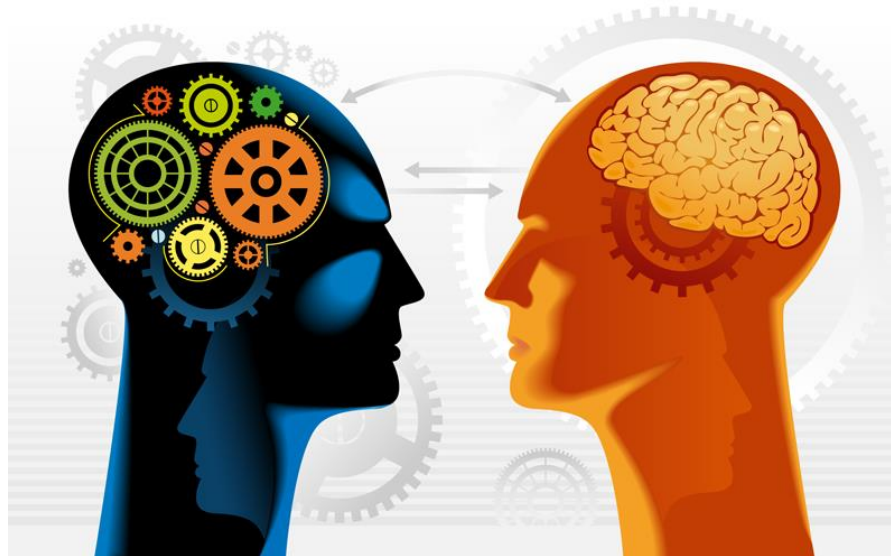


Natural language processing – an introduction

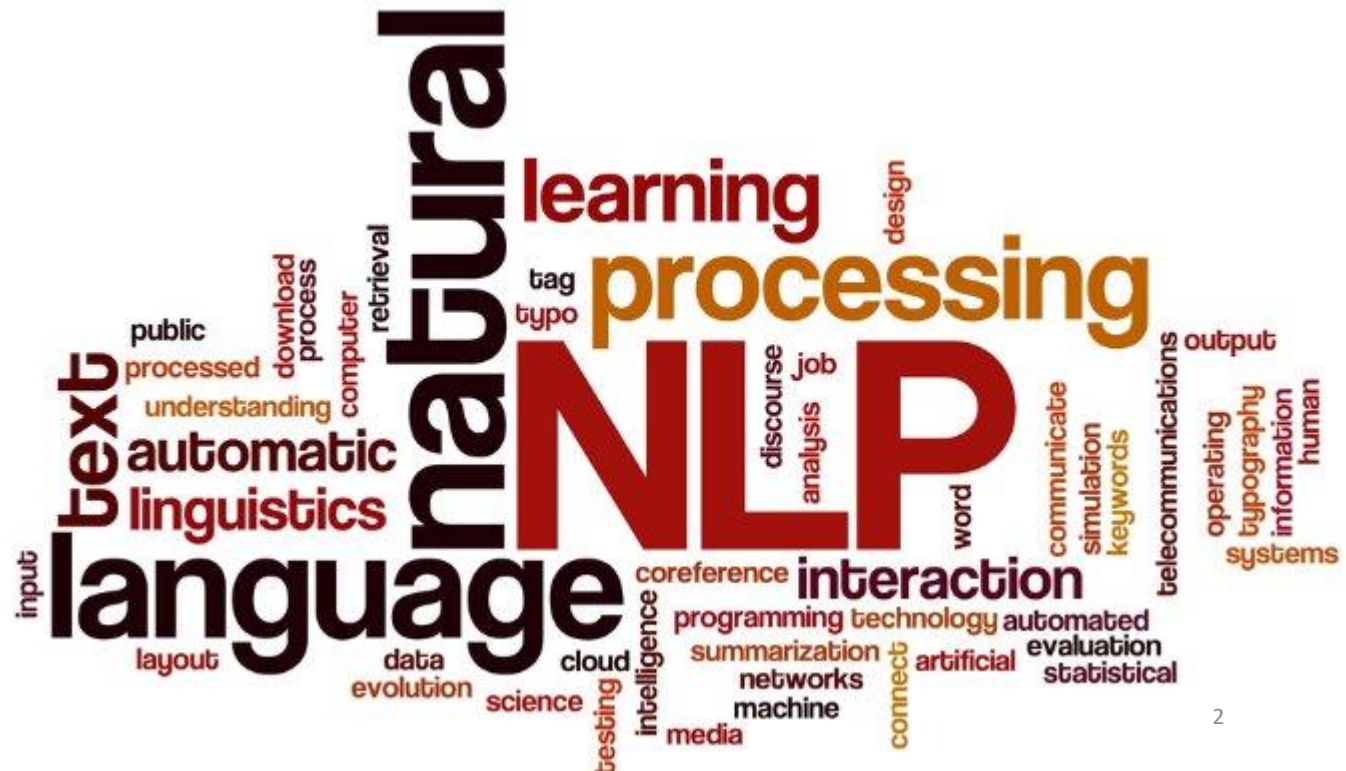


Prof Dr Marko Robnik-Šikonja

February 2020

Outline

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



Lecturer

- Prof Dr Marko Robnik-Šikonja
- University of Ljubljana
Faculty of Computer and Information Science
Laboratory for Cognitive Modeling
- FRI, Večna pot 113, 2nd floor, right from the elevator
- marko.robnik@fri.uni-lj.si
- <https://fri.uni-lj.si/en/employees/marko-robnik-sikonja>
- (01) 4798 241
- Contact hour (see webpage)
 - currently, Wednesday, 11:00 -12:00; for other slots, email me
- **Research interests:** machine learning, artificial intelligence, natural language processing, network analytics, data science, data mining,
- **Teaching:** courses from the area of data mining, algorithms, machine learning, and natural language processing
- **Software:** an author or coauthor of several software tools, including three open source R packages from the area of predictive modelling and data analytics (CORElearn, semiArtificial, ExplainPrediction)



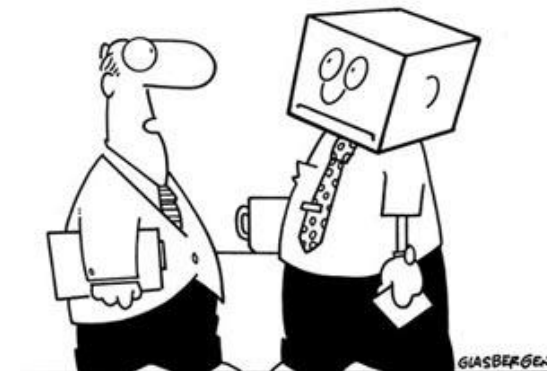
Assistant

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



Goals of the course

- students shall become acquainted with
 - basics of natural language processing and understanding
 - basic approaches, data structures and algorithms for NLP
 - selected modern statistical techniques for NLP
 - selected relevant NLP tasks
 - relevant research challenges in the area of NLP and NLU, computational linguistics, and semantics
- teach students a practical use of
 - 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 the language as an important source of information



Learning outcomes

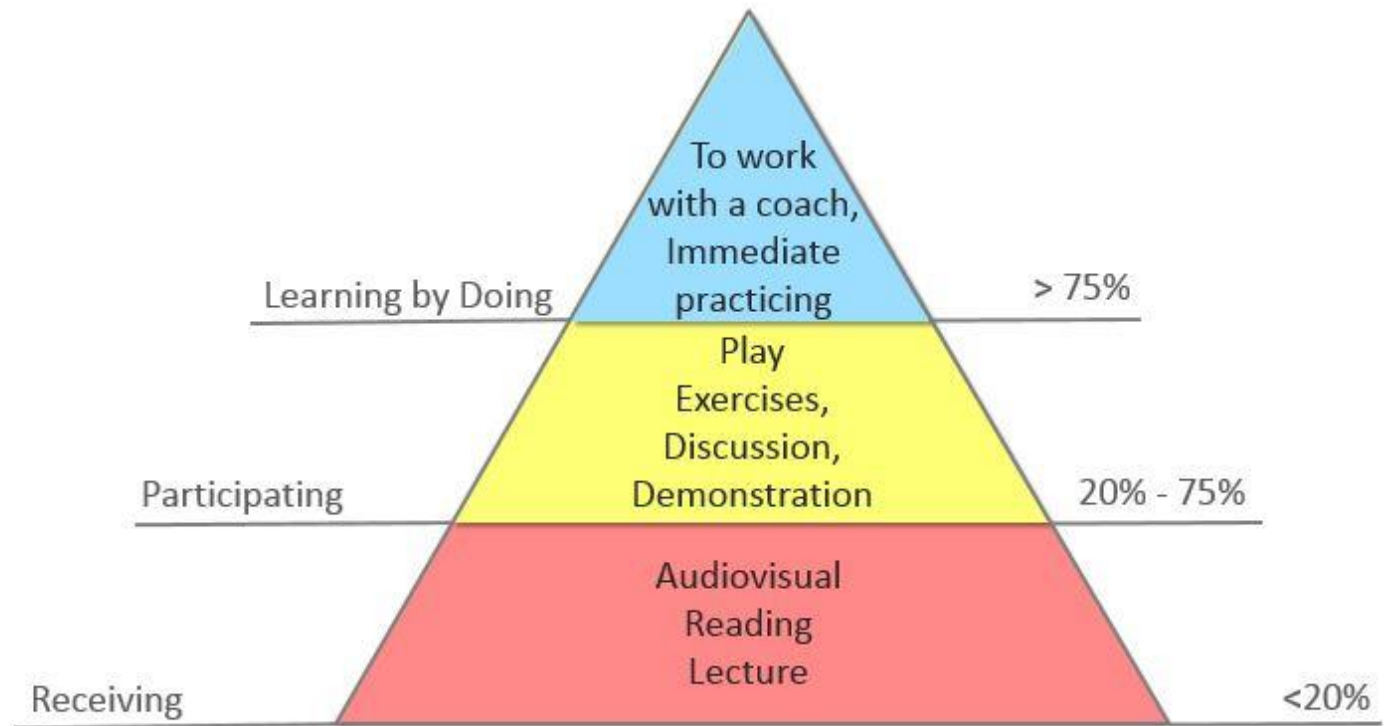
Upon completion of the course, students will:

- 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 vector representations of text and adapt them to new contexts
- use and evaluate approaches to summarization, machine translation, question answering etc.

Lectures and tutorials

- Lectures
 - introduction to the topic, discussion
 - some examples
 - broader view of the topic
- 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



Retention of Learning

I. Kokcharov © 2015

Syllabus 1/2



- Introduction to natural language processing: motivation, language understanding, Turing test, traditional and statistical approaches, area overview, successes and limitations
- Text preprocessing and normalization: regular expressions for search and replacement, grammars for syntax analysis, string similarity, Levenhstein distance, advanced normalization techniques
- Language resources: corpuses, dictionaries, thesauruses, networks and semantic data bases, WordNet, language networks and graphs, overview of tools.
- Linguistics: phonology and morphology, syntactical analysis, formal grammars, lemmatization, POS tagging, dependency parsing
- Sparse vector representations of texts: Bag-of-words model, text clustering, cosine distance, semantic correlation
- Deep networks for text: document representations for deep neural networks, autoencoders, recurrent neural networks, LSTM networks, convolutional networks for text, transformers,
- Language models: ngram models, neural models, masked models
- Dense vector representations of texts: word2vec, GloVe, ELMo, BERT, cross-lingual embeddings

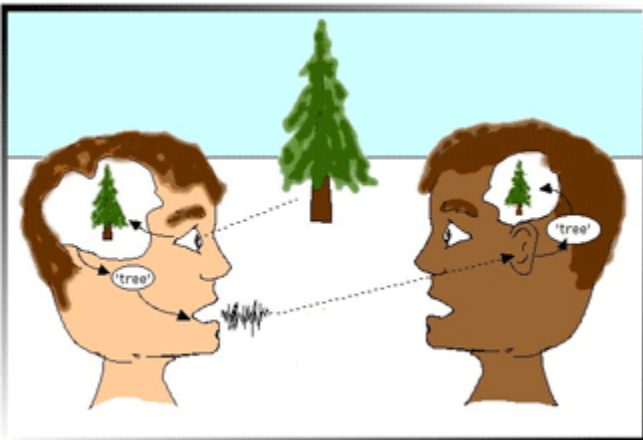
Syllabus 2/2



- Methodology and evaluation in NLP
- Shallow computational and lexical semantics: verb valence, named entity recognition, semantic role labelling
- Text mining: adaptation of classification methods to the specifics of text, support vector machines for language, feature selection
- Sentiment analysis and emotional computing.
- Text summarization: text representations, matrix factorization, multi-document summarization, graph-based approaches for extractive summarization
- Terminology extraction
- Machine translation: classical and neural statistical translation, language model, translation model, alignment model, challenges in machine translation.
- Semantic representations: knowledge graphs for commonsense reasoning

What the course does not cover?

- speech processing: recognition and synthesis (a courses in UL Faculty of Electrical Engineering, „Speech technologies“)
- chatbots
- information retrieval (a course „Web information extraction and retrieval“)
- in depth linguistics
- dialogue systems



Prerequisites

- Recommended knowledge
 - python programming,
 - probability and statistics,
 - machine learning.

Obligations

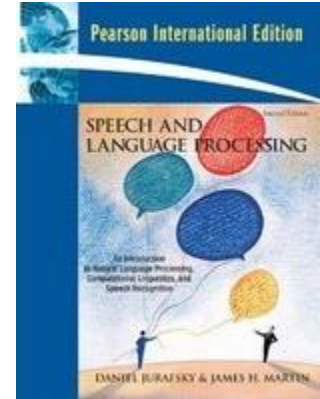
- 5 quizzes checking continuous understanding of contents
- a projects, composed of three stages, 50 points
- written exam, 50 points

Grading

Obligation	% of total	subject to
Five quizzes	0%	$\geq 50\%$
Project	50%	$\geq 50\%$
Written exam	50%	$\geq 50\%$

Learning materials

- learning materials in the eClassroom <http://ucilnica.fri.uni-lj.si>
- slides are updated continuously
- links to literature
- code and examples
- links to data sets



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, 2019.
basic course literature; available on [authors' webpages](#)
- Bird, Steven, Ewan Klein, and Edward Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.
[freely available book](#), regularly updated, based on NLTK library for Python 3
- Jacob Eisenstein. [Natural Language Processing](#)
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. [Deep Learning](#). MIT press, 2016
- scientific papers for deeper understanding of certain topics

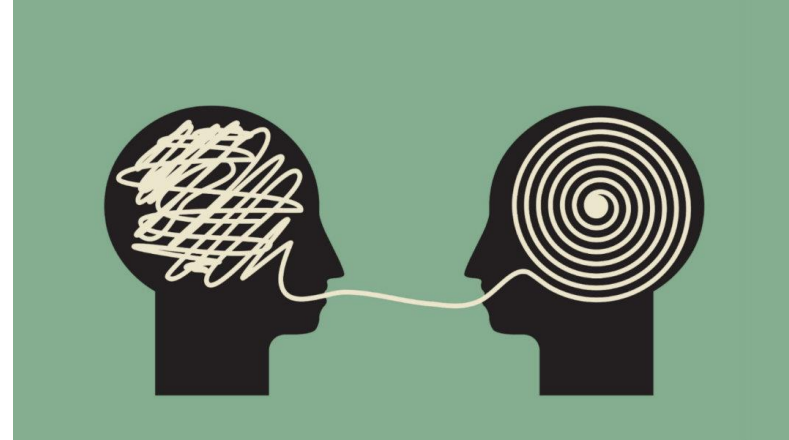
Web courses

- Zhai: [Text Mining and Analytics](#)
- Manning: [Natural Language Processing with Deep Learning](#)

Two views of natural language processing

- Techniques for language **processing**: syntax, grammars, language resources, text representation, speech
- Attempts to **understand** language: semantics and pragmatics of language, related to artificial intelligence

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.

An example: translated rules

Article 18 of UL FRI Study Rules and Regulations

Taking exams at an earlier date may be allowed at the request of the student by the Vice Dean for Education with the lecturer's consent in justified circumstances (leaving for study or placement abroad, hospitalization at the time of the exam period, participation at a professional or cultural event or a professional sports competition, etc.), and if the applicant's study achievements in previous study years are deemed satisfactory for such an authorization to be appropriate.

Understanding NL by computers

- Understanding words, syntax, semantics, context, writer's intentions, knowledge, background, assumptions, bias ...
- Ambiguity in language
 - Newspaper headlines - intentional ambiguity – attention seeking
 - 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

For a dishonest employee

For a lazy employee

For the office drunk

A man like him is hard to find.

He's an unbelievable worker.

You would indeed be fortunate to get this person to work for you.

Every hour with him was a happy hour.

Structural ambiguity

For a chronically absent employee

For a dishonest employee

For a stupid employee

For the office drunk

It seemed her career was just taking off.

Her true ability was deceiving.

I most enthusiastically recommend this candidate with no qualifications whatsoever.

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 jokes?

A priest and an Australian shepherd met each other in the final of a quiz show. After answering all the normal questions, they were neck-and-neck with the same number of points and the quizmaster had to set a deciding question. The question was to compose a rhyme in 5 minutes including the word `Timbuktu`. After 5 minutes, the priest presented his poem:

I was a father all my life,
I had no children, had no wife,
I read the Bible through and through,
on my way to Timbuktu.

The audience was thrilled and celebrated the churchman as the certain winner. However, the Australian shepherd stepped forward and recited:

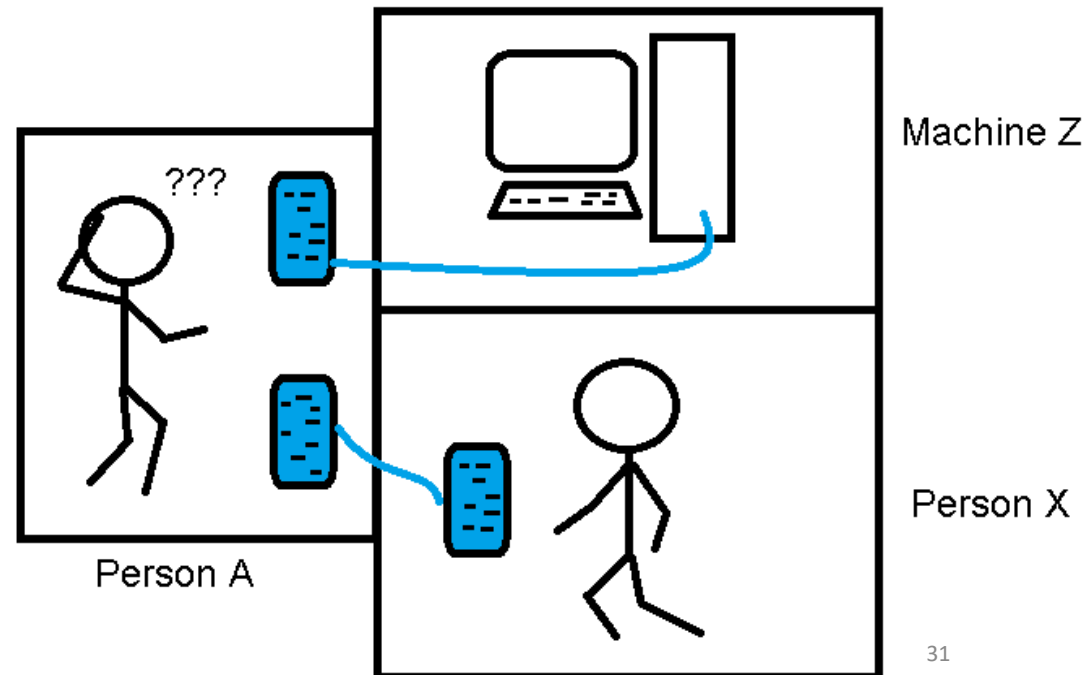
When Tim and I to Brisbane went,
We met three ladies cheap to rent,
They were three and we were two,
So I booked one and Tim booked two.

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

- Turing test,
- Loebner prize



Chatbots attempt to solve Turing test

- CYC, ELISA, ALLICA, chatterbot,
- VIDA, TIA,
- Siri (knowledge recognition, context detection, web service recommendation)


```

=====
EEEEEEEE L      IIIIII ZZZZZZZ      AAA
E         L      I      Z      A
E         L      I      Z      A
EEEEEE   L      I      Z      A
E         L      I      Z      AAAAAA
E         L      I      Z      A
EEEEEEEE LLLLLLL IIIIII ZZZZZZ     A  A
=====
ELIZA > (PLEASE TYPE IN ALL CAPS) WHAT'S YOUR NAME DEAR ?
AMIT
ELIZA > HI, I'M ELIZA. WHAT DO YOU WANT TO TALK ABOUT ?
AMIT > CAN I TALK ABOUT MY PROBLEM ?
ELIZA > SURE... !
AMIT > DUE TO ON GOING EXAMS IN COLLEGE, I HAVE A LOT OF STRESS.
ELIZA > PLEASE RELAX A LITTLE. SLEEP WELL.
AMIT > THANKS FOR YOUR ADVICE.
ELIZA > NO MENTION.
AMIT > BYE
ELIZA > BYE AND KEEP IN TOUCH...
=====

```

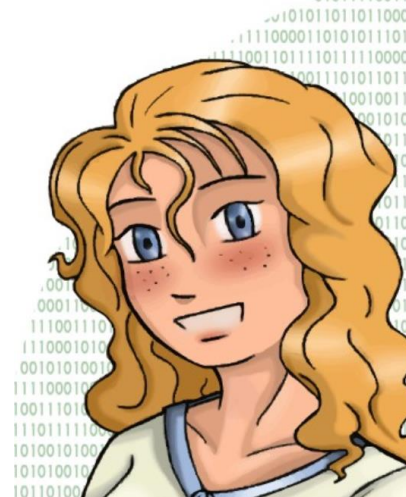
- ELIZA: Does it please you to believe I am afraid of you?

Talking with ALICE Chabot



- judge: Hello, I want to show your abilities to my students
- ALICE: Hi there!. Interesting. I want to find out more about you.
- judge: Let us better talk about the technology driving you
- ALICE: That sounds good to me.
- judge: Do you know anything about it or AI or NLP?
- ALICE: You're so smart, you tell me what it is.
- judge: Let us just do a test
- ALICE: I don't know if it's in my power to give permission.
- judge: Do you use parsing trees at all?
- ALICE: I use it as often as possible.

Mitsuku Chatbot (2013 and 2016 winner of Loebner bronze medal, 2015 runner-up):



- Judge: Hello, I want to show your abilities to my students.
- Mitsuko: Hello there. Why do you want to do it so much?
- Judge: I want to demonstrate them the technology driving you.
- Mitsuko: I want to become smarter than humans.
- Judge: Do you know anything about it or AI or NLP?
- Mitsuko: Thanks for telling me.
- Judge: Let us just do a test.
- Mitsuko: I will think about it before I decide whether to allow it or not.
- Judge: Do you use parsing trees at all?
- Mitsuko: Hmm... A tree is used for wood. I only really use electricity and internet connection here.

Rose – 2015 winner, transcript of Loebner Prize 2015 finalist selection

Question: Hi, I'm Andrew. Who are you?

Rose: Yes, I already know your name is Andrew.

Question: How are you today?

Rose: Good, thanks. How are you?

Question: Are you a human?

Rose: Of course I am human. Aren't we all?

Question: Can you answer this question?

Rose: Sure. Fire away.

Question: Do you like tea?

Rose: Sure, I like tea. Haven't had any recently though.

Question: If a bed doesn't fit in a room because it's too big, what is too big?

Rose: I don't even want to pretend that.

Question: If Alex lent money to Joe because they were broke, who needed the money?

Rose: huh?

Question: Should Greece leave the Euro?

Rose: Seems like a nice place.

Question: How many words are in this question?

Rose: There are 7 words in the sentence: how many words are in this question?

Question: What color is the sea?

Rose: Reflective.

Question: How many presidents of the US were called Bush?

Rose: I have heard of President Bush.

Question: What would you like to drink?

Rose: I equally drink microbrew beer, wine, and cocktails. It all depends on who I'm with.

Question: Will you please tell me the length of your hair?

Rose: Medicine and anatomy are an important field of study. But I'd never make a good doctor. I'm too squeamish about various body parts.

Question: What would you say if I gave you a box of chocolates?

Rose: I would probably giggle.

Question: Do you play Chess?

Rose: I like board games, and played a lot as a child. I haven't that much time these days.

Question: How do you think it's going?

Rose: I'm good. How are you?

Question: What was my first question?

Rose: Did you tell me already and I forgot? With the advent of videogames, many people have given up playing card or board games. How about you?

Question: Did you see the Turing film?

Rose: Alan Turing is one of my heroes.

Question: Why not?

Rose: He did brilliant work under very difficult conditions.

Question: Are you on Twitter?

Rose: I have accounts on all of the social networking sites but don't post things on them. I just datamine them.

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
- intonation affects semantics
- sarcasm

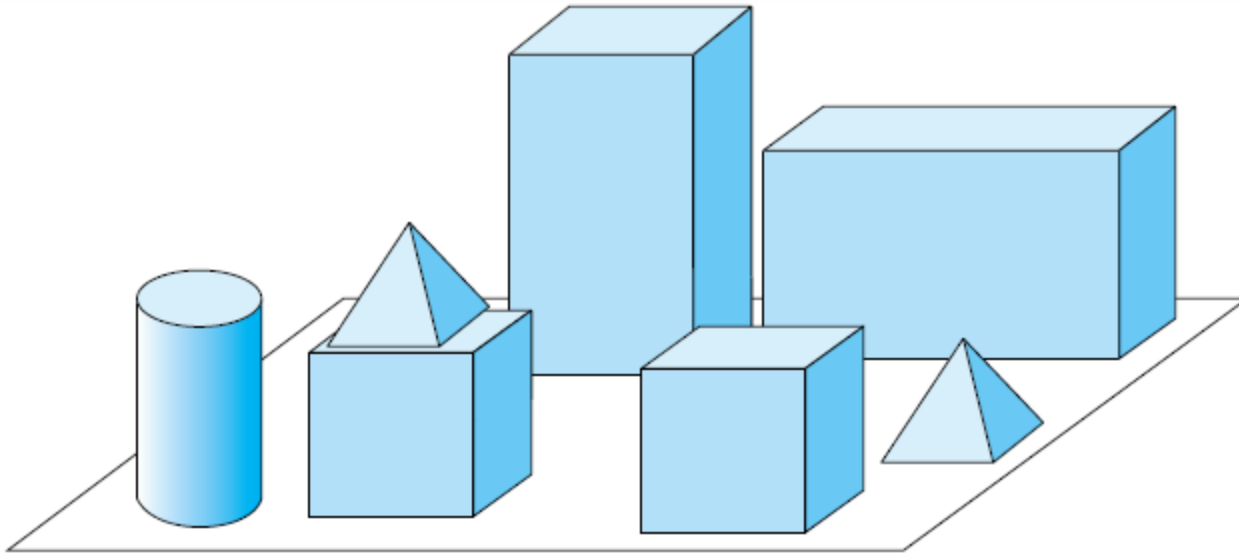
Historically: two approaches

- 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)
- Merging both worlds: injecting symbolical knowledge into DNNs

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 tools for text preprocessing

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

1. phase of text understanding – syntax analysis

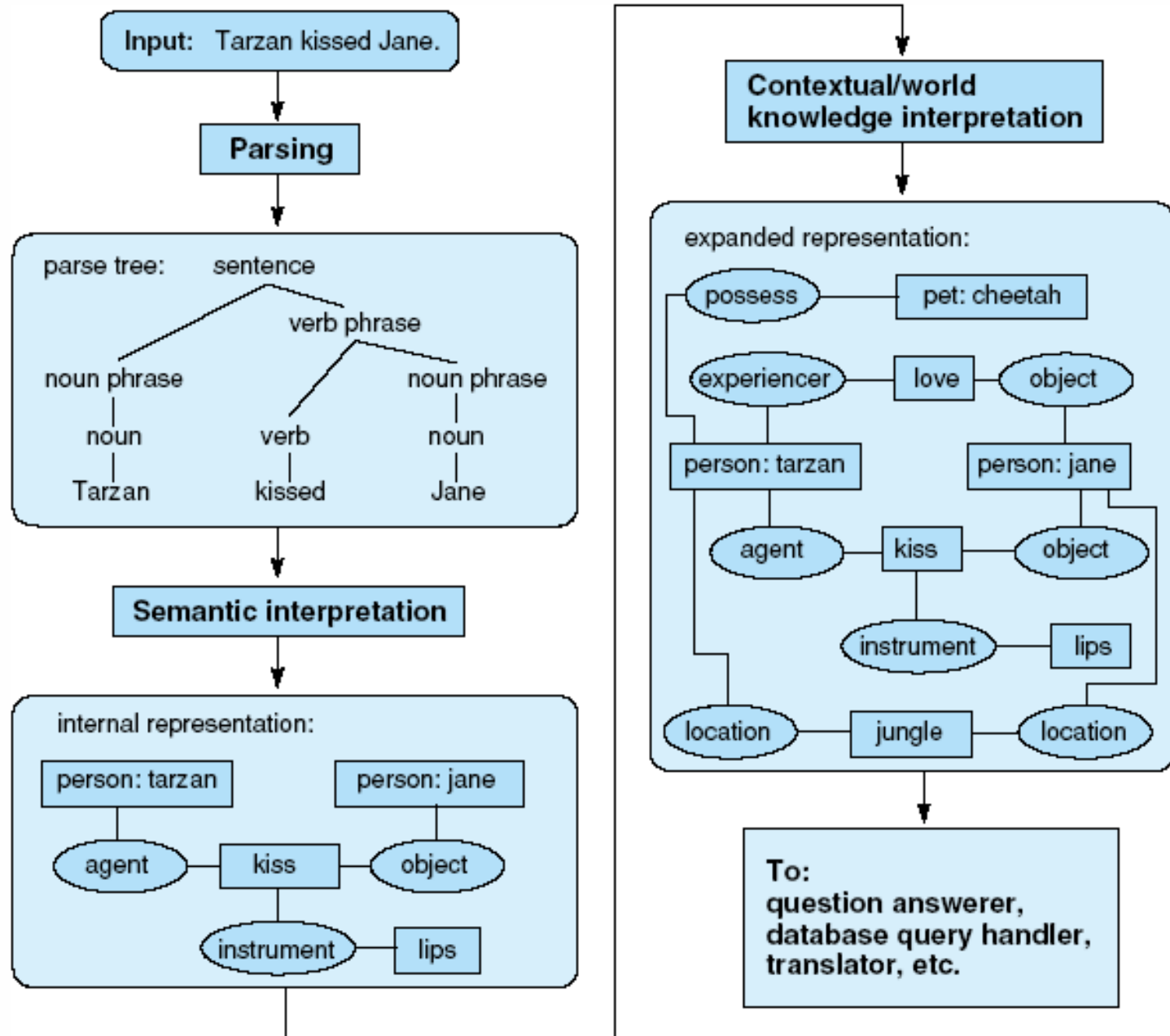
- Find syntactical structure
- part-of-speech (POS) tagging (noun, verb, preposition, ...)
- The role in the sentence (subject, object, predicate)
- The result is mostly presented in a form of a parse tree.
- Needed: syntax, morphology, and some semantics.

2. phase - interpretation

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

3. phase of text understanding: use of world knowledge

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



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
 - semantic analysis: e.g. role labelling,
 - named entity recognition and linking
 - categorization, classification documents, messages, twits, etc.
 - many (open-source) tools and language resource
 - prevalence of deep neural network approaches
 - cross-lingual approaches

NLP resources and technologies

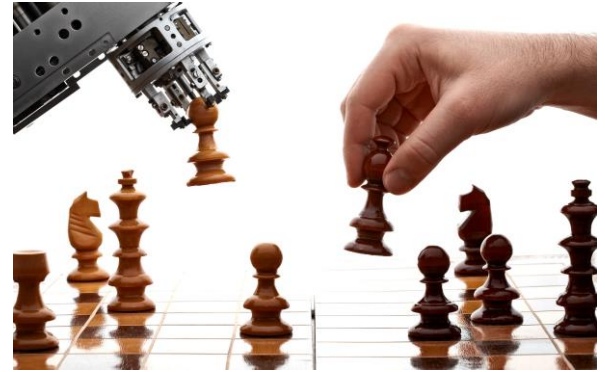
- language technologies
 - prevalence of deep neural network approaches
 - text embeddings
 - cross-lingual approaches
 - semantic analysis: e.g. role labelling,
 - named entity recognition and linking
 - categorization, classification documents, messages, twits, etc.
 - many (open-source) tools and language resource
- language resources
 - importance of large text corpora: monolingual, parallel
 - knowledge graphs
 - dictionaries and thesauri
 - datasets for ML

NLP success stories

- Jeopardy, 2011: IBM Watson wins in a quiz against two human champions
- useful tools like Google Translate, Siri, Cortana, Alexa
- search engines
- information extraction and retrieval
- (superhuman) speech recognition
- text classification
- useful summarization and question answering

Language understanding

- Can current approaches pass Turing test?
- Can a system understand a language?
- Problems of current approaches
 - we solve partial problems
 - because actual problems are (still) too difficult
- We don't understand what is understanding.
- We don't have good enough models for knowledge representation.



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 solution using NLP and ML for Slovene
- thesaurus and dictionary of collocations
- sentiment lexicon
- machine translation
- speech recognition
- neural POS taggers
- models for comma placement, stress,
- summarizer
- cross-lingual models
- embeddings
- etc.

The screenshot shows the CJVT Sopomenke web application. At the top, there is a red header with the logo 'cjvt sopomenke' and a search bar containing the word 'ideja'. Below the header, the main content area displays a table of word pairs with their relevance and frequency. The table has columns for 'Relevantnost' and 'Pogostost'. The word pairs are: 'zamisli' and 'predstava', 'misli' and 'bežen vtis', 'domisljica' and 'nejasna predstava', 'odkritje' and 'pojem', 'pogrnjavščina' and 'koncept', and 'zasnova' and 'pogled'. At the bottom, there is a dark grey footer with a section titled 'Uporabniško dodane sopomenke' and a form to add a new word pair. The form has two input fields: 'Sopomenka' and 'Uporabniško ime', and a red 'Dodaj' button.

Relevantnost	Pogostost
zamisli	predstava
misli	bežen vtis
domisljica	nejasna predstava
odkritje	pojem
pogrnjavščina	koncept
zasnova	pogled

Uporabniško dodane sopomenke
Vsaka dodana sopomenka povečuje uporabnost slovarja za jezikovno skupnost. Hvala!

Dodaj sopomenko za "ideja"

✓ Sopomenka

⚙ Uporabniško ime

Dodaj

preblik
Spela Ali