

IN1140: Introduksjon til språkteknologi

Forelesning #1

Lilja Øvrelid

Universitetet i Oslo

22. august 2019



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



- ▶ Tar opp **screencast** for hver forelesning (lyd + foiler).
- ▶ Egen **YouTube-kanal**:
<https://www.youtube.com/channel/UCE1IhV-Q-PuAkg2Fb350MIQ>
- ▶ Ment som et supplement, for repetisjon.



Forelesere

- ▶ **Samia Touileb** (samiat@ifi.uio.no)
- ▶ **Lilja Øvrelid** (liljao)
- ▶ Fra språkteknologigruppa (LTG)

Forelesere

- ▶ **Samia Touileb** (samiat@ifi.uio.no)
- ▶ **Lilja Øvrelid** (liljao)
- ▶ Fra språkteknologigruppa (LTG)

Gruppelærere

- ▶ **Tania-Adelina Bulz** (taniaadb)
- ▶ **Annika Willoch Olstad** (annikaol)

Forelesere

- ▶ **Samia Touileb** (samiat@ifi.uio.no)
- ▶ **Lilja Øvrelid** (liljao)
- ▶ Fra språkteknologigruppa (LTG)

Gruppelærere

- ▶ **Tania-Adelina Bulz** (taniaadb)
- ▶ **Annika Willoch Olstad** (annikaol)

Tid & sted

- ▶ Gruppe 1: man. 10:15–12:00, Datastue Limbo.
- ▶ Gruppe 2: ons. 08:15–10:00, Datastue Limbo.
- ▶ Forelesninger: tors. 12:15–14:00 i **Cam1** (Ole-Johan Dahls hus / IFI).
- ▶ **NB!** Første gruppetime er mandag **2 september**

- ▶ **Gruppetimene:** Gruppelærerene er der for å hjelpe og veilede.
- ▶ **Piazza** (diskusjonsforum):
<https://piazza.com/uio.no/fall2019/in1140/>
NB! litt ventetid på svar
- ▶ **in1140-hjelp [at] ifi.uio.no:** Felles adresse til fag-/gruppelærere.



- ▶ Husk å sjekke **UiO-eposten** din og **beskjedlisten** på semestersiden.
- ▶ <http://www.uio.no/studier/emner/matnat/ifi/IN1140/h19/>



Hva er språkteknologi?



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





Eksempler på språkteknologi?



amazon.com

Recommended for You

Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own.



[The Little Big Things: 163 Ways to Pursue EXCELLENCE](#)



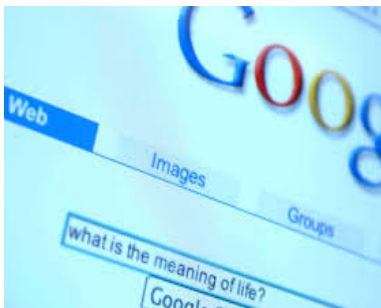
[Fascinate: Your 7 Triggers to Persuasion and Captivation](#)



[Sherlock Holmes \[Blu-ray\]](#)



[Alice in Wonderland \[Blu-ray\]](#)



Google Translate interface showing the translation of 'kammer' (German) to 'chamber' (English). The interface includes language selection options (German, English, Spanish, Detect language) and a list of related terms for 'chamber' such as 'room', 'cell', 'professional association', and 'boxroom'.

Screenshot of a social media feed. A circular callout highlights a 'Trending' section with the following items:

- Golden Globes:** The 27 Best Moments from the Golden Globe Awards
- Cristiano Ronaldo:** Cristiano Ronaldo wins Fifa Ballon d'Or after stellar year at Real Madrid
- 24:** Fox Sets May 5 Premiere for '24: Live Another Day'

Below the trending items is a 'People You May Know' section featuring a profile for Sanjeet Hajarnis, with 21 mutual friends and an 'Add Friend' button.

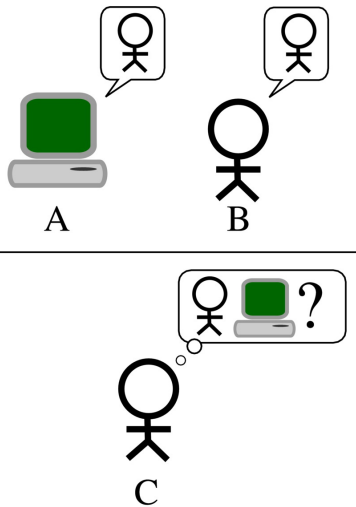
NLP er et tverrfaglig felt

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



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

- ▶ Alan Turing i 1950:
- ▶ *I propose to consider the question, 'Can machines think?'*
- ▶ Definisjonsspørsmål. Skulle avgjøres ved Turingtesten.



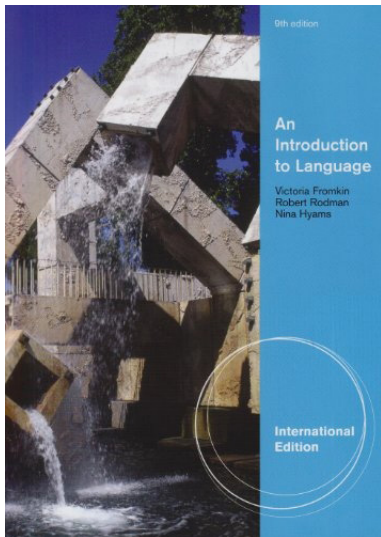


- ▶ Stoffet vi dekker i IN1140 tar også for seg stoff fra flere ulike felt.
- ▶ Innføring i **lingvistikk**,
- ▶ grunnleggende **sannsynlighetsregning**,
- ▶ **programmering**, og
- ▶ språkteknologiske **anvendelser**.

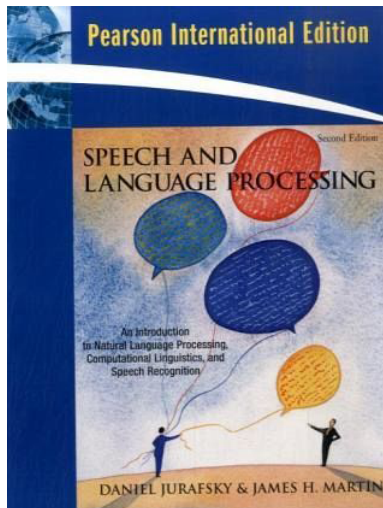


- ▶ Stoffet vi dekker i IN1140 tar også for seg stoff fra flere ulike felt.
- ▶ Innføring i **lingvistikk**,
- ▶ grunnleggende **sannsynlighetsregning**,
- ▶ **programmering**, og
- ▶ språkteknologiske **anvendelser**.

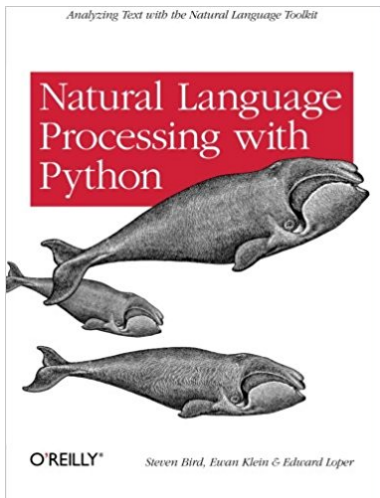
- ▶ Gjør deg godt rustet for flere viderekommende emner, f.eks
 - ▶ IN2110 – Språkteknologiske metoder
 - ▶ IN3050 – Kunstig intelligens og maskinlæring
 - ▶ IN3120 – Søketeknologi
 - ▶ og mange flere!



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



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

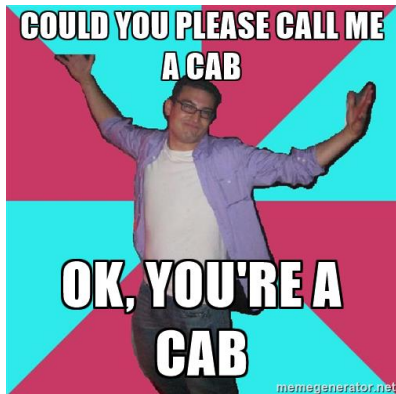


- ▶ *Natural Language Processing with Python*, av Bird, Klein & Loper
- ▶ Oppdatert for Python 3 og NLTK 3 (Natural Language Toolkit)
- ▶ Utvalgte deler
- ▶ Gratis nettbok:
<http://www.nltk.org/book/>

- ▶ Programmering 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.



- ▶ 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

We gave the monkeys₁ the bananas₂
... because they₁ were hungry.
... because they₂ were ripe.






Jeg spiser sushi med pinner .



Jeg spiser sushi med pinner .





Jeg spiser sushi med pinner .

The diagram illustrates the ambiguity of the sentence "Jeg spiser sushi med pinner ." by using two curved arrows. A green arrow starts above the word "spiser" and points to "pinner", indicating the interpretation "I eat sushi with sticks." A red arrow starts below the word "sushi" and points to "pinner", indicating the interpretation "I eat sushi sticks with sticks."



Jeg spiser sushi med pinner .

Jeg spiser sushi med laks .



Jeg spiser sushi med pinner .

A diagram illustrating ambiguity in the sentence "Jeg spiser sushi med pinner .". A green arrow starts at the word "spiser" and points to "pinner", indicating the subject-verb relationship. A red arrow starts at "sushi" and points to "pinner", indicating the object-prepositional phrase relationship. This visualizes how the prepositional phrase "med pinner" can be interpreted as either modifying the verb "spiser" or the noun "sushi".

Jeg spiser sushi med laks .

A diagram illustrating ambiguity in the sentence "Jeg spiser sushi med laks .". A green arrow starts at the word "spiser" and points to "laks", indicating the subject-verb relationship. This visualizes how the prepositional phrase "med laks" can be interpreted as modifying the verb "spiser".



Jeg spiser sushi med pinner .

The diagram illustrates the ambiguity of the sentence "Jeg spiser sushi med pinner ." by showing two possible syntactic structures. A green arrow connects the verb "spiser" to the noun "pinner", and a red arrow connects the noun "sushi" to the noun "pinner".

Jeg spiser sushi med laks .

The diagram illustrates the ambiguity of the sentence "Jeg spiser sushi med laks ." by showing two possible syntactic structures. A green arrow connects the verb "spiser" to the noun "laks", and a red arrow connects the noun "sushi" to the noun "laks".



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*



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.

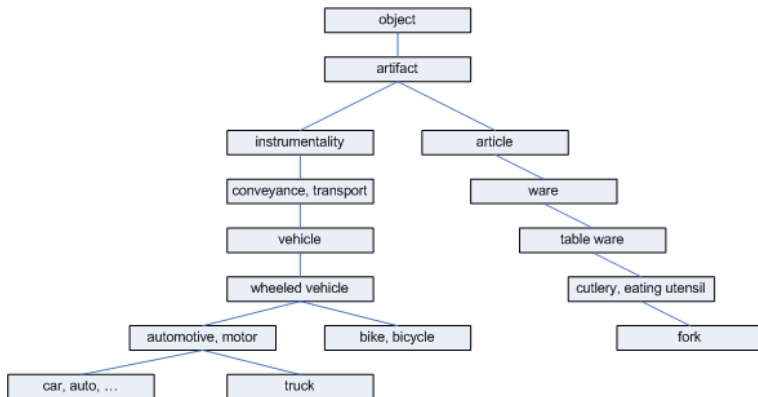
Steven Pinker, *The language instinct*

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

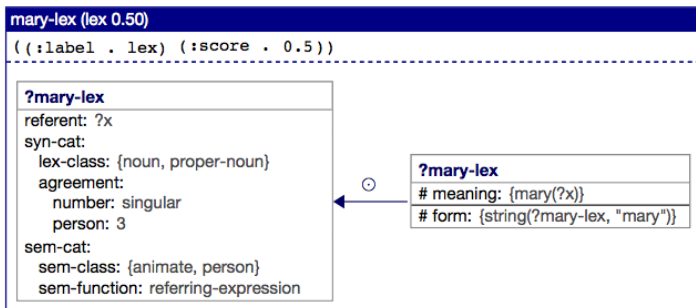


- ▶ Vi **mennesker** tolker språklige uttrykk basert på delt **bakgrunnskunnskap** og gjensidige **forventninger** i en gitt **kontekst**.
- ▶ **Språkforståelse** handler mye om **entydiggjøring**.
- ▶ Språkteknologi, og **IN1140**, handler i stor grad om strategier for hvordan maskiner kan takle dette.

→ 2000-tallet: manuelt utformede regler og leksikon



→ 2000-tallet: manuelt utformede regler og leksikon



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



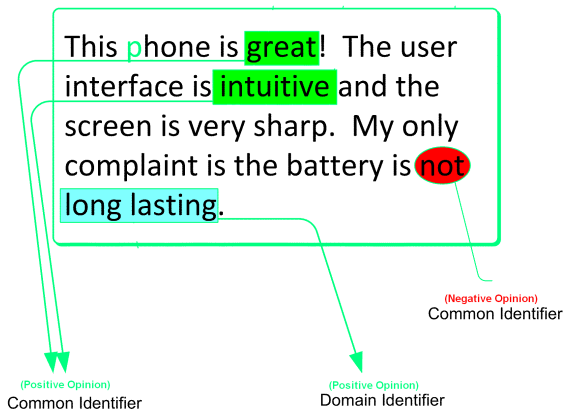




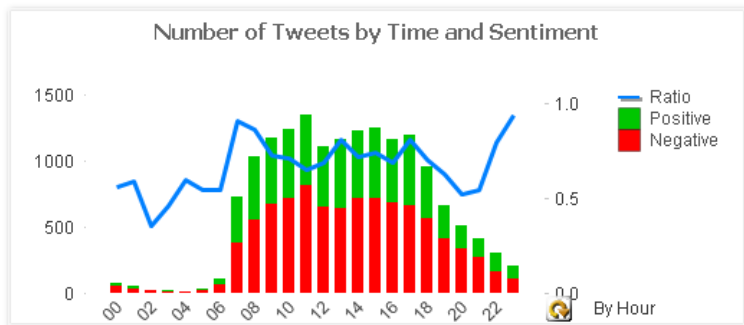
Fed Chairman
Ben Bernanke
said the U.S.
economy...
The euro rose to
\$1.2008.
compared to
\$1.1942
on Tuesday.



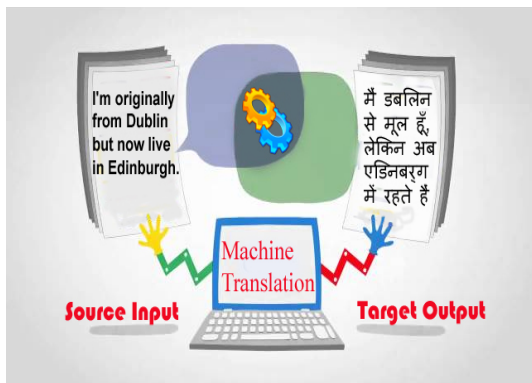
Automatisk analyse av subjektivt språk



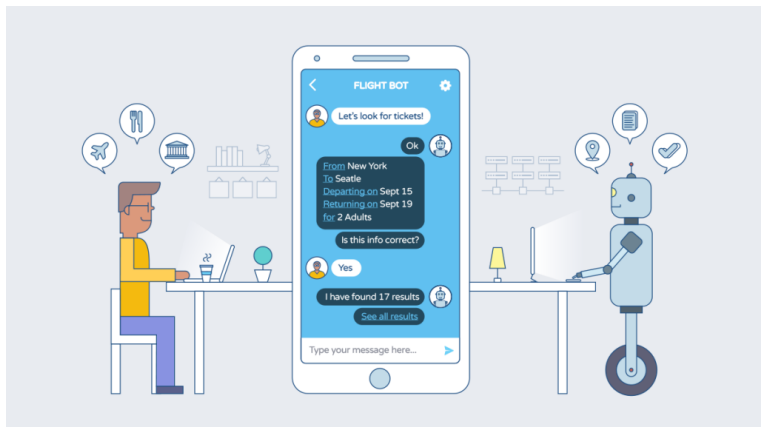
Medieovervåking



Maskinoversettelse



Dialogsystemer





- ▶ 3 obliger.
- ▶ Oblig 1 har to deler ($a + b$).
- ▶ Dvs. 4 innleveringer tilsammen: 1a + 1b, 2, 3.
- ▶ Alle obligene må bestås for å kunne ta eksamen.
- ▶ Ingen omlevering.

Poengsystemet

- ▶ Man kan oppnå opptil 100 poeng per innlevering
- ▶ For å bestå kreves minst 100 poeng (av 200 mulige) for oblig 1(a+b), og 50 poeng (av 100 mulige) for oblig 2 og 3.
- ▶ Eksempel:
 - ▶ 37 poeng på 1a
 - ▶ 68 poeng på 1b
 - ▶ = 105 poeng på oblig 2 (= bestått).



- ▶ **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
- ▶ **Konkurransen:** den/de som får flest poeng tilsammen på obligene gjennom semesteret får en premie (overraskelse)!



- ▶ Skriftlig (digital) eksamen på fire timer
 - ▶ 27 november kl. 14:30
- ▶ Pensumlitteratur + forelesningsnotater
- ▶ **NB! Ikke en programmeringseksamen.**
- ▶ Fokus på teoretiske konsepter.

Suksessoppskrift

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

Suksessoppskrift

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

