IN1140: Introduksjon til språkteknologi

Forelesning #1

Lilja Øvrelid

Universitetet i Oslo

19 august 2024



Tema for i dag

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



Hei

Forelesere

- Lilja Øvrelid
- Sondre Wold
- Egil Rønningstad
- Fra språkteknologigruppen (LTG), ved seksjon for maskinlæring (MLS)

Hei

Forelesere

- Lilja Øvrelid
- Sondre Wold
- Egil Rønningstad
- Fra språkteknologigruppen (LTG), ved seksjon for maskinlæring (MLS)

Gruppelærere

- Ada-Marie Sneve (Gr. 1)
- Håkon Hyrve (Gr. 2)
- Silje Helgesen (Gr. 3)
- Wona Lee (retter)

Hei

Forelesere

- Lilja Øvrelid
- Sondre Wold
- Egil Rønningstad
- Fra språkteknologigruppen (LTG), ved seksjon for maskinlæring (MLS)

Gruppelærere

- Ada-Marie Sneve (Gr. 1)
- Håkon Hyrve (Gr. 2)
- Silje Helgesen (Gr. 3)
- Wona Lee (retter)

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)



Språkteknologi og tverrfaglighet

NLP er et tverrfaglig felt

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



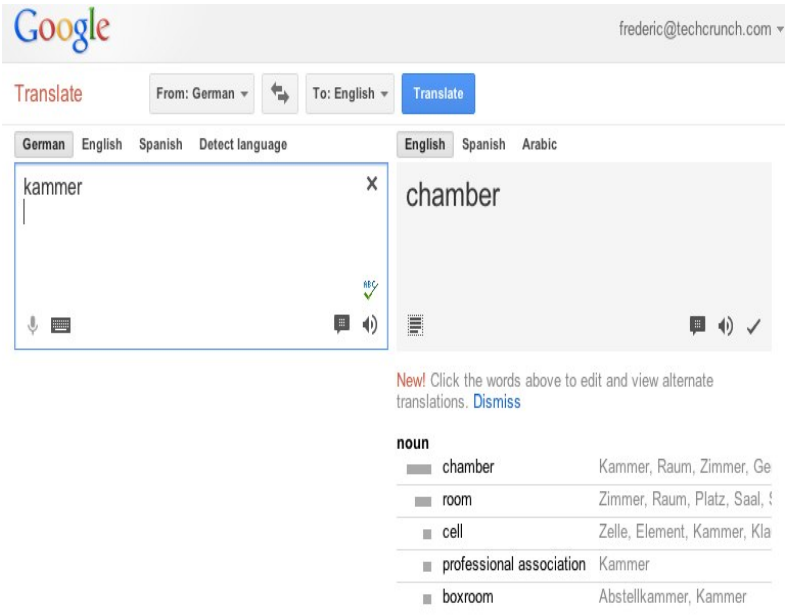
- Del av det bredere feltet kunstig intelligens (AI).

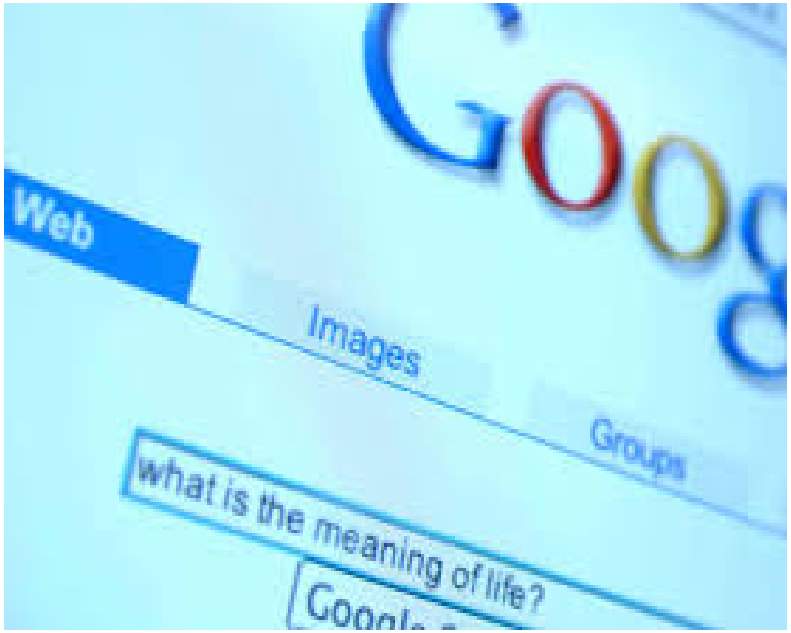
Eksempler på språkteknologi?

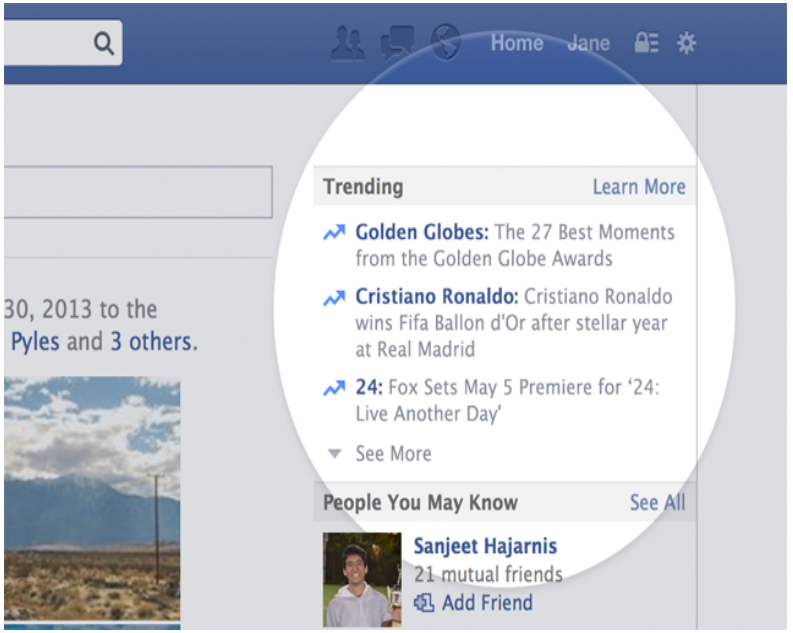


Eksempler på språkteknologi?

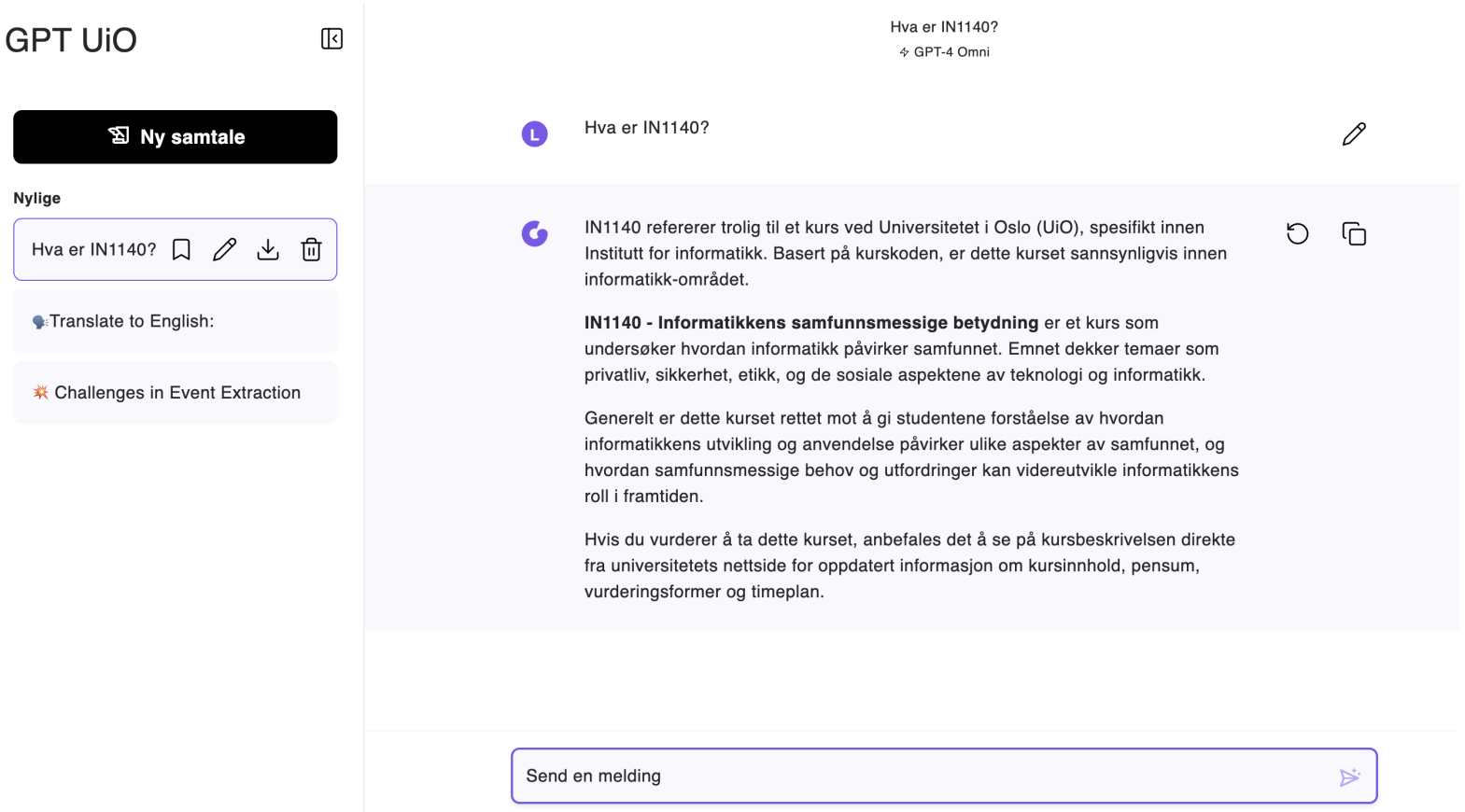






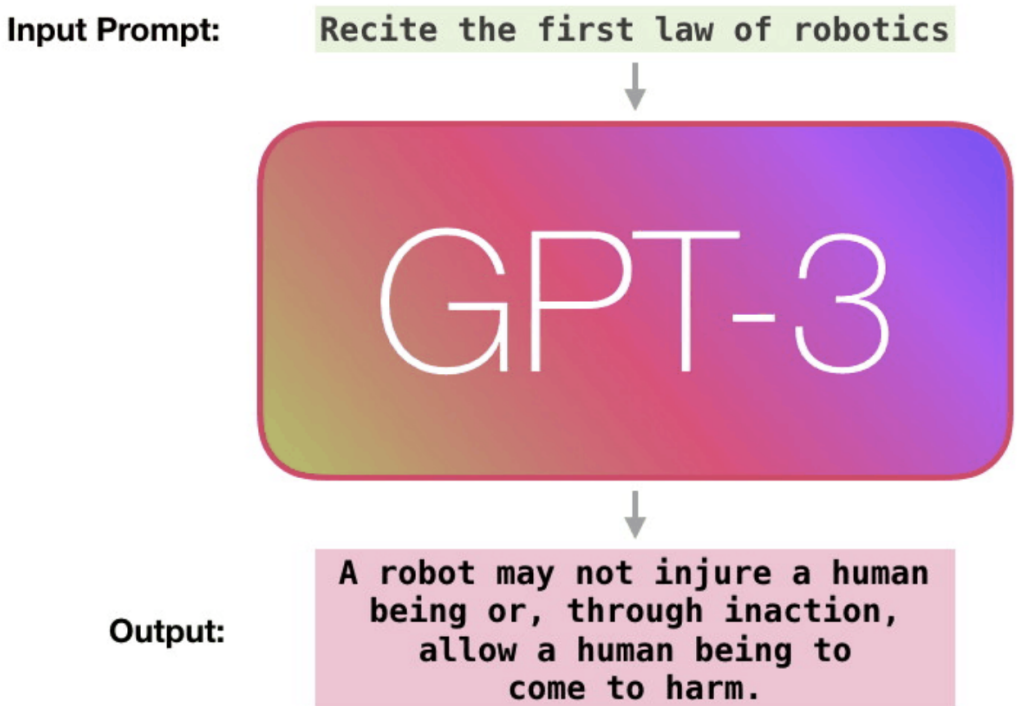


Tekstgenerering



(https://www.uio.no/tjenester/it/ki/gpt-uio/)

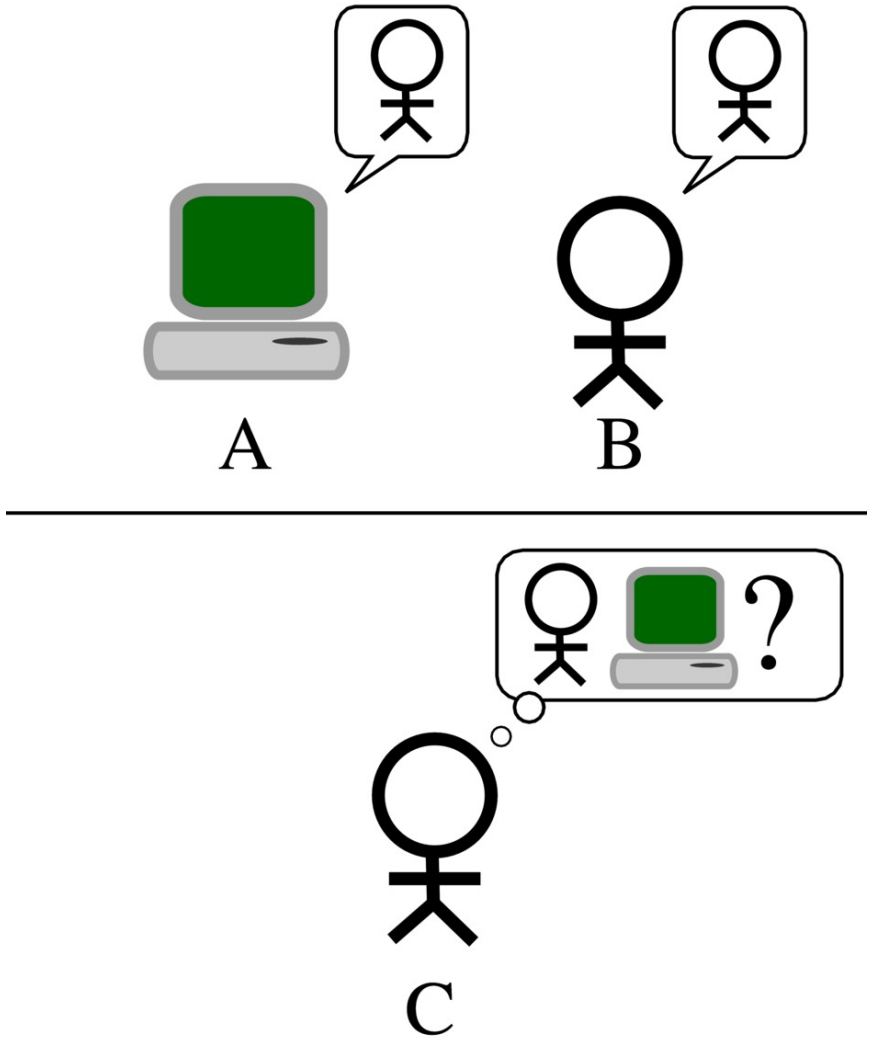
Tekstgenerering



(https://jalammar.github.io/how-gpt3-works)

Turingtesten

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



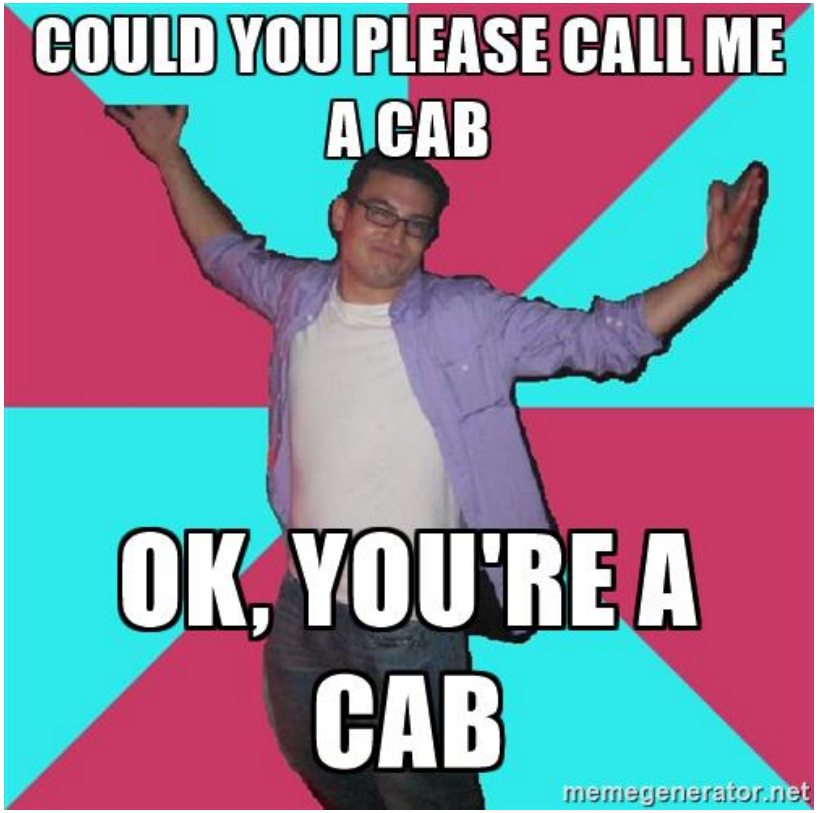
Språkforståelse er vanskelig!

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.

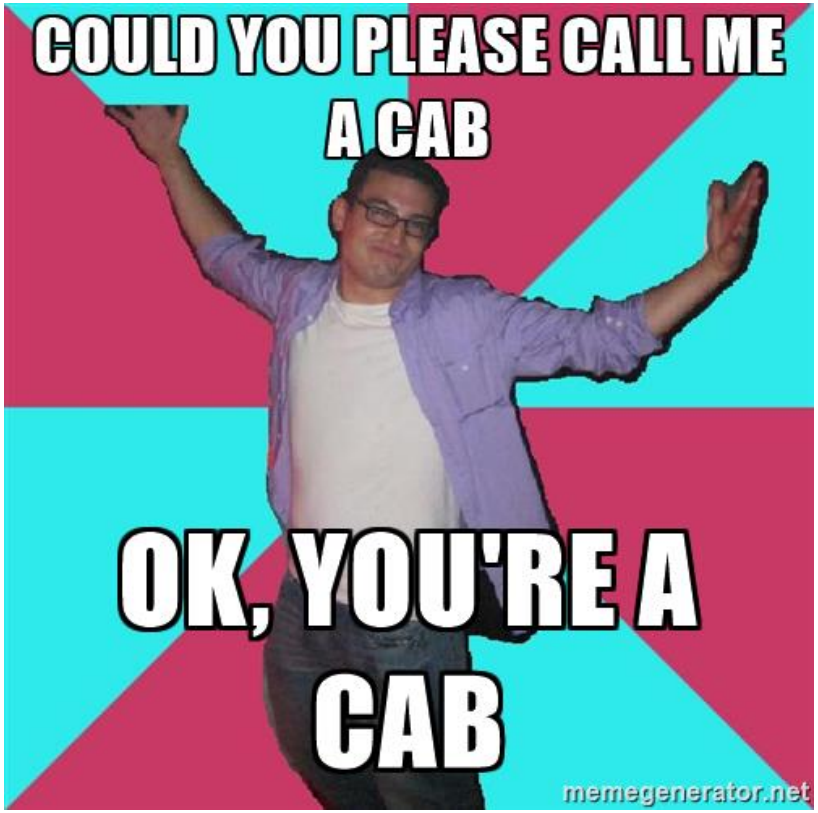
Steven Pinker, *The language instinct*, 1994

Hvorfor er språkforståelse utfordrende?

Hvorfor er språkforståelse utfordrende?

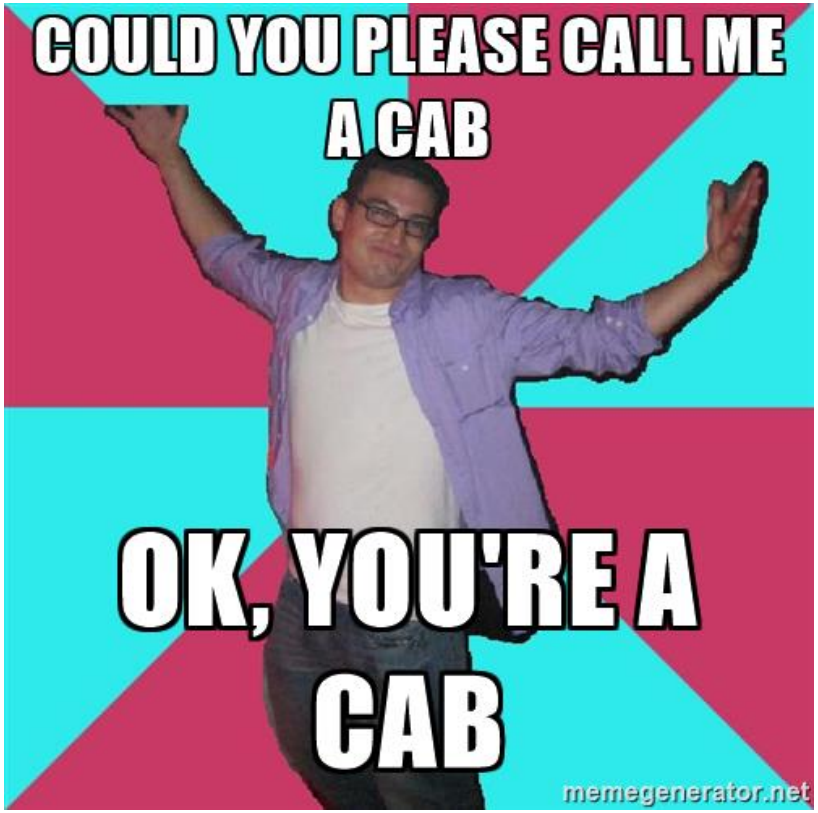


Hvorfor er språkforståelse utfordrende?



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

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: *?*

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

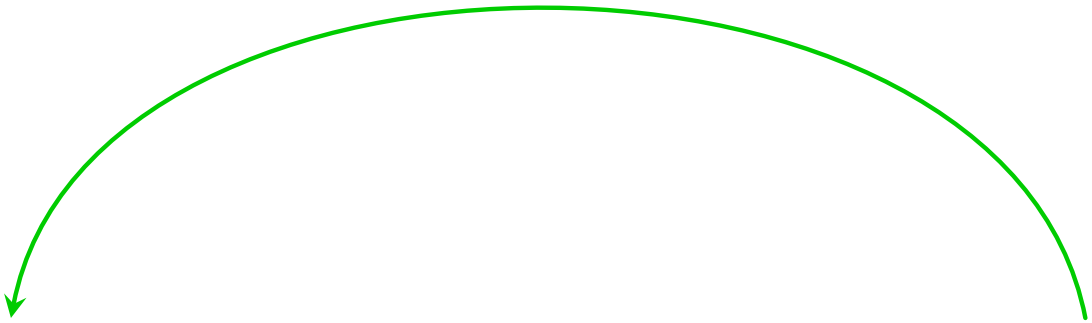
- Norsk: *rett*.
- Engelsk: *?*
- Flertydig ift betydning + ordklasse (verb, subst., adj., adv.).
- Vi trenger kontekst for å avgjøre.

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

Flertydighet i referanse

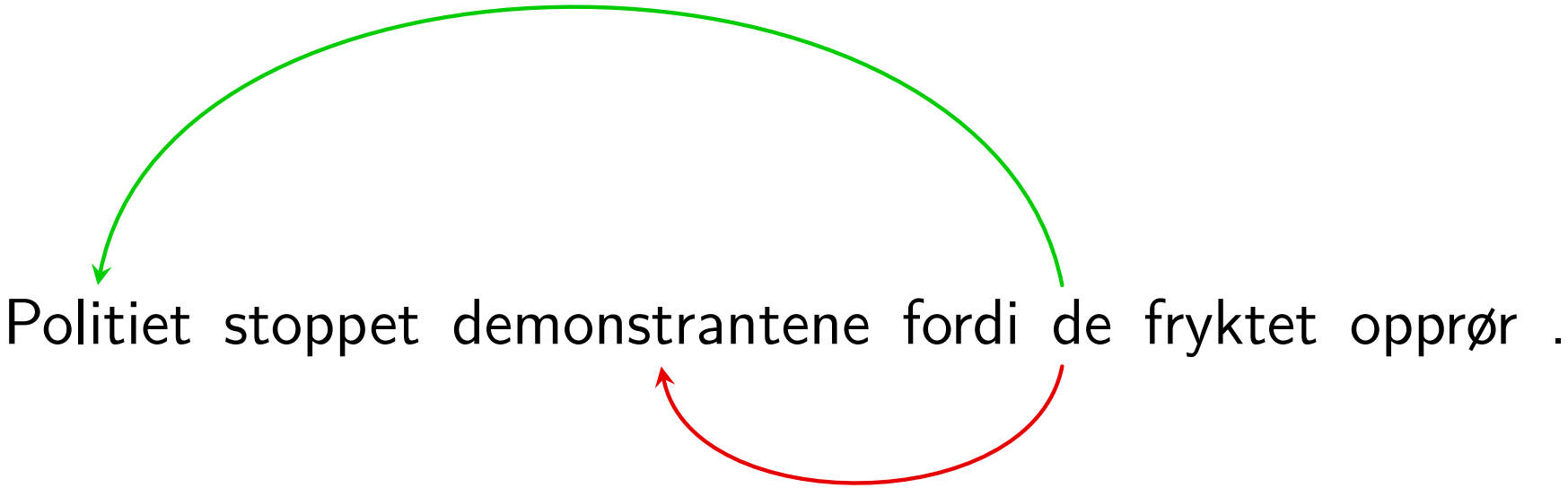
Politiet stoppet demonstrantene fordi de fryktet opprør .

Flertydighet i referanse

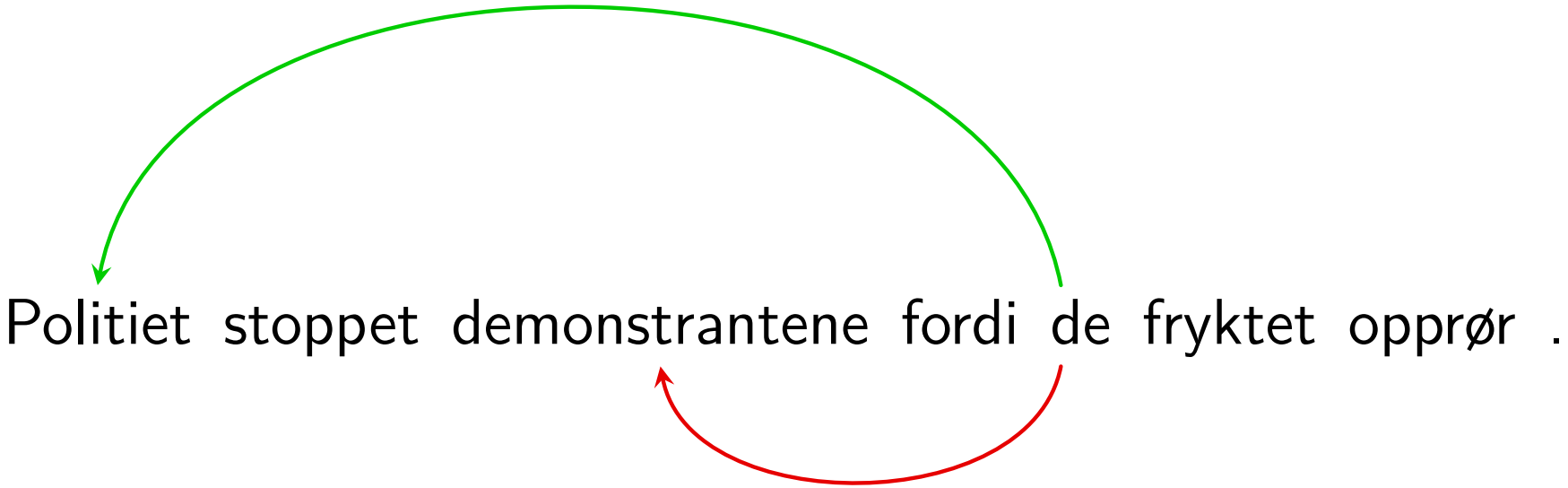


Politiet stoppet demonstrantene fordi de fryktet opprør .

Flertydighet i referanse

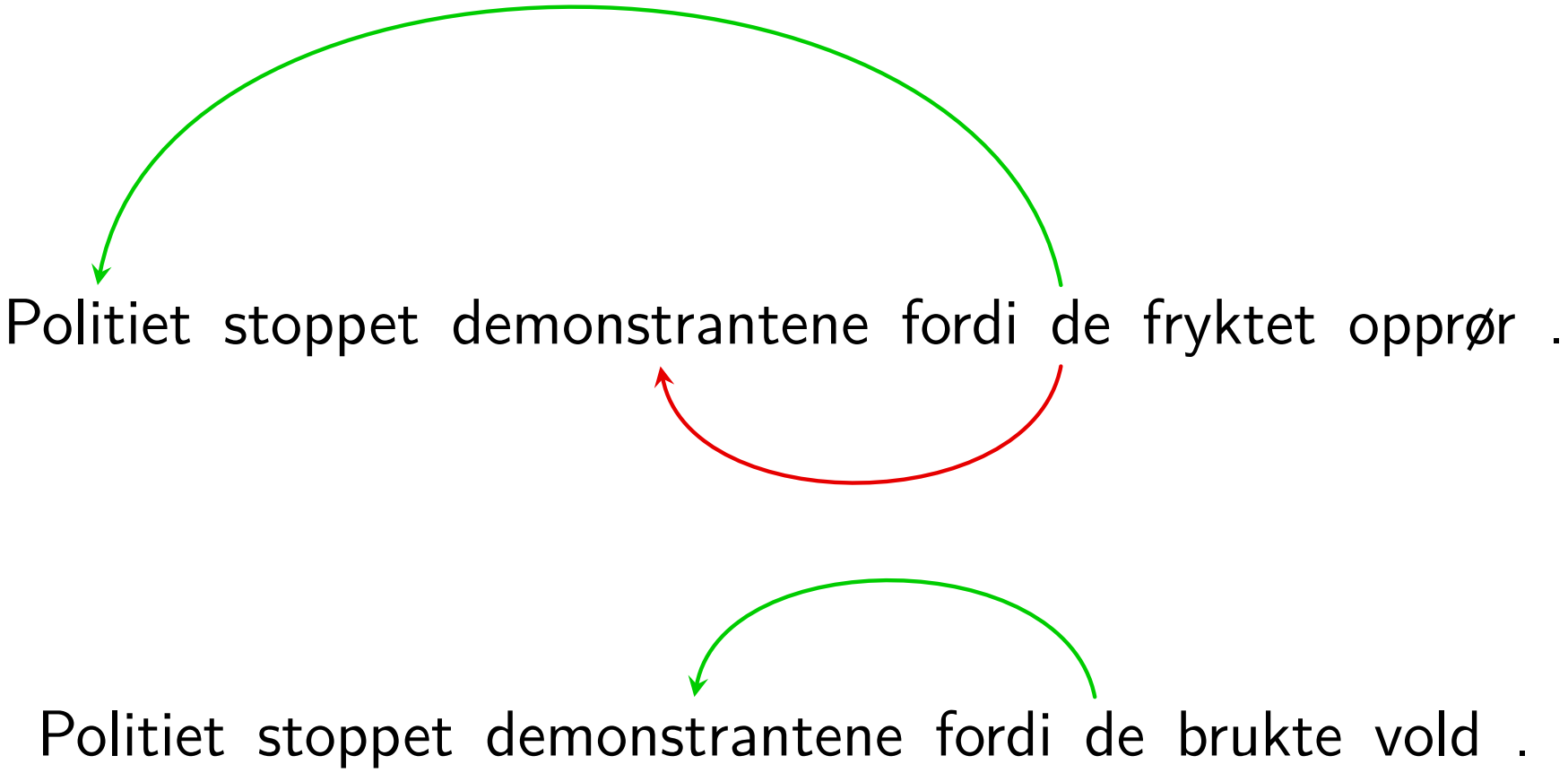


Flertydighet i referanse

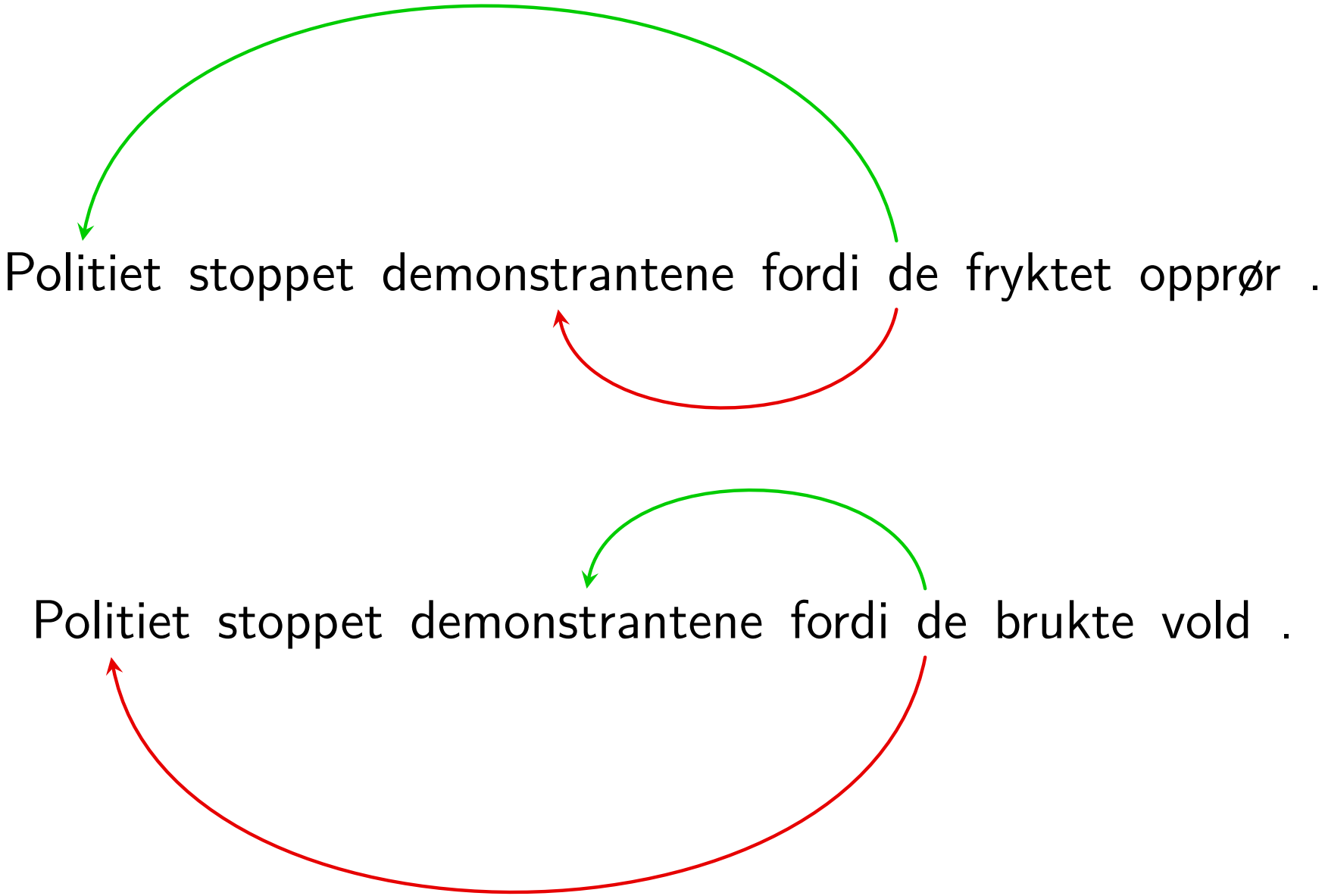


Politiet stoppet demonstrantene fordi de brukte vold .

Flertydighet i referanse



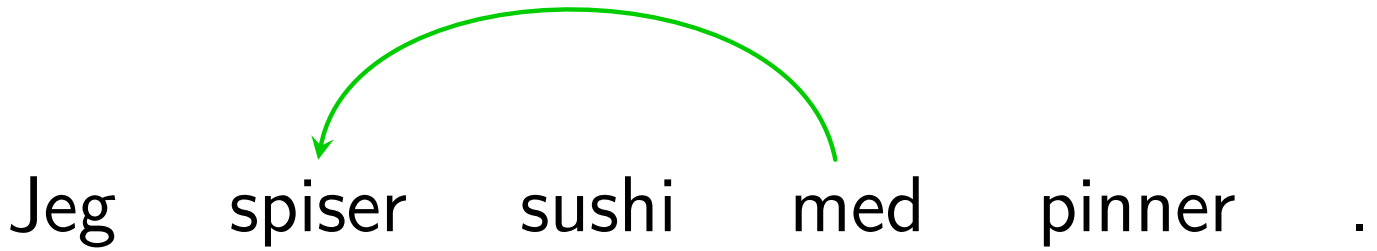
Flertydighet i referanse



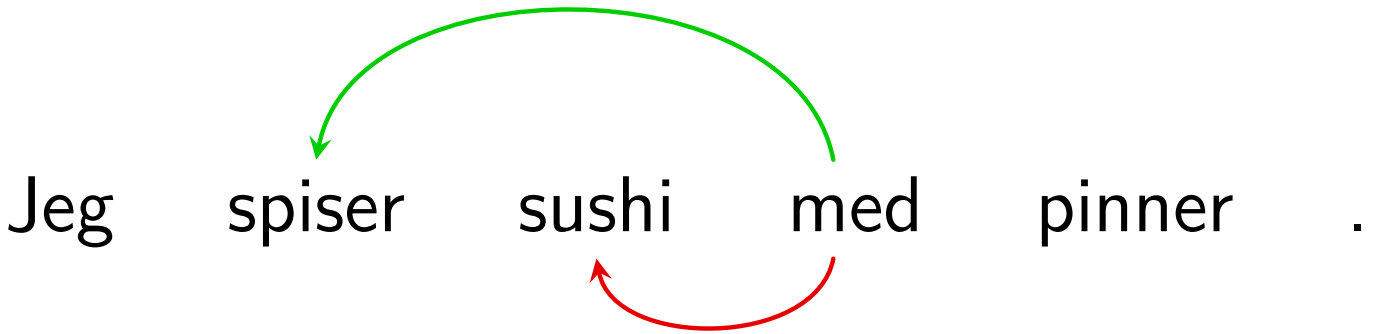
Flertydighet på setningsnivå

Jeg spiser sushi med pinner .

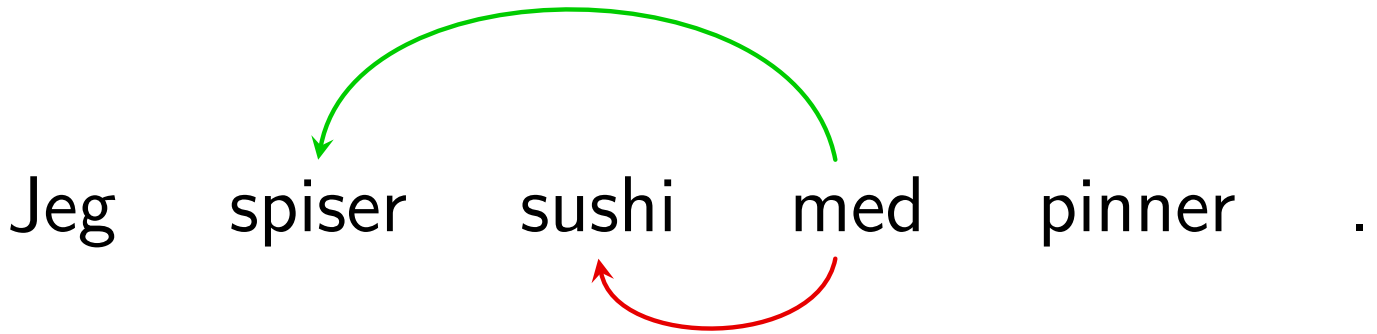
Flertydighet på setningsnivå



Flertydighet på setningsnivå

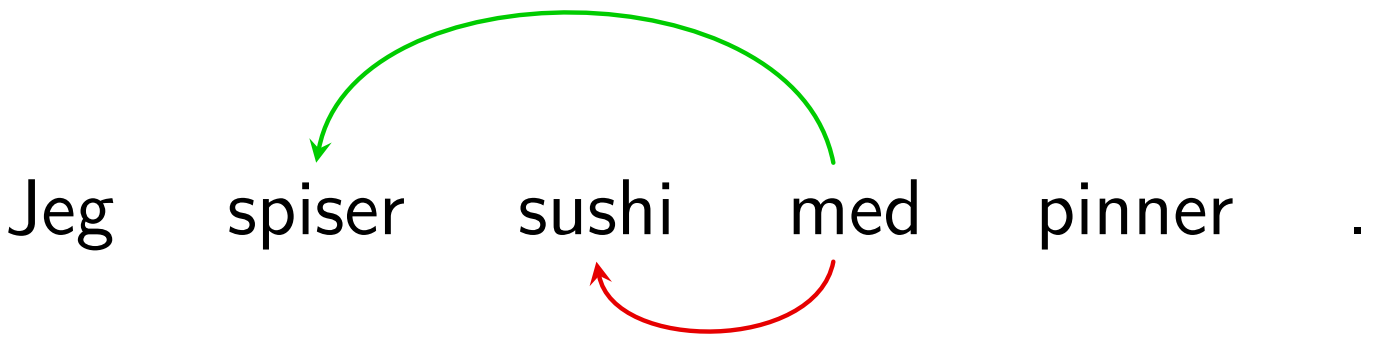


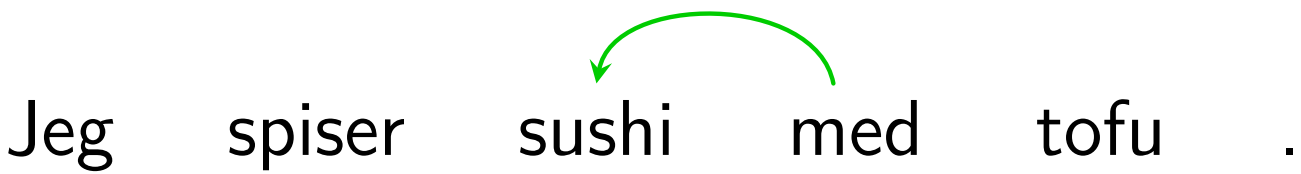
Flertydighet på setningsnivå



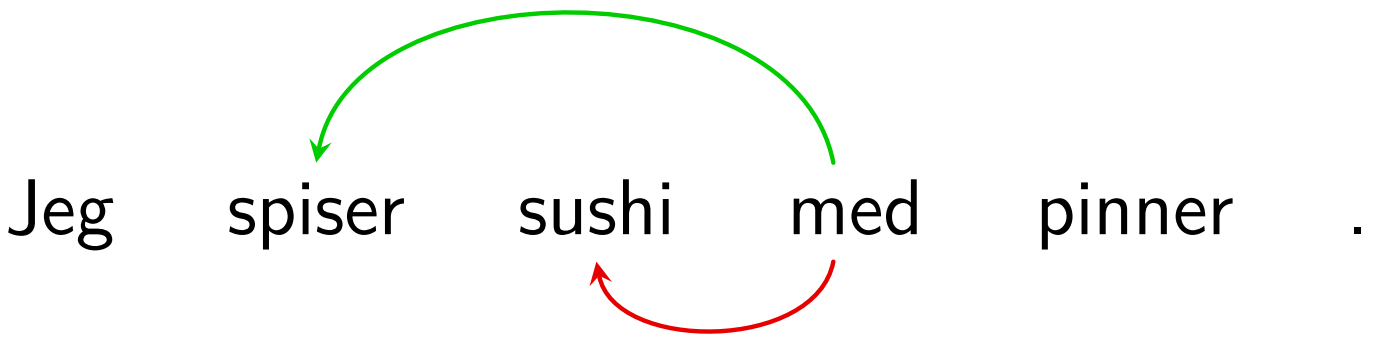
Jeg spiser sushi med tofu .

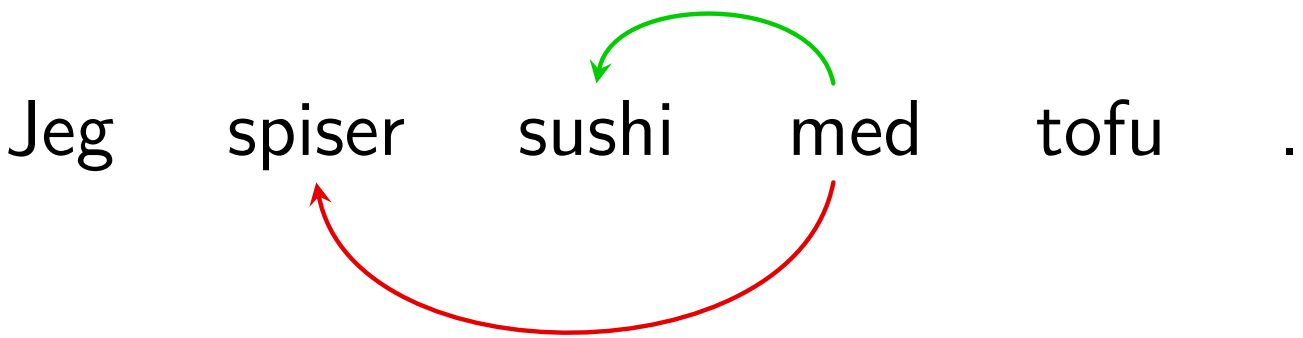
Flertydighet på setningsnivå





Flertydighet på setningsnivå





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.

Historisk: To paradigmer

- Regel-baserte metoder
- Data-drevne metoder / maskinlæring
- Følg instrukser vs. lær fra erfaring

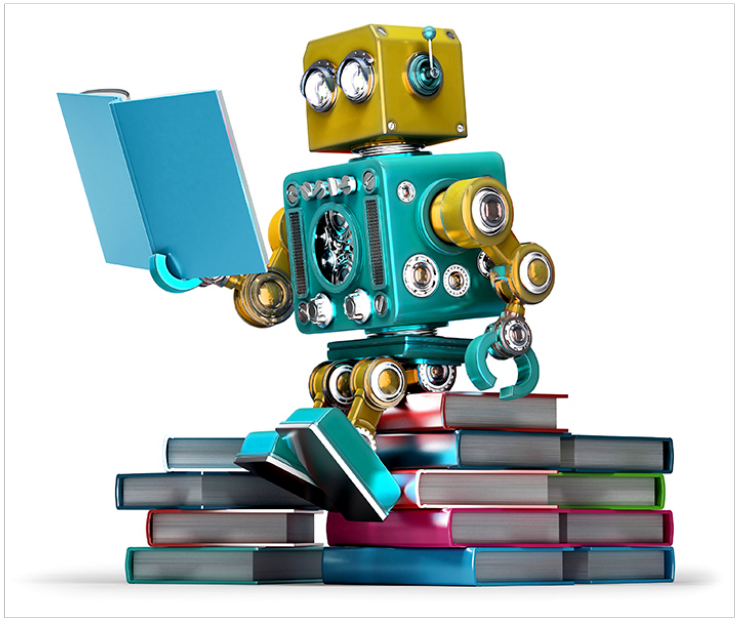
Regel-baserte metoder

- NLP/AI på 1950–90-tallet
- 'Symbolske metoder'
- Formelle regler for å analysere språk og beskrive verden.
- F.eks. manuelt utformede regler og leksikon
- Manuell innsats fra en ekspert.
- Problem: skalerer ikke.

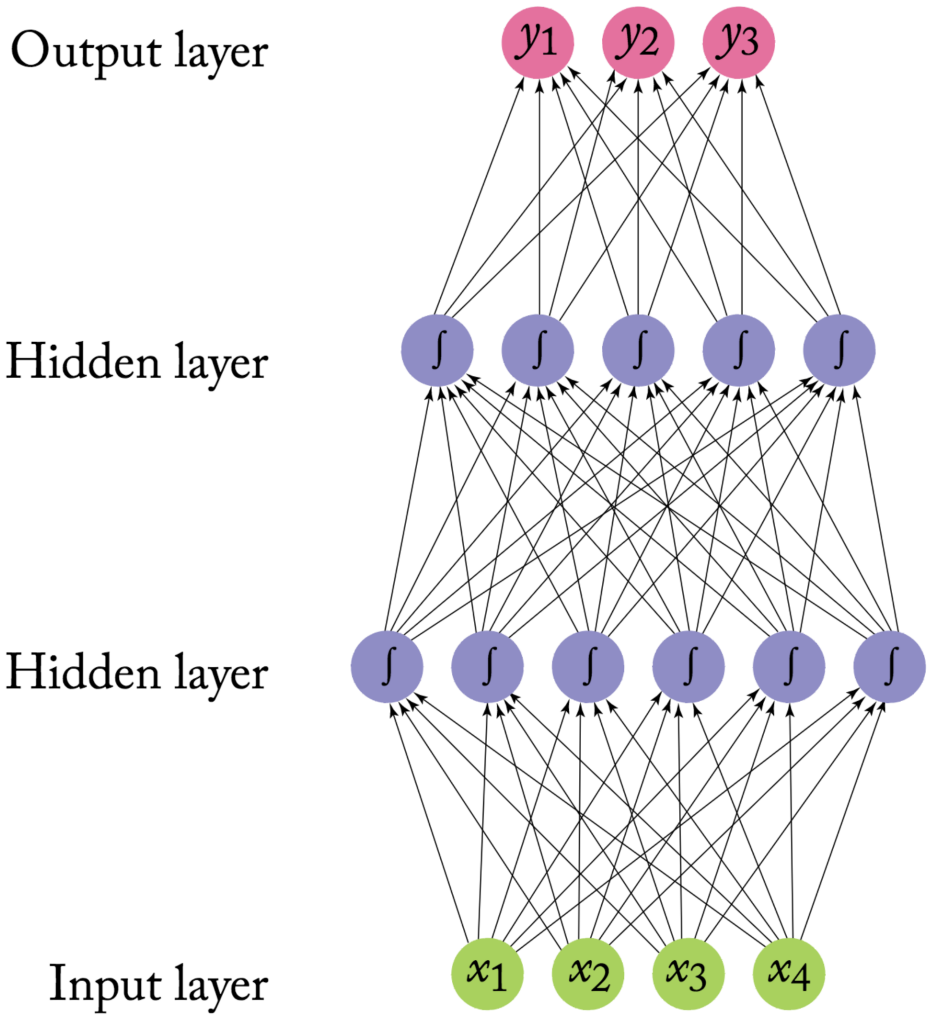


Data-drevne metoder

- NLP/AI på 2000-tallet
- 'Empiriske metoder'
- Maskinlæring: algoritmer som selv lærer fra data.
- I stedet for manuell koding av kunnskap:
- Finn statistiske mønstre i observert språkbruk.



Et andre paradigmeskifte



- NLP/AI fra 2011/2015–
- Kunstige nevrale nett
- En familie kraftige maskinlæringmodeller
- Har revolusjonert mange felt, inkludert språkteknologi
- 'Dyp Læring': NN med flere skjulte lag.

IN1140 Praktiske detaljer



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 BSc-emner, f.eks
  - IN2110 – Språkteknologiske metoder
  - IN3050 – Introduksjon til kunstig intelligens og maskinlæring
  - IN3120 – Søketeknologi

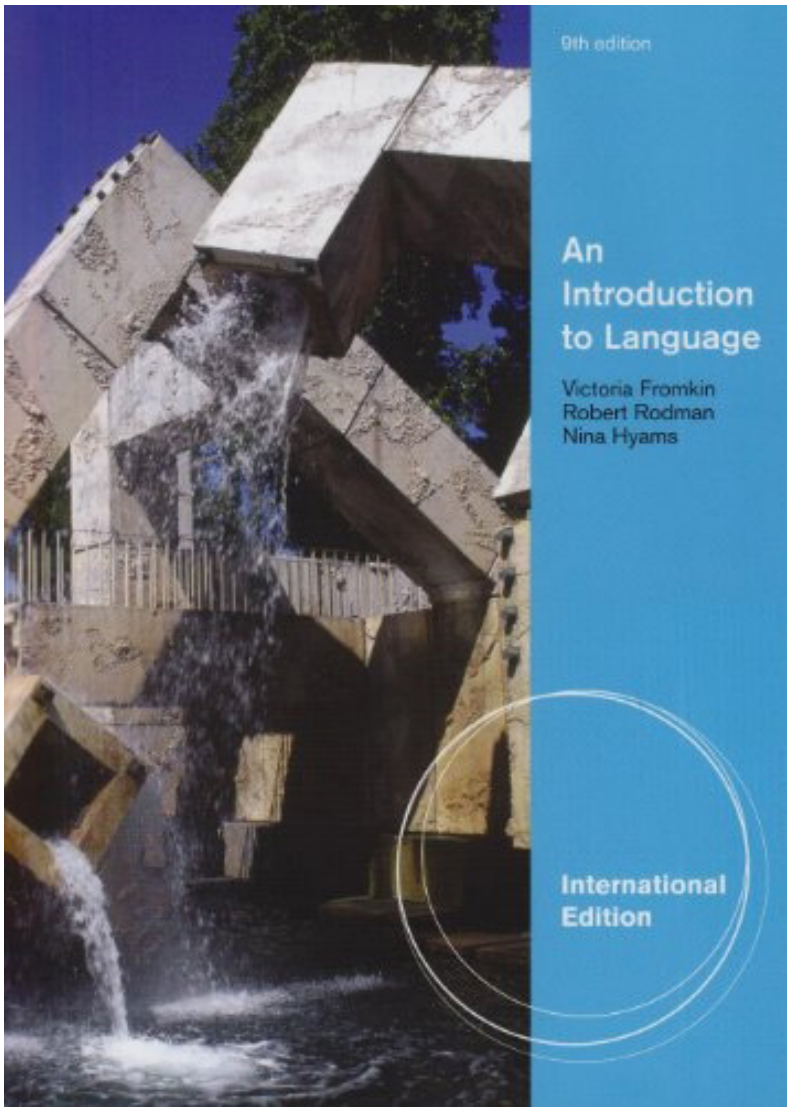
IN1140 og tverrfaglighet

- 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 BSc-emner, f.eks
  - IN2110 – Språkteknologiske metoder
  - IN3050 – Introduksjon til kunstig intelligens og maskinlæring
  - IN3120 – Søketeknologi

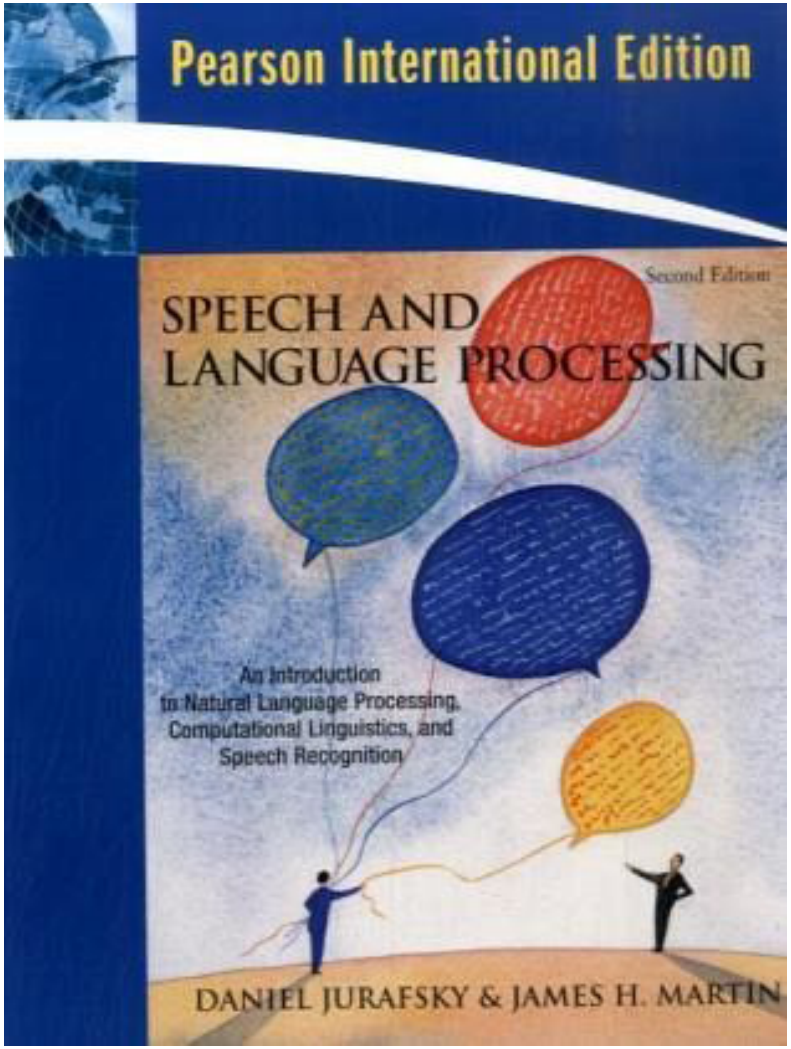
- Mulig å fortsette på MSc-nivå, f.eks:
  - IN4080 – Natural Language Processing
  - IN5550 – Neural Methods in NLP

Pensumlitteratur



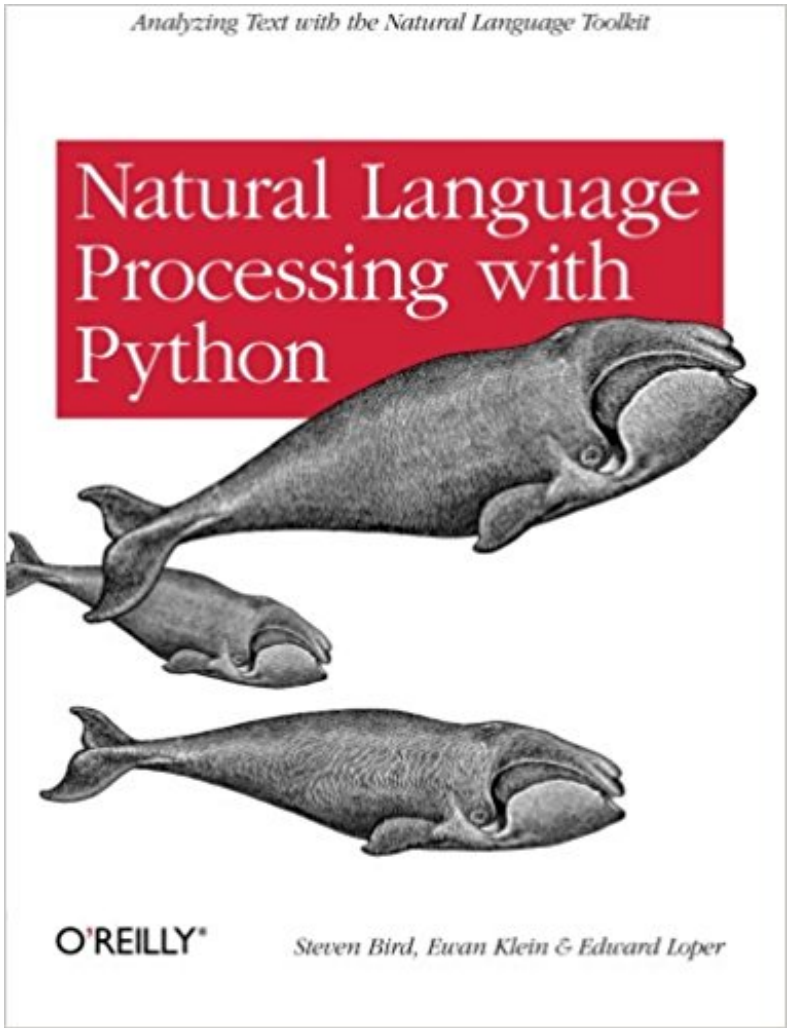
- *An Introduction to Language* av Fromkin, Rodman & Hyams
- Utvalgte deler (ca 5 kapitler)

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.



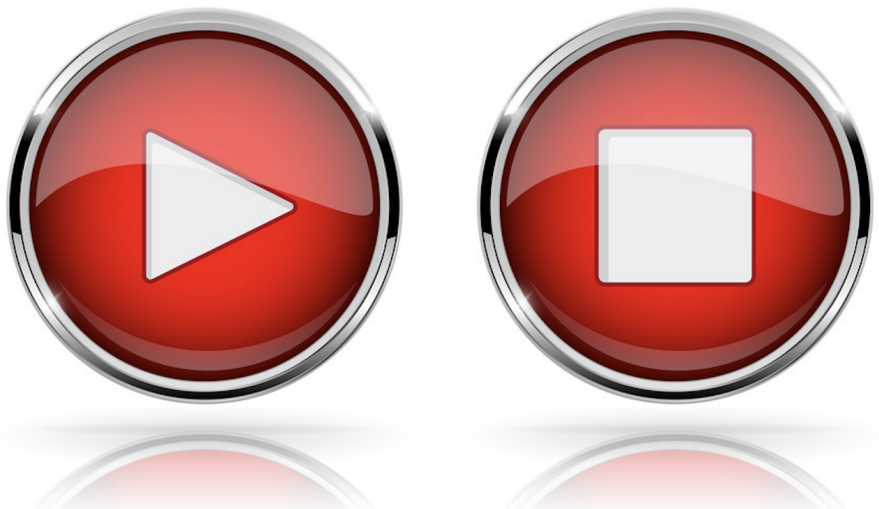
Forelesninger og grupper

Tid & sted

- Forelesninger: man. 14:15–16:00 i Lille Aud. (Kristen Nygaards hus).
- Gruppe 1: tirs. 14:15–16:00, Datastue Modula (OJD/IFI)
- Gruppe 2: tors. 14:15–16:00, Seminarrom Sed (OJD/IFI)
- Gruppe 3: tors. 12:15–14:00, Seminarrom Sed (OJD/IFI)
- **NB!** Første gruppetime er neste uke.

Screencasting

- Tar opp screencast for hver forelesning (lyd + foiler).
- Men emnet er ikke tilrettelagt for digital undervisning:
- Ment som et supplement, for repetisjon.
- Alternativ ved sykdom.



Spørsmål og hjelp

- Gruppetimene: Gruppelærerene er der for å hjelpe og veilede.
- Diskusjonsforum: `https://astro-discourse.uio.no/c/in1140-24h`
- `in1140-hjelp [at] ifi.uio.no`: Felles adresse til fag-/gruppelærere.
- Send alltid henvendelser til lista framfor individuelle adresser.
- Faglige spørsmål skal rettes til forumet.



Beskjeder

- Husk å sjekke UiO-eposten din og beskjedlisten på semestersiden.
- `http://www.uio.no/studier/emner/matnat/ifi/IN1140/h24/`



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)!

Eksamen

- Skriftlig (digital) skoleeksamen på fire timer
  - 28 november kl. 15:00
- Pensumlitteratur + forelesningsnotater
- NB! Ikke en programmeringseksamen.
- Fokus på teoretiske konsepter.

Suksessoppskrift

- Emnesiden: timeplan, pensum, lesehenvisninger, beskjeder etc.
- Lesehenvisninger: selvstudier i tillegg til forelesninger/grupper
- Still spørsmål
- Gå på gruppetimene:
  - 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