

IN1140: Introduksjon til språkteknologi

Forelesning #12

Lilja Øvrelid

Universitetet i Oslo

14 november 2019





I dag

- ▶ Repetisjon
- ▶ Digital prøveeksamen
- ▶ Obligkonkurransen

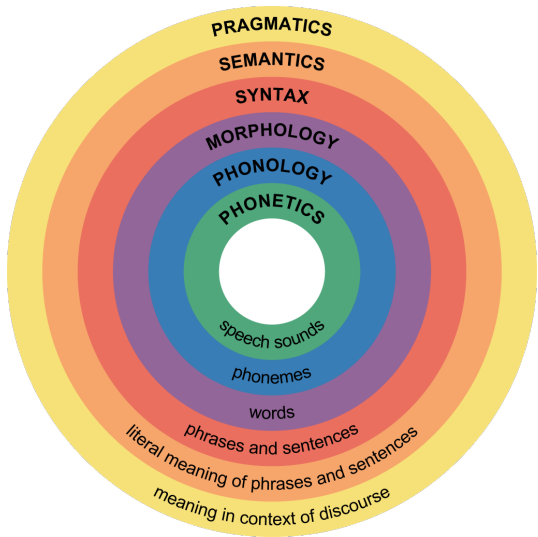
Neste uke

- ▶ Gjennomgang av eksamensoppgave



Emnet gir en innføring i språkteknologi: metoder for automatisk analyse av språklige data. Det vil videre gi en innføring i *lingvistisk teori* og relatere denne til *språkteknologiske problemområder*. Det vil gis et første møte med noen språkteknologiske oppgaver som eksempelvis *tokenisering*, *n-grammodeller*, *tagging* og *klassifisering*.

Lingvistikk



- ▶ Vitenskapelige studiet av språk
- ▶ Vitenskapelig? Systematisk studie av regler, systemer og prinsipper i menneskelige språk
- ▶ Kunnskapen om enheter og regler i et språk

Nivåer

- ▶ Fonologi: lyder \Rightarrow ord
- ▶ **Morfologi**: morfemer \Rightarrow ord
- ▶ **Syntaks**: ord \Rightarrow fraser, fraser \Rightarrow setninger
- ▶ **Semantikk**: ord \Rightarrow mening, setninger \Rightarrow mening



- ▶ De fleste språkteknologiske applikasjoner må håndtere **flertydighet** (“ambiguity”)
- ▶ Kjennetegner naturlige språk, på alle nivåer
 - ▶ *I saw her duck*
 - ▶ *Krasjet med rådyr på moped* (Agderposten)



- ▶ Menneskelig språkprosessering:
 - ▶ afasistudier, hjernescanning
- ▶ Språkteknologi:
 - ▶ **korpusdata** helt sentralt (**representativitet**)



Handler om **ord**

- ▶ Hvordan ord er bygd opp (morfemer)
- ▶ Hvordan nye ord **dannes** (avledning, sammensetning)
- ▶ Hvordan ord **bøyes**



- ▶ Morfemet er den elementære (minste) lingvistiske enheten
- ▶ To hovedtyper:
 - ▶ **Frie** morfemer: ord. *boy, desire, gentle, man*
 - ▶ **Bundne** morfemer: affikser.
 - ▶ prefikser: *un-, pre-, bi-*
 - ▶ suffikser: *-ing, -ish, -ness*
- ▶ Morfologisk komplekse ord består av:
 - ▶ **Rot** + en eller flere affikser (*hus+lig*)
 - ▶ En rot er et ordelement som ikke kan deles opp i mindre (meningsbærende) deler

- ▶ En avledning er et ord som er dannet fra et annet ord ved hjelp av et avledningsaffiks (prefiks eller suffiks),
 - ▶ Avledningsbasen kan være et rotord (*barn*) eller en avledning (*barnslig*)
 - ▶ Avledningsaffiksene har ofte et klart **semantisk** innhold
- ▶ *u-* - negasjon: *umulig, uvel, urolig*
 - ▶ *for-* - foran: *forelese, forbokstav, formann*
 - ▶ *-er* - den som utfører handlingen: *fisker, baker*

- ▶ Markerer kategorier som tid (tempus), tall (numerus), bestemthet, etc.
- ▶ Noen eksempler:

- ▶ **Tid (tempus)**: angir tidspunktet for handlingen. Presens (nåtid) og preteritum (fortid) *liker-likte, likes-liked*
- ▶ **Tall**: entall og flertall *bil-biler, car-cars*
- ▶ **Bestemthet**: uttrykkes i hovedsak ved suffiks (*bilen, huset*)



- ▶ Bindeledd mellom ordet og setningen (syntaks):
 - ▶ Sier noe om hva slags kontekster et ord forekommer i
 - ▶ Sier noe om uttale (*record*, *content*)
- ▶ Viktig i en rekke språkteknologiske oppgaver:
 - ▶ Talesyntese
 - ▶ “Chunking”, syntaktisk parsing
 - ▶ Informasjonsekstraksjon
- ▶ Vi trenger **kriterier** for ordklasseinndeling

Tre slags kriterier:

1. **Formelle** (morfologiske) kriterier
 - ▶ hvilke bøyningsformer har ordet?
 - ▶ *hare - haren* og *redd - reddere*
 - ▶ **harere* og **redde*
2. **Funksjonelle** (syntaktiske) kriterier
 - ▶ hvordan kan ordet kombineres med andre ord?
 - ▶ *en redd hare* og *redd for ilden*
 - ▶ **en redd* og **en hare for ilden*
3. **Betydningsmessige** (semantiske) kriterier
 - ▶ hva er typiske betydninger hos ord i ordklassen?
 - ▶ *hare* - dyr, levende vesen
 - ▶ *redd* - egenskap

åpne vs. lukkede ordklasser

- ▶ nye medlemmer?
- ▶ **åpne**: substantiv, verb og adjektiv inneholder mange tusen ord, kan enkelt fylle på med nye
- ▶ **lukkede**: inneholder mange færre ord enn de åpne kan ikke fritt skape nye ord gjennom orddannelse

innholdsord vs. funksjonsord

- ▶ semantisk innhold?
- ▶ **innholdsord**: substantiv, verb, adjektiv rikt betydningsinnhold, refererer utenfor språket
- ▶ **funksjonsord**: mer allment betydningsinnhold, refererer ikke utenfor språket. Finns fremst i de lukkede ordklassene.
- ▶ Ikke helt én-til-én, feks hjelpeverb.

“Studiet av hvordan setninger bygges opp av ord og ordkombinasjoner”



- ▶ **Konstituenter** – grupperinger av ord i en setning, fungerer som en enhet

Lingvistiske tester:

- ▶ “stå alene”-testen
- ▶ “erstattes med pronomen”
- ▶ “Flyttes som enhet”

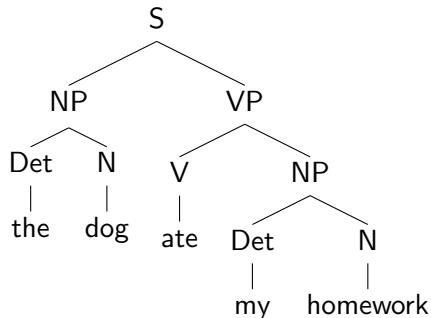
▶ Norsk eksempel: *Den lille hunden lekte **i hagen***

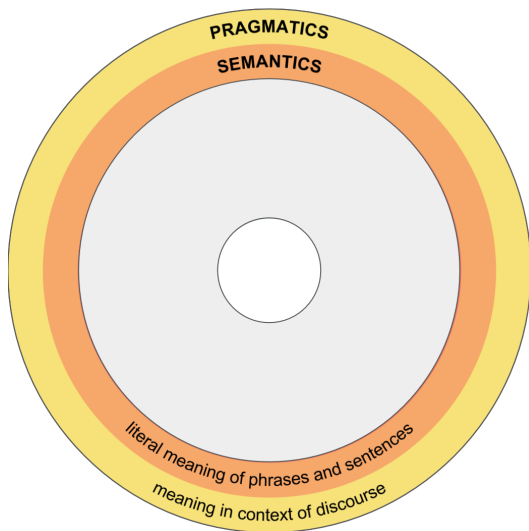
- ▶ (*Hvor lekte hunden?*) *I hagen* (stå alene)
- ▶ *Hunden lekte **der*** (erstattes med pronomen)
- ▶ *I hagen lekte hunden* (flytter som enhet)



- ▶ **Fraser** - bygger opp setningen eller andre fraser og navngis etter hodet
 - ▶ NP (noun phrase), f.eks. *det pene huset*
 - ▶ VP (verb phrase) f.eks. *liker fotball*
 - ▶ PP (prepositional phrase), f.eks. *i skogen*
etc.

- ▶ Frasale kategorier: NP, VP, AdjP, PP
- ▶ Leksikale kategorier (ordklasser): N, V, P, Adj, Adv
- ▶ Frasestrukturtre (Phrase Structure (PS) tree)





- ▶ Studiet av betydning slik det uttrykkes gjennom språk
- ▶ Betydning til morfemer, ord, fraser og setninger
 - ▶ Leksikal semantikk
 - ▶ Setningssemantikk
 - ▶ (Pragmatikk: hvordan konteksten påvirker betydning)

- ▶ Representerer betydningen til **ord**
- ▶ Vise hvordan betydningene er relatert

Leksikale relasjoner:

- ▶ **Homonymi** (not, knot) og **polysemi** (flere betydninger, men betydningene er relatert)
- ▶ **Synonymi**
- ▶ **Antonymi** (enkle, gradérbare, taksonomiske søstre ...)
- ▶ **Hyponymi** (hund, katt er hyponymer av dyr)



- ▶ Beregner **sannhetsverdien** for setninger basert på betydningen til mindre deler (ord, fraser)
- ▶ Semantisk analyse ved oversettelse fra engelsk (eller norsk eller ...) til et universelt metaspråk (førsteordenslogikk)



- ▶ Semantiske roller beskriver den semantiske relasjonen som argumenter har til handlingen beskrevet av verbet
- ▶ *Nina* hevet *bilen* med *jekken*
 - ▶ *Nina* – deltageren som er ansvarlig for å utføre handlingen beskrevet av verbet
 - ▶ *bilen* – blir påvirket av handlingen
 - ▶ *jekken* – middelet som Nina bruker til å utføre handlingen

Formelle modeller



- ▶ Hentet fra matematikk, statistikk og (generell) informatikk
- ▶ Representere lingvistisk kunnskap
 - ▶ Regulære uttrykk
 - ▶ Formelle regelsystemer
 - ▶ Statistiske modeller



- ▶ Et regulært uttrykk er en beskrivelse av en mengde strenger
- ▶ Brukes til tekstsøk

Regulære uttrykk består av

- ▶ **Strenger** bestående av tegn: b, INF1820, informatikk
- ▶ **Disjunksjon**: penge(r|ne), [Dd]en, [A-Z], [0-9]
- ▶ **Negasjon**: [^b] [^A-Z0-9]
- ▶ **Tellere**: Kleenes *, +, ?
 - ▶ opsjonalitet (0 eller 1): ? woodchucks?
 - ▶ hvilket som helst antall (0 eller flere): Kleenes * baaa! [0-9][0-9]*
 - ▶ minst en: +: baa+! [0-9]+ kroner
- ▶ **"wildcard"** for et hvilket som helst tegn: .
beg.n beltedyr.*beltedyr

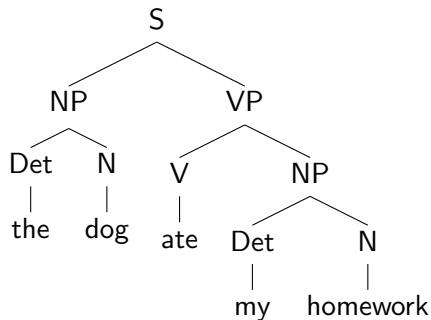


- ▶ Formell modell som fanger inn **syntaktisk** konstituentstatus og rekkefølge
- ▶ Formelt: en CFG er en 4-tupel $\langle N, \Sigma, P, S \rangle$, der
 - ▶ N er en mengde **ikke-terminale** symboler (syntaktiske kategorier)
 - ▶ Σ er en mengde **terminale** symboler (ord)
 - ▶ R er en mengde **regler** (produksjoner) på formen $A \rightarrow \alpha$, der
 - ▶ A er en ikke-terminal
 - ▶ α er en streng av symboler hentet fra mengden $(\Sigma \cup N)^*$, dvs både terminaler og ikke-terminaler
 - ▶ Et særskilt startsymbol S

Eksempel CFG

- ▶ La $G = \langle N, \Sigma, R, S \rangle$ der
 - ▶ $N = \{S, NP, VP, DT, N, V, N\}$
 - ▶ $\Sigma = \{the, my, dog, homework, ate\}$
 - ▶ $R = \{S \rightarrow NP VP,$
 $NP \rightarrow Det N,$
 $VP \rightarrow V NP,$
 $DT \rightarrow the, my,$
 $N \rightarrow dog, homework,$
 $V \rightarrow ate,$
 $\}$
 - ▶ $S = S$

- ▶ En rekke applikasjoner av reglene kan visualiseres som **trær**





Direkte rekursjon:

NP \rightarrow NP PP

VP \rightarrow VP PP

Indirekte rekursjon:

S \rightarrow NP VP

VP \rightarrow V CP

CP \rightarrow C S

said that the flight was late

- ▶ Lar oss kvantifisere **usikkerhet**.
- ▶ Uttrykk for sjanse eller odds.
- ▶ **Sannsynlighet** for en hendelse: en verdi mellom 0 og 1, gitt ved P :
 - ▶ P (seks på terningen)
 - ▶ Betinget sannsynlighet:
 $P(\text{sol i morgen} \mid \text{regn i dag})$





- ▶ **Felles** sannsynlighet (“joint probability”)
 - ▶ $P(A, B)$: sannsynligheten for at både A og B skjer.
 - ▶ Skrives også: $P(A \cap B)$.
- ▶ **Betinget** sannsynlighet (“conditional probability”)
 - ▶ Vi har ofte *delvis kunnskap* om en hendelse: $P(A|B)$

- ▶ En **språkmodell** tilskriver sannsynligheter $P(x)$ til alle ordsekvenser x i et språk L .
- ▶ Gitt en sekvens $w_1, w_2 \dots w_k = w_1^k$ så ønsker vi å estimere den felles sannsynligheten for ordene $P(w_1^k)$.
- ▶ F.eks: $P(\text{jeg, vil, drikke, kaffe, nå})$

1. Vi bruker **produktsetningen**:

$$P(w_1 \dots w_k) = P(w_1) P(w_2|w_1) P(w_3|w_1, w_2) \dots P(w_k|w_1^{k-1})$$

2. Vi forenkler med **Markovantagelsen**:

De $n - 1$ siste elementene lar oss tilnærme effekten av å betrakte hele sekvensen.

Eksempel for $n = 2$:

$$P(w_1^k) \approx \prod_{i=1}^k P(w_i|w_{i-1})$$



- ▶ Et n -gram er en delsekvens på n ord:
- ▶ n -grammer i *jeg vil drikke kaffe nå*:

- ▶ **unigrammer** ($n = 1$): $\langle \text{jeg} \rangle$, $\langle \text{vil} \rangle$, $\langle \text{drikke} \rangle$, $\langle \text{kaffe} \rangle$, $\langle \text{nå} \rangle$
- ▶ **bigrammer** ($n = 2$): $\langle \text{jeg, vil} \rangle$, $\langle \text{vil, drikke} \rangle$, $\langle \text{drikke, kaffe} \rangle$, $\langle \text{kaffe, nå} \rangle$
- ▶ **trigrammer** ($n = 3$): $\langle \text{jeg, vil, drikke} \rangle$, $\langle \text{vil, drikke, kaffe} \rangle$, $\langle \text{drikke, kaffe, nå} \rangle$
- ▶ **4-grammer** ($n = 4$): $\langle \text{jeg, vil, drikke, kaffe} \rangle$, $\langle \text{vil, drikke, kaffe, nå} \rangle$



- ▶ Vi estimerer P ved å telle n -grammer og se på relativ frekvens:

$$P(w_i | w_{i-n+1}, \dots, w_{i-1}) = \frac{\text{Count}(w_{i-n+1}, \dots, w_i)}{\text{Count}(w_{i-n+1}, \dots, w_{i-1})}$$

- ▶ F.eks, for bigrammer:

$$P(\text{kaffe} | \text{drikke}) = \frac{\text{Count}(\text{drikke}, \text{kaffe})}{\text{Count}(\text{drikke})}$$

- ▶ Kalles **Maximum Likelihood Estimation** (MLE).
- ▶ I praksis legger vi gjerne også til markører for start og slutt for sekvensen: $\langle s \rangle$ og $\langle /s \rangle$.



$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- ▶ Bayes regel/teorem viser hvordan vi kan beregne inverse sannsynligheter, dvs dersom vi vet $P(A|B)$, hvordan kan vi beregne $P(B|A)$?
- ▶ Brukes i Naive Bayes klassifisering, gjør enklere å estimere sannsynligheter fra et korpus

Statistisk klassifisering

- ▶ Sentral metode innenfor maskinlæring
- ▶ Automatisk avgjøre hvilken kategori en observasjon tilhører
- ▶ Basert på **treningsdata**: observasjoner der kategorien er kjent
 - ▶ e-post \rightarrow {spam, ikke-spam}
 - ▶ dokument \rightarrow {positiv, negativ}
 - ▶ ord \rightarrow {betydning₁, betydning₂}
- ▶ **veiledet** ('supervised') klassifisering: klassifisering som benytter treningsdata
- ▶ Evaluerer på usette **testdata**



- ▶ En enkel men effektiv **statistisk klassifiserer**

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c) \prod_{j=1}^n P(f_j | c)$$

- ▶ Vi **trener** klassifisereren ved å beregne sannsynligheter fra et korpus (MLE)

2 sannsynligheter:

1. prior-sannsynligheten for **klassen** $P(c)$
2. sannsynligheten for **individuelle trekk** $P(f_j | c)$



► To sannsynligheter:

1. **prior**-sannsynligheten for klassen $P(c)$

$$P(c) = \frac{N_c}{N_{doc}}$$

N_c = antall dokumenter i treningsdataen som er i klassen c

N_{doc} = total antall dokumenter

2. sannsynligheten for **individuelle trekk** $P(f_j|c)$

$$P(f_j|c) = \frac{\text{count}(f_j, c)}{\text{count}(c)}$$



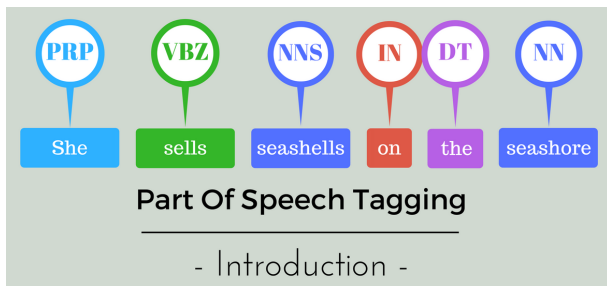
Språkteknologiske Oppgaver & Applikasjoner

- ▶ Dele opp en tekst i løpende ord
- ▶ Første skritt i nesten alle språkteknologiske oppgaver

Problematiske tilfeller:

- ▶ Forkortelser: *f.eks.*
- ▶ Bindestrek: *Oslo-borgeren*
- ▶ Mellomrom: *New York*
- ▶ URL'er
- ▶ ...
- ▶ Tokens og typer

- ▶ Input: streng av ord og en spesifisert mengde tagger (taggsett)
- ▶ Output: en tagg per ord



- ▶ **Flertydighet**
- ▶ **Ekstrinsisk** og **intrinsisk** evaluering

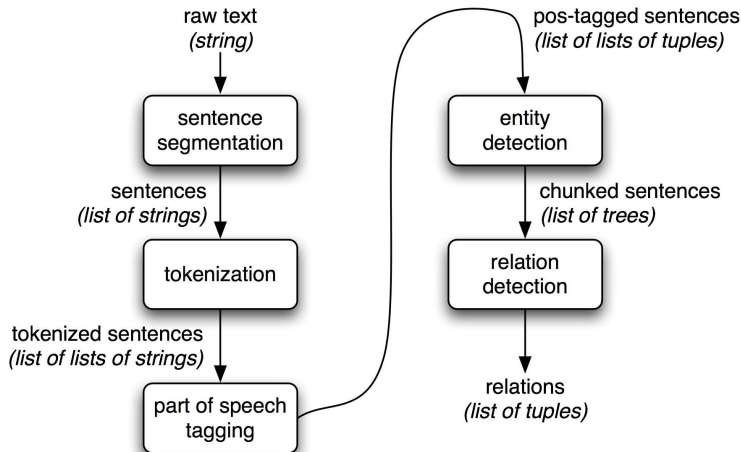


- ▶ **Chunking:**
 - ▶ Dele setningen inn i en sekvens “**chunks**”
 - ▶ En chunk inneholder et **hode**, muligens med noen funksjonsord/modifikatorer først
[walk] [straight past] [the lake]
- ▶ Forenklete konstituenter (“fram til hodet”)
- ▶ Ikke komplett syntaktisk beskrivelse, men tilstrekkelig for mange applikasjoner
- ▶ **Ikke-rekursive:** en chunk kan ikke inneholde en chunk av samme kategori



- ▶ Gitt en setning med et spesifikt ord (“target word”) og en liste med betydninger (f.eks. fra WordNet), angi den betydningen som passer best for målordet i den setningen
- ▶ Klassifisering fra annotert datasett: **veiledet** klassifisering
 - ▶ Hente ut trekk som er sentrale for disambiguering, f.eks. ord i konteksten, ordklasser, lemma, etc.
- ▶ Naive Bayes-klassifisering

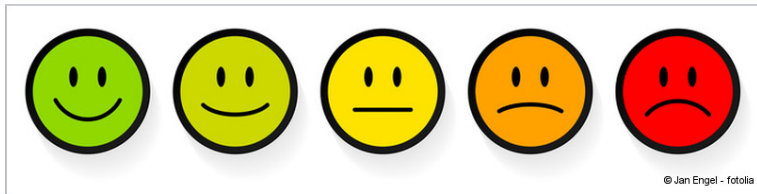
Kombinerer mange språkteknologiske moduler:





Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY \$6] per round trip on flights to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim Wagner] said. [ORG United], a unit of [ORG UAL Corp]. said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].

- ▶ BIO-klassifisering



[https:](https://www.clickworker.de/wp-content/uploads/2017/02/Sentiment-Analyse.jpg)

[//www.clickworker.de/wp-content/uploads/2017/02/Sentiment-Analyse.jpg](https://www.clickworker.de/wp-content/uploads/2017/02/Sentiment-Analyse.jpg)

- ▶ På dokument/setningsnivå
- ▶ Klassifisering (f.eks. Naive Bayes)
- ▶ Leksikon (manuelt/automatisk)

A graphic featuring a series of concentric, glowing circles in shades of red and orange, creating a spiral effect that draws the eye towards the center. The text 'That's all Folks!' is written in a white, cursive script across the middle of the spiral.

That's all Folks!



- ▶ Tenk eksamen
 - ▶ pensum: les **og** jobb aktivt med stoffet
 - ▶ gruppeoppgaver, teoretiske oblig-oppgaver, tidligere eksamensoppgaver
 - ▶ få hjelp: gruppetimer, medstudenter, digitalt
- ▶ Hva funker for deg? (Skrive, snakke og/eller lytte? Gjerne en kombinasjon)
- ▶ Møt opp neste uke!