IN1140: Introduksjon til språkteknologi

Forelesning #1

Lilja Øvrelid

Universitetet i Oslo

25 august 2021



Tema for i dag

- Introduksjon
- Hva er språkteknologi?
- Hva er IN1140?
- Praktiske detaljer
  - Grupper
  - Obliger
  - Lærebøker
  - Kontakt
  - m.m.



Korona

- Smittevern er fortsatt viktig
  - Hold avstand
  - God håndhygiene
  - Bli hjemme hvis du er syk
  - Gode digitale alternativer
- Oppdatert informasjon om UiO og korona:
  `https://www.uio.no/om/hms/korona/`

Hei

Forelesere

- Lilja Øvrelid (liljao@ifi.uio.no)
- Fra språkteknologigruppen (LTG)
- Samia Touileb (Samia.Touileb@uib.no), UiB

Hei

Forelesere

- Lilja Øvrelid (liljao@ifi.uio.no)
- Fra språkteknologigruppen (LTG)
- Samia Touileb (Samia.Touileb@uib.no), UiB

Gruppelærere

- Fredrik Aas Andreassen (fredaan@ifi.uio.no)
- Nikolas Martin (nikolhm@student.ikos.uio.no)
- Lilja Charlotte Storset (liljacs@ifi.uio.no)

Hei

Forelesere

- Lilja Øvrelid (liljao@ifi.uio.no)
- Fra språkteknologigruppen (LTG)
- Samia Touileb (Samia.Touileb@uib.no), UiB

Gruppelærere

- Fredrik Aas Andreassen (fredaan@ifi.uio.no)
- Nikolas Martin (nikolhm@student.ikos.uio.no)
- Lilja Charlotte Storset (liljacs@ifi.uio.no)

Hvem er dere?

Fyll inn nettskjema på emnesiden (under beskjeder)!

Hva er språkteknologi?

- Mål: å få datamaskiner til å 'forstå' naturlige språk.
- Aka:
  - computational linguistics (datalingvistikk)
  - language technology
  - language engineering
  - natural language processing (NLP)

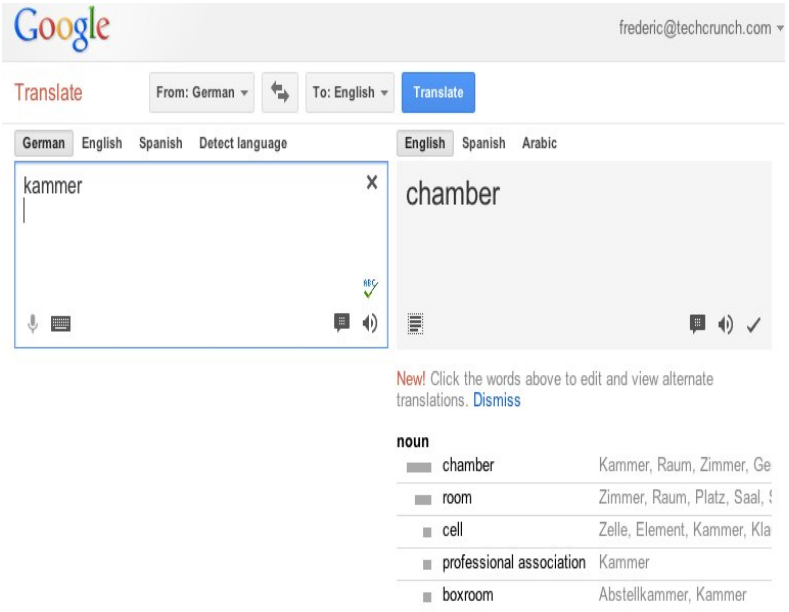


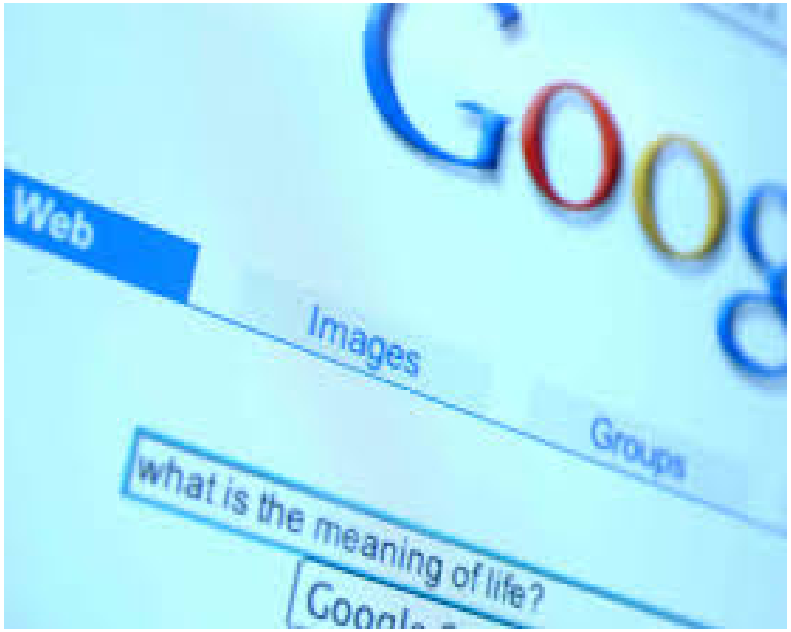
Eksempler på språkteknologi?

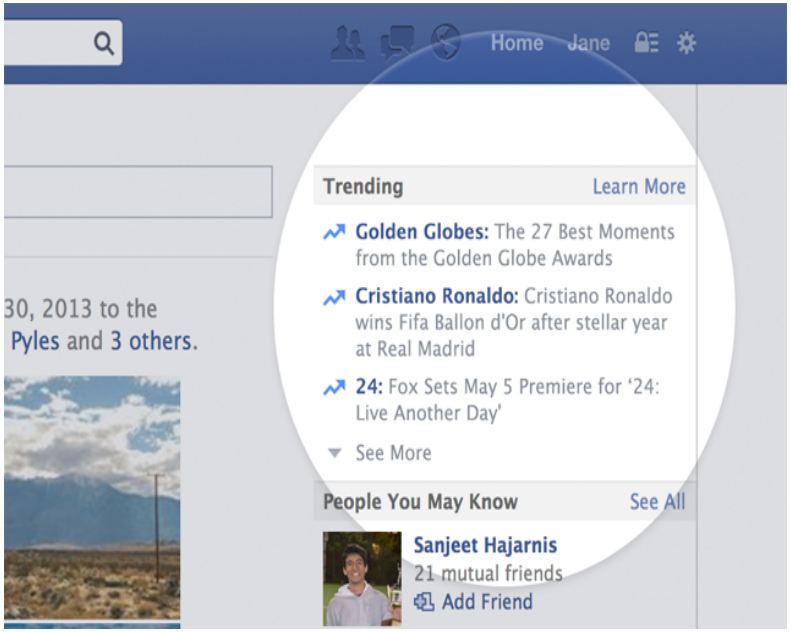


Eksempler på språkteknologi?









Språkteknologi og tverrfaglighet

NLP er et tverrfaglig felt

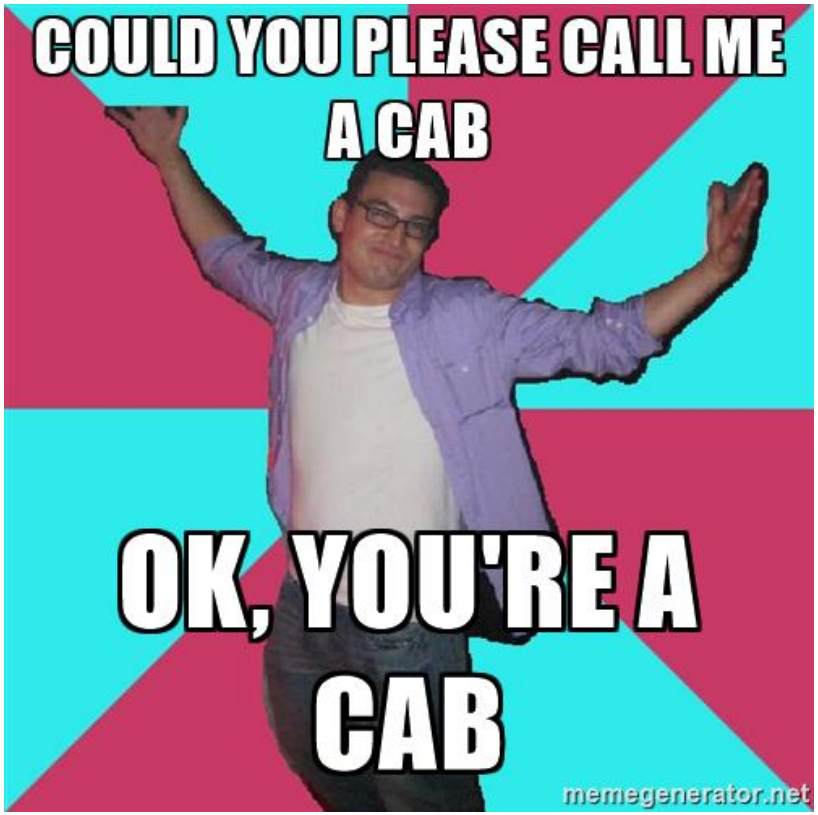
- Lingvistikk
- Informatikk
- Statistikk
- Maskinlæring
- Logikk, Filosofi, Psykologi, ...



- Del av det bredere feltet kunstig intelligens (AI).

Hvorfor er språkforståelse utfordrende?

- Språk er vagt, ulike tolkninger mulig.
- Flertydighet overalt.
- Gir kompakt kommunikasjon:
- Samme uttrykk kan brukes i ulike kontekster.



- Flertydighetene er stort sett usynlige for oss, vi finner den intenderte tolkningen nærmest ubevisst.
- For maskiner er det motsatt: lett å finne alle mulige tolkninger, men vanskelig å se hvilken som er riktig.

Eksempel: Flertydighet på ordnivå

- Norsk: *rett*.
- Engelsk: *?*
- Flertydig ift betydning + ordklasse (verb, subst., adj., adv.).
- Vi trenger kontekst for å avgjøre.

Eksempel: Flertydighet på ordnivå

- Norsk: *rett*.
- Engelsk: *?*
- Flertydig ift betydning + ordklasse (verb, subst., adj., adv.).
- Vi trenger kontekst for å avgjøre.

avgrenset av en rett linje tvers over kanalen	*straight*
Hva er rett svar?	*correct, right*
lovbestemt rett til innsyn	*right*
Denne rett avsa enstemmig dom i saken 4. juli 1980	*court*
Norsk rett tilpasses EUs regelverk	*law*
Vennligst rett disse prøvene!	*grade, correct*
Det bar rett i fengsel	*directly, straight*
De spiste en deilig rett av grønnsaker.	*meal, dish*
han var rett utenfor, rett nå	*just*
Slikt skjer rett som det er.	må omskrives

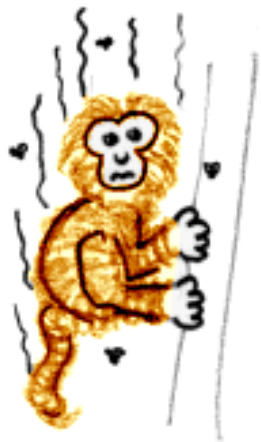
Eksempel: Flertydighet i referanse

We gave the monkeys$_1$ the bananas$_2$

... because they$_1$ were hungry.

... because they$_2$ were ripe.

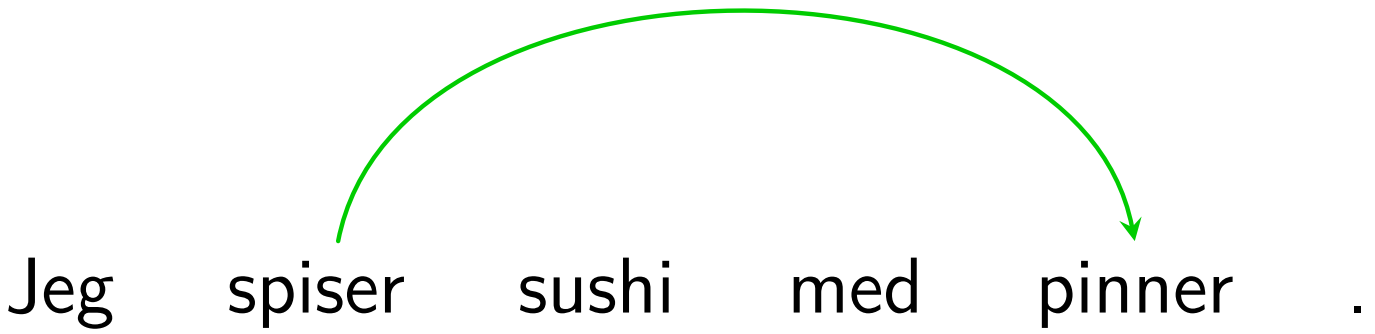




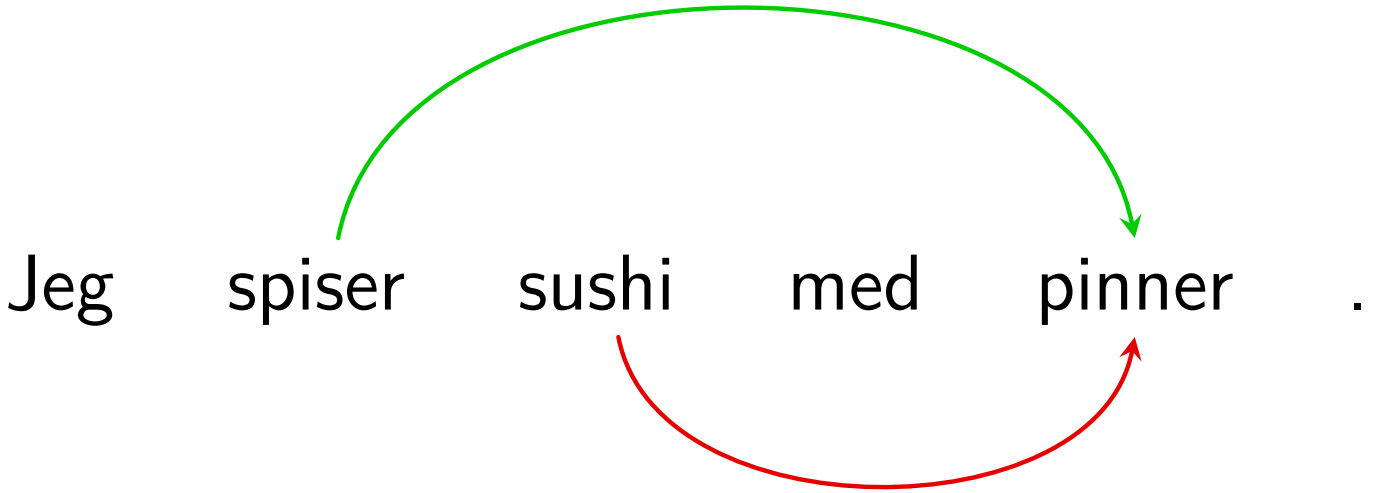
Flertydighet på setningsnivå

Jeg spiser sushi med pinner .

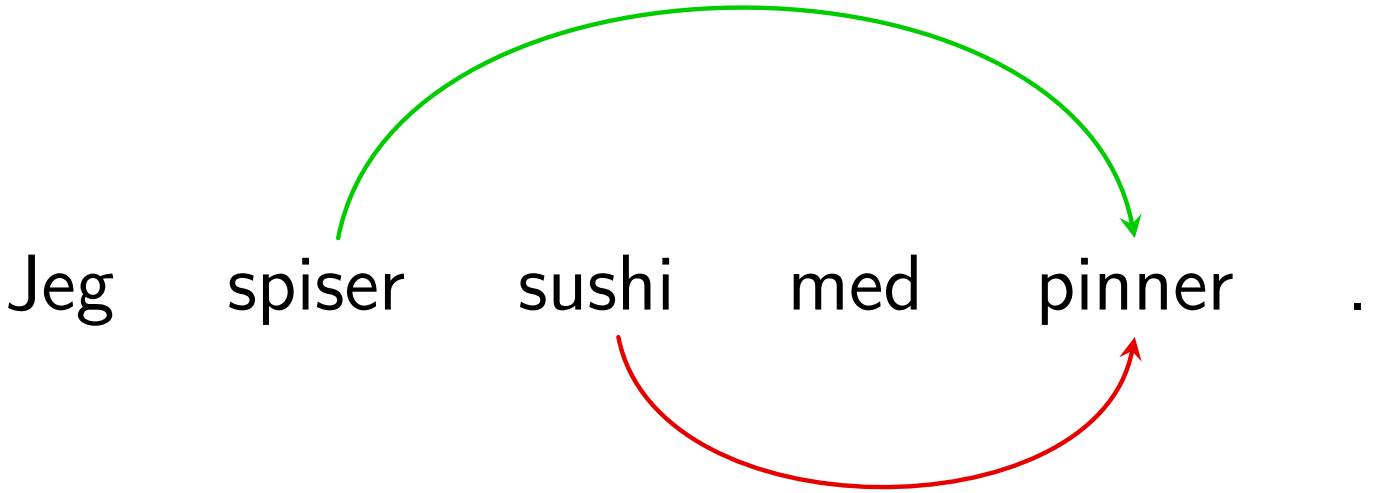
Flertydighet på setningsnivå



Flertydighet på setningsnivå

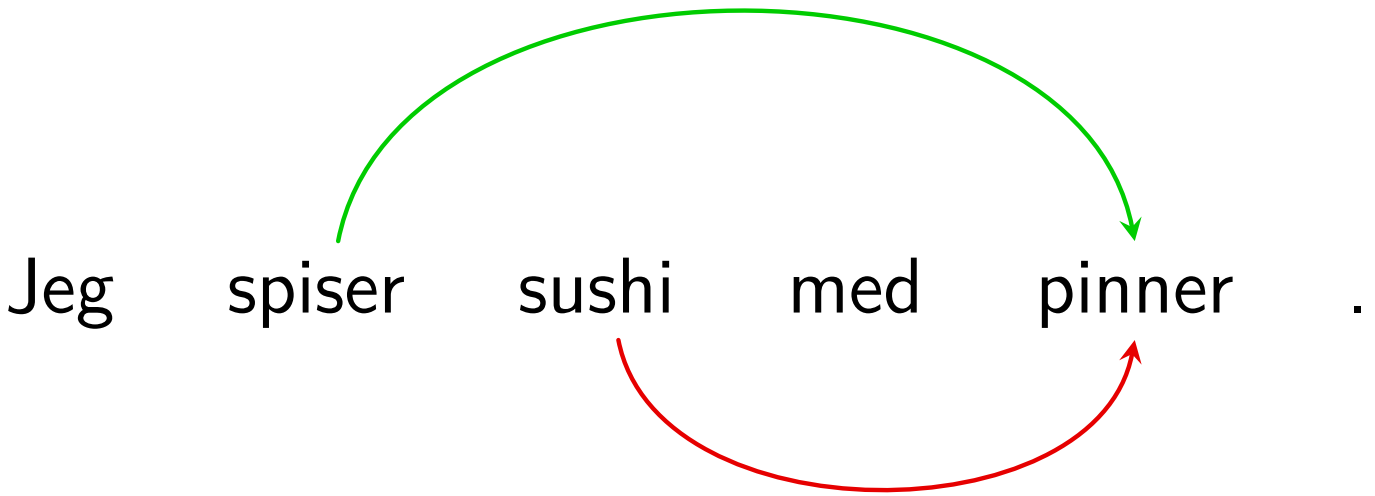


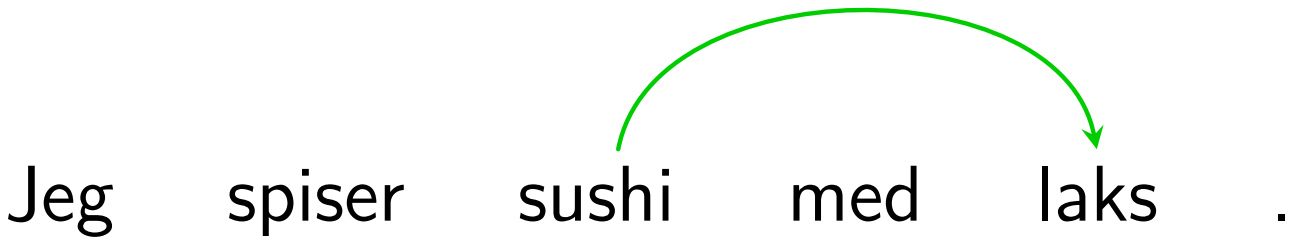
Flertydighet på setningsnivå



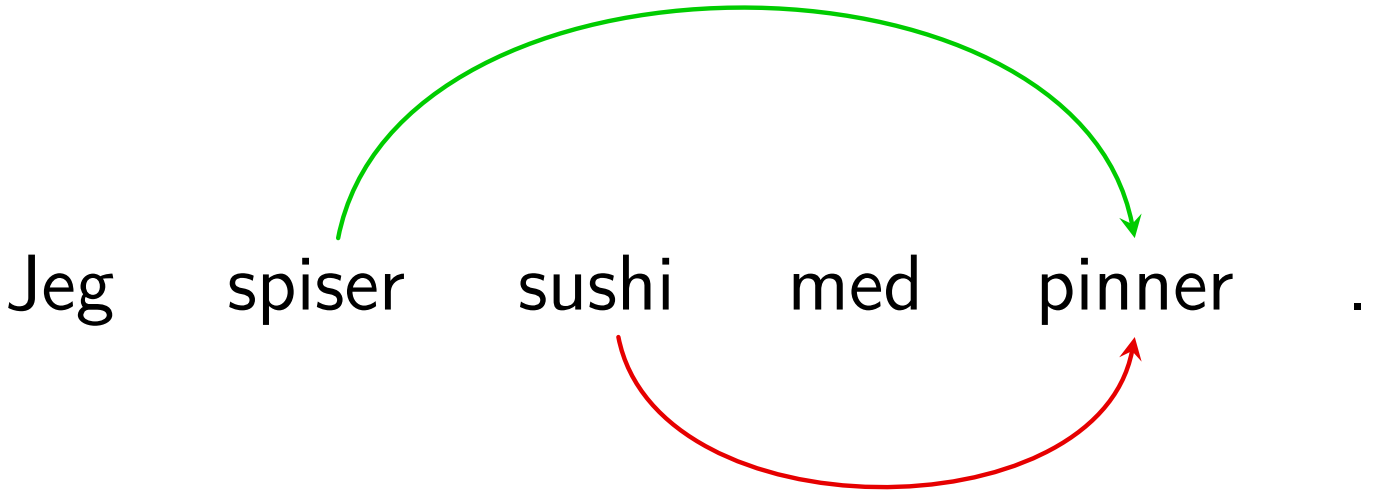
Jeg spiser sushi med laks .

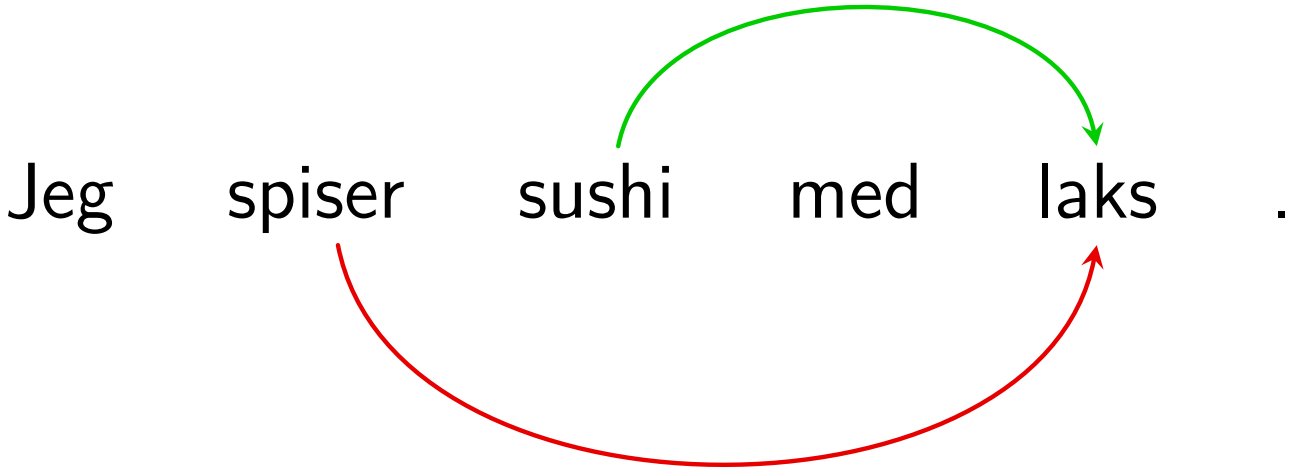
Flertydighet på setningsnivå





Flertydighet på setningsnivå





Språkforståelse er vanskelig!

The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard. The mental abilities of a four-year-old that we take for granted — recognizing a face, lifting a pencil, walking across a room, answering a question — in fact solve some of the hardest engineering problems ever conceived. . . As the new generation of intelligent devices appears, it will be the stock analysts and petrochemical engineers and parole board members who are in danger of being replaced by machines. The gardeners, receptionists, and cooks are secure in their jobs for decades to come.

Steven Pinker, *The language instinct*

Språkforståelse er vanskelig!

The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard. The mental abilities of a four-year-old that we take for granted — recognizing a face, lifting a pencil, walking across a room, answering a question — in fact solve some of the hardest engineering problems ever conceived. . . As the new generation of intelligent devices appears, it will be the stock analysts and petrochemical engineers and parole board members who are in danger of being replaced by machines. The gardeners, receptionists, and cooks are secure in their jobs for decades to come.

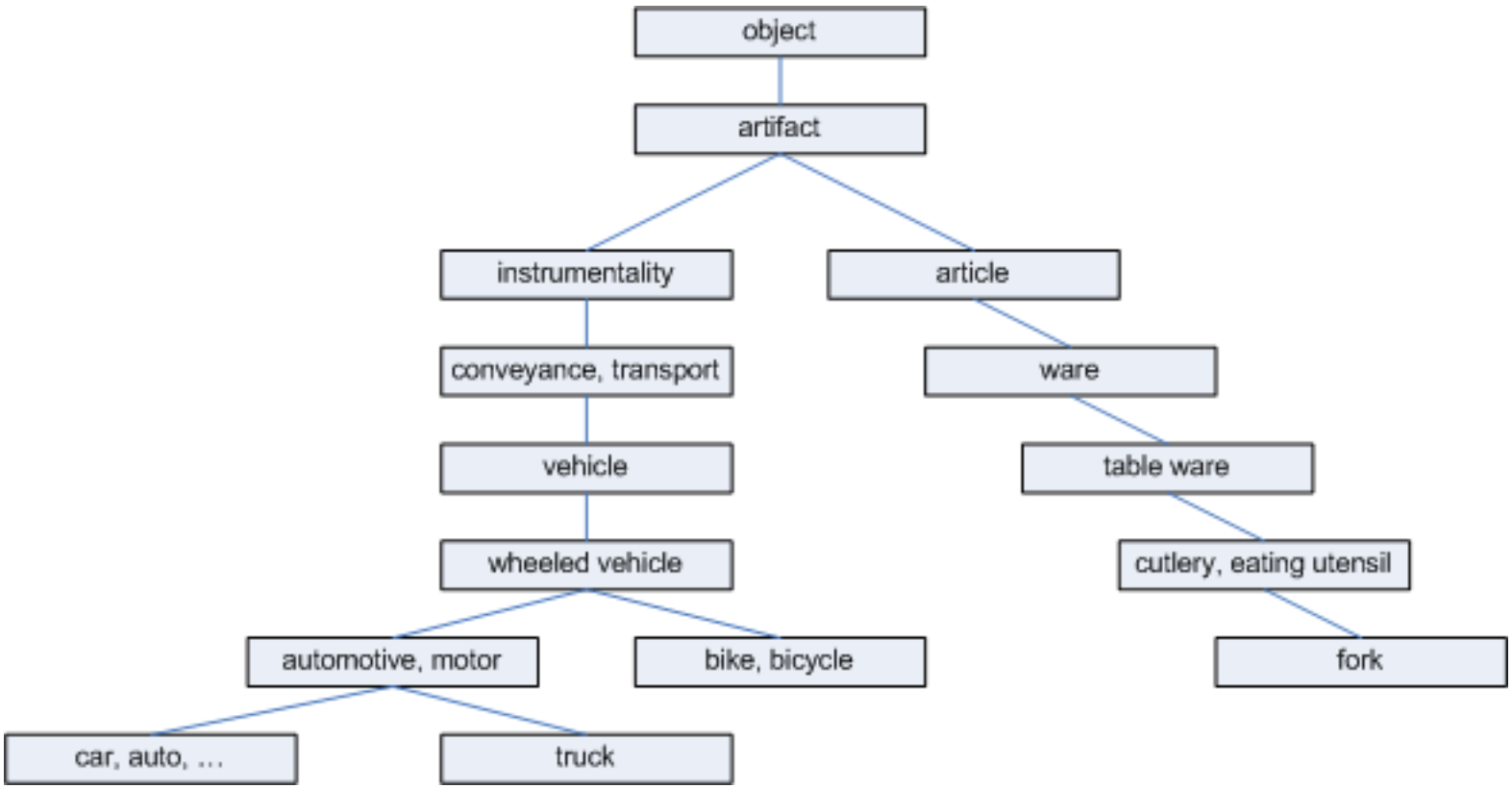
Steven Pinker, *The language instinct*

- En robot som bretter et håndkle (videoen er 50 ganger normal hastighet): `http://www.youtube.com/watch?v=gy5g33S0Gzo`

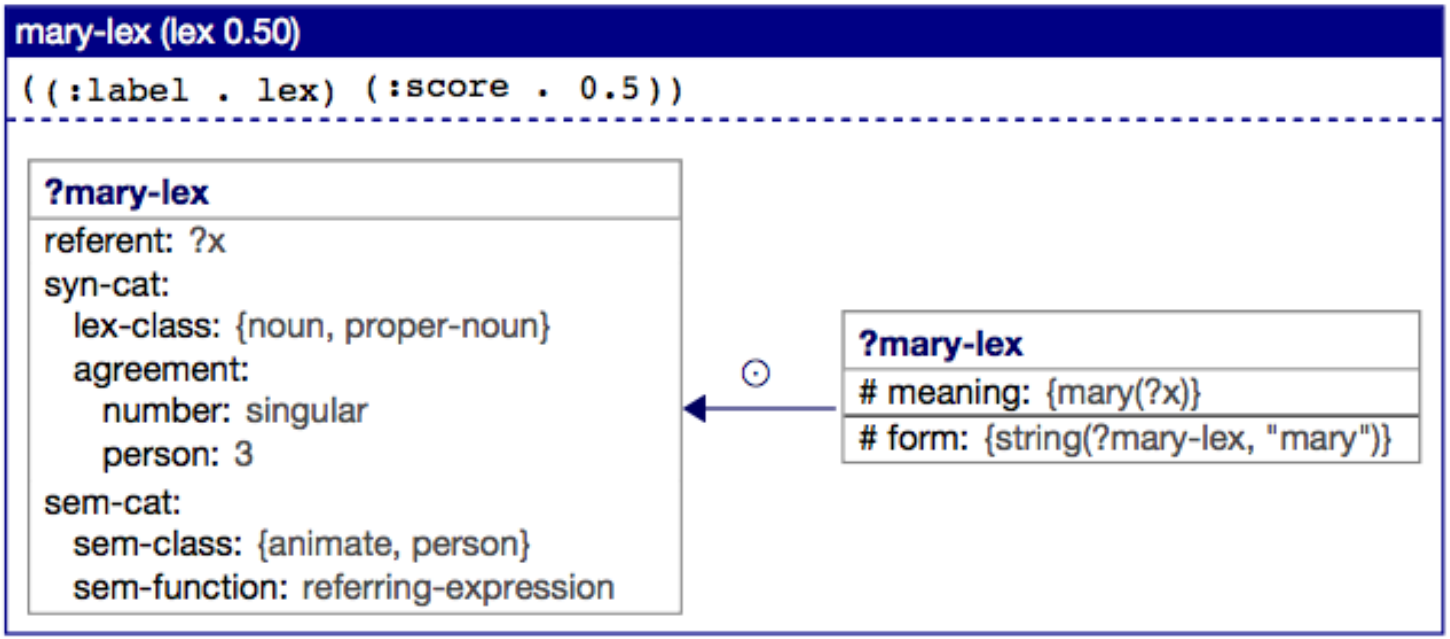
Entydiggjøring

- Vi mennesker tolker språklige uttrykk basert på delt bakgrunnskunnskap og gjensidige forventninger i en gitt kontekst.
- Språkforståelse handler mye om entydiggjøring.
- Språkteknologi, og IN1140, handler i stor grad om strategier for hvordan maskiner kan takle dette.

→ 2000-tallet: manuelt utformede regler og leksikon



→ 2000-tallet: manuelt utformede regeler og leksikon

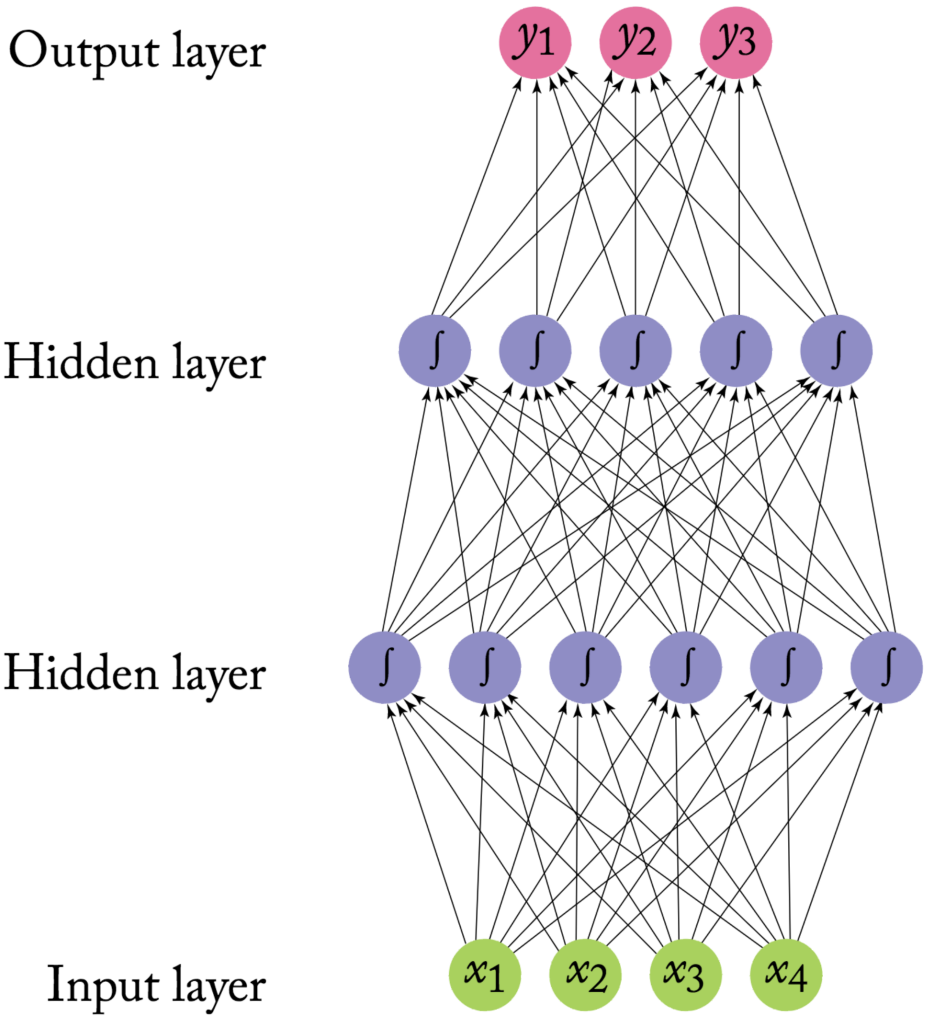


Språkteknologiske metoder

- 2000-tallet $\rightarrow$: empirisk revolusjon
- **Maskinlæring**
  - Datamaskiner kan lære fra data: fange opp mønstre og generalisere til nye eksempler



- 2014- $\rightarrow$: Nevrale Nettverk (NN), også kalt "dyp læring"



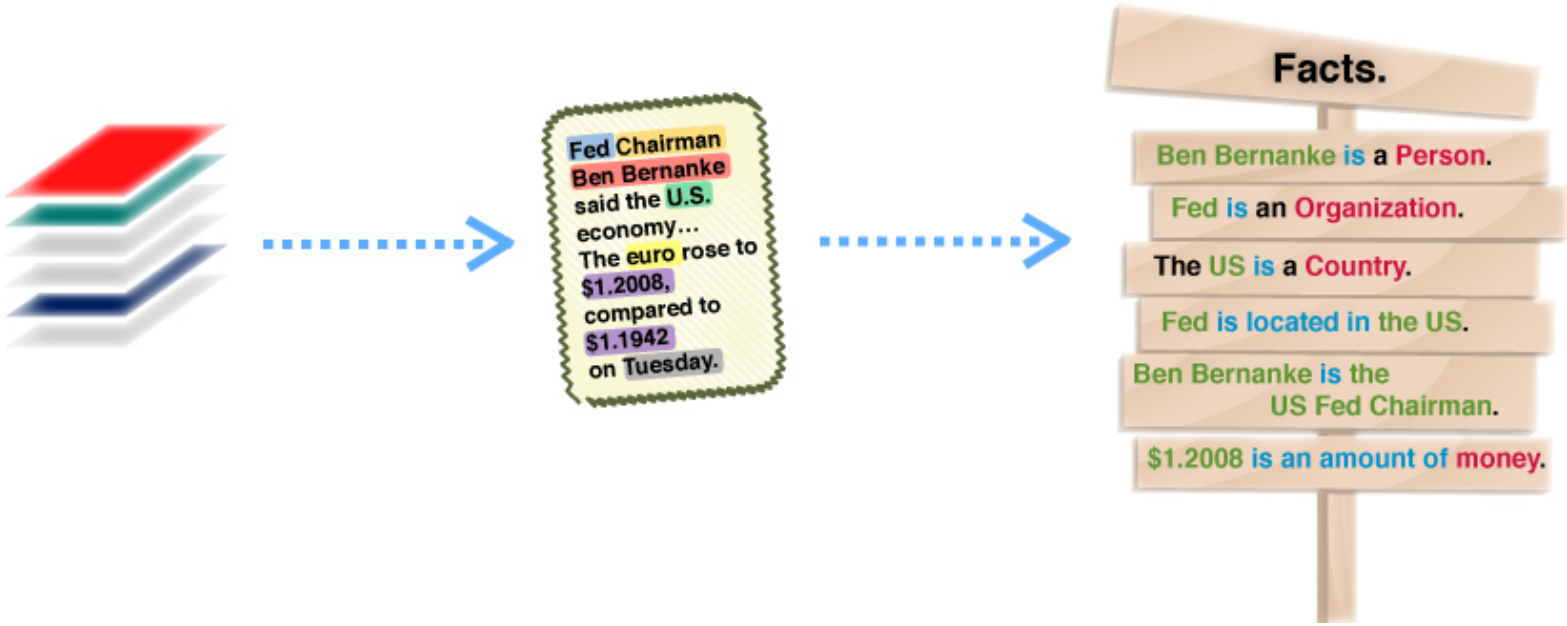
- NN er en familie kraftige maskinlæringmodeller
- Har revolusjonert språkteknologi som felt
- 'Dyp Læring': NN med flere skjulte lag.

Hva kan vi bruke språkteknologi til?



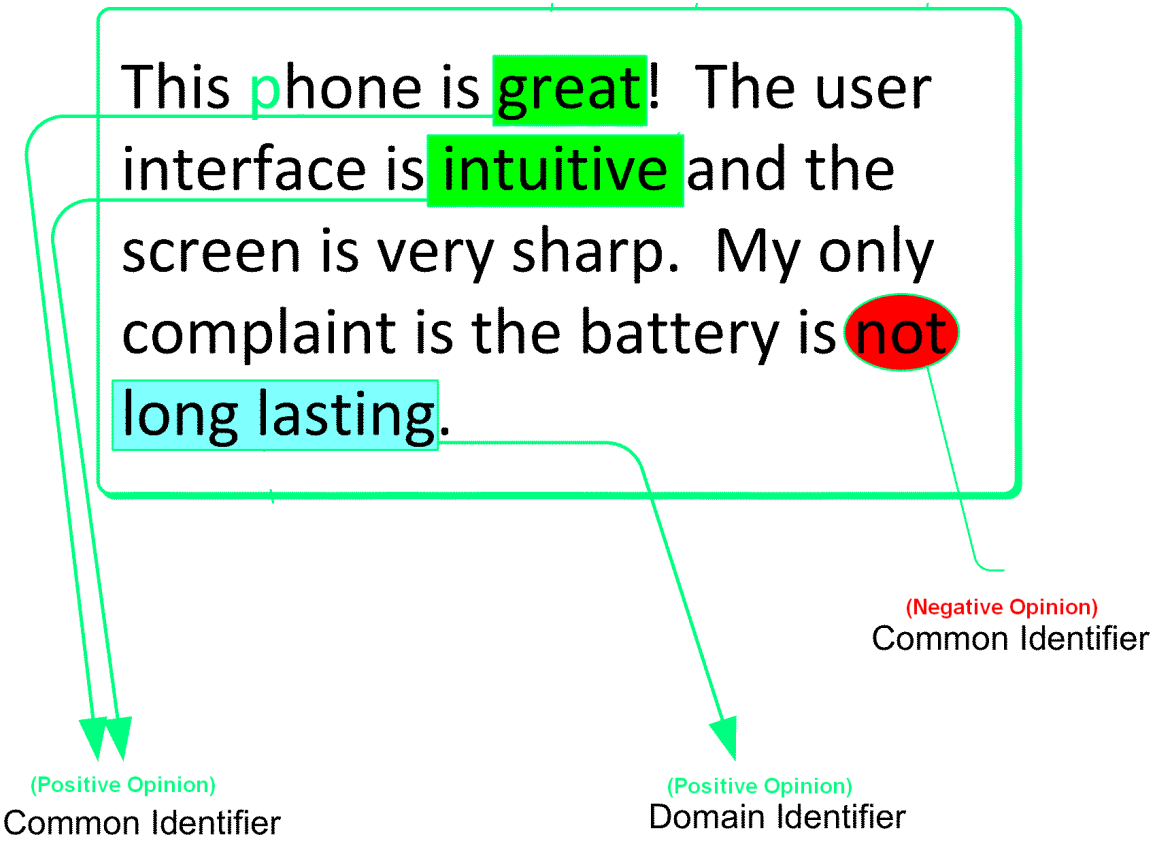
Finne relevant informasjon

Informasjonsekstraksjon

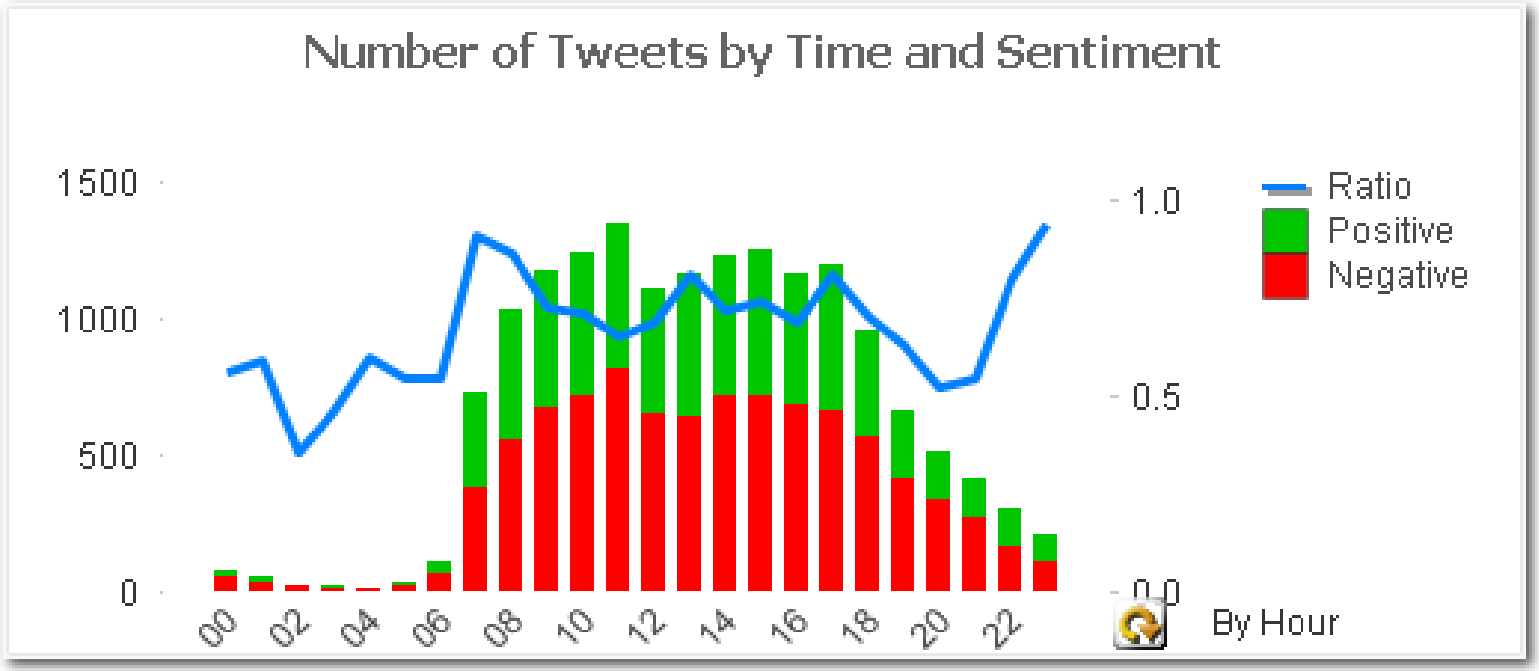


Analysere meninger

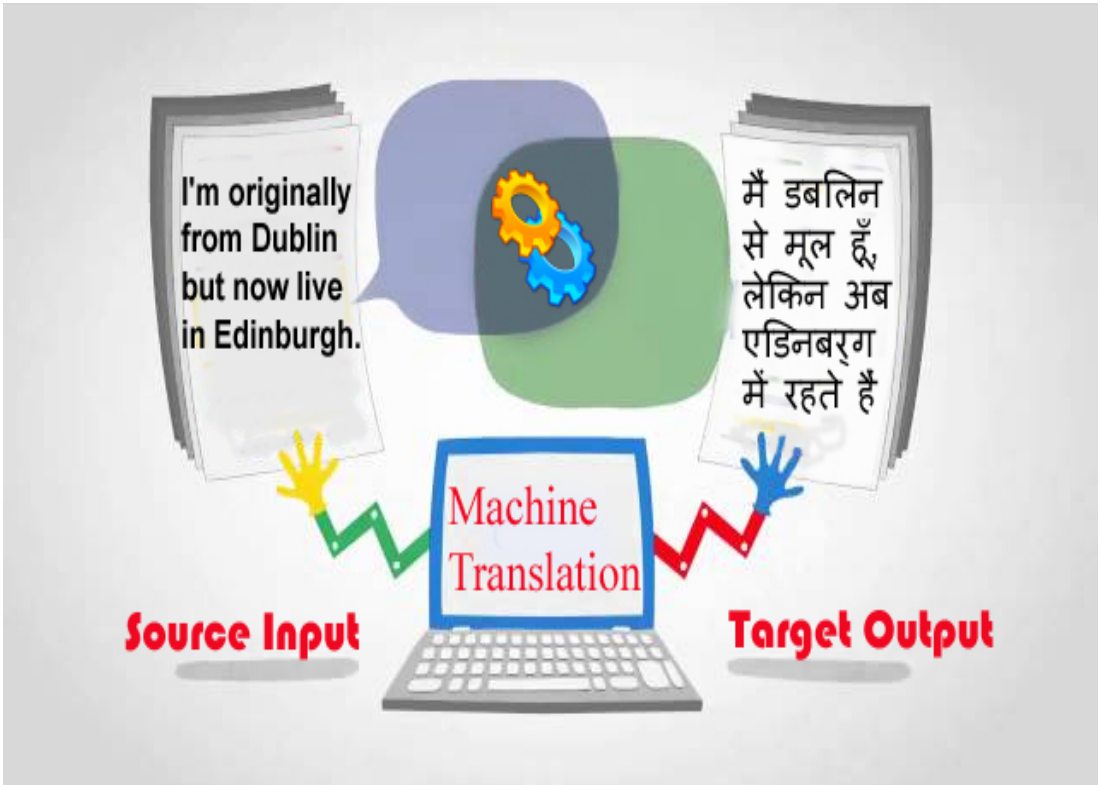
Sentiment Analyse: automatisk analyse av subjektivt språk



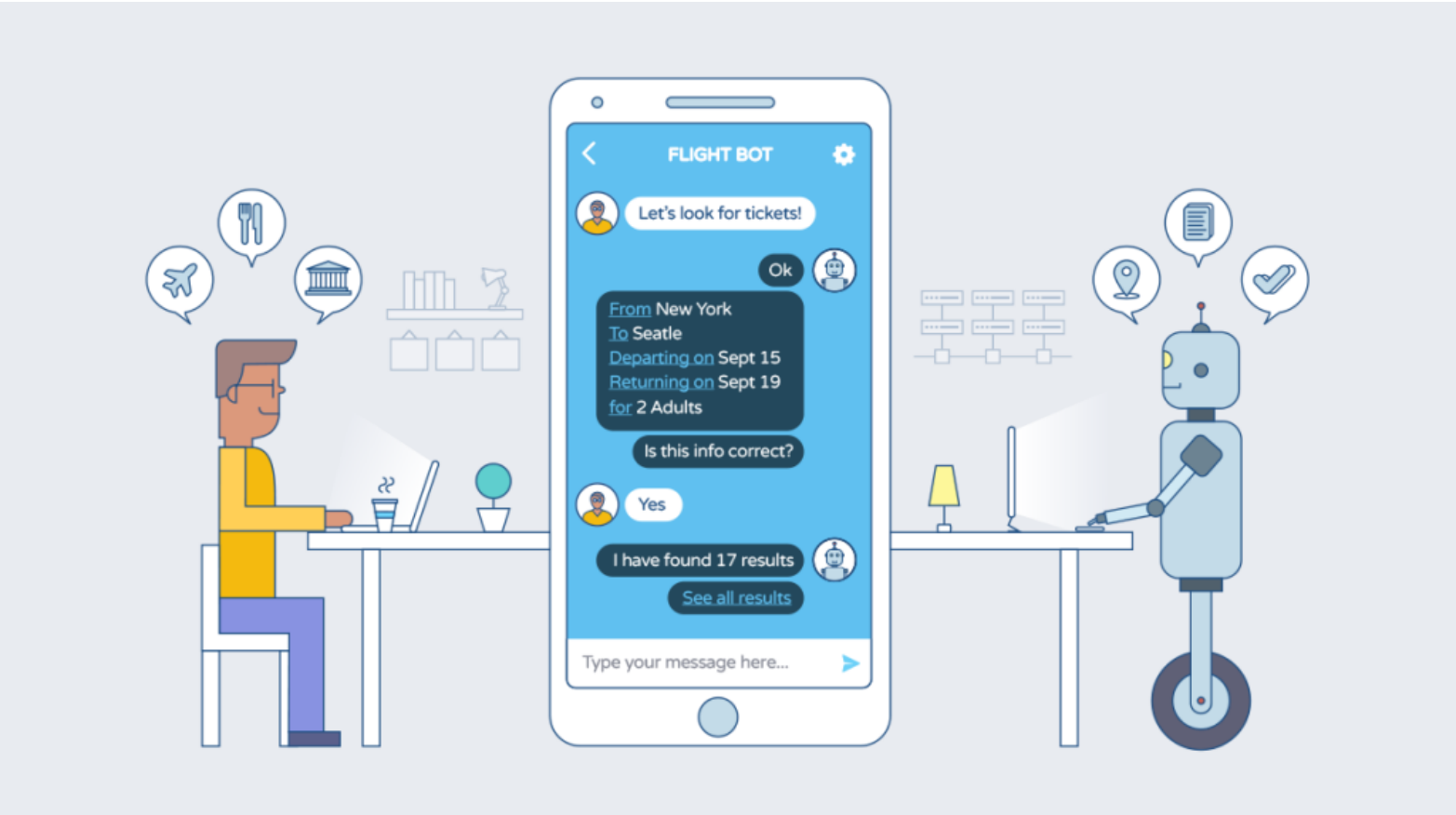
Medieovervåkning



Maskinoversettelse



Dialogsystemer



IN1140 og tverrfaglighet

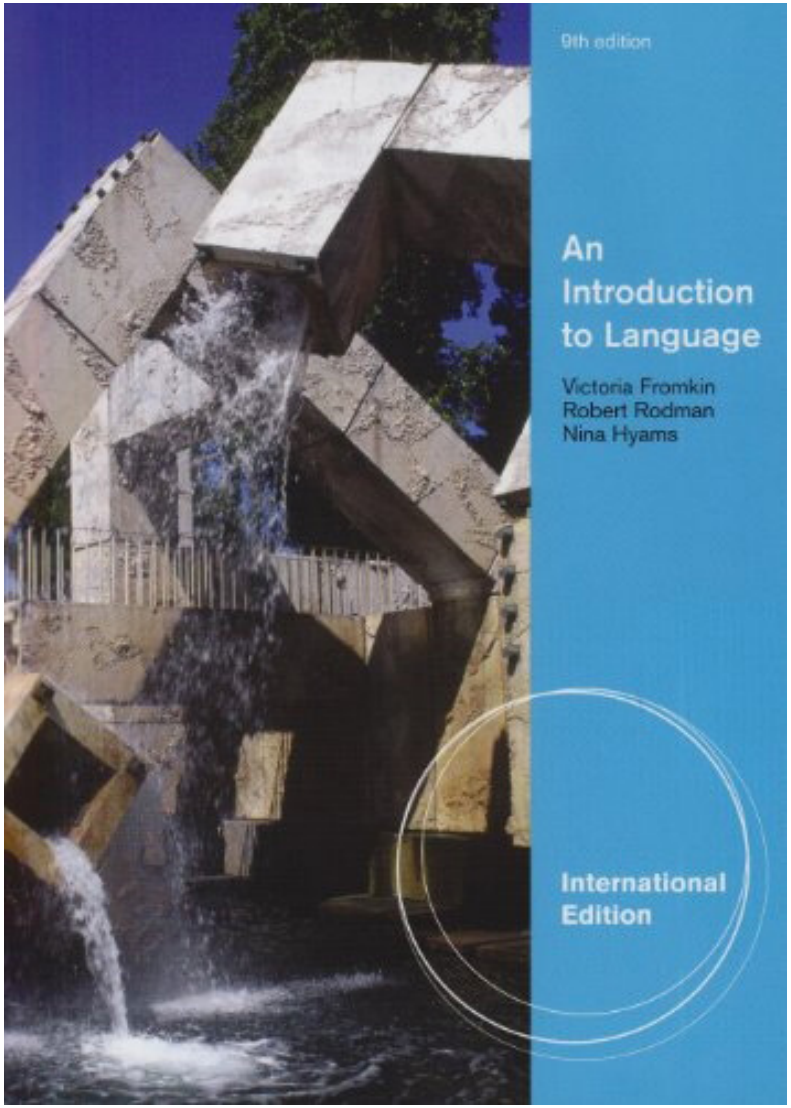
- Stoffet vi dekker i IN1140 tar også for seg stoff fra flere ulike felt.
- Innføring i lingvistikk,
- grunnleggende sannsynlighetsregning,
- programmering, og
- språkteknologiske anvendelser.

IN1140 og tverrfaglighet

- Stoffet vi dekker i IN1140 tar også for seg stoff fra flere ulike felt.
- Innføring i lingvistikk,
- grunnleggende sannsynlighetsregning,
- programmering, og
- språkteknologiske anvendelser.

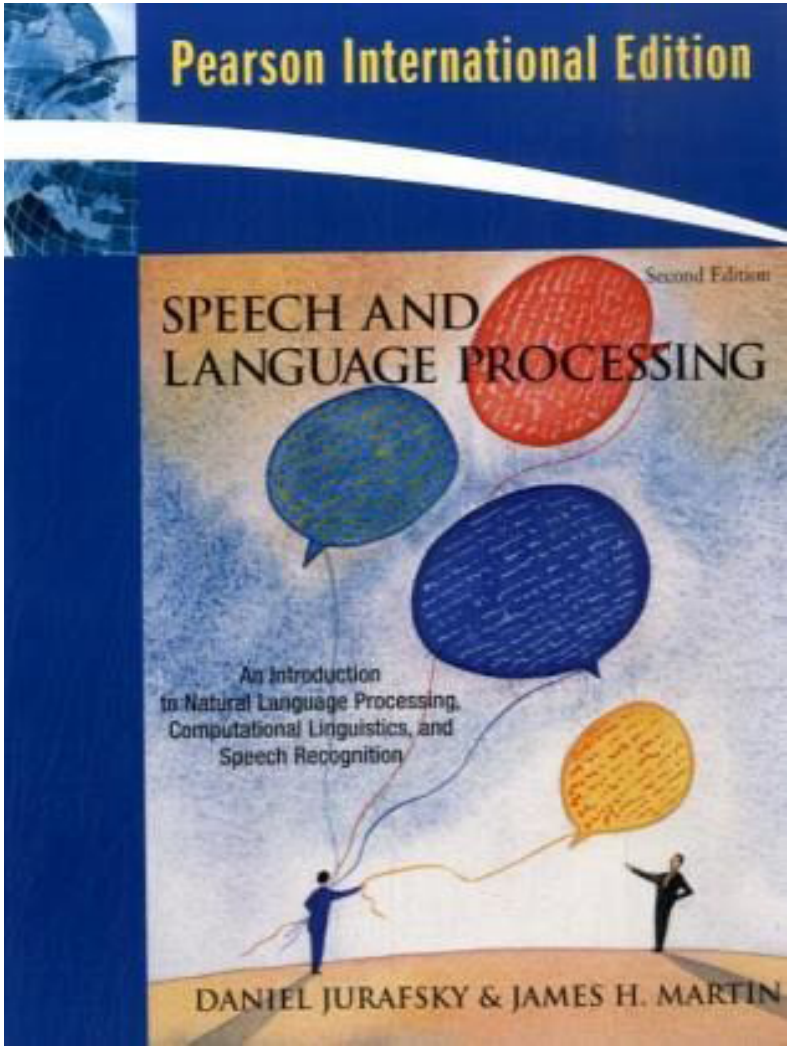
- Gjør deg godt rustet for flere viderekommende emner, f.eks
  - IN2110 – Språkteknologiske metoder
  - IN3050 – Kunstig intelligens og maskinlæring
  - IN3120 – Søketeknologi
  - og mange flere!

Pensumlitteratur



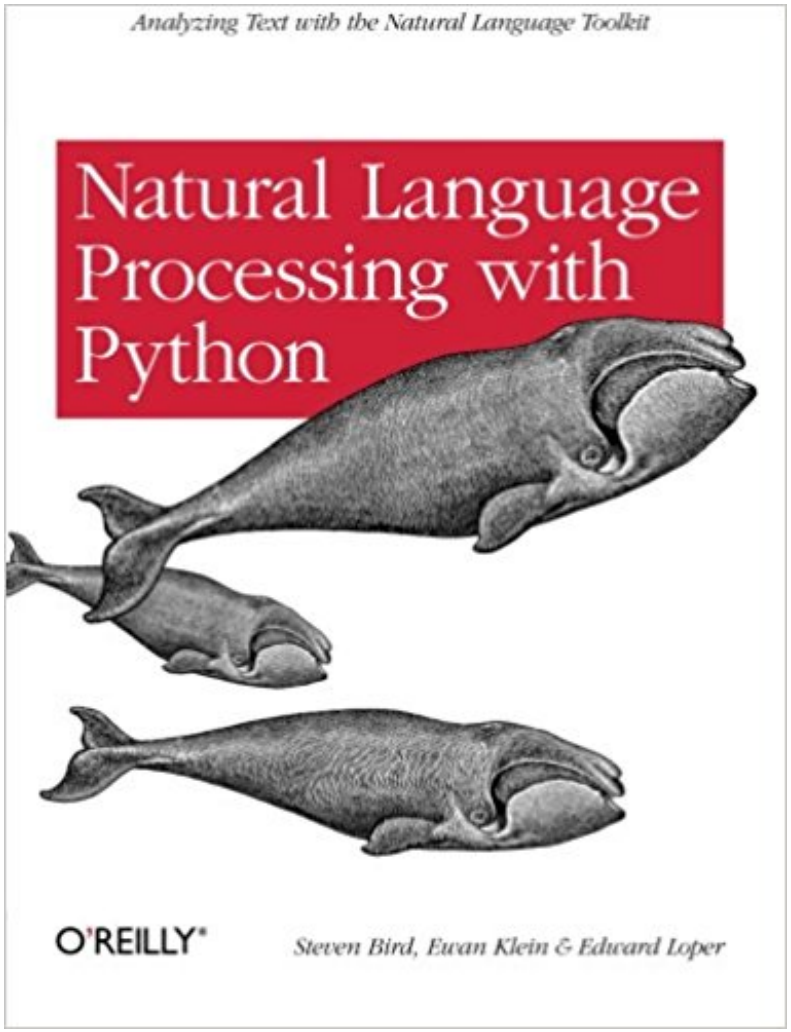
- *An Introduction to Language* av Fromkin, Rodman & Hyams
- Utvalgte deler (ca 5 kapitler)

Pensumlitteratur



- *Speech and Language Processing* av Jurafsky & Martin
- Utvalgte deler
- Gratis nettbok:
  `https://web.stanford.edu/~jurafsky/slp3/`

Pensumlitteratur



- *Natural Language Processing with Python*,
  av Bird, Klein & Loper
- Oppdatert for Python 3 og NLTK 3 (Natural Language Toolkit)
- Utvalgte deler
- Gratis nettbok:
  `http://www.nltk.org/book/`

Python

- Progammering lærer dere først og fremst i IN1000, ikke IN1140.
- Forelesningene i IN1140 kommer til fokusere på teori.
- Samtidig ønsker vi å implementere stoffet i praksis, i Python.
- Implementasjon blir fokus på gruppene og innleveringene.
- Kræsjkurs i Python-programmering på de første gruppetimene.
- Viktig med en del egeninnsats i starten for å henge med.



Litt om årets IN1140-studenter

Gjennomgang av nettskjema

IN1140 Praktiske detaljer



Forelesninger og grupper

Tid & sted

- Forelesninger: ons. 14:15–16:00 i Simula (Ole-Johan Dahls hus / IFI).
- Gruppe 1: tor. 08:15–10:00, Datastue Assembler.
- Gruppe 2: fre. 12:15–14:00, IT-seminarrom Sed.
- Gruppe 3: tors. 14:15–16:00, Datastue Assembler.
- Gruppe 4: fre. 14:15-16:00, Digital gruppe i Zoom (åpen for alle)
- **NB!** Første gruppetime er neste uke

Screencasting

- Tar opp screencast for hver forelesning (lyd + foiler).
- Ment som et supplement, for repetisjon.
- Alternativ ved sykdom.
- Noen få forelesninger vil bli gitt digitalt (via Zoom)



Spørsmål og hjelp

- Gruppetimene: Gruppelærerene er der for å hjelpe og veilede.
- Mattermost (chat-basert diskusjonsforum)
- `in1140-hjelp [at] ifi.uio.no`: Felles adresse til fag-/gruppelærere.



Beskjeder

- Husk å sjekke UiO-eposten din og beskjedlisten på semestersiden.
- `http://www.uio.no/studier/emner/matnat/ifi/IN1140/h21/`



Obligatoriske innleveringer

- 2 obliger.
- Oblig 1 har to deler (a + b).
- Oblig 2 har to deler (a + b).
- Dvs. 4 innleveringer tilsammen: 1a + 1b, 2a + 2b.
- Begge obligene må bestås for å kunne ta eksamen.
- Ingen omlevering.

Obligatoriske innleveringer

- 2 obliger.
- Oblig 1 har to deler (a + b).
- Oblig 2 har to deler (a + b).
- Dvs. 4 innleveringer tilsammen: 1a + 1b, 2a + 2b.
- Begge obligene må bestås for å kunne ta eksamen.
- Ingen omlevering.

Poengsystemet

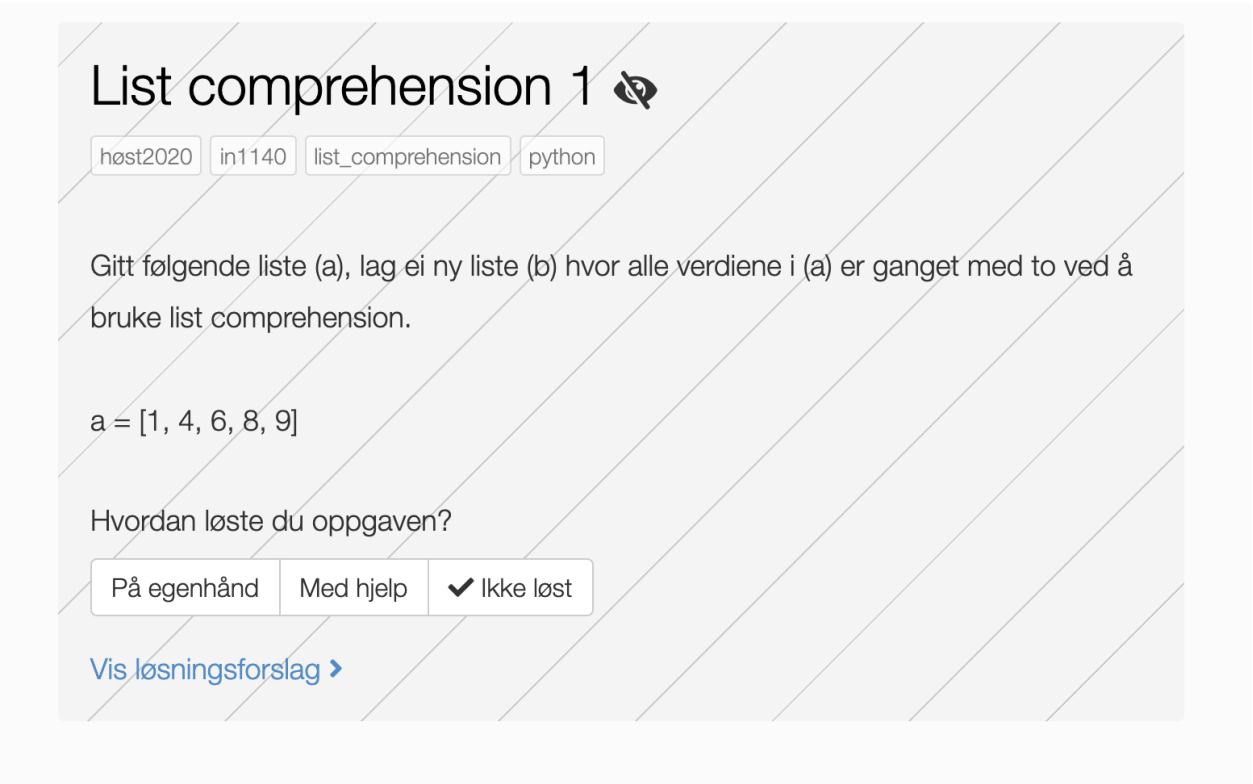
- Man kan oppnå opptil 100 poeng per innlevering
- For å bestå en oblig kreves minst 100 poeng (av 200 mulige).
- Eksempel:
  - 37 poeng på 1a
  - 68 poeng på 1b
  - = 105 poeng på oblig 1 (= bestått).

Mer om obligene

- Absolutte frister:
- Utsettes *kun* ved egenmelding (opptil 3 dager) eller legeerklæring.
- Kopiering/plagiat godtas ikke. Sett deg inn i reglene.
- Husk at hvis du distribuerer løsningsforslaget ditt på nett (f.eks via Github), kan du bidra til juks. Styr unna.
- Benytt deg av gruppeundervisningen, og planlegg tiden din.
- Tidsregnskap:
  - Arbeidsinnsats (minimum): 37,5 / 3 = 12,5 timer
  - Etter forelesning+gruppe: 9,5 timer
- Konkurranse: den/de som får flest poeng tilsammen på obligene gjennom semesteret får en premie (overraskelse)!

Ekstraoppgaver

- Oppgaver til gruppetimene fokuserer ofte på teori
- En del ekstraoppgaver som Trix-oppgaver



Eksamen

- Skriftlig (digital) hjemmeeksamen på fire timer
  - 29 november kl. 09:00
- Pensumlitteratur + forelesningsnotater
- NB! Ikke en programmeringseksamen.
- Fokus på teoretiske konsepter.

Suksessoppskrift

- Emnesiden: timeplan, pensum, lesehenvisninger, beskjeder etc.
- Lesehenvisninger: forbered deg til forelesning
- Still spørsmål
- Gruppetimer:
  - forbered deg
  - delta aktivt
  - gjør oppgaver (også de ikke-obligatoriske!)
- Benytt deg av medstudentene dine

Neste uke

- Hva lærer jeg i IN1140?
- Hva er lingvistikk?
- Språkteknologiske komponenter
- Metoder