IN1140: Introduksjon til språkteknologi

Forelesning #2

Sondre Wold

Universitetet i Oslo

31 August 2022



Tema forrige uke

- Introduksjon
- Hva er språkteknologi?
- Hva er IN1140?
- Praktiske detaljer

Tema for i dag

- Hva lærer jeg i IN1140?
- Hva er lingvistikk?
  - Til i dag: *An Introduction to Langauge*, kapittel 1: What is Language?
- Språkteknologiske komponenter
- Metoder

Men først...

Resultater fra spørreundersøkelsen!

Etter å ha tatt IN1140, kan du ...



Etter å ha tatt IN1140, kan du . . .

- . . . skrive enkle programmer for å manipulere store tekstmengder i Python

Etter å ha tatt IN1140, kan du . . .

- . . . skrive enkle programmer for å manipulere store tekstmengder i Python
- . . . trekke ut alle ord i en tekst, dvs. utføre såkalt **tokenisering**

Etter å ha tatt IN1140, kan du . . .

- . . . skrive enkle programmer for å manipulere store tekstmengder i Python
- . . . trekke ut alle ord i en tekst, dvs. utføre såkalt **tokenisering**
- . . . lage frekvenslister
  - Hva er "årets ord"?

Etter å ha tatt IN1140, kan du . . .

- . . . skrive enkle programmer for å manipulere store tekstmengder i Python
- . . . trekke ut alle ord i en tekst, dvs. utføre såkalt **tokenisering**
- . . . lage frekvenslister
  - Hva er "årets ord"?
    2021: *å sportsvaske*

Etter å ha tatt IN1140, kan du . . .

- . . . skrive enkle programmer for å manipulere store tekstmengder i Python
- . . . trekke ut alle ord i en tekst, dvs. utføre såkalt **tokenisering**
- . . . lage frekvenslister
  - Hva er "årets ord"?
    2021: *å sportsvaske*
    2020: *koronaen*,

Etter å ha tatt IN1140, kan du . . .

- . . . skrive enkle programmer for å manipulere store tekstmengder i Python
- . . . trekke ut alle ord i en tekst, dvs. utføre såkalt **tokenisering**
- . . . lage frekvenslister
  - Hva er "årets ord"?
    2021: *å sportsvaske*
    2020: *koronaen*,
    2019: *klimabrøl*,

Etter å ha tatt IN1140, kan du . . .

- . . . skrive enkle programmer for å manipulere store tekstmengder i Python
- . . . trekke ut alle ord i en tekst, dvs. utføre såkalt **tokenisering**
- . . . lage frekvenslister
  - Hva er "årets ord"?
    2021: *å sportsvaske*
    2020: *koronaen*,
    2019: *klimabrøl*,
    2018: *skjebnelandsmøte*,

Etter å ha tatt IN1140, kan du . . .

- . . . skrive enkle programmer for å manipulere store tekstmengder i Python
- . . . trekke ut alle ord i en tekst, dvs. utføre såkalt **tokenisering**
- . . . lage frekvenslister
  - Hva er "årets ord"?
    2021: *å sportsvaske*
    2020: *koronaen*,
    2019: *klimabrøl*,
    2018: *skjebnelandsmøte*,
    2017: *falske nyheter*,
    2016: *hverdagsintegrering*,
    2015: *det grønne skiftet*,
    2014: *fremmedkriger*,
    2013: *sakte-tv*,
    2012: *å nave (naving)*,
    2011: *rosetog*,
    2010: *askefast*

Etter å ha tatt IN1140, kan du . . .

. . . beregne sannsynligheten for ord i en viss kontekst ved hjelp av enkle språkmodeller

Eksempel 1

ja takk, det vil jeg . . .

- *gjerne*?
- *hjerne*?

Eksempel 2

Jeg har ikke . . . det!

- *hjort*?
- *gjort*?

Etter å ha tatt IN1140, kan du . . .

- . . . automatisk merke opp ("tagge") en tekst med ordklasser:

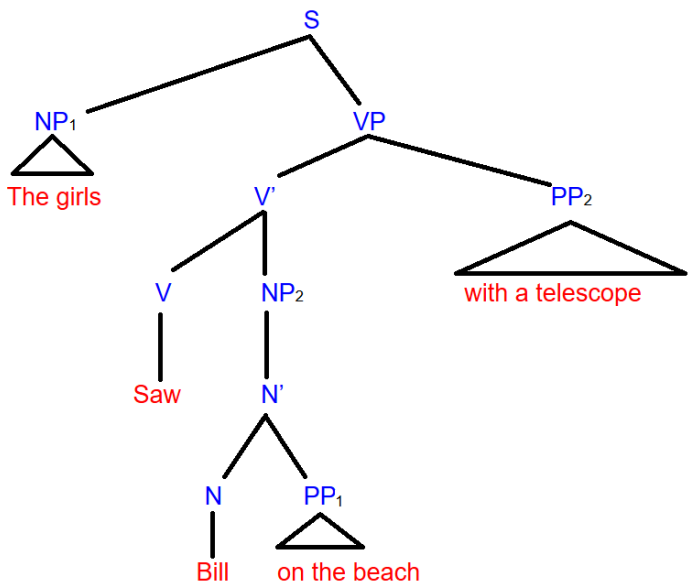
Eksempel

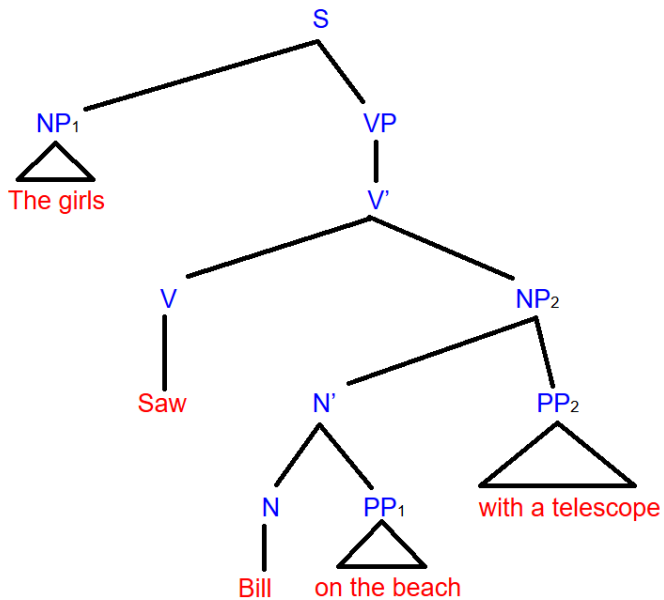
After the social browser launched two weeks earlier, talk about it exploded

1	After	after	IN
2	the	the	DT
3	social	social	JJ
4	browser	browser	NN
5	launched	launch	VVD
6	two	two	JJ
7	weeks	week	NN
8	earlier	earlier	RBR
9	,	,	,
10	talk	talk	NN
11	about	about	IN
12	it	it	PP
13	exploded	explode	VVD

Etter å ha tatt IN1140, kan du . . .

- . . . forklare hva som gir opphav til flertydighet i språk og illustrere forskjeller, feks ved hjelp av syntaktiske trær:
  - *The girls saw Bill on the beach with a telescope*





Etter å ha tatt IN1140, kan du ...

- ...forstå og anvende en enkel maskinlæringsalgoritme (Naive Bayes) til automatisk tekstklassifisering, f.eks:
  - Sentimentanalyse
  - Spam/ikke-spam
  - Tema
- ...gi en enkel redegjørelse for konsepter som supervised og unsupervised læring.
- ...forklare hvilke utfordringer automatisk analyse av naturlige språk gir opphav til.

Hva er lingvistikk?

Hva er lingvistikk?

- Det vitenskapelige studiet av menneskelige språk
- Sentrale spørsmål for en lingvist kan være:
  - hva har ulike språk til felles?
  - hvordan fungerer språk?
  - hvordan forandrer språk seg over tid?
  - hvordan lærer barn språk?
  - hvordan lagres språk i hjernen?

Hva er lingvistikk?

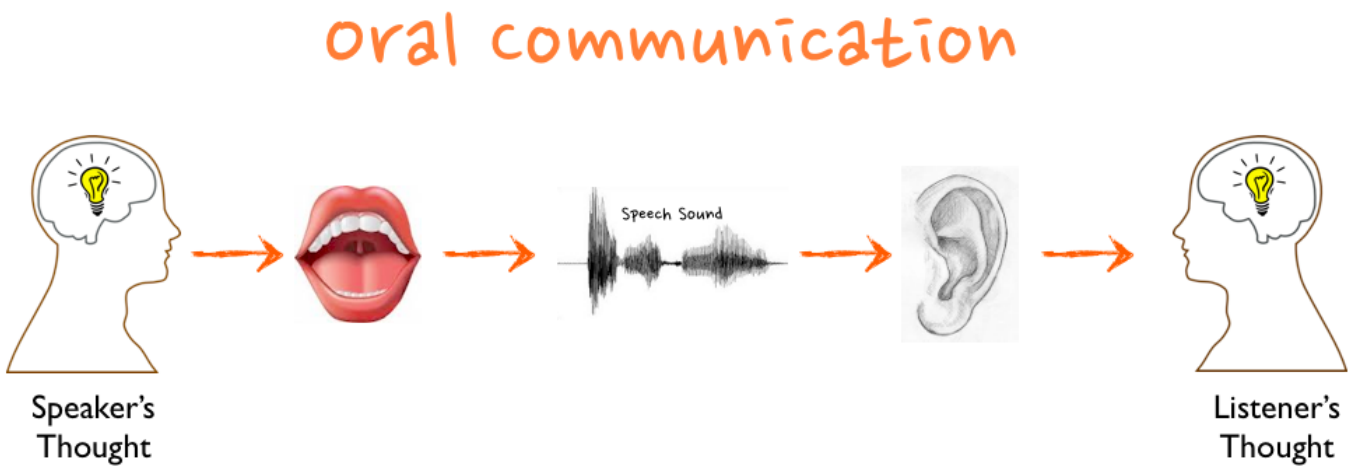
- Det vitenskapelige studiet av menneskelige språk
- Sentrale spørsmål for en lingvist kan være:
  - hva har ulike språk til felles?
  - hvordan fungerer språk?
  - hvordan forandrer språk seg over tid?
  - hvordan lærer barn språk?
  - hvordan lagres språk i hjernen?
  - Hva vil det si å kunne et språk?

Lingvistisk kunnskap

Lingvistisk kunnskap

- Kunnskap om lyd:
  - lydsystemet for et språk
  - rekkefølgen på lyder
- Kunnskap om tegn (tegnspråk)

Lingvistisk kunnskap





Lingvistisk kunnskap

- Kunnskap om ord:
  - Visse lydsekvenser korresponderer til et visse konsepter, eller betydning
  - Vilkårlig (arbitrær) kobling mellom form og betydning

- tree
- sabah
- ciel

Lingvistisk kunnskap

- Kunnskap om ord:
  - Visse lydsekvenser korresponderer til et visse konsepter, eller betydning
  - Vilkårlig (arbitrær) kobling mellom form og betydning

- tree : "tre" – engelsk
- sabah
- ciel

Lingvistisk kunnskap

- Kunnskap om ord:
  - Visse lydsekvenser korresponderer til et visse konsepter, eller betydning
  - Vilkårlig (arbitrær) kobling mellom form og betydning

- tree : "tre" – engelsk
- sabah : "morgen" – arabisk
- ciel

Lingvistisk kunnskap

- Kunnskap om ord:
  - Visse lydsekvenser korresponderer til et visse konsepter, eller betydning
  - Vilkårlig (arbitrær) kobling mellom form og betydning

- tree : "tre" – engelsk
- sabah : "morgen" – arabisk
- ciel : "himmel" – fransk

Lingvistisk kunnskap

- Kunnskap om ord:
  - Visse lydsekvenser korresponderer til et visse konsepter, eller betydning
  - Vilkårlig (arbitrær) kobling mellom form og betydning

- tree : "tre" – engelsk
- sabah : "morgen" – arabisk
- ciel : "himmel" – fransk

Lingvistisk kunnskap

- Kunnskap om ord:
  - Visse lydsekvenser korresponderer til et visse konsepter, eller betydning
  - Vilkårlig (arbitrær) kobling mellom form og betydning

- tree : "tre" – engelsk
- sabah : "morgen" – arabisk
- ciel : "himmel" – fransk

  - Konvensjonell sammenheng: må læres

Lingvistisk kunnskap

- Kunnskap om ord:
  - Visse lydsekvenser korresponderer til et visse konsepter, eller betydning
  - Vilkårlig (arbitrær) kobling mellom form og betydning

- tree : "tre" – engelsk
- sabah : "morgen" – arabisk
- ciel : "himmel" – fransk

  - Konvensjonell sammenheng: må læres

- Er det alt?

Lingvistisk kunnskap

- Kunnskap om hvordan ord settes sammen til setninger
  - Clinton slo Trump
  - Trump slo Clinton



Lingvistisk kunnskap

Mengden av mulige setninger er i teorien uendelig

Lingvistisk kunnskap

Mengden av mulige setninger er i teorien uendelig

- Dette er en setning

Lingvistisk kunnskap

Mengden av mulige setninger er i teorien uendelig

- Dette er en setning
- Dette er en setning som jeg skriver akkurat nå

Lingvistisk kunnskap

Mengden av mulige setninger er i teorien uendelig

- Dette er en setning
- Dette er en setning som jeg skriver akkurat nå
- Dette er en setning som jeg tror at jeg skriver akkurat nå
- Dette er en setning som dere mener at jeg tror at jeg skriver akkurat nå
- osv.

Lingvistisk kunnskap

Mengden av mulige setninger er i teorien uendelig

- Dette er en setning
- Dette er en setning som jeg skriver akkurat nå
- Dette er en setning som jeg tror at jeg skriver akkurat nå
- Dette er en setning som dere mener at jeg tror at jeg skriver akkurat nå
- osv.

- Dette er en kjedelig setning
- Dette er en kjedelig kjedelig setning
- Dette er en kjedelig kjedelig kjedelig setning
- osv.

Lingvistisk kunnskap

Mengden av mulige setninger er i teorien uendelig

- Dette er en setning
- Dette er en setning som jeg skriver akkurat nå
- Dette er en setning som jeg tror at jeg skriver akkurat nå
- Dette er en setning som dere mener at jeg tror at jeg skriver akkurat nå
- osv.

- Dette er en kjedelig setning
- Dette er en kjedelig kjedelig setning
- Dette er en kjedelig kjedelig kjedelig setning
- osv.

- Hvor setter vi grensen?
- Evne til å forstå og skape nye setninger, språkbruk er kreativ – universell egenskap ved språk

Lingvistisk kunnskap

- Grammatikalitet

- Norske sykehus bruker for mye antibiotika
- *Sykehus norske bruker for mye antibiotika
- *Norske sykehus for mye antibiotika bruker
- *Norsk sykehus bruker for mye antibiotika

Lingvistisk kunnskap

- Grammatikalitet

- Norske sykehus bruker for mye antibiotika
- *Sykehus norske bruker for mye antibiotika
- *Norske sykehus for mye antibiotika bruker
- *Norsk sykehus bruker for mye antibiotika

- Kunnskap om *regler* for hvordan man danner setninger i et språk
  - en endelig mengde regler, med et endelig vokabular $\rightarrow$ en uendelig mengde setninger

Grammatikk

- Deskriptiv vs. preskriptiv grammatikk

Grammatikk

- Deskriptiv vs. preskriptiv grammatikk
  - Hvordan det er vs. hvordan det burde være

Grammatikk

- Deskriptiv vs. preskriptiv grammatikk
  - Hvordan det er vs. hvordan det burde være
- Skolegrammatikker
  - Kan være nyttige verktøy

Grammatikk

- Deskriptiv vs. preskriptiv grammatikk
  - Hvordan det er vs. hvordan det burde være
- Skolegrammatikker
  - Kan være nyttige verktøy
- Universell grammatikk
  - Engelsk har et sett med regler, det samme med norsk, arabisk og japansk.
  - Men har de ikke også noe til felles?

Grammatikk

- Deskriptiv vs. preskriptiv grammatikk
  - Hvordan det er vs. hvordan det burde være
- Skolegrammatikker
  - Kan være nyttige verktøy
- Universell grammatikk
  - Engelsk har et sett med regler, det samme med norsk, arabisk og japansk.
  - Men har de ikke også noe til felles?
    - Tid, spørsmål, kommandoer, negasjon osv..

Grammatikk

- Deskriptiv vs. preskriptiv grammatikk
  - Hvordan det er vs. hvordan det burde være
- Skolegrammatikker
  - Kan være nyttige verktøy
- Universell grammatikk
  - Engelsk har et sett med regler, det samme med norsk, arabisk og japansk.
  - Men har de ikke også noe til felles?
    - Tid, spørsmål, kommandoer, negasjon osv..
  - Biologisk forankret?

Språk og tenkning

- Det virker å være en intuitiv kobling mellom det vi sier og det vi tenker.

Språk og tenkning

- Det virker å være en intuitiv kobling mellom det vi sier og det vi tenker.
  - Denne intuisjonen ser ut til å være universell og er derfor også en sitatmaskin:

Språk og tenkning

- Det virker å være en intuitiv kobling mellom det vi sier og det vi tenker.
  - Denne intuisjonen ser ut til å være universell og er derfor også en sitatmaskin:
    - "The limits of my language mean the limits of my world" - Ludwig Wittgenstein, *Tractatus Logico-Philosophicus, 1922*

Språk og tenkning

- Det virker å være en intuitiv kobling mellom det vi sier og det vi tenker.
  - Denne intuisjonen ser ut til å være universell og er derfor også en sitatmaskin:
    - "The limits of my language mean the limits of my world" - Ludwig Wittgenstein, *Tractatus Logico-Philosophicus, 1922*
    - "Langauge is the dress of thought" - Samuel Johnson, *Lives of the English Poets, 1891 ed.*
    - "Without language, thought is a vauge, uncharted nebula" - Ferdinand de Saussure, *Course in General Linguistics, 1916*

Språk og tenkning

- Det virker å være en intuitiv kobling mellom det vi sier og det vi tenker.
  - Denne intuisjonen ser ut til å være universell og er derfor også en sitatmaskin:
    - "The limits of my language mean the limits of my world" - Ludwig Wittgenstein, *Tractatus Logico-Philosophicus, 1922*
    - "Langauge is the dress of thought" - Samuel Johnson, *Lives of the English Poets, 1891 ed.*
    - "Without language, thought is a vauge, uncharted nebula" - Ferdinand de Saussure, *Course in General Linguistics, 1916*
- Sapir-Whorf-hypotesen: språket vårt påvirker hvordan vi oppfatter verden

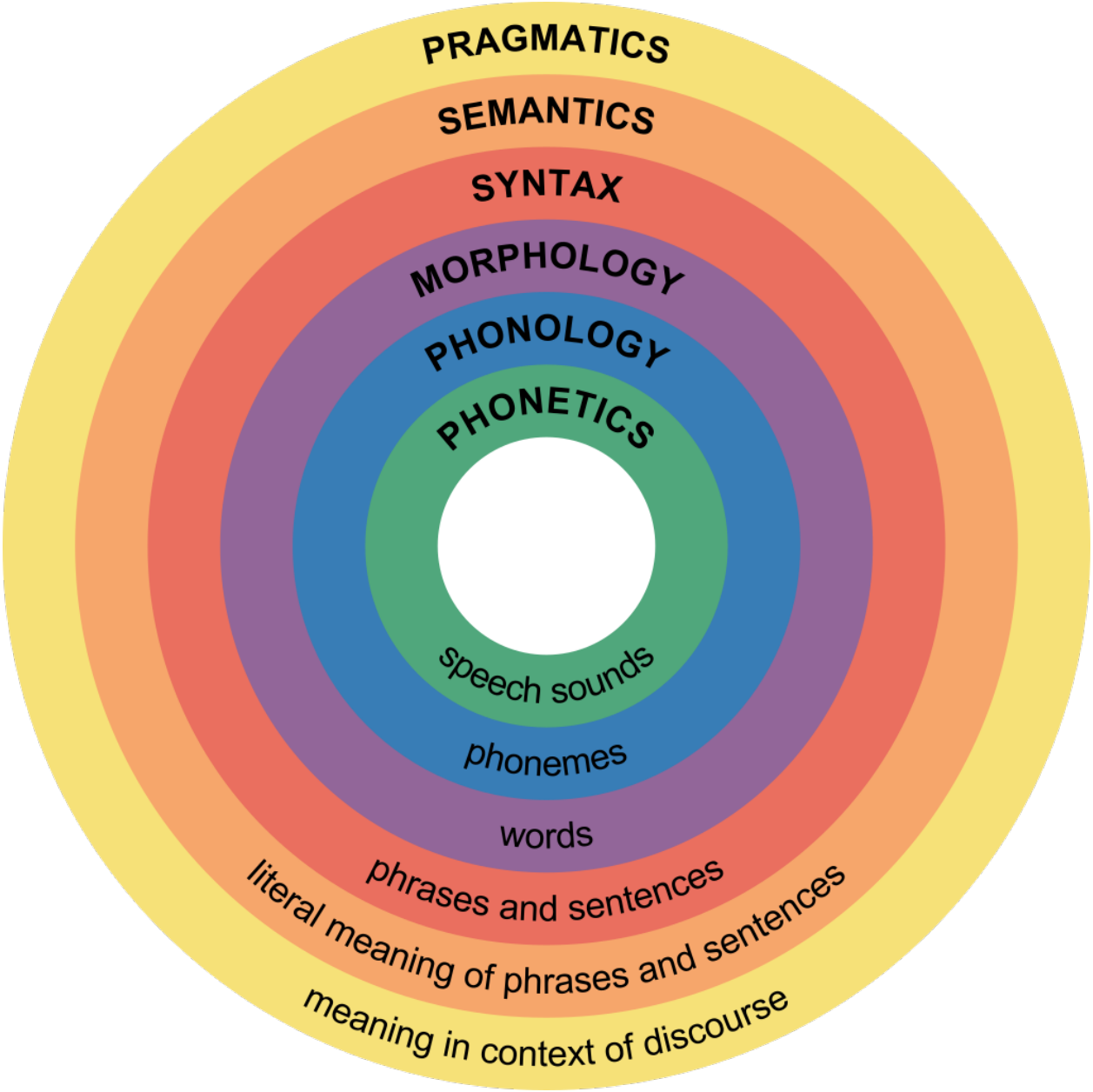
Språk og tenkning

- Det virker å være en intuitiv kobling mellom det vi sier og det vi tenker.
  - Denne intuisjonen ser ut til å være universell og er derfor også en sitatmaskin:
    - "The limits of my language mean the limits of my world" - Ludwig Wittgenstein, *Tractatus Logico-Philosophicus, 1922*
    - "Langauge is the dress of thought" - Samuel Johnson, *Lives of the English Poets, 1891 ed.*
    - "Without language, thought is a vauge, uncharted nebula" - Ferdinand de Saussure, *Course in General Linguistics, 1916*
- Sapir-Whorf-hypotesen: språket vårt påvirker hvordan vi oppfatter verden
  - Sterk variant:
    - Lingvistisk determinisme, språket **bestemmer** verdensbildet vårt.

Språk og tenkning

- Det virker å være en intuitiv kobling mellom det vi sier og det vi tenker.
  - Denne intuisjonen ser ut til å være universell og er derfor også en sitatmaskin:
    - "The limits of my language mean the limits of my world" - Ludwig Wittgenstein, *Tractatus Logico-Philosophicus, 1922*
    - "Langauge is the dress of thought" - Samuel Johnson, *Lives of the English Poets, 1891 ed.*
    - "Without language, thought is a vauge, uncharted nebula" - Ferdinand de Saussure, *Course in General Linguistics, 1916*
- Sapir-Whorf-hypotesen: språket vårt påvirker hvordan vi oppfatter verden
  - Sterk variant:
    - Lingvistisk determinisme, språket **bestemmer** verdensbildet vårt.
  - Svak variant:
    - Lingvistisk relativisme, språk varierer og derfor varierer også måten mennesker med ulike språk tenker på verden.

Lingvistiske nivåer



Hva vi vet om språk

- Der vi finner mennesker, finner vi språk

Hva vi vet om språk

- Der vi finner mennesker, finner vi språk
- Det finnes ingen ”primitive” språk

Hva vi vet om språk

- Der vi finner mennesker, finner vi språk
- Det finnes ingen "primitive" språk
- Alle språk forandrer seg over tid

Hva vi vet om språk

- Der vi finner mennesker, finner vi språk
- Det finnes ingen "primitive" språk
- Alle språk forandrer seg over tid
- Forholdet mellom lyd og betydning er (stort sett) vilkårlig

Hva vi vet om språk

- Der vi finner mennesker, finner vi språk
- Det finnes ingen "primitive" språk
- Alle språk forandrer seg over tid
- Forholdet mellom lyd og betydning er (stort sett) vilkårlig
- Alle menneskelige språk bruker endelig (finitt) mengde lyder og ord til å danne uendelig mengde mulige setninger

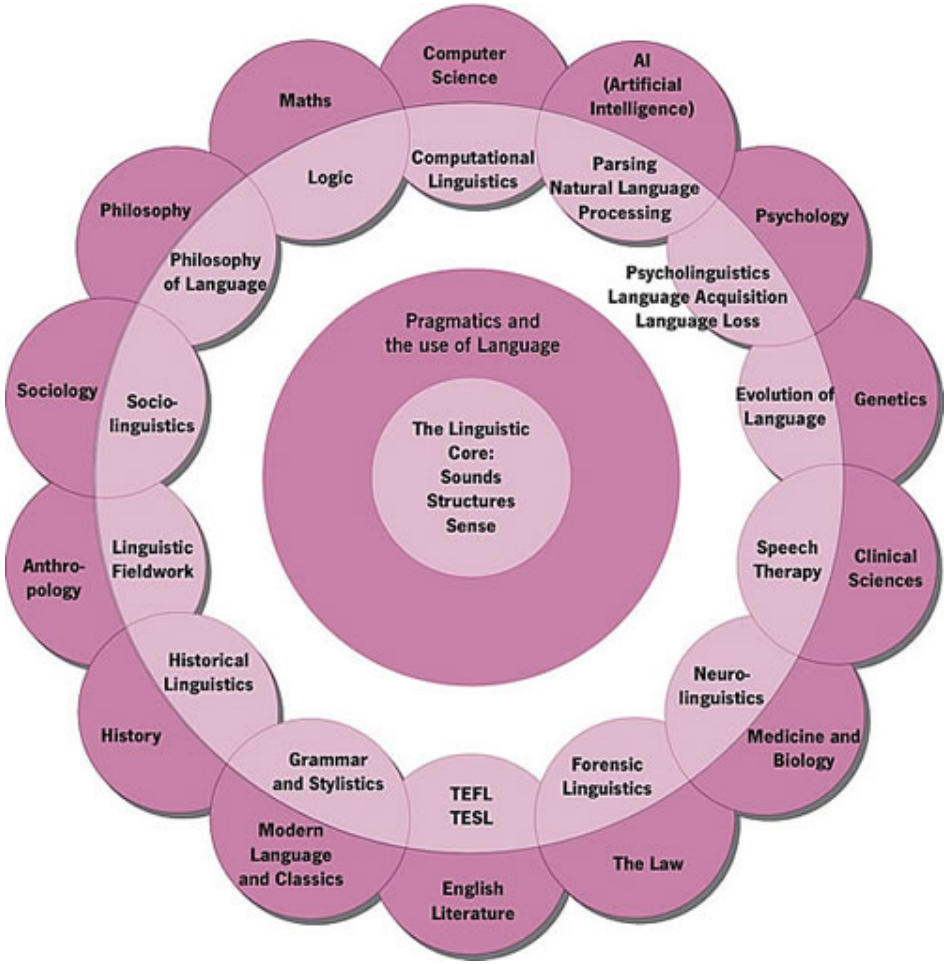
Hva vi vet om språk

- Der vi finner mennesker, finner vi språk
- Det finnes ingen "primitive" språk
- Alle språk forandrer seg over tid
- Forholdet mellom lyd og betydning er (stort sett) vilkårlig
- Alle menneskelige språk bruker endelig (finitt) mengde lyder og ord til å danne uendelig mengde mulige setninger
- Alle språk kan uttrykke negasjon, spørsmål, gi kommandoer, snakke om fortid/framtid, hypotetiske situasjoner

Hva vi vet om språk

- Der vi finner mennesker, finner vi språk
- Det finnes ingen "primitive" språk
- Alle språk forandrer seg over tid
- Forholdet mellom lyd og betydning er (stort sett) vilkårlig
- Alle menneskelige språk bruker endelig (finitt) mengde lyder og ord til å danne uendelig mengde mulige setninger
- Alle språk kan uttrykke negasjon, spørsmål, gi kommandoer, snakke om fortid/framtid, hypotetiske situasjoner
- Ethvert normalt barn er i stand til å lære morsmålet sitt

- Lingvistiske disipliner



Språkteknologiske komponenter

Språkteknologiske komponenter

- *Fonetikk/fonologi:* kunnskap om lingvistiske lyder
- Talegjenkjenning/talesyntese:
  - tale $\Rightarrow$ tekst
  - tekst $\Rightarrow$ tale

Språkteknologiske komponenter

- *Fonetikk/fonologi:* kunnskap om lingvistiske lyder
- Talegjenkjenning/talesyntese:
  - tale $\Rightarrow$ tekst
  - tekst $\Rightarrow$ tale

Eksempel problem

Homofoner (homonymer) – ord som uttales likt men har forskjellig betydelse

- weak — week
- to — too — two

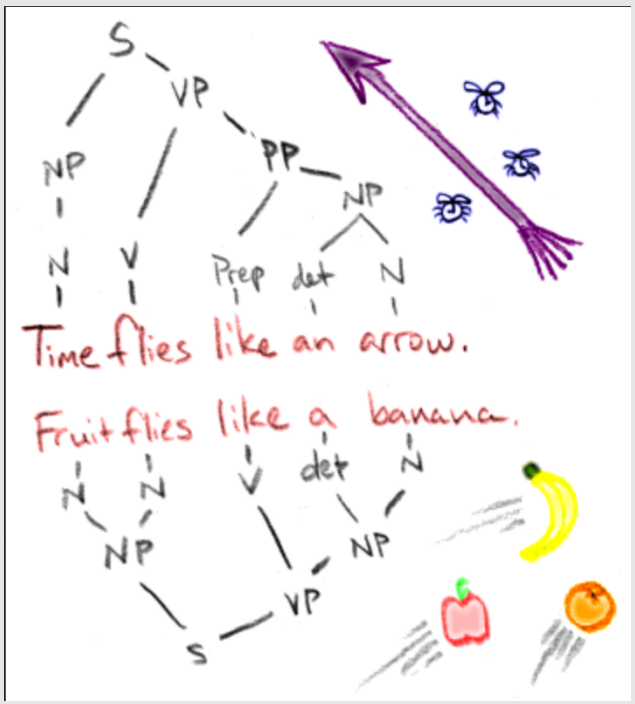
Språkteknologiske komponenter

- *Morfologi:* kunnskap om ordstruktur
  Morfologisk analyse, ordklassetagging

Språkteknologiske komponenter

- *Morfologi:* kunnskap om ordstruktur
  Morfologisk analyse, ordklassetagging

Eksempel problem



Time flies like an arrow

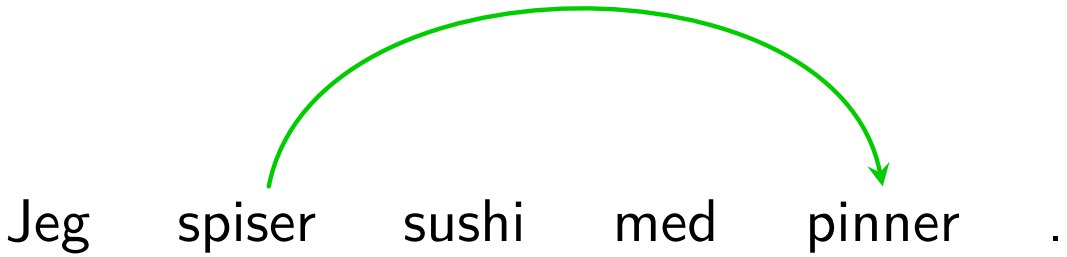
Flertydighet av *flies* og *like* gir opphav til ulike tolkninger

Språkteknologiske komponenter

- *Syntaks:* kunnskap om relasjoner mellom ord
  Chunking, parsing

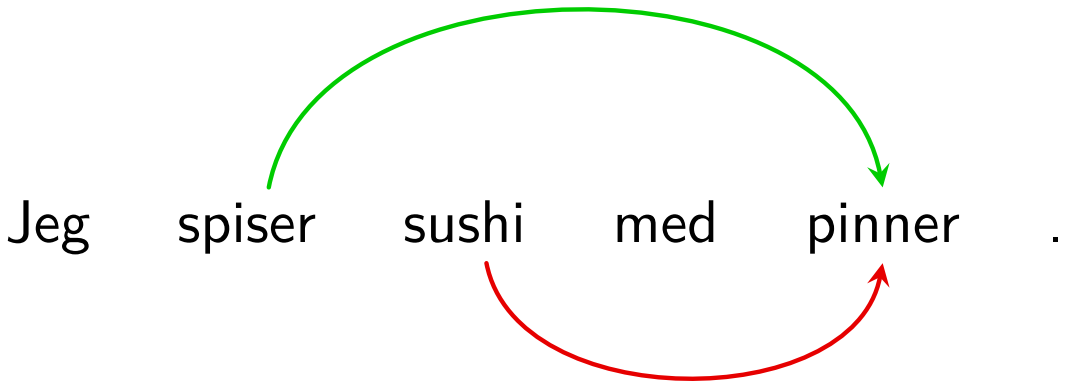
Språkteknologiske komponenter

- *Syntaks:* kunnskap om relasjoner mellom ord
  Chunking, parsing



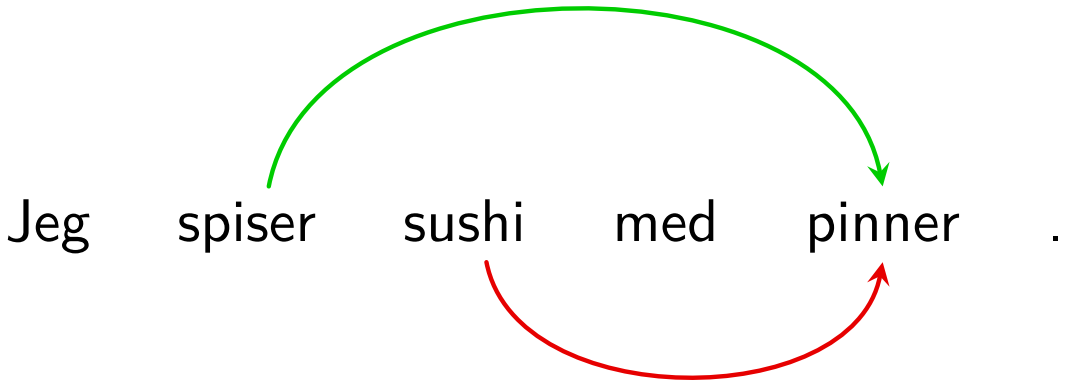
Språkteknologiske komponenter

- *Syntaks:* kunnskap om relasjoner mellom ord
  Chunking, parsing



Språkteknologiske komponenter

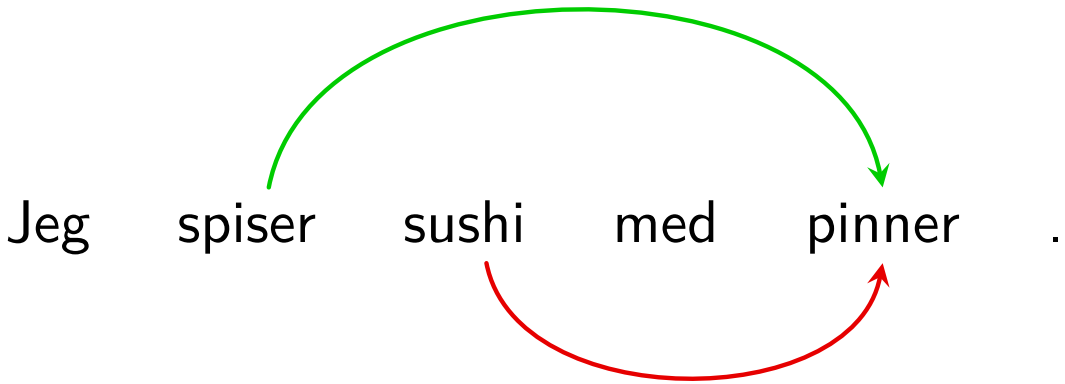
- *Syntaks:* kunnskap om relasjoner mellom ord
  Chunking, parsing

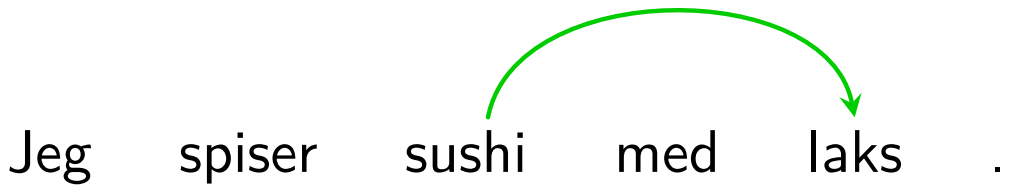


Jeg spiser sushi med laks .

Språkteknologiske komponenter

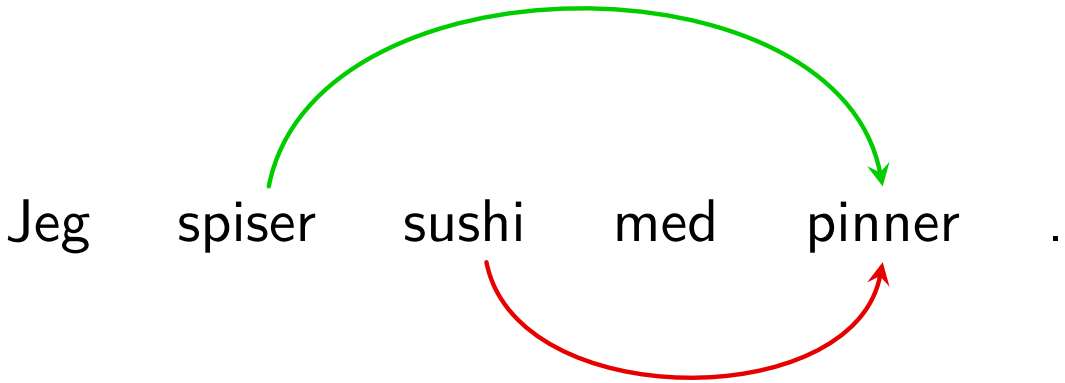
- *Syntaks:* kunnskap om relasjoner mellom ord
  Chunking, parsing

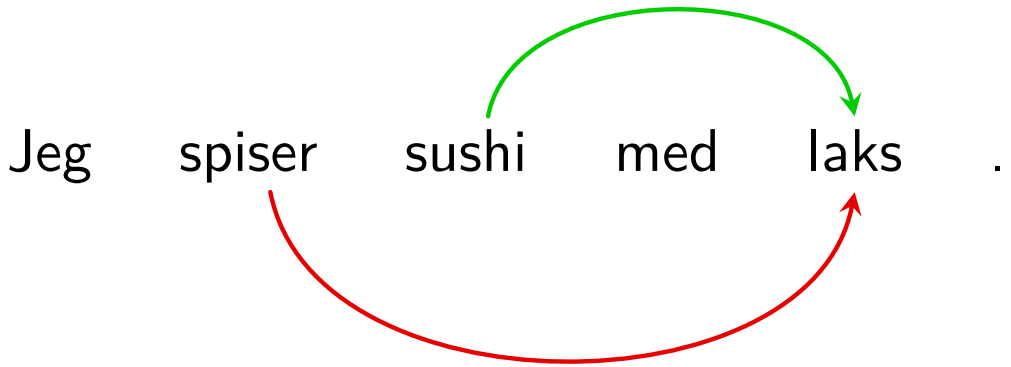




Språkteknologiske komponenter

- *Syntaks:* kunnskap om relasjoner mellom ord
  Chunking, parsing





Språkteknologiske komponenter

- *Semantikk:* kunnskap om betydning — ord, setninger
  "Word Sense Disambiguation" (WSD)

Språkteknologiske komponenter

- *Semantikk:* kunnskap om betydning — ord, setninger
  "Word Sense Disambiguation" (WSD)

Eksempel problem

En form – flere meninger

- Mine mål er egentlig ganske forskjellige

Språkteknologiske komponenter

- *Semantikk:* kunnskap om betydning — ord, setninger
  "Word Sense Disambiguation" (WSD)

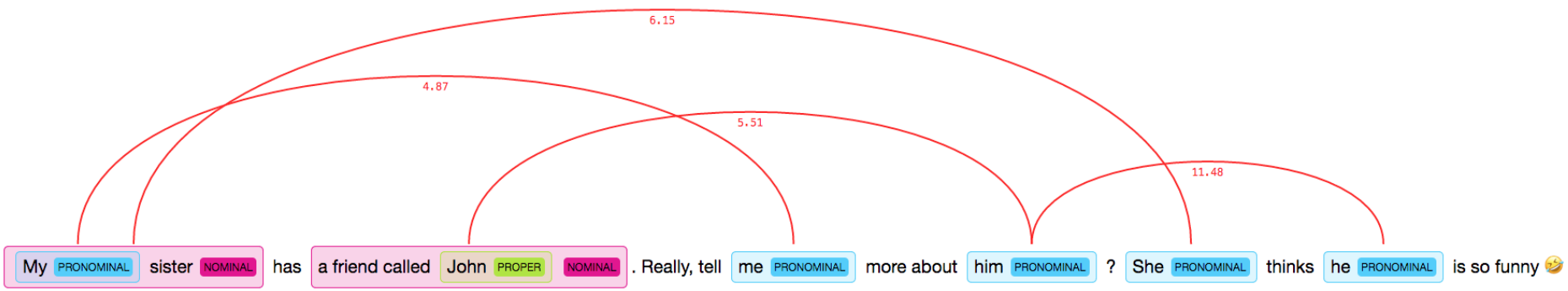
Eksempel problem

En form – flere meninger

- Mine mål er egentlig ganske forskjellige
  - uttalt av en fotballspiller
  - uttalt av en ivrig hobbystrikker
  - uttalt av en IN1140 student

Språkteknologiske komponenter

- *Pragmatikk:* kunnskap om enheter ut over enkelte ytringer
  Anaforresolusjon, dialogsystemer



Språkteknologiske komponenter

- NLP-systemer: moduler som representerer forskjellige lingvistiske nivåer
- "Pipeline"-arkitektur
- "Høyere" nivåer avhenger typisk av "lavere"

- Hvorfor blir resultatene (noen ganger) dårlige?

- Hvorfor blir resultatene (noen ganger) dårlige?
  - Språkforståelse er komplisert
  - Den nødvendige kunnskapen er enorm
  - De fleste stadier viser flertydighet

Flertydighet

- De fleste språkteknologiske applikasjoner må håndtere *flertydighet* ("ambiguity")
- Kjennetegner naturlige språk, på alle nivåer
  - Lindsay told Jessica that she has cancer.
  - Krasjet med rådyr på moped (Agderposten)
  - Heng deg på DNT-tur (Ringerikes blad)

Antagelser

- I møte med språk gjør vi en haug med nødvendige antagelser som baserer seg på tidligere erfaringer.

Antagelser

- I møte med språk gjør vi en haug med nødvendige antagelser som baserer seg på tidligere erfaringer.

Eksempel problem

Mary was invited to Jack's party.
She wondered if he would like a kite.

Antagelser

- I møte med språk gjør vi en haug med nødvendige antagelser som baserer seg på tidligere erfaringer.

Eksempel problem

Mary was invited to Jack's party.
She wondered if he would like a kite.

- The party is a birthday party
- Jack and Mary are children
- "She" is Mary
- "He" is Jack
- She is considering giving Jack a kite
- She wonders if he would like the kite

(Eksempel fra boken The Society of Mind, av Marvin Minsky (1986))



Metoder

Metoder

- $\rightarrow$ 2000-tallet: manuelt utformede regler og leksikon

Regler

En regel er definert over "pattern" og "action"

- (token = "Mr." orthography type = FirstCap) $\rightarrow$ person name
- Mr. Darcy is perfectly imperfect.

Metoder

- 2000-tallet $\rightarrow$: empirisk revolusjon
- Maskinlæring
  - Datamaskiner kan lære fra data: fange opp mønstre og generalisere til nye eksempler.
    - Supervised
    - Weakly supervised
    - Unsupervised

Maskinlæring

Kunstig intelligens: delområde innen informatikk (fra 60-tallet)

- fokus på oppgaver som er lette for mennesker, men vanskelige for maskiner
- språkforståelse er en slik oppgave

Maskinlæring:

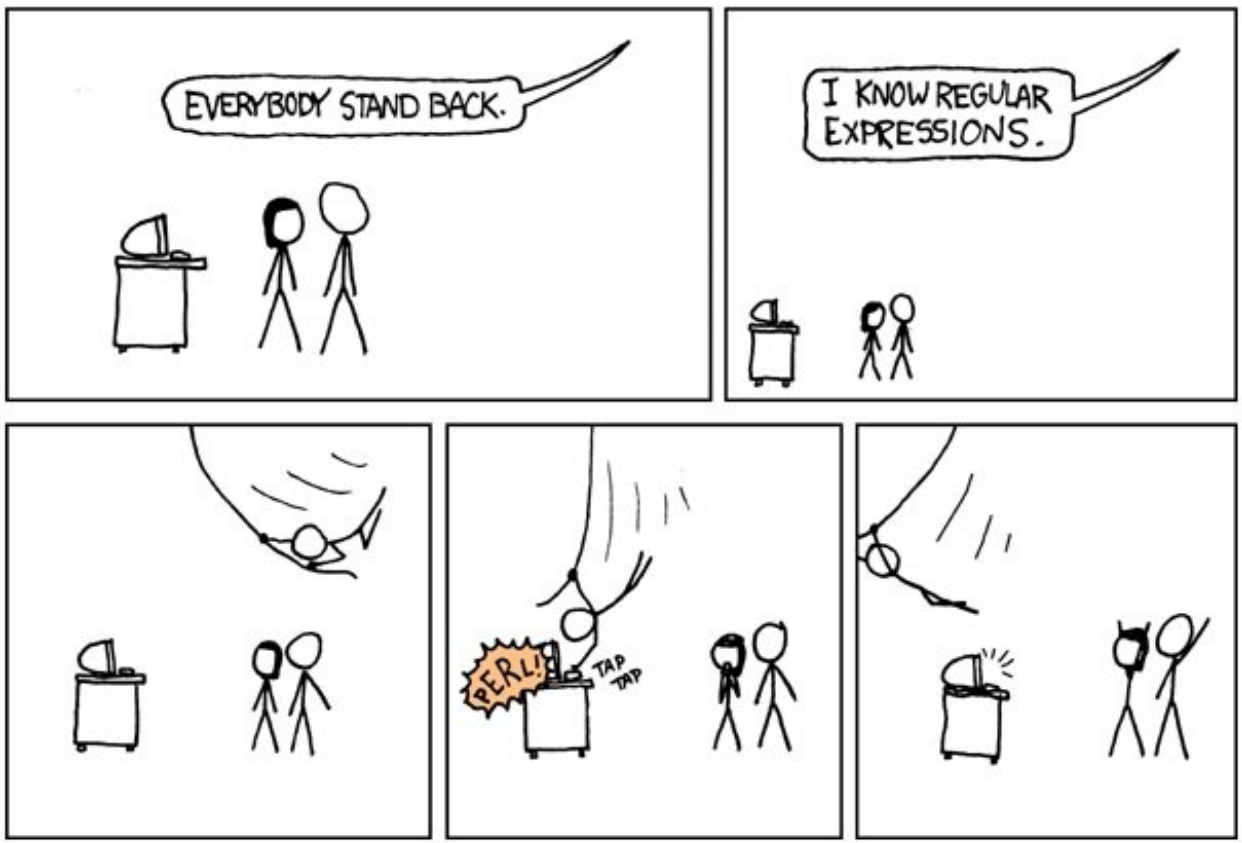
- gitt et AI problem og en masse data om verden (eksempler): la datamaskinen finne riktig svar
- unngår hardkoding av svarene

Formelle modeller

- Formelle modeller hentet fra matematikk, logikk, statistikk
- Beskrive språklige fenomener
- Disse modellene kan prosesseres ved et lite antall kjente algoritmer
- Maskinlæring brukes for å håndtere flertydighet

Formell språkteori

- Regulære språk og regulære uttrykk
- Sekvens av tegn som beskriver et mønster
- Mye brukt til tekstsøk, normalisering av tekst



Formell språkteori

- Formelle regelsystemer
- feks kontekstfrie grammatikker
- Syntaktiske trær

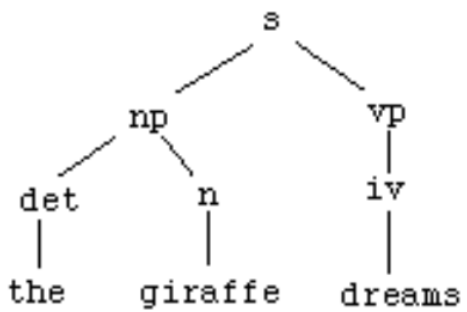
```
s   → np vp

np  → det n
vp  → tv np
    → iv

det → the
    → a
    → an

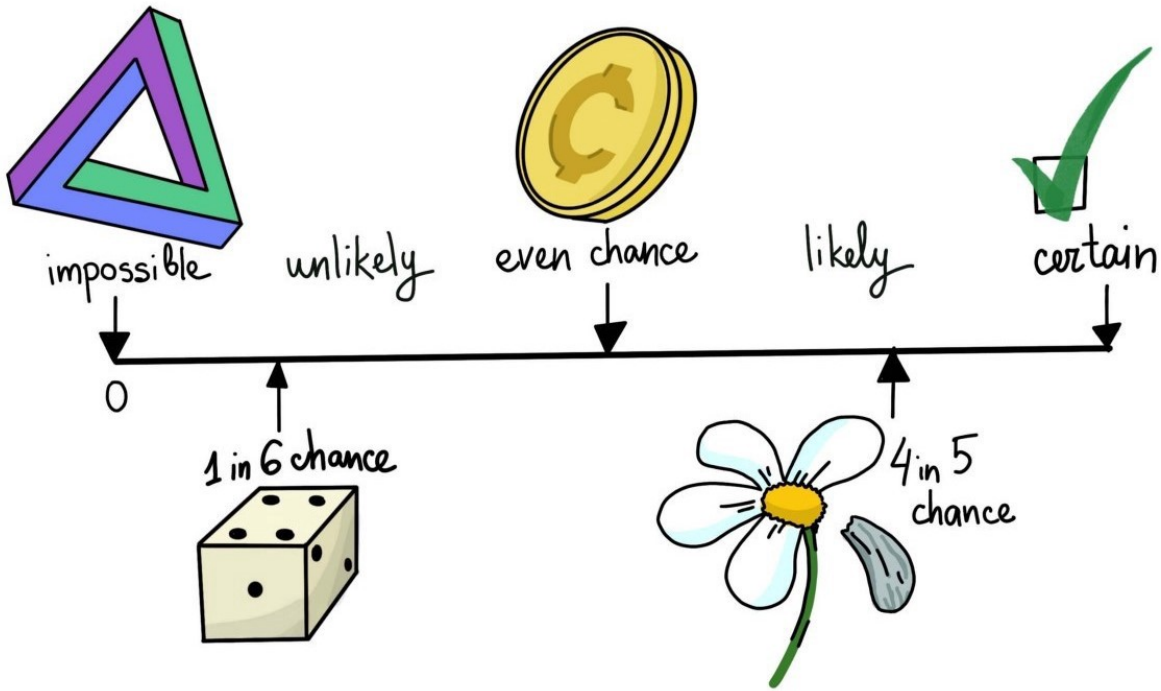
n   → giraffe
    → apple

iv  → dreams
tv  → eats
    → dreams
```



Statistiske modeller

- Sannsynlighetsteori: matematiske modeller som kvantifiserer usikkerhet
- Grunnleggende for alle maskinlæringsalgoritmer



Statistiske modeller

- Språkmodeller: sannsynlighetsfordeling over ord i en sekvens
- Meget sentrale i NLP



- Naive Bayes-klassifisering

Neste uke

- Språklige data
- Første lingvistiske nivå: morfologi
- NB! Husk gruppetime (I morgen for gruppe en og tre, på fredag for gruppe to): Tekst i Python