

# TKO 7095 Intro to human language technology

---

## HLT

### What is HLT

Natural language equals human language, different from formal and artificial languages such as programming languages

#### 2 key challenges

- Variety:
  - Same meaning can be expressed in many different ways
  - You don't get it - You don't understand it
- Ambiguity
  - Same form can express different meanings based on context
  - different meanings for eg. tie

## NLP

Computational methods to analyze, understand or generate human language

### Applications

Language is the most efficient way we have to communicate *meaning*

- Reasonably easy to describe a picture or music using words
- Very hard to describe music using a picture, or a book using dance

A large portion of what is presented as AI is actually build on NLP models:

- DALL-E
- MusicLM

#### Text correction and generation

- Spelling and grammar check
- Autocorrect and predictive text
- ChatGPT

#### Web search includes a lot of NLP

- Disambiguation
- Text classification
- Information retrieval
- Question answering
- Information extraction

#### Machine Translation

- DeepL

Chatbots and generative AI

Text to speech

- Eg. speak written books with ai machine voice

Speech recognition and subtitling

Filterin and moderation support via:

- Sentiment analysis
- Text filters

## Language as data

### Methods

- Simple statistics == counting
  - How many words / chars
  - word frequencies
- Segmentation == Divide bigger units into smaller ones
  - Almost always part of raw text preprocessing in NLP pipeline
  - *Tokenization / word segmentation*
    - Segment text into individual tokens == word-like elements
    - *Sentence splitting / sentence segmentation* == Segment text into individual sentences