Aalto University

Statistical Natural Language Processing: an introduction

+contents of the 2024 course

Presented by: Mikko Kurimo

Interaction during the course

- Lectures can be participated only **at campus** (no zoom)
- Lectures are recorded and will be available for the course participants (target: one week after each lecture). Lecture slides available immediately (target: before the lecture)
- Participate in lecture activities and provide lecture and assignment feedback using **MyCourses**
- Submit answers to weekly assignments using **JupyterHub & Nbgrader**
- Use the course **Zulip** space for questions and discussions before and after lectures and about the weekly assignments
- Provide peer feedback for the project works and vote for the best video using **MyCourses**

Part I: Statistical natural language processing

1. Introduction
2. Applications
3. Why is it so hard?
    - Challenges of natural language data

Language is processed in our phones and homes

Including televisions, phones, new assistance devices, toys

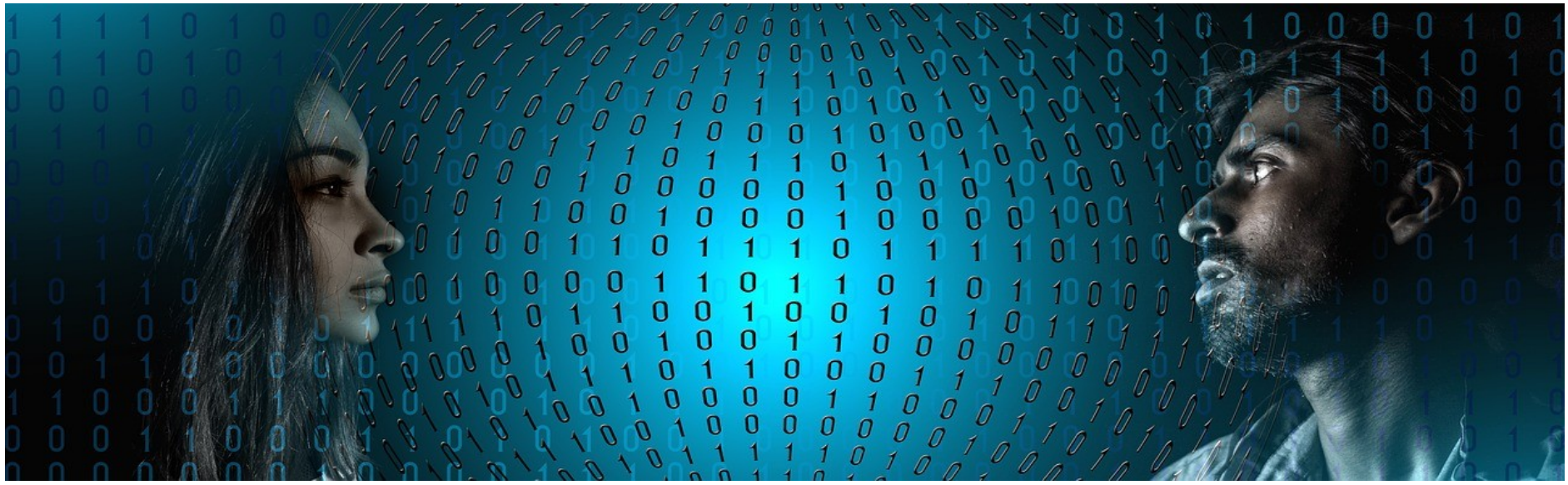
Language is used for several everyday tasks

Including dictation, captioning, translation, interpretation,
information retrieval, conversational assistants, language learning

skype translator

Language is human communication

- Rich communication signal **between humans**
- Human speech is the most complex of all biosignals
- speech => text + emotion,loudness,speed,emphasis,...
- text + *emotion,loudness,speed,emphasis,...* => speech
- How much language “understanding” is needed?
- People perceive the use of language as a sign of “intelligence”



Modeling of language

- Language is complex, adaptive system
  - Storing and processing text and speech
  - Large datasets
- We want to make systems that 'understand'
  - Take into account language related phenomena
- Building models about natural language using large data sets

Statistical Natural Language Processing

Red text is A! Aalto University research

Methodological basis:

- machine learning
- pattern recognition
- probability theory
- statistics
- signal processing

Related fields:

- computational linguistics
- corpus linguistics
- Phonetics
- speech processing
- discourse analysis
- cognitive science
- artificial intelligence

What is in a language?

Phonetics and phonology:

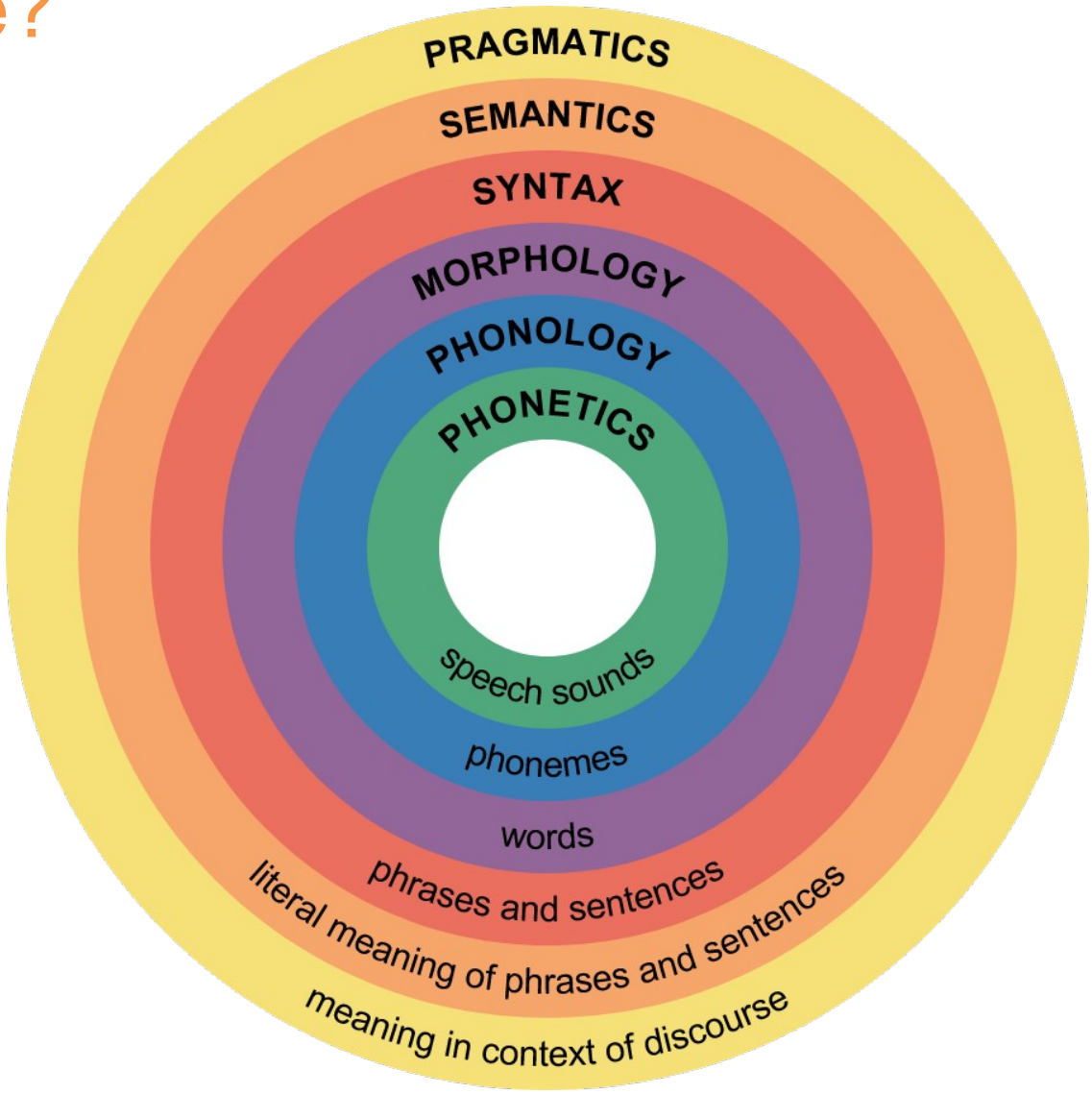
- physical sounds
- patterns of sounds

Morphology: building blocks of words

Syntax: grammatical structure

Semantics: meaning of words

Pragmatics, discourse, spoken interaction...



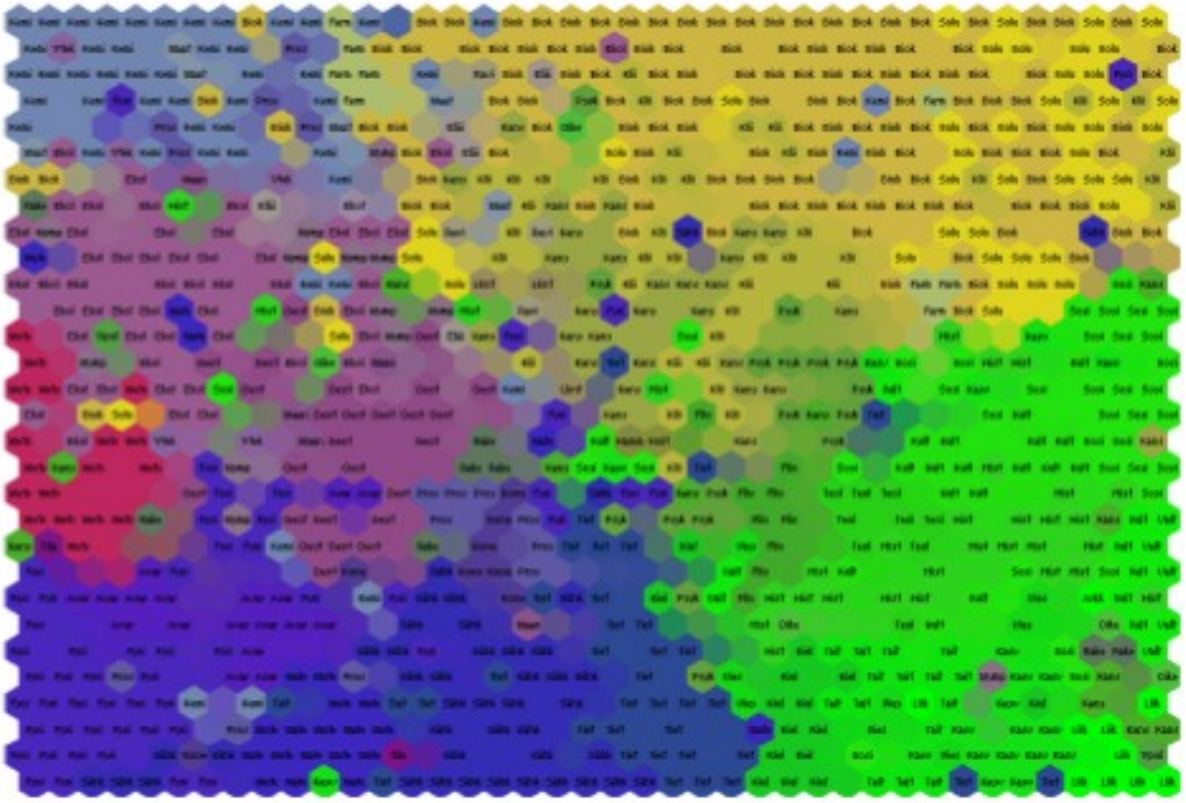
Application areas

Red text is A! Aalto University research

- Information retrieval
- Text clustering and classification
- Automatic speech recognition
- Natural language interfaces
- Automatic question answering, chatbots
- Text and image generation
- Machine translation
- ...

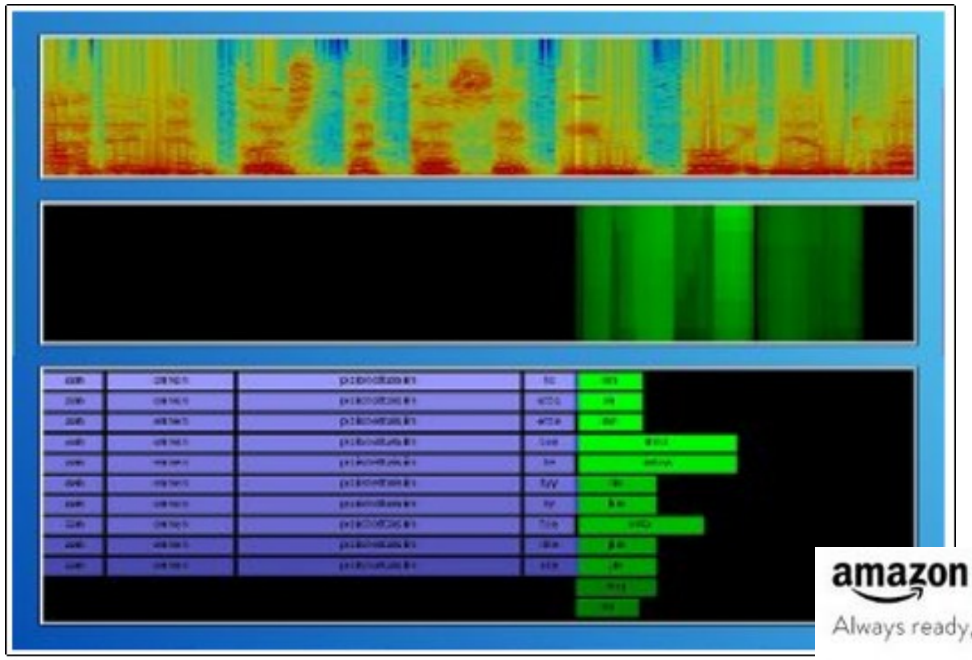
Text clustering and classification

This is A! Aalto University research



WEBSOM (Honkela, Kaski, Kohonen & Lagus, 1996, etc.)

Speech recognition



This is A! Aalto University research

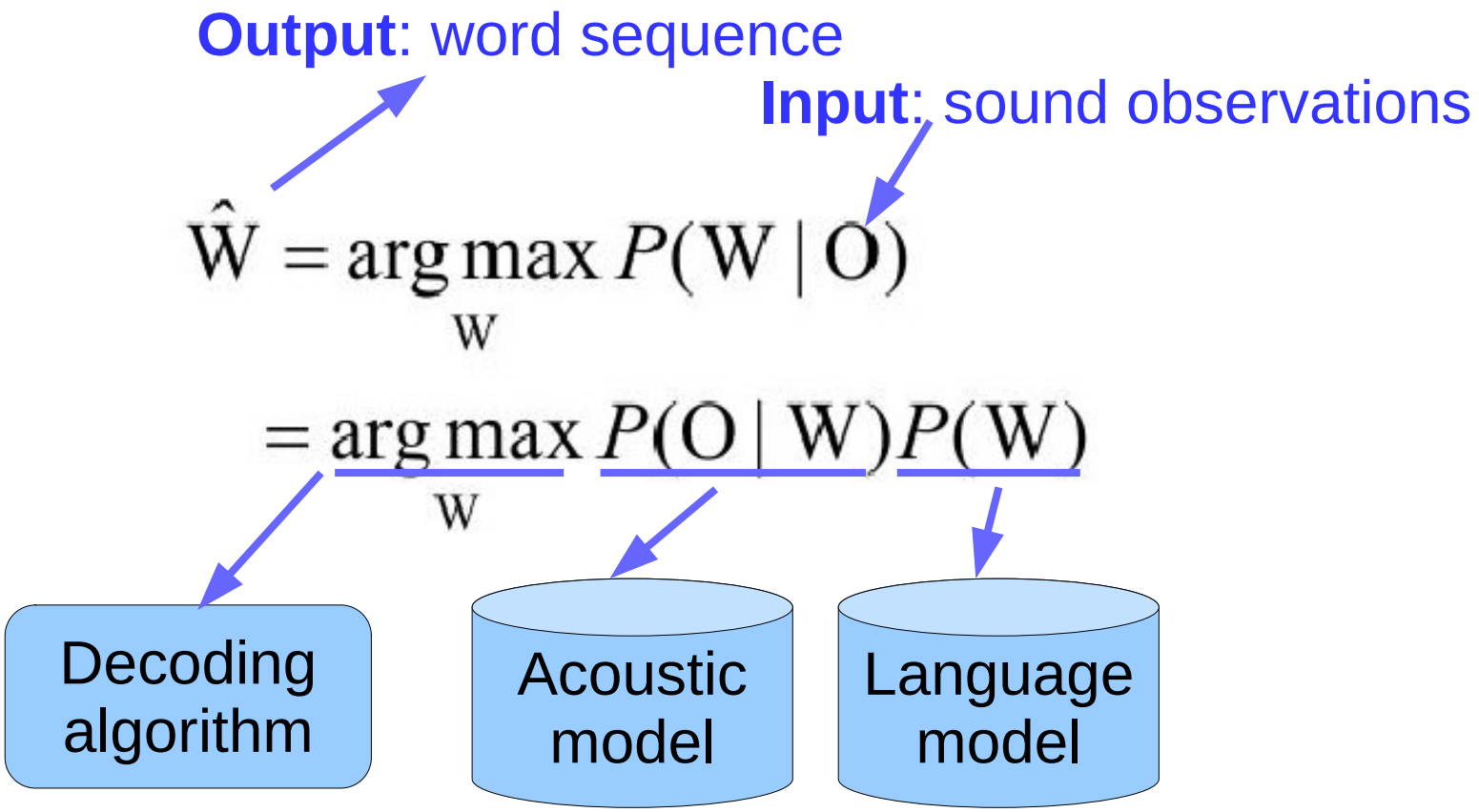
Start using Cortana

Cortana is your personal assistant on your Windows Phone 8.1, ready to kinds of tasks and keep you up to date on the things that you care about.

Siri.
Your wish is its command.

Siri lets you use your voice to send messages, schedule meetings, place phone calls, and more. Ask Siri to do things just by talking the way you talk.

Speech recognition: large probabilistic models

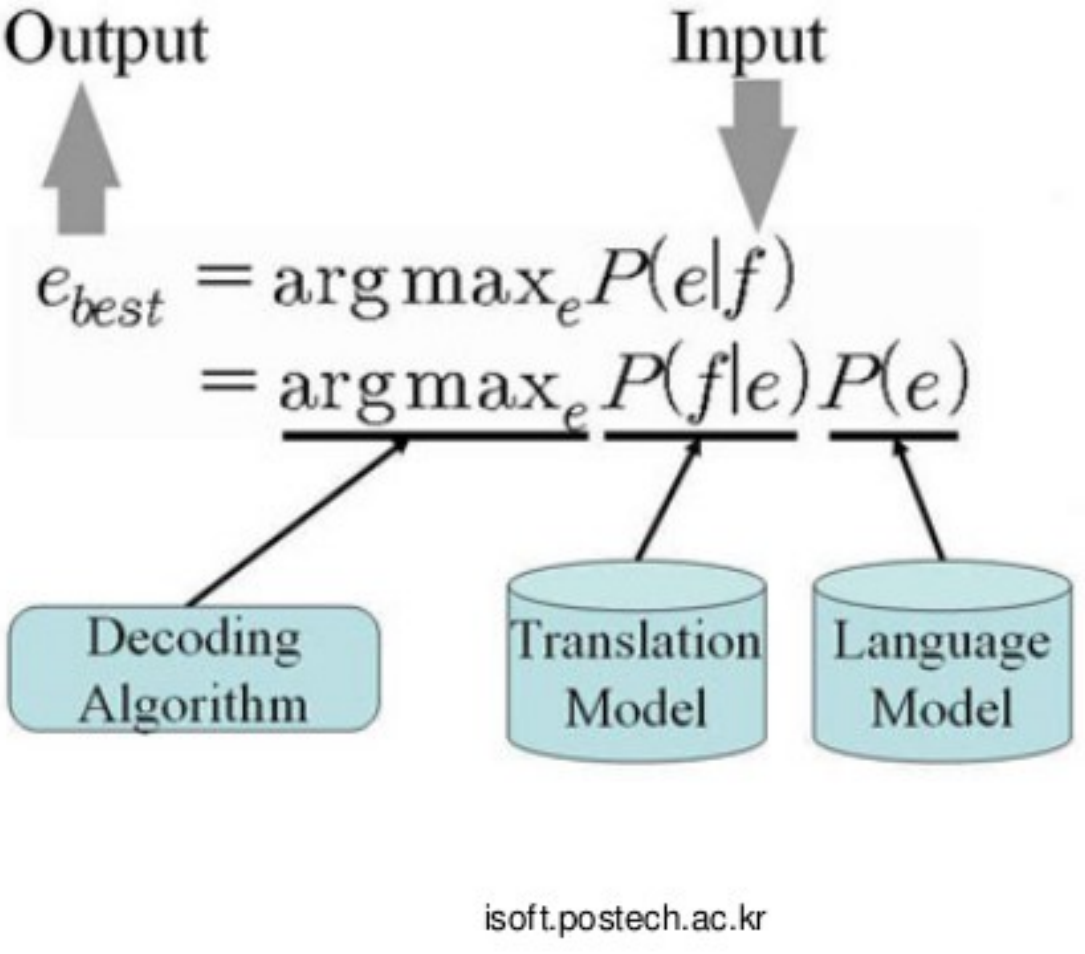


Machine translation

Google translate:

- Jos ei osaa suomen kieltä, on vaikea arvata sanojen merkityksiä.
- If you do not speak Finnish, it is difficult to guess the meanings of words.
- Wenn Sie nicht sprechen Finnisch, ist es schwierig, die Bedeutung der Worte erraten.
- Если вы не говорите по фински, трудно угадать смысл слов.
- 如果你不会讲芬兰语，很难猜测词的含义。

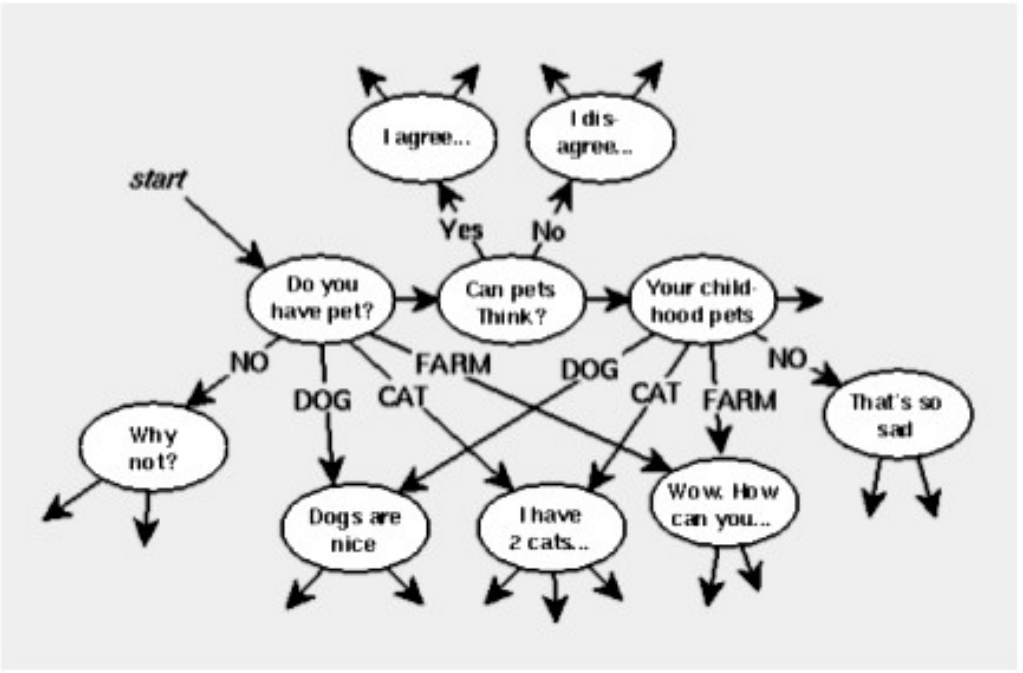
Machine translation: large probabilistic models



isoft.postech.ac.kr

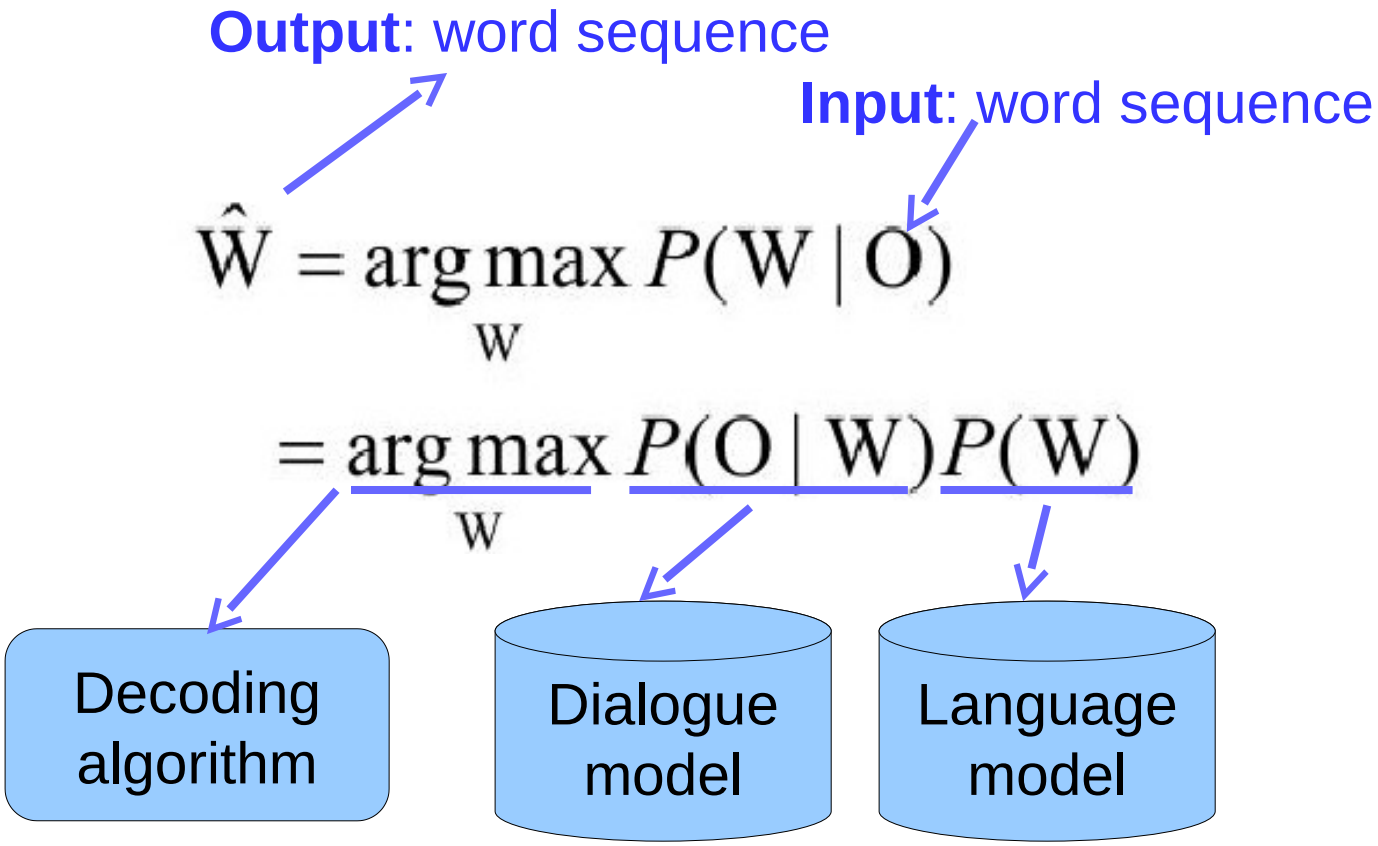
Natural language interfaces

www.zabaware.com



robot-club.com

Dialogue generation: a large probabilistic models point of view



Discussion

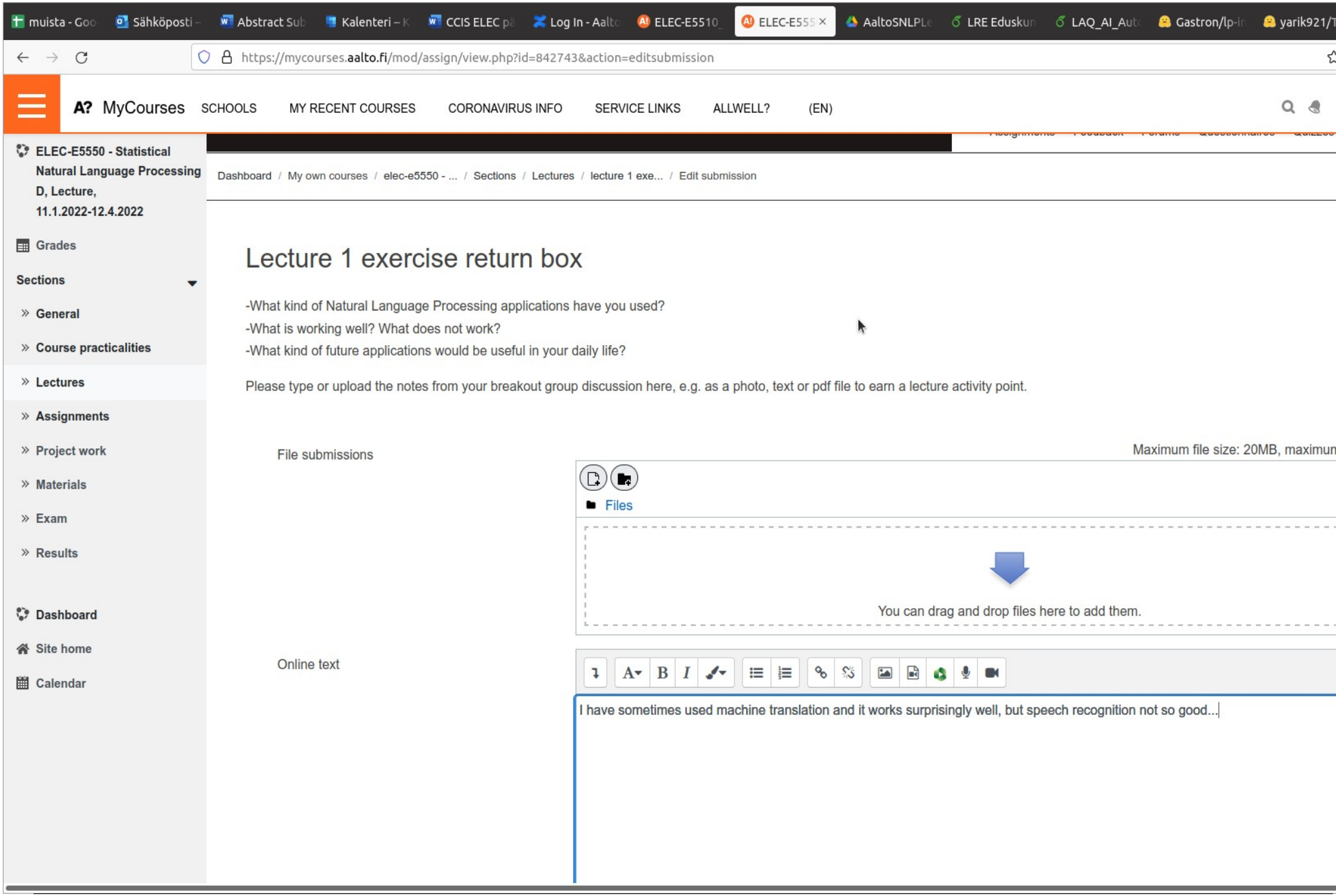
Discuss 10 mins in groups and write notes:

1. Introduce yourself to the group
2. What kind of Natural Language Processing applications have you used?
3. What is working well? What does not work?
4. What kind of future applications would be useful in your daily life?

To receive an activity point, submit your notes (photo, text or pdf):

- In MyCourses => Lectures => Lecture 1 exercise return box
- Hint: Write the notes directly in the text box while you discuss and submit

More about how to earn activity points and how they affect course grading will be discussed at the end of this lecture.


https://mycourses.aalto.fi/mod/assign/view.php?id=842743&action=editsubmission
MyCourses
SCHOOLS
MY RECENT COURSES
CORONAVIRUS INFO
SERVICE LINKS
ALLWELL?
(EN)
ELEC-E5550 - Statistical Natural Language Processing D, Lecture, 11.1.2022-12.4.2022
Grades
Sections
General
Course practicalities
Lectures
Assignments
Project work
Materials
Exam
Results
Dashboard
Site home
Calendar
Dashboard / My own courses / elec-e5550 - ... / Sections / Lectures / lecture 1 exe... / Edit submission
Lecture 1 exercise return box
-What kind of Natural Language Processing applications have you used?
-What is working well? What does not work?
-What kind of future applications would be useful in your daily life?
Please type or upload the notes from your breakout group discussion here, e.g. as a photo, text or pdf file to earn a lecture activity point.
File submissions
Maximum file size: 20MB, maximum
Files
You can drag and drop files here to add them.
Online text
I have sometimes used machine translation and it works surprisingly well, but speech recognition not so good...

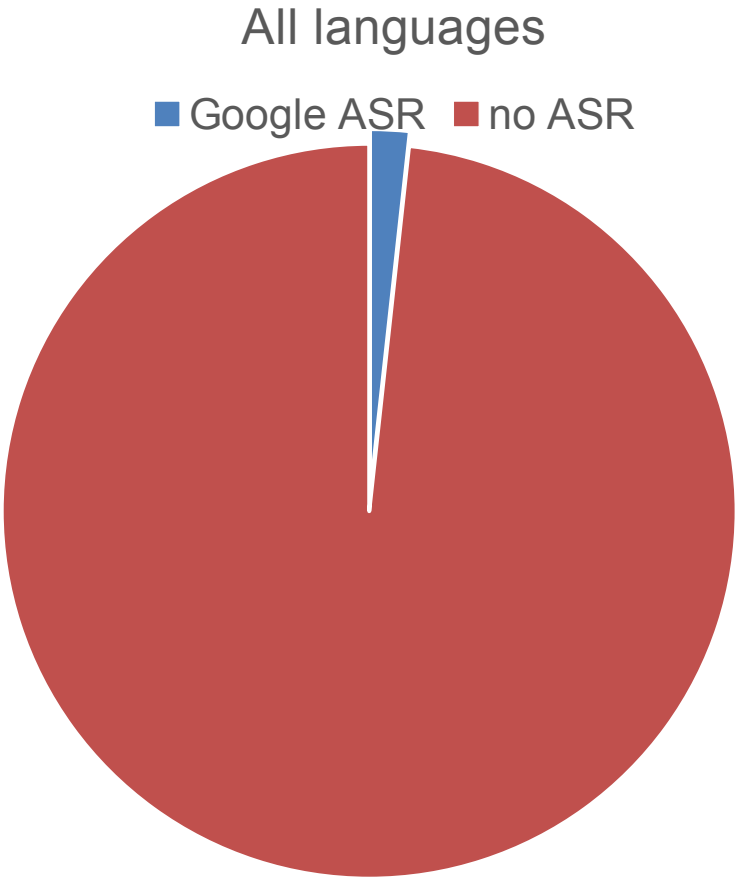
More application areas

Red text is A! Aalto University research

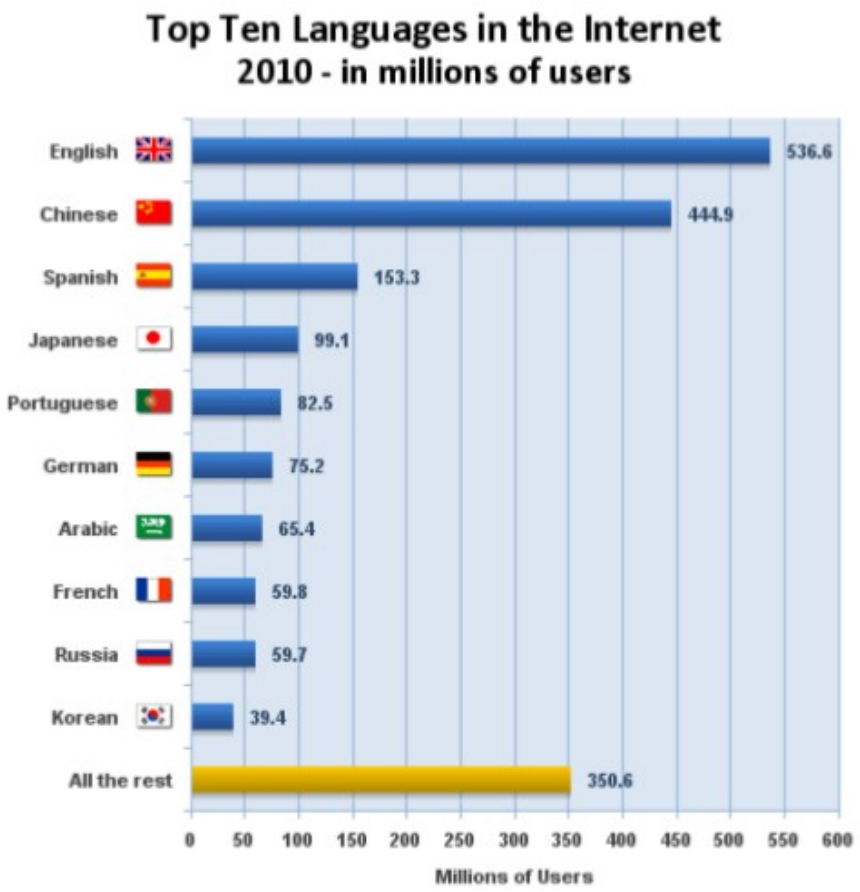
- Information retrieval
- Text clustering and classification
- Automatic speech recognition
- Natural language interfaces
- Automatic question answering, chatbots
- Text and image generation
- Machine translation
- Topic detection
- Sentiment analysis
- Word sense disambiguation
- Syntactic parsing
- Image, audio and video description
- Text-to-speech synthesis
- ...

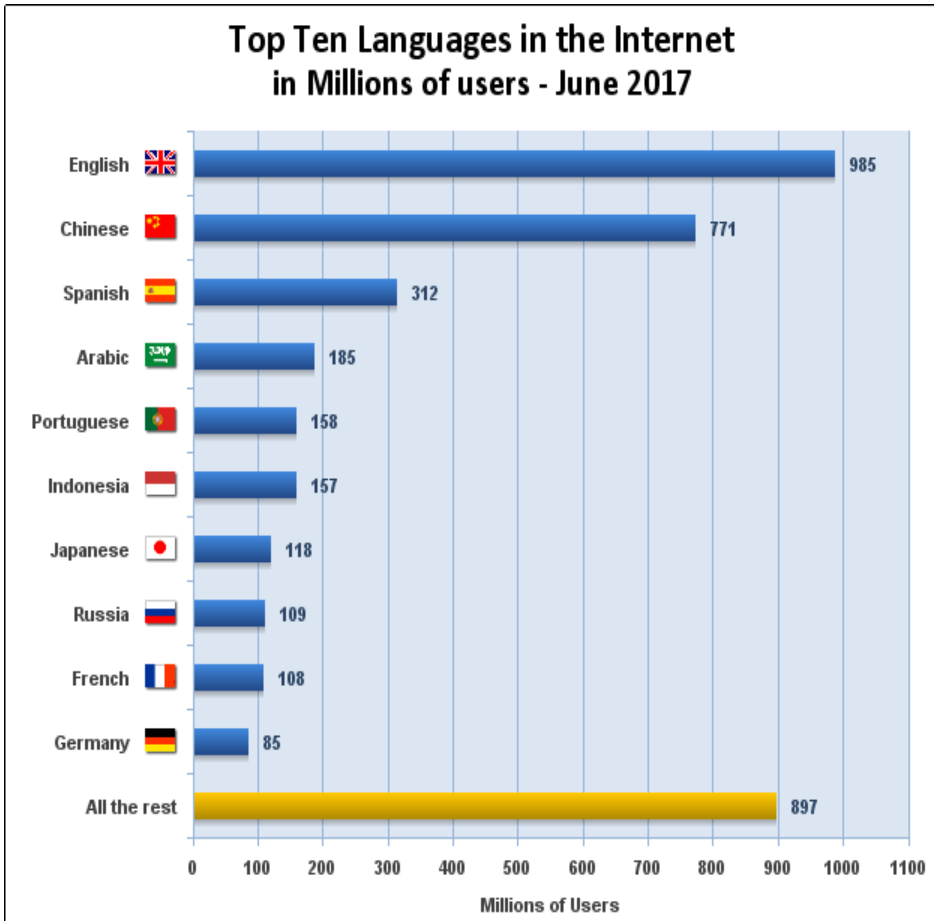
Complexity of natural languages

- 6000+ languages, many dialects
- Each has many words
- Each word is understood slightly differently by each speaker
- Large variety of sentence structures



Languages in the internet





www.internetworldstats.com

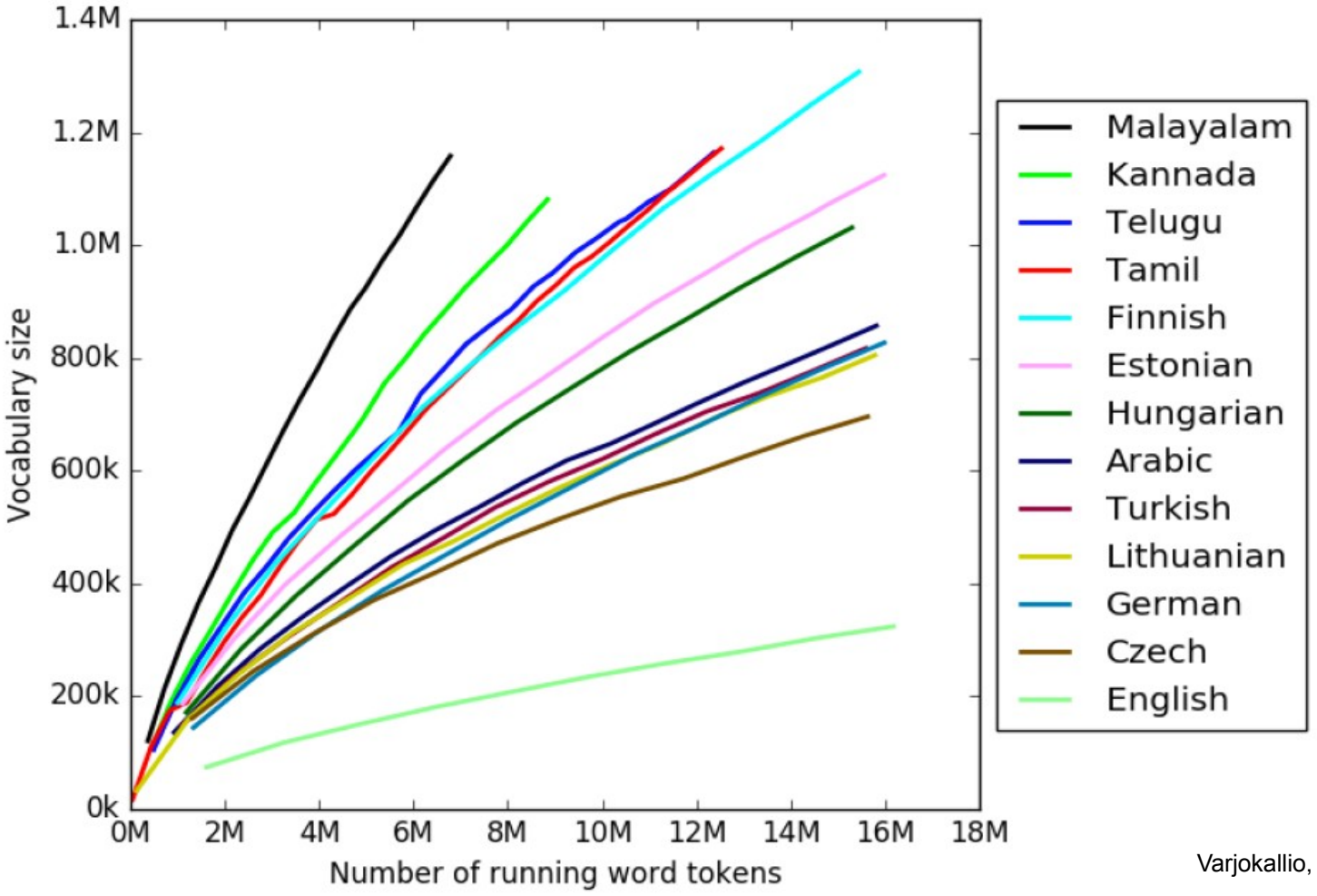
EU languages

Table 2. An example phrase from each of the EU languages.

cs	Smlouva o ústavě pro evropu
da	Traktat om en forfatning for europa
de	Vertrag über eine Verfassung für Europa
el	Συνθήκη για τη θέσπιση Συνταγμάτος Ευρώπης
en	Treaty establishing a Constitution for Europe
es	Tratado por el que se establece una constitución para Europa
et	Euroopa põhiseaduse leping
fi	Sopimus euroopan perustuslaista
fr	Traité établissant une Constitution pour l'Europe
ga	Conradh ag bunú Bunreachta don eoraip
hu	Szerződés európai alkotmány létrehozásáról
it	Trattato che adotta una Costituzione per l'Europa
lt	Sutartis dėl Konstitucijos Europai
lv	Līgums par konstitūciju eiropai
mt	Trattat Li Jistabbilixxi kostituzzjoni għall-Ewropa
nl	Verdrag tot vaststelling van een grondwet voor europa
pl	Traktat ustanawiają Konstytucję dla europy
pt	Tratado que estabelece uma Constituição para a Europa
sl	Zmluva o ústave pre Európu
sk	Pogodba o ustavi za evropo
sv	Fördrag om upprättande av en konstitution för europa

Effect of morphology: vocabulary size as function of corpus size

This is from A! Aalto University research



Varjokallio, Kurimo, Virpioja (2016)

Challenges of segmentation

- Modeling morphology –- segmenting words
  - istua "to sit", istuutua “to sit down”,
  - Istun "I sit", istahdan "I sit down for a while"
  - istahtaisin "I would sit down for a while"
  - istahtaisinko? "should I sit down for a while?"
  - istahtaisinkohan? "I wonder if I should sit down for a while?"
- Where are the word boundaries?

Hello World

周公吐哺

Challenge of modeling syntax

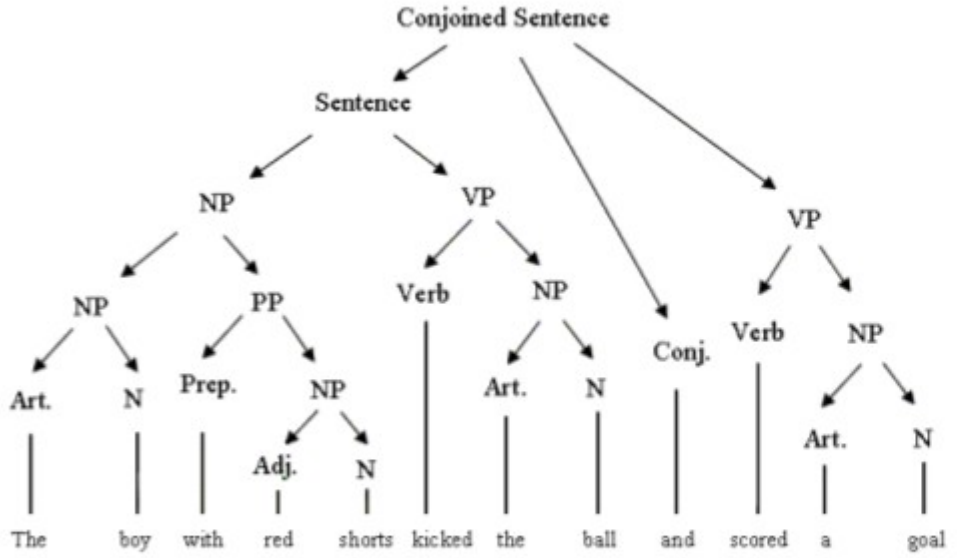


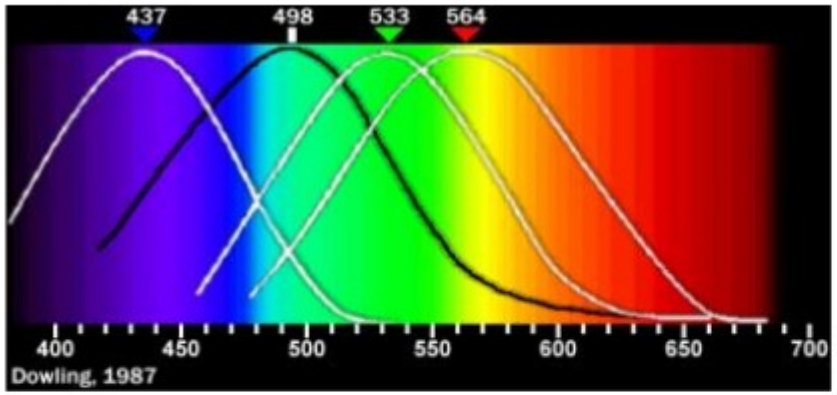
Figure 1.1.3.

"White House"
versus
"white house"

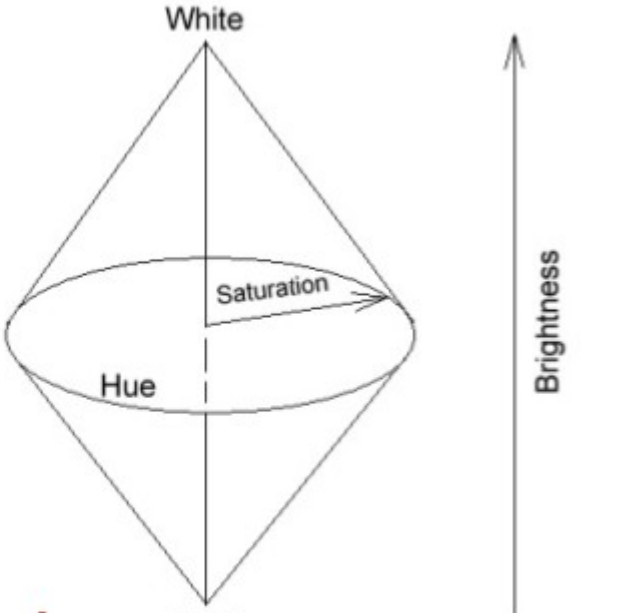
Challenges of natural language

- Understanding the meaning of words is subjective:
  - learning language through individual life paths
  - end up having different ways of understanding and producing language
- Many words have several meanings:
  - E.g. “play”, “game”, “window”
- Sentences have several interpretations:
  - E.g. “Big children and adults saw a man with a telescope”

Example: color naming



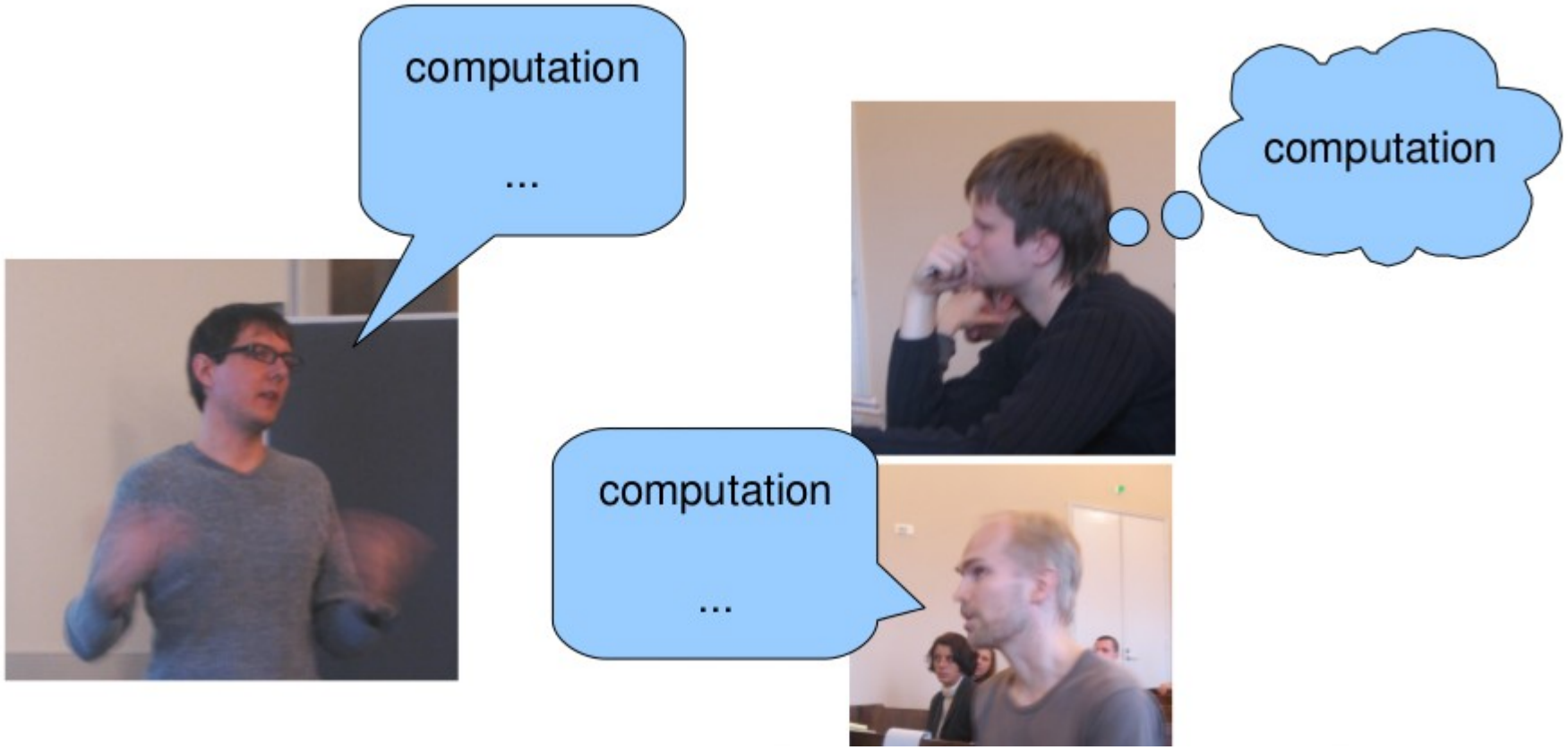
Human vision: rods, cones,...
Physical reasons for color
Contextuality of naming



red wine
red skin
red shirt

→ Hardin
→ Gärdenfors

Complex concepts: e.g. computation or modeling



Different cultural contexts

? ?

Shakespeare's sonnet:
“Shall I compare thee to a summer's day?”

Challenge of encoding world knowledge

- For good performance, world knowledge is needed
  - Quantitatively this is challenging
  - Qualitatively there are also many problems (mapping between language and the world is complex, cf. examples above)
- Note: world is essentially dynamic, continuous and multimodal, symbolic systems are not

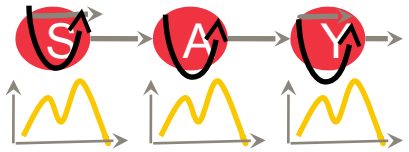
Corpus-based methods

- Corpora are large collections of text
  - Annotated: add knowledge about words or structure into corpus
  - Or just plain text
- Statistical information on
  - Distribution of words and parts of words
  - Structure
  - Word similarity
- Allow us to build models and **test** hypotheses
- Allow us to explore
- Choose the best models based on statistics

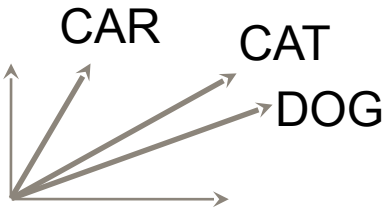
Natural language processing

METHODS

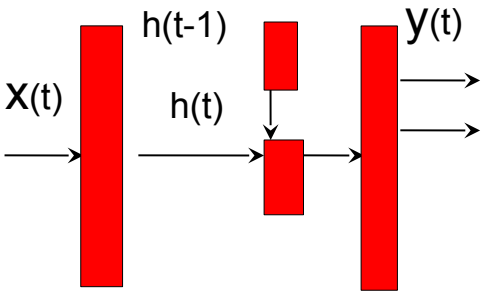
Hidden Markov model



Vector space model



Recurrent neural network



TOOLS

- Speech-to-text
- Text-to-speech
- Machine translation
- Information retrieval
- Named entity recognition
- Sentence parsing
- Topic detection

Natural language modeling: basic tasks

Red text is research

Aalto University

Word level

1. Vector space models
2. Text preprocessing
3. Bag of words models
4. Modeling morphology

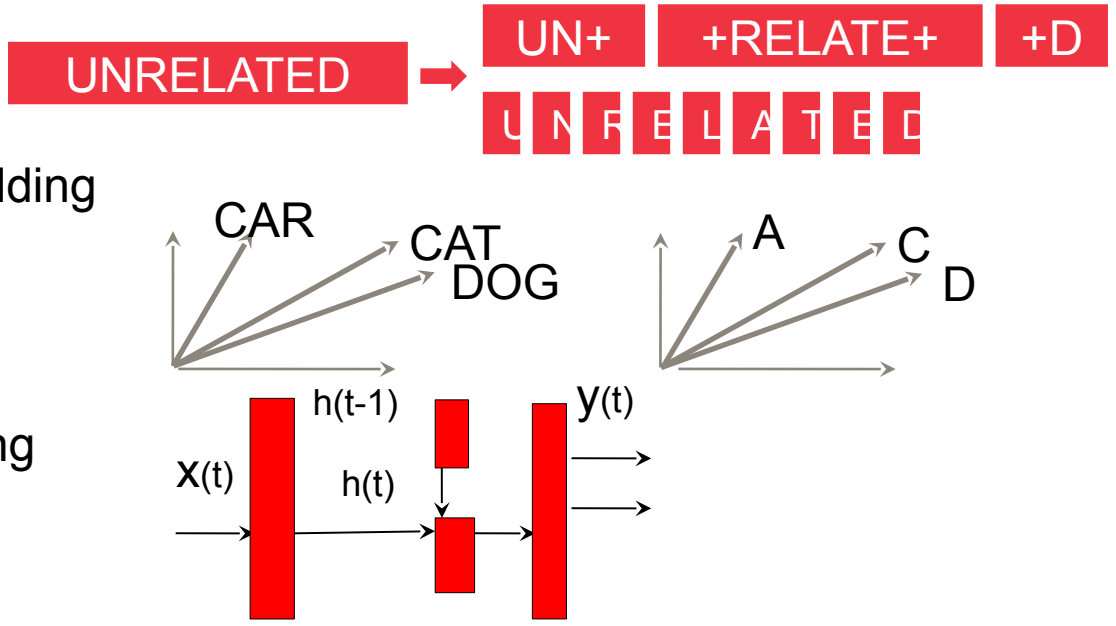
Sentence level

1. Part-of-speech tagging
2. Named entity recognition
3. Large language models

A recent revolution in the language modeling approach

- Split language into tokens
- Vector space modeling, embedding
- Representation learning
- Deep & recurrent learning
- Sequence to sequence mapping

=> artificial intelligence



Read more

- Manning & Schütze: Foundations of Statistical Natural language processing
  - Chapter 1: Introduction
  - Chapter 2: Probability and Information Theory basics

Part II: Course details

1. Goals
2. Materials and tools
3. Lectures
4. Exercises
5. Course project
6. Grading
7. Submission DLs

1. Goals

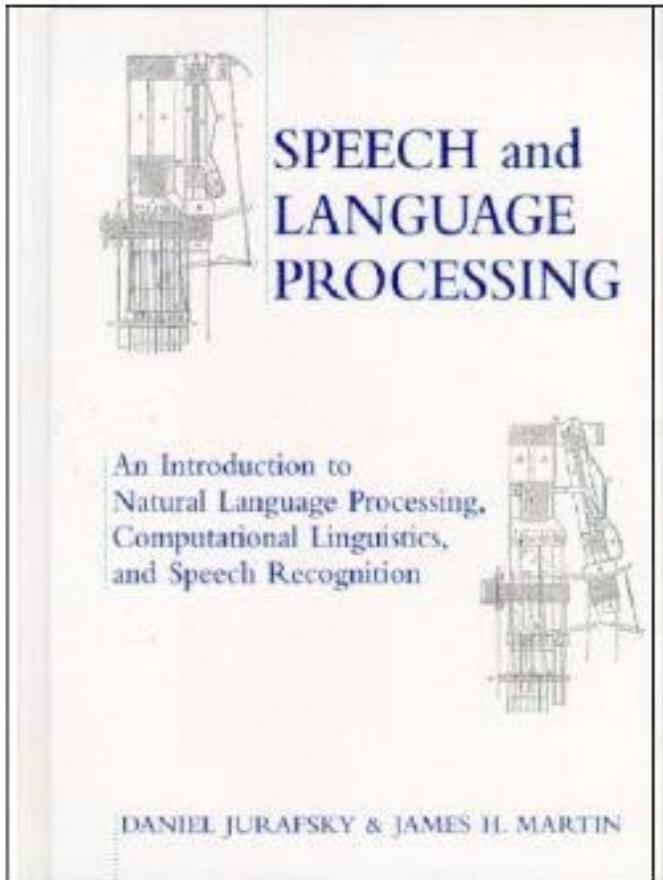
- To learn **how statistical and adaptive methods** are used in information retrieval, machine translation, text mining, speech processing and related areas **to process natural language data**
- To learn how to apply the basic methods and techniques for clustering, classification, generation and recognition **by natural language modeling**

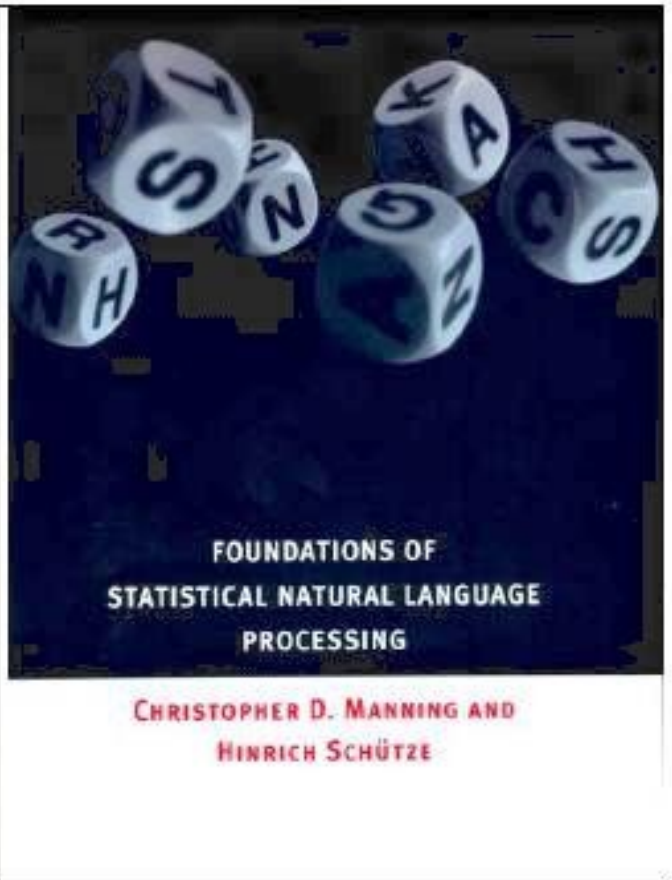
Course personnel

- Responsible professor & lecturer: *Mikko Kurimo*
- Assistant & exercises: *Ekaterina (Katja) Voskoboinik, Nhan Phan, Dejan Porjazovski, Anssi Moisio*
- Project work: *all the above + research group members*
- Visiting lecturers: *Mathias Creutz, Tiina Lindh-Knuutila, Sami Virpioja, Jaakko Väyrynen, Mittul Singh, Tamas Grosz, Shantiprya Parida*

2. Materials: Text books

1. **C. Manning, H. Schütze.** Foundations of Statistical Natural Language Processing.MIT Press,1999.*http://nlp.stanford.edu/fsnlp/*
2. **D. Jurafsky, J. H. Martin**. Speech and Language Processing (3rd ed. Draft, 2023) *http://web.stanford.edu/~jurafsky/slp3/*





2. Other materials and tools

- **MyCourses**: Lecture slides and recordings, assignments and project work
- **Zulip**: For questions and discussions before and after the lectures and exercise sessions, project groups can make their own public or private channels
- **Internet**: Plenty of books, articles, demos and course material from all over the world

3. Lectures

- **11 lectures: January 10 – March 28**
- Tue 12:15 – 14:00 in Rakentajanaukio 4, Hall R1 (Period 3)
- Tue 12:15 – 14:00 in Otakaari 3 Auditorio (Period 4)
- Visiting experts who have PhD and industrial experience in their topic. Typically they are the best experts available in Finland.
- Slides, links and other material provided by lecturers
- The lectures will be recorded and available for the course participants (target: one week after each lecture)
- Active attendance to the weekly lectures, studying the material and taking the exam corresponds to 2 cr amount of work.
- Participation to the lectures is not mandatory, but recommended for reaching the learning outcomes of the course.

Lectures in the course (changes possible)

09 jan 1 Introduction / Mikko Kurimo

16 jan 2 Statistical language models / Mikko Kurimo

23 jan 3 Sentence level processing / Mikko Kurimo

30 jan 4 Word2vec / Tiina Lindh-Knuutila

06 feb 5 Neural language modeling and large language models / Mittul Singh

13 feb 6 Morpheme-level processing / Mathias Creutz

20 feb Exam week, no lecture

27 feb 7 Speech recognition / Tamas Grosz

05 mar 8 Chatbots and dialogue agents / Mikko Kurimo

12 mar 9 Statistical machine translation / Jaakko Väyrynen

19 mar 10 Neural machine translation / Sami Virpioja

26 mar 11 LLMs in industry / Shantipriya Parida

02 apr 12 LLM discussion and course conclusion / Mikko Kurimo

See Mycourses for updates

4. Home assignments

- **9 weekly assignments** including first an introduction to JupyterHub + 7 autograded assingments + forum discussion
- The assignments are released on Tuesdays and the submission DL is in 10 days (until next week Friday). Exception: The first introduction DL is in 7 days (until next lecture)
- The workload of completing the home exercises corresponds to 1 cr
- There is one responsible TA for each home exercises and assistance is available via Zulip (only during working hours)

5. Course project

- Done in **groups of 3 students**
  - smaller groups only in exceptional cases
- **Two options:**
  - your own NLP topic (e.g. testing word2vec, BERT, GPT or other LLM for your own language or task), or
  - participate in a shared task (topic TBA)
- **Goal:** Learn to use word embeddings and/or apply pre-trained models by doing and reporting your own experiments

An example BERT task

- Try pre-trained BERT models for General Language Understanding Evaluation (GLUE) tasks https://gluebenchmark.com/tasks
- Tasks like sentiment analysis, document classification etc.
- Various pre-trained models are available
  - Bert-base-uncased
  - Bert-base-cased etc
- Analyse for your task which model performs best
- Experiments with dataset size and sequence length, batch size, learning rate

word2vec example task

1. Try existing word2vec embeddings

- Embeddings in various languages available

- Analyze: Are the nearest neighbors semantically meaningful? How? Why not?

2. Build your own model: English, Finnish, or any other language you are fluent with

3. Experiment with parameters: Window size, CBOW vs. skipgram, downsampling

4. Evaluate the model with evaluation set(s) of your choice: Nearest neighbor, noun categorization, Analogical reasoning

5. Evaluate the non-English model with a similar task, translate part of the evaluation set

6. EXTRA: Compare results between languages

word2vec programming

- Minimal programming skills needed
- Word2vec available in Python in Gensim package
  - available on Aalto machines
  - Can be also installed on your own computer
- For usage, see: https://radimrehurek.com/gensim/index.html
- original Word2vec package available in C
  - https://code.google.com/archive/p/word2vec
- For analysis, use whatever you want: Python, Matlab, R, CLUTO...

Course project grading

- See Mycourses for the requirements of an acceptable project report
- Peer grading will be performed for some parts to get more feedback, but that is separate from the final project grade
- Best projects typically include additional work such as
  - Exceptional analysis of the data
  - Application of the method to a task or several
  - Algorithm development
  - Own data set(s) (with preprocessing etc to make them usable)

Entrance survey

- The course includes a mandatory **entrance survey**.
- The purpose is to help in forming the project groups.
- It will also filter the students who aim at doing the project work, completing the course and getting the credits.
- It will also be used to find out expectations, preferences and background skills of the students by self-evaluation.

6. Course grading

- 20% of the grade comes from the **optional exam.** The exam will be organized in 16 April. Exam points are counted on top of the exercise points (see below) which are then cut to 2/3. Examples:
  - 40/60 exercises + 10/20 exam = 50/60 points (40/60 without exam)
  - 50/60 exercises + 15/20 exam = 55/60 points (50/60 without exam)
- 60% (or 40% + exam) of the grade is from the weekly **home assignments and lecture activities.**
  - The lecture activities (10/60) include pen&paper tasks, quizzes, discussions and feedback. To get the points return your solutions during the lecture or the day after, at the latest.
  - The home assignments (50/60) include manually and automatically graded assignments
- 40% of the grade is from the **course project**. It depends on experiments, literature study, short (video) presentation and final report. Course projects accepted in previous years are still a valid for completing the course.

7. Submission DLs

- First home assignment: **Jan 16**
- Entrance test is **Jan 22**
- Each weekly home assignment is released on Tuesdays and submitted by **Friday** in the following week (see MyCourses for info)
- The project group should select and register their topic by submitting a short abstract of what they plan to do by **Feb 9**
- Detailed project plan and Literature survey: 22 March
- Final project report: **April 19**
- Project Presentation video (5 min): **May 3**

See Mycourses for updates

How to achieve the learning goals (and pass the course)?

- *Participate actively in each lecture, read the corresponding material and ask questions* to learn the basics, take part in discussions, complete the lecture exercises
- *Solve each home assignment by yourself after each lecture* to learn how to solve the problems, in practice (you may collaborate, but everyone must write their own solution independently)
- *Participate actively in project work* to learn to apply your knowledge
- *Prepare well for the examination*

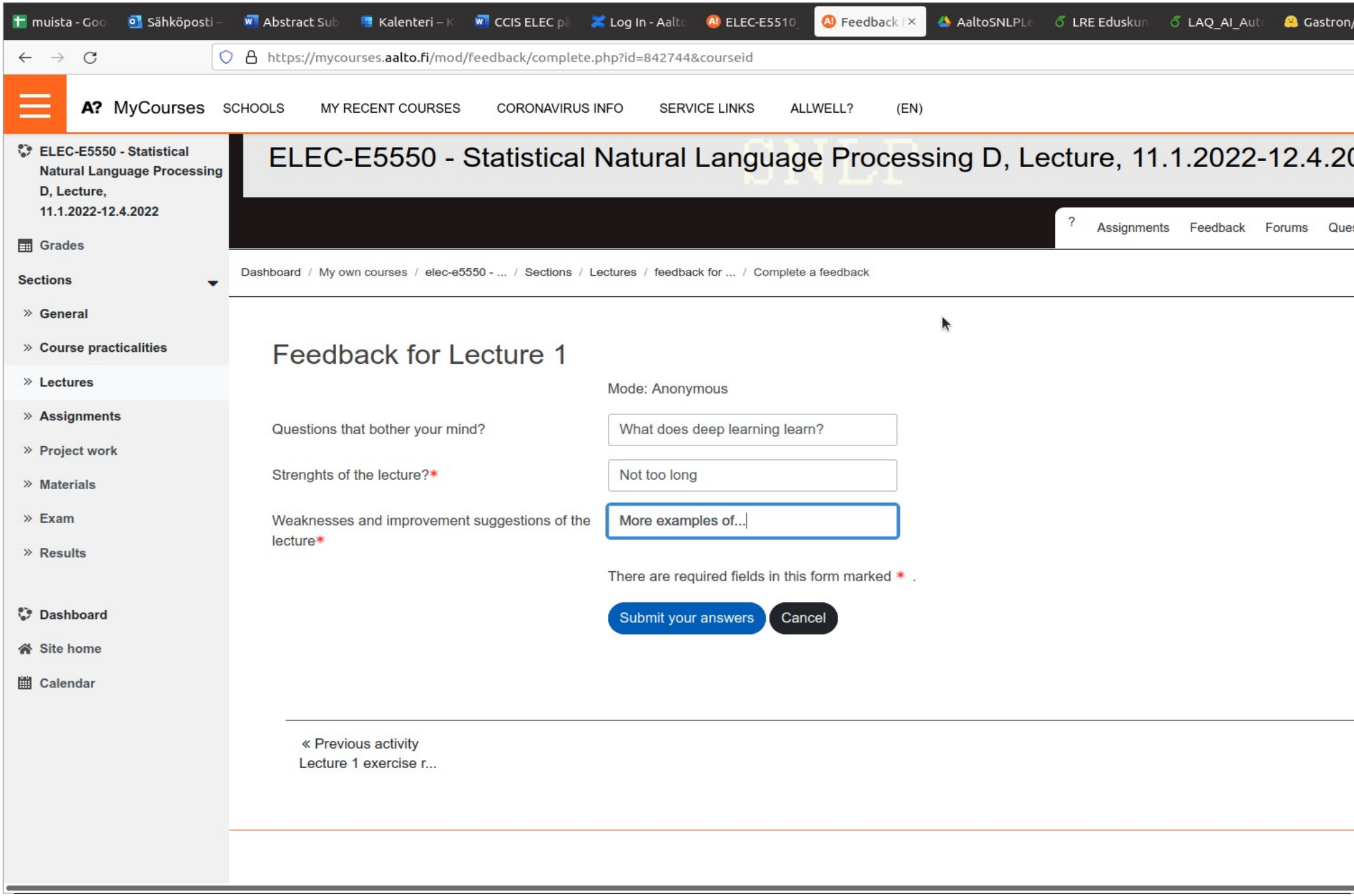
Feedback

Go to **MyCourses > Lectures > Lecture 1 feedback**

- To get an activity point submit the form before Thursday.
- Write down questions from the lecture that troubled your mind
- Suggest some strengths and weaknesses of the lecture

Idea's taken from last years' feedback:

- Replace demo exercises and exercise sessions by returned home assignments
- Replace the exam by collecting points from lecture activity and home assignments
- Present the results of the group works to other students
- Make lecture recordings available


muista - Goo
Sähköposti –
Abstract Sub
Kalenteri – K
CCIS ELEC pä
Log In - Aalto
ELEC-E5510_
Feedback
AaltoSNLPLe
LRE Eduskun
LAQ_AI_Aut
Gastron/
https://mycourses.aalto.fi/mod/feedback/complete.php?id=842744&courseid
A? MyCourses
SCHOOLS
MY RECENT COURSES
CORONAVIRUS INFO
SERVICE LINKS
ALLWELL?
(EN)
ELEC-E5550 - Statistical Natural Language Processing D, Lecture, 11.1.2022-12.4.2022
Grades
Sections
General
Course practicalities
Lectures
Assignments
Project work
Materials
Exam
Results
Dashboard
Site home
Calendar
ELEC-E5550 - Statistical Natural Language Processing D, Lecture, 11.1.2022-12.4.20
Assignments
Feedback
Forums
Dashboard / My own courses / elec-e5550 - ... / Sections / Lectures / feedback for ... / Complete a feedback
Feedback for Lecture 1
Mode: Anonymous
Questions that bother your mind?
What does deep learning learn?
Strenghts of the lecture?*
Not too long
Weaknesses and improvement suggestions of the lecture*
More examples of...
There are required fields in this form marked * .
Submit your answers
Cancel
« Previous activity
Lecture 1 exercise r...

Questions?

- Responsible professor & lecturer: *Mikko Kurimo*
- Projects & exercises: *Ekaterina Voskoboinik, Nhan Phan, Dejan Porjazovski, Anssi Moisio*
- Emails: *firstname.lastname@aalto.fi*
- Home page*:* *https://mycourses.aalto.fi/course/view.php?id=39341*