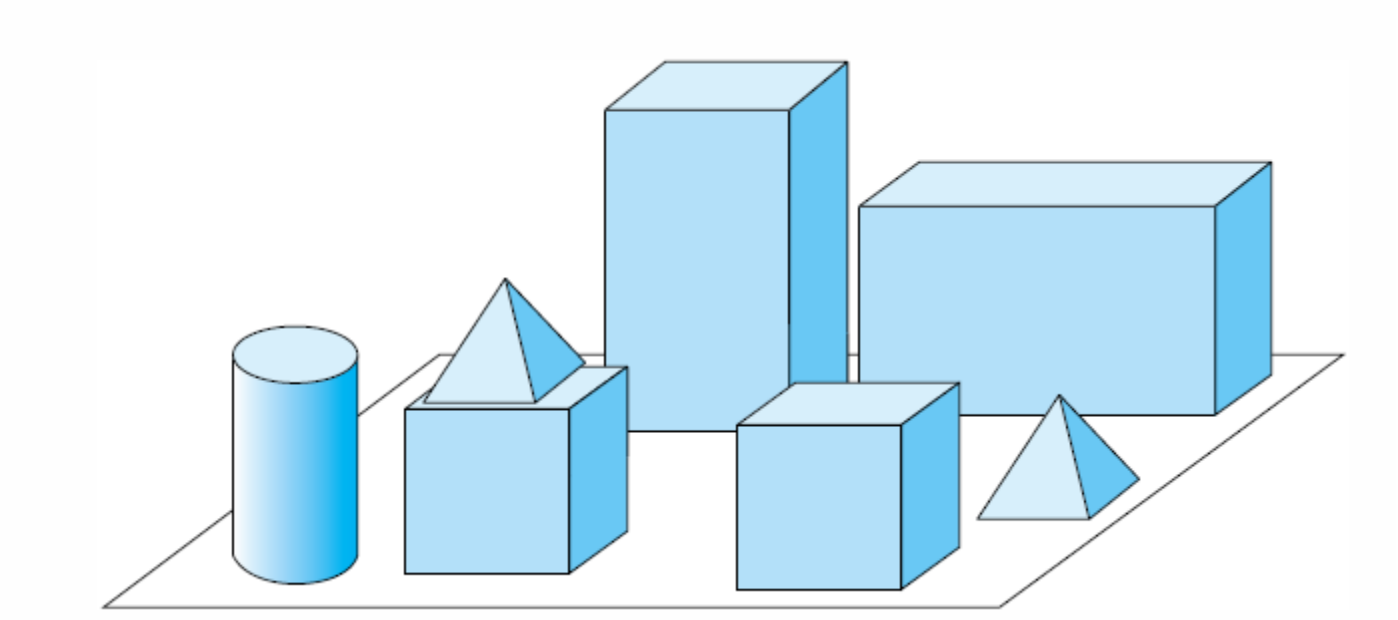
- statistical
  - knowledge is extracted from large corpora
  - bottom-up from texts, learning patterns and links, probabilistic reasoning (possibly syntactically or semantically wrong)
  - large pretrained language models: BERT, GPT, T5, ChatGPT, LLaMa, Gemini, Claude, DeepSeek, GaMS, Qwen, EuroLLM, Lamma, Gemma

- Merging both worlds: injecting (symbolical) knowledge into LLMs, let the LLMs call external services

How it all started?

- micro worlds
- example: SHRDLU, world of simple geometric objects
  - What is sitting on the red block?
  - What shape is the blue block on the table?
  - Place the green pyramid on the red brick.
  - Is there a red block? Pick it up.
  - What color is the block on the blue brick? Shape?

Micro world: block world, SHRDLU (Winograd, 1972)



Classical approach to text understanding

- text preprocessing
- 1. phase: syntactic analysis
- 2. phase: semantic interpretation
- 3. phase: use of world knowledge

- Hmm, what is text understanding, actually?

Basic text preprocessing – the classical pipeline

- document → paragraphs → sentences → words
- words and sentences ← POS tagging
- sentences ← syntactical and grammatical analysis

- still partially used in neural text processing

1. phase of classical text understanding: Syntactic analysis

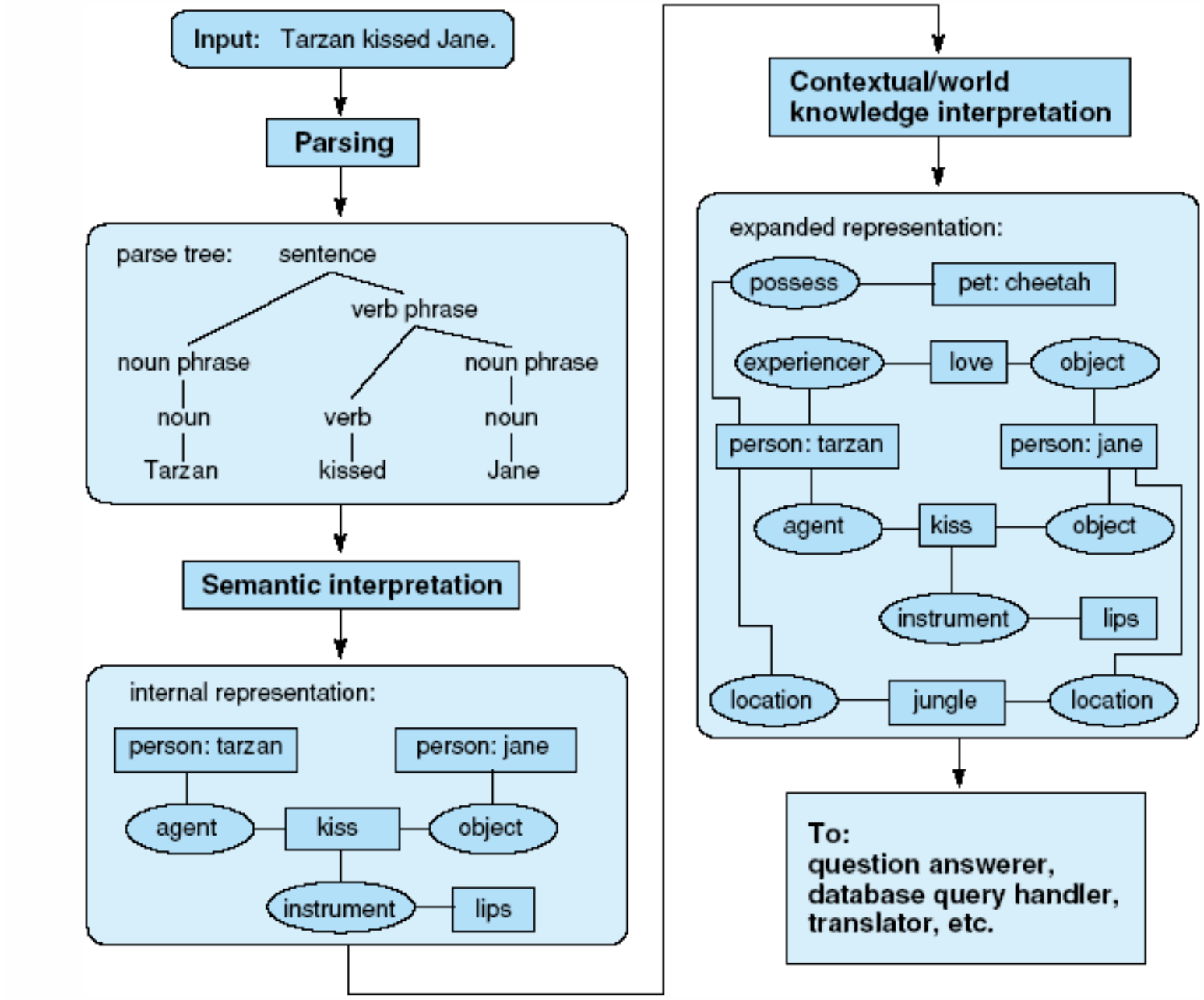
- Find syntactical structure
- part-of-speech (POS) tagging (noun, verb, preposition, …)
- The role in the sentence (subject, object, predicate)
- The result is mostly presented in a form of a parse tree.
- Needed: syntax, morphology, and some semantics.

2. phase: Interpretation

- Knowledge of word meaning and their language use
- Result: conceptual graphs, frames, logical program
- Check semantics

3. phase: Use of world knowledge

- Extend with background knowledge
- Consider the purpose of the system: summarization, database interface, etc.
- E.g., Cyc and openCyc knowledge bases present ontology and knowledge base of everyday common-sense knowledge, e.g.,
  "Every tree is a plant" and "Plants die eventually”
- process incrementally, adding the meaning of previous sentences


Input: Tarzan kissed Jane.
Parsing
parse tree: sentence
noun phrase
verb phrase
noun phrase
noun
verb
noun
Tarzan
kissed
Jane
Semantic interpretation
internal representation:
person: tarzan
person: jane
agent
kiss
object
instrument
lips
Contextual/world knowledge interpretation
expanded representation:
possess
pet: cheetah
experiencer
love
object
person: tarzan
person: jane
agent
kiss
object
instrument
lips
location
jungle
location
To:
question answerer,
database query handler,
translator, etc.

NLP success stories

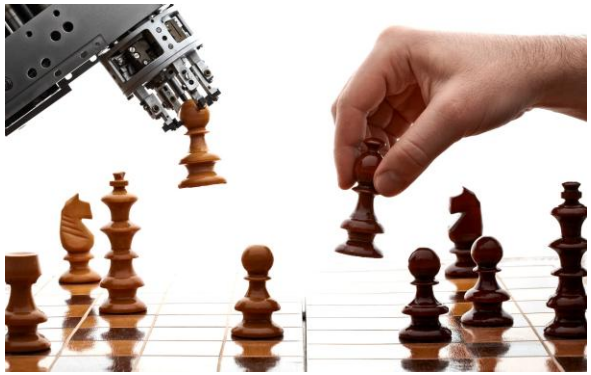
- Jeopardy, 2011: IBM Watson wins in a quiz against two human champions
- After 2010 steady progress
- useful tools like Google Translate, Siri, Cortana, Alexa
- search engines
- information extraction and retrieval
- speech recognition
- text classification
- useful summarization and question answering
- ChatGPT
- LLMs as a general approach to AI

Where is NLP today?

- active research area with many commercial applications
  - speech recognition and synthesis
  - automatic reply engines
  - machine translation
  - text summarization
  - question answering
  - language generation
  - interface to databases
  - intelligent search and information extraction
  - sentiment detection
  - named entity recognition and linking
  - categorization, classification documents, messages, tweets, etc.
  - many (open-source) tools and language resource
  - prevalence of deep neural network approaches
  - cross-lingual approaches
  - LLMs are everywhere, LLMs for everything

Language understanding

- Can current approaches pass the Turing test?
- Can a system understand a language?
- Do large language models understand language?



- Do we understand what is understanding?
- We don‘t have good enough models for knowledge representation.
- Injecting knowledge into deep neural networks: factual, linguistic, common-sense, domain-specific

CJVT UL: Center for language resources and technologies of University of Ljubljana (Center za jezikovne vire in tehnologije Univerze v Ljubljani)

- many practical open-source technologies and solutions using NLP and ML for Slovene
- corpora and datasets
- thesaurus, dictionary of collocations
- lexicon of wordforms, lexical database
- sentiment lexicon
- machine translation
- speech recognition
- neural POS taggers
- models for comma placement, stress, readability
- summarizer
- cross-lingual models
- embeddings
- LLLMs, such as GaMS.
- www.cjvt.si
- www.slovenscina.eu
- https://povejmo.si/