IN1140: Introduksjon til språkteknologi

Forelesning #2

Lilja Øvrelid

Universitetet i Oslo

1 september 2021



Tema forrige uke

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

Tema for i dag

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

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"?
    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"?
    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"?
    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"?
    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

Eksempel

ja takk, det vil jeg . . .

- *gjerne*?
- *hjerne*?

Etter å ha tatt IN1140, kan du . . .

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

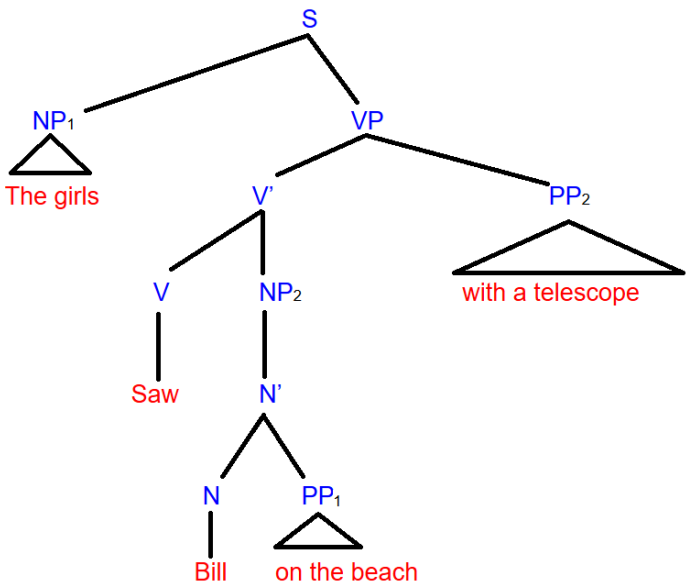
Eksempel

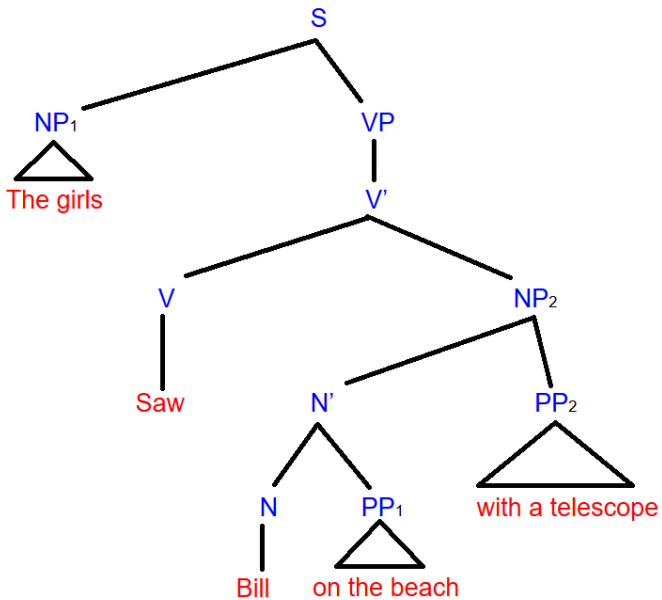
After the social browser launched two weeks earlier, talk about it exploded

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



Hva er lingvistikk?

Hva er lingvistikk?

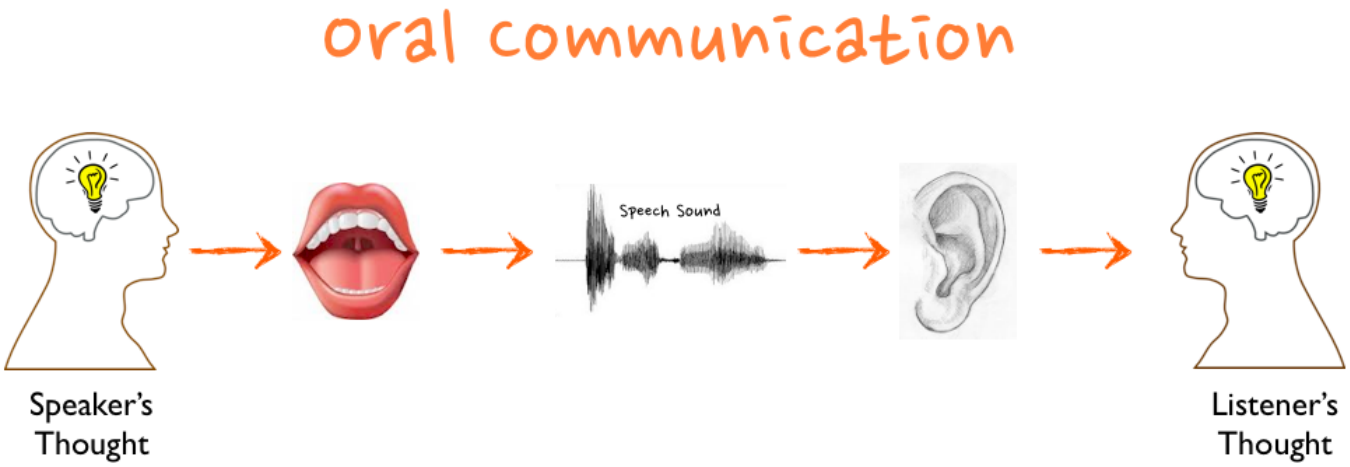
- Det vitenskapelige studiet av menneskelige språk
- Regler, systemer og prinsipper i språk
  - hva har ulike språk til felles?
  - hvordan fungerer språk?
  - hvordan forandrer språk seg over tid?
  - hvordan lærer barn språk?
  - hvordan lagres språk i hjernen?

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

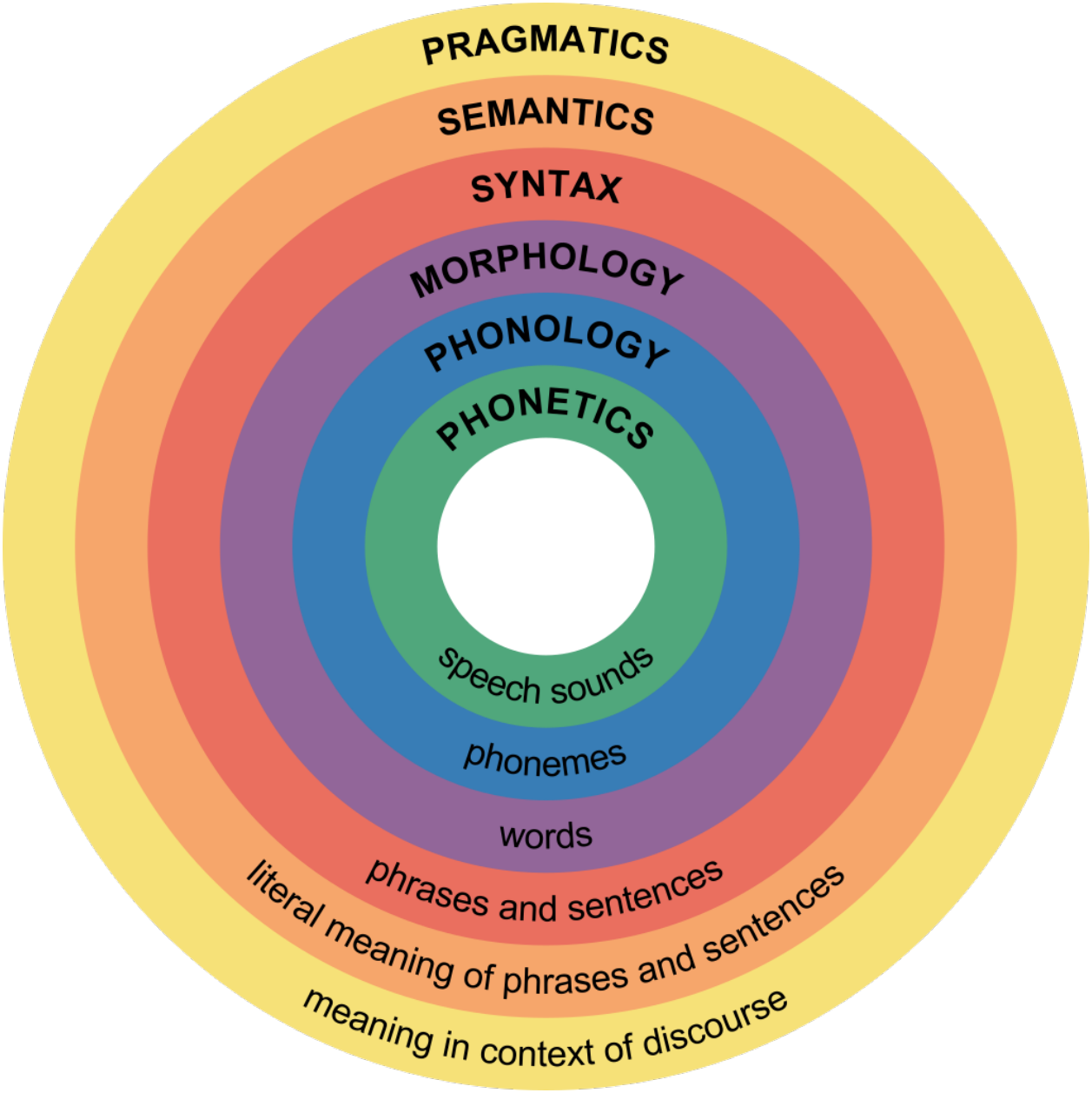
Lingvistisk kunnskap

- 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

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

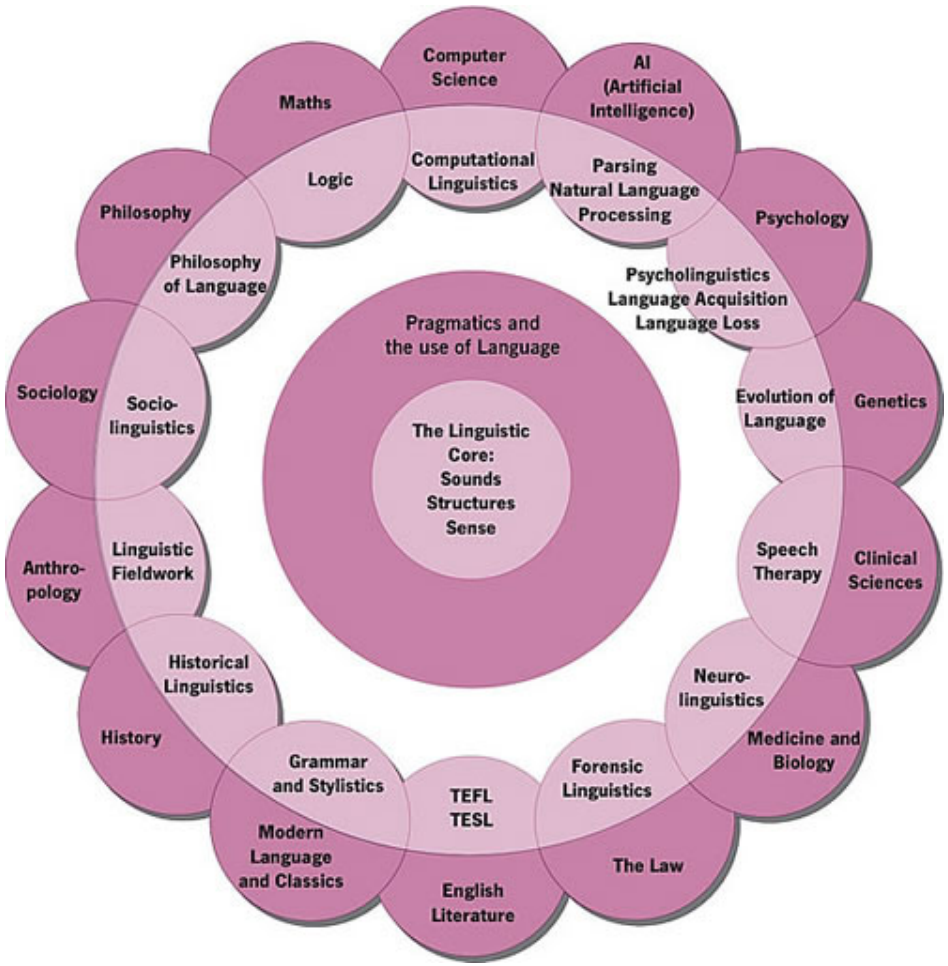
Hva vi vet om språk

- 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

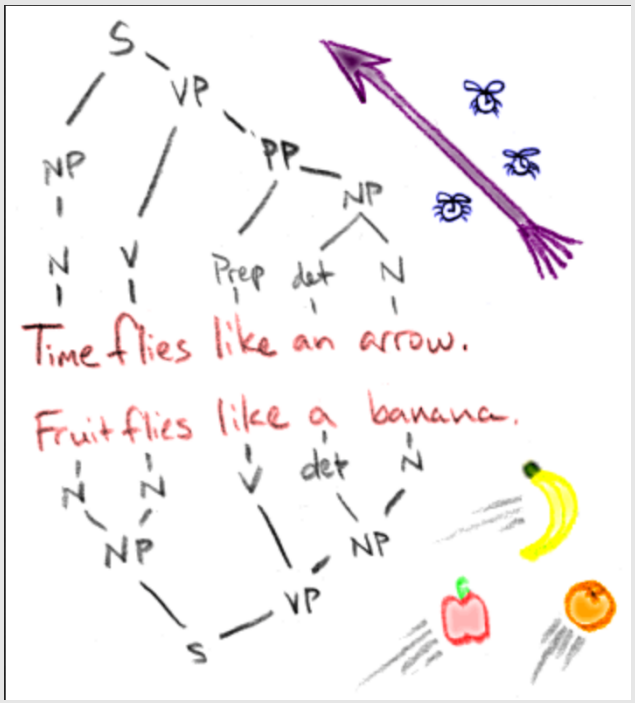
Språkteknologiske komponenter

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

Språkteknologiske komponenter

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

Eksempel problem



Time flies like an arrow

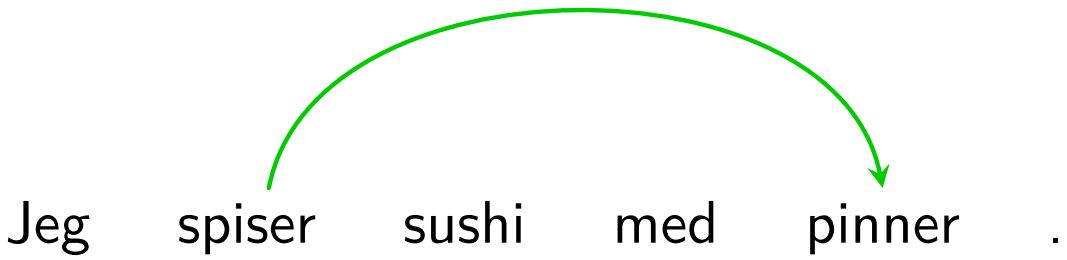
Flertydighet av *flies* og *like* gir opphav til ulike tolkninger

Språkteknologiske komponenter

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

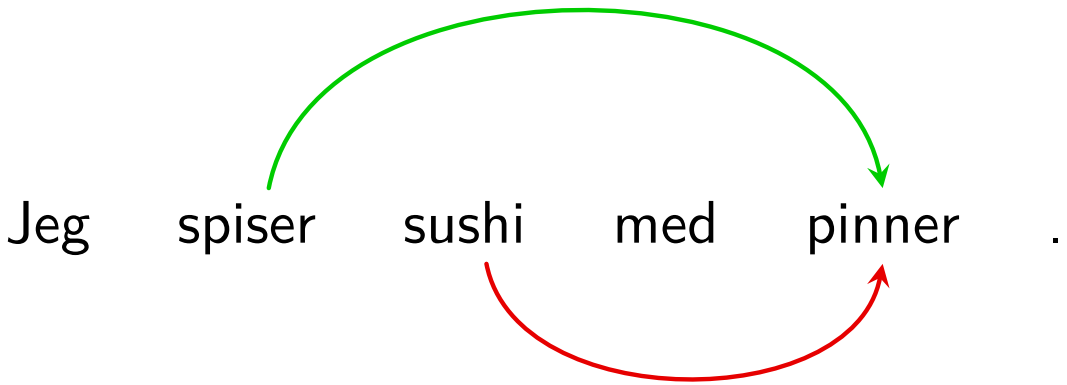
Språkteknologiske komponenter

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



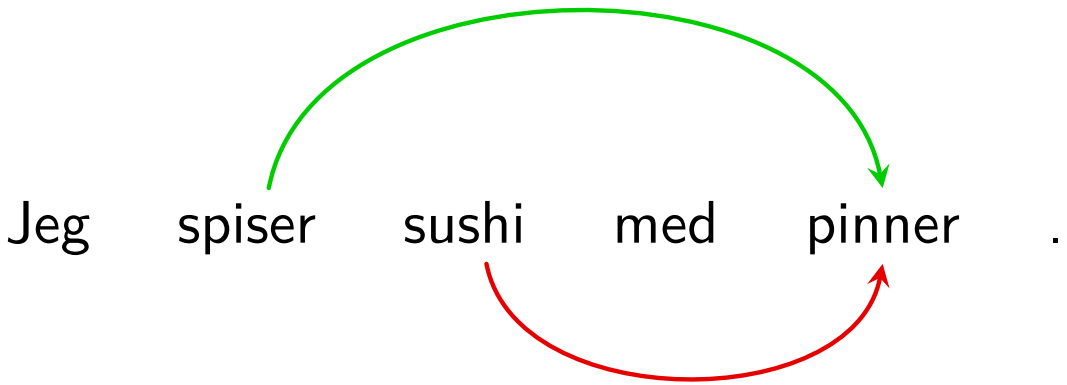
Språkteknologiske komponenter

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



Språkteknologiske komponenter

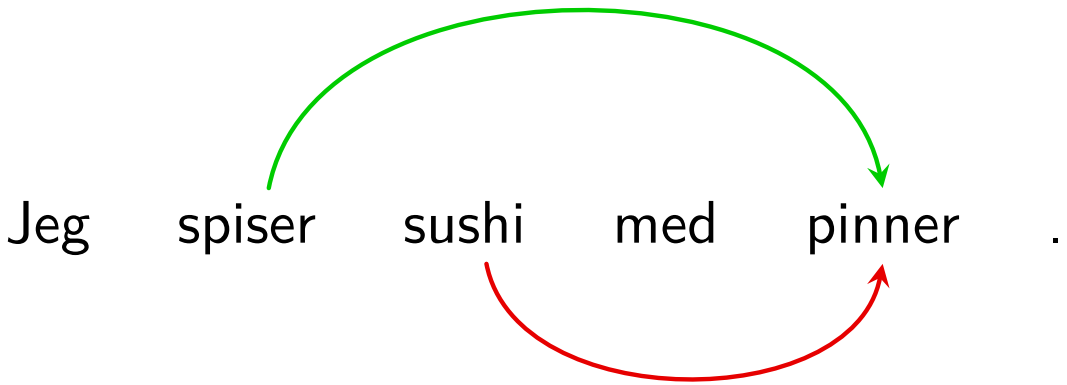
- *Syntaks:* kunnskap om relasjoner mellom ord
  Chunking, parsing

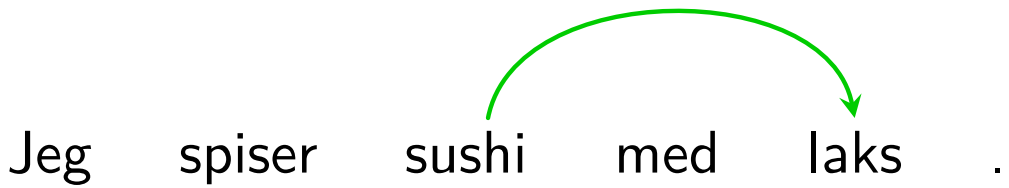


Jeg spiser sushi med laks .

Språkteknologiske komponenter

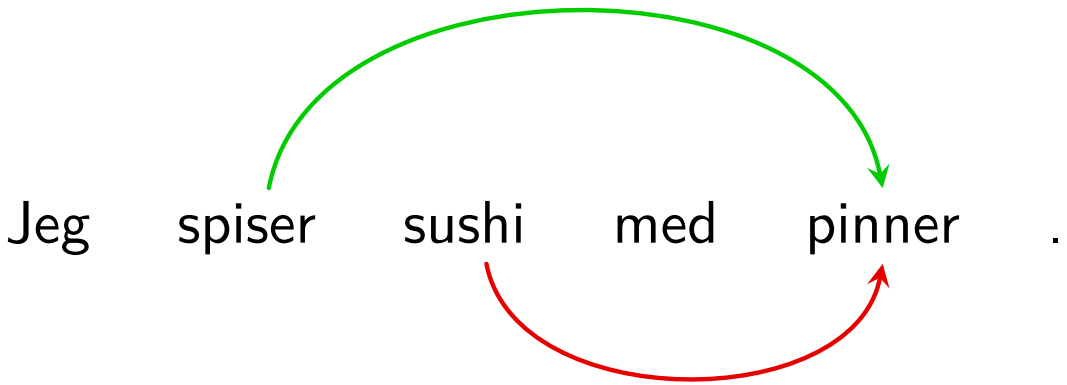
- *Syntaks:* kunnskap om relasjoner mellom ord
  Chunking, parsing

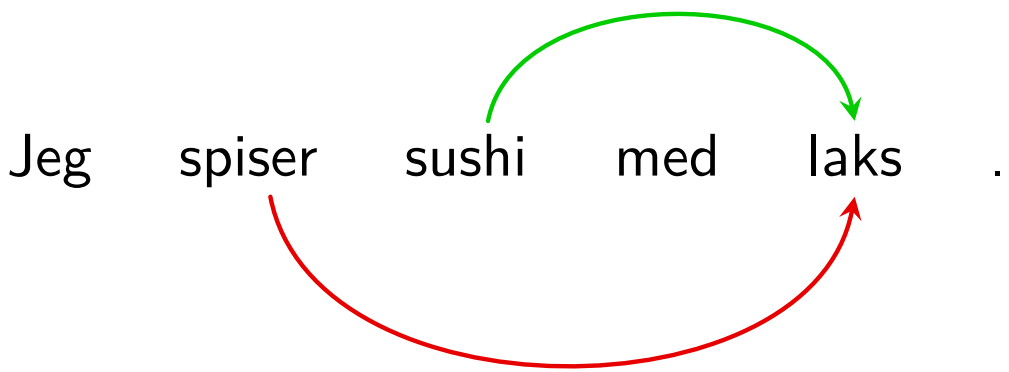




Språkteknologiske komponenter

- *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)

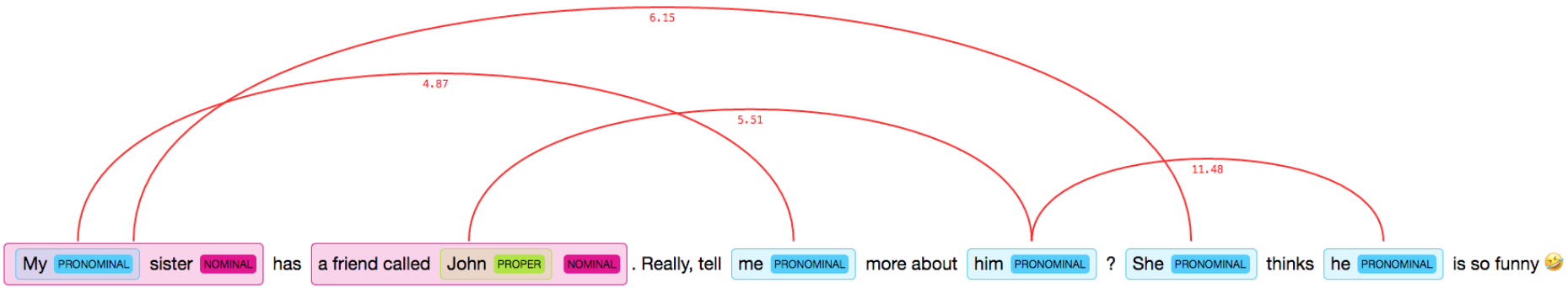
Eksempel problem

En form – flere meninger

- Mine mål er egentlig ganske forskjellige
  - uttalt av en fotballspiller
  - uttalt av en modell som sammenligner seg med Kate Moss
  - 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
  - I saw her duck
  - Krasjet med rådyr på moped (Agderposten)

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:

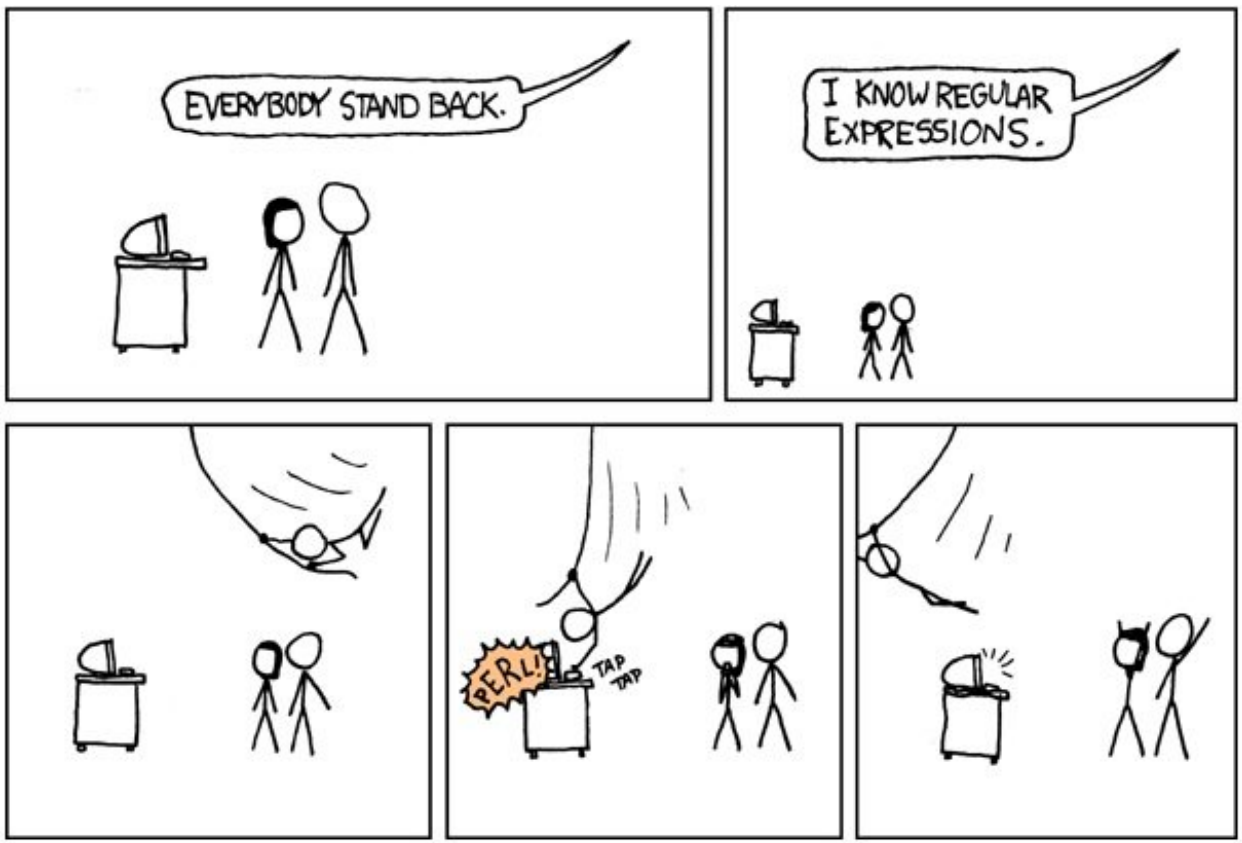
- gitt et AI problem og en masse data om verden (eksempler): la datamaskinen finne riktig svar
- unngår hardkoding av svarene

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

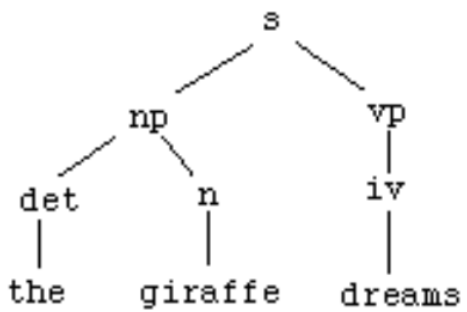
```
s    ⟶ np vp

np   ⟶ det n
vp   ⟶ tv np
     ⟶ iv

det  ⟶ the
     ⟶ a
     ⟶ an

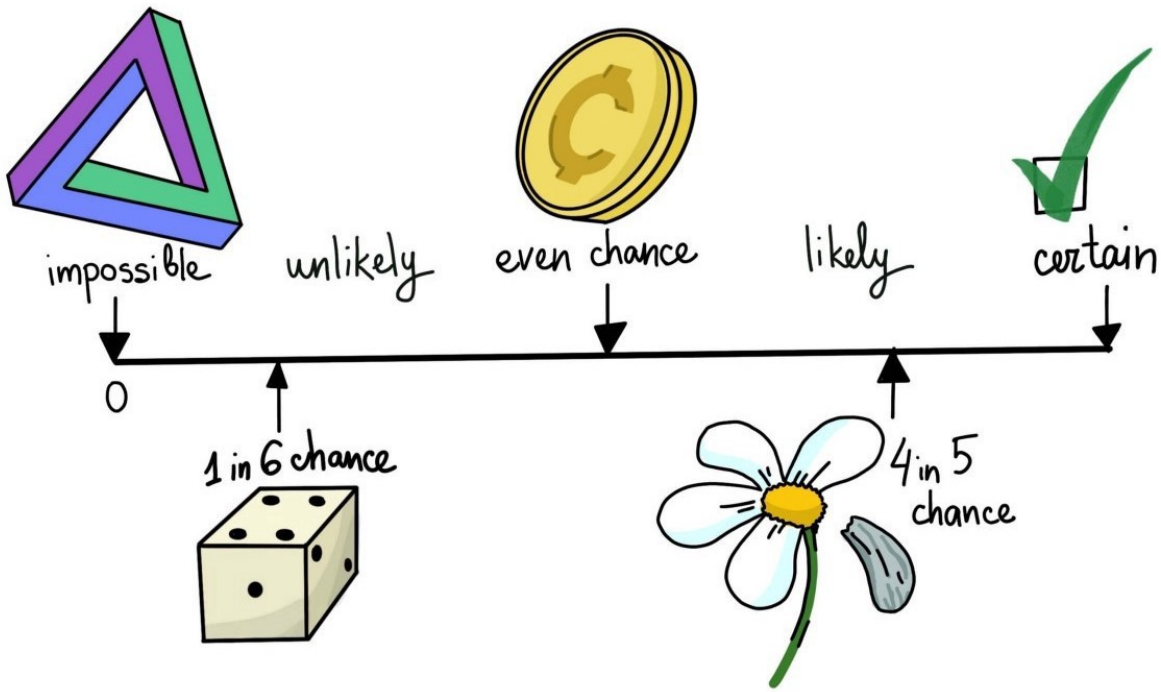
n    ⟶ giraffe
     ⟶ apple

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: Tekst i Python