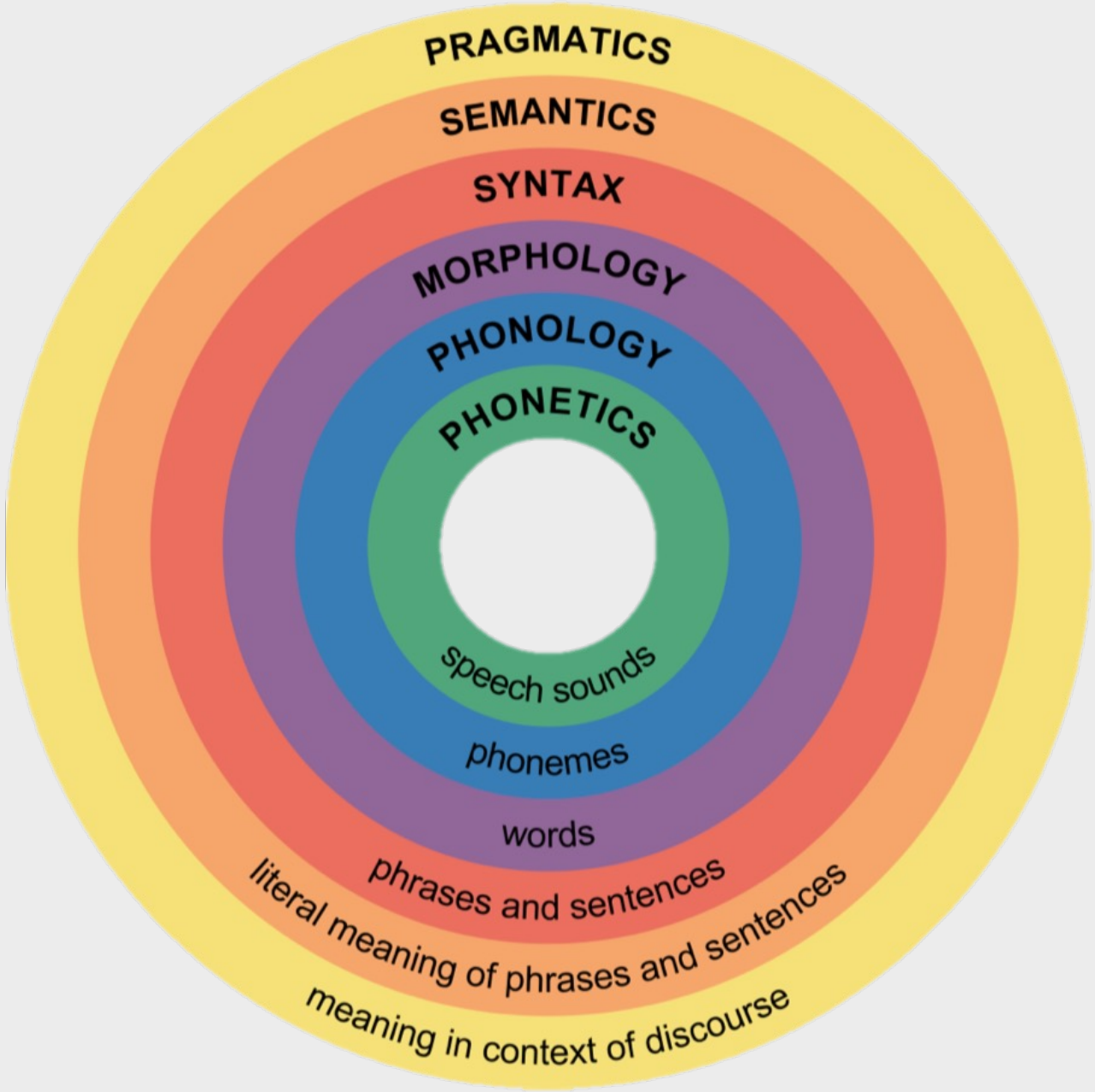
IN1140

Repetisjon // eksamensoppgaver

Annika Willoch Olstad

semantikk syntaktiske trær * morfemer avledning syntaks
fonetikk flertydighet regex VP innholdsord AdvP funksjonsord
lemmatisering stemming korpus naive bayes AdjP betinget
sannsy lerende språk typologi suffiks lingvistikk
sta språkmodell affiks sentimentanalyse
p sparseness n-gram regelbasert tagger
Språkteknologi morfologi
nser maximum likelihood estimation
features grammatikk agent NP
siv maskinlæring informasjonsekstraksjon
setningssemantikk leksikal semantikk

2 Lingvistikk



3 Morfologi

- Morfologi – læren om ord og ordenes oppbygning
- Innholdsord vs. funksjonsord
- Morfemet – minste meningsbærende enhet
  - Frie morfemer vs. bundne morfemer
    - Affikser: prefikser, suffikser, (infikser, sirkumfikser)
- 3 typer orddannelse:
  1. Avledning – affikser, ny ordklasse
  2. Bøyning – affikser, markerer genus, tempus, kasus, etc.
  3. Sammensetning

Oppgave 2 Morfologi (vekt 25%)

1. Vi skiller mellom såkalte åpne og lukkede ordklasser. Hva skiller de to typene ordklasse fra hverandre?

2. Av de ti ordklassene vi opererer med på norsk, hvilke regner vi som åpne?

4 Regulære uttrykk

Tegn	Funksjon
?	0 eller 1
*	0 eller flere
+	1 eller flere (minst en)
.	wildcard, fanger alt.
[^...]	negasjon, eks. [^b] = alt utenom 'b'
\spesialtegn	selve tegnet, ikke dets funksjon.
{n}	n forekomster (skal gjentas n ganger)
{n,m}	fra n til m forekomster (må gjentas minst n ganger, max m ganger)
{n,}	minst n forekomster
|	disjunksjon 'eller'
[Aa]	disjunksjon 'eller' – her: enten 'A' eller 'a'
[...-...]	ranges/rekker: f.o.m., t.o.m.
[A-Z]	alle store bokstaver fra og med A til og med Z
[a-z]	alle små bokstaver fra og med a til og med z
[aA-zZ] [A-Za-z]	alle store og små bokstaver fra og med A/a til Z/z.
[0-9]	tall (mellom 0 og 9)
\d	= [0-9] = digit, tall
\D	non-digit, alt som ikke er \d
\w	word, store/små bokstaver og tall
\W	non-word, alt som ikke er \w
\s	alt 'tomrom': \n \r \t
\S	non-\s, alt som ikke er \s
\n	new line
\t	tab
\b	word boundary, på slutten/starten
\B	non-boundary, alt som ikke er på 'kanten' av et ord
$	på slutten av linjen
^	(utenfor []) begynnelsen av linjen

5 Sannsynlighet Språkmodeller

- Statistisk språkmodell: bruker korpusfrekvenser for å beregne sannsynligheter for ordsekvenser
- N-grammodell: delsekvenser på n ord (n-gram)

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

- Bigram:

$$P(w_n|w_{n-1}) = \frac{count(w_{n-1}, w_n)}{count(w_{n-1})}$$

- «Data sparseness»-problemet → smoothing

$$P(w_n|w_{n-1}) = \frac{count(w_{n-1}, w_n) + 1}{count(w_1) + V}, \quad \text{der V er antall ordtyper}$$

6 Ordklasser
POS-tagging

- Ordklasser: hva slags kontekst forekommer ordet i?
- 3 kriterier:
  1. Formelle/morfologiske kriterier
  2. Funksjonelle/syntaktiske kriterier
  3. (Semantiske/betydningsmessige kriterier)
- Åpen klasse vs. lukket klasse
- Ordklassetaggere – 2 hovedkategorier:
  1. Regelbasert tagger – database med regler
  2. Statistisk tagger – treningskorpus → sannsynlighet

! Oppgave

Tabell 1

CC	konjunksjon
DET	determinativ
JJ	adjektiv
NN	substantiv
PR	preposisjon
PO	pronomen
RB	adverb
SB	subjunksjon
VB	verb

Gitt ordklassene i tabell 1 til venstre, tildel ordklasser til alle ordene i setningen:

«Jeg ser mannen med kikkerten»

Oppgave

Tabell 1

CC	konjunksjon
DET	determinativ
JJ	adjektiv
NN	substantiv
PR	preposisjon
PO	pronomen
RB	adverb
SB	subjunksjon
VB	verb

PO VB NN PR NN

Jeg ser mannen med kikkerten

7 Syntaks

- Syntaks - studiet av prinsipper og regler for setningsdannelse
- Skiller mellom:
  1. Syntaktisk form – hvordan en språklig enhet ser ut
  2. Syntaktisk funksjon – rollen enheten spiller ift. andre
- Konstituenter – grupperinger av ord
  - Tester:
    - «Stå alene»
    - «Erstattes med pronomen»
    - «Flyttes som enhet»
- Fraser: (adledd) hode (adledd)
  - Leksikale vs. frasale kategorier
- Strukturell vs. leksikal flertydighet

7 Kontekstfrie grammatikker

- CFG – formell modell: konstituentstatus og rekkefølge
- Formelt: 4-tuppel $< N, \Sigma, R, S >$, der:
  - N = {ikke-terminale symboler}
  - Σ = {terminale symboler}
  - R = {regler på formen A → B}
  - S = startsymbol
- Rekursivitet – uendelige strukturer
  - XP → XP ...
- Chunking – «fattigmannssyntaks» –
  - Ikke-overlappende ordsekvenser
  - Ikke-rekursiv (XP kan ikke inneholde XP)

! Oppgave

Under ser du en liten kontekstfri grammatikk for norsk:

```
S -> NP VP
VP -> V
VP -> V NP
NP -> julenissen, snekkeren, ungene, kjelken
V -> danset, snekret, så, trakk
```

1) Grammatikken er langt fra noen komplett grammatikk for norsk. Gi ett eksempel på en grammatisk, norsk setning som ikke gis noen analyse av denne grammatikken, samt ett eksempel på en ugrammatisk setning som gis en analyse.

2) Last opp det syntaktiske treet som grammatikken tildeler den ugrammatiske setningen fra forrige oppgave.

8 Naive Bayes

1. Hovedantagelse: $\hat{k} = \underset{k \in K}{argmax}\, P(k|d)$

2. Betinget sannsynlighet: $P(k|d) = \frac{P(k \cap d)}{P(d)}$ («og»)

3. Produktsetningen: (sier at: $P(A \cap B) = P(A|B)P(B) = P(B \cap A) = P(B|A)P(A)$)

$$P(k \cap d) = P(k|d)P(d) = P(d|k)P(k)$$ (symmetrisk)

4. Bayes teorem: $P(k|d) = \frac{P(k \cap d)}{P(d)} = \frac{P(d|k)P(k)}{P(d)}$

5. Fjern trekkvektor: $P(k|d) = \frac{P(d|k)P(k)}{P(d)} = P(d|k)P(k)$ (fordi konstant)

6. Uavhengighetsantagelse: $P(d|k) = \prod_{j=1}^{n} P(d_j|k)$ (n = antall trekk, j = økningsverdi)

Hvert trekk uavhengig → multiplikasjonsregelen

$$P(k|d) = P(d|k)P(k)$$

$$P(k|d) = \prod_{j=1}^{n} P(d_j|k)P(k)$$

7. Fullfør: $\hat{k} = \underset{k \in K}{argmax}\, P(k) \cdot \prod_{j=1}^{n} P(d_j|k)$

argmax fordi vi skal ha argumentet, ikke P

9 Semantikk (pragmatikk)

- Semantikk – studiet av betydning uttrykt i språk
- Leksikal semantikk vs. setningssemantikk
- Pragmatikk – studiet av språklig forståelse i kontekst

- Leksikale relasjoner – hvordan betydninger relateres:
  - Homonymi – urelatert homofon/-graf
    - Ex.: «jul», «hjul»
  - Polysemi – relatert homofon/-graf
    - Ex.: «vill», «vill»
  - Synonymi – lik betydning, ulikhet ortografisk
    - Ex.: «gøy», «morsomt»
  - Antonymi – motsatt betydning (alt likt, unntatt *noe)*
    - *Ex.: «treffe», «bomme»*
  - Hyponymi – hyponymet har betydningen til et mer generelt ord (hypernym)
    - Ex.: «hund» er hyponym av «dyr»
  - Meronymi – relasjon mellom del og helhet
    - Ex.: «finger» er en del av «hånd»

9 Semantikk (pragmatikk)

- Semantiske roller – *«å forstå en setningsbetydning er å vite hvilke roller de ulike deltagerne innehar»:*
  - Agent – handler med viten og vilje
  - Patient – påvirkes av handlingen
  - Theme – beveges av handlingen
  - Experiencer – bevisst på, men ikke i kontroll av handlingen
  - Beneficiary – den som handlingen utføres for
  - Instrument – middelet for å utføre handlingen
  - Goal – den som noe beveger seg mot
  - Source – den som noe beveger seg fra

Oppgave 4 **Semantikk** (vekt 25%)

1. Vi sier at *Erna Solberg* og *statsministeren* har lik referanse, men forkjellig betydning. Hva mener vi med dette?

2. Forklar kort hva en semantisk rolle er, og hvordan vi bruker semantiske roller til å analysere setninger.

3. Analyser setningen Den søte babyen kaster grøt. ved hjelp av semantiske roller. Beskriv de forskjellige rollene du har tilordnet.

1) Referanse og betydning er to ulike aspekter ved mening. Der referanse peker ut objekter i verden, representerer betydning et mer varig meningsaspekt. F.eks. i 2016 er Erna Solberg statsminister, ordene “Erna Solberg” og “statsministeren” referer til det samme, men statsminister inneholder og mer informasjon, f.eks. landets politiske overhode, etc. som er uavhengig av referanse.

2) Semantiske roller beskriver de ulike rollene deltagerne i handlingen som beskrives av hovedverbet innehar. Det er i hovedsak verbets argumenter som analyseres og tildeles en rolle og dette gjøres på konstituentnivå.

3) *Den søte babyen* – **AGENT**
grøt – **THEME**
AGENT-rollen brukes for å beskrive deltageren som utfører handlingen beskrevet av verbet med viten og vilje.
THEME-rollen brukes for å beskrive deltageren som påvirkes av handlingen og som forflyttes som følge av at handlingen finner sted.

10 Semantikk i språkteknologi

- **Betydningsdisambiguering(WSD)** – angi korrekt betydning for et målord i en gitt setning
  - *Naive Bayes (!)*
    - c (klassen) → s (betydningen)
- **Named entity recognition (NER)** – automatisk gjenkjenning og kategorisering av egennavn:
  - BIO-klassifisering: **B**egynnelsen, **I**nnenfor, **O**utside et egennavn, + kategori av egennavn
  - Ser på trekk: ordklasse, affikser, stor/liten, chunk, ++
- **Semantic Role Labeling** – gitt en setning, finn predikatets semantiske roller

11 Språkteknologiske applikasjoner

- **Informasjonsekstraksjon(IE)** – hente ut semantisk informasjon
  - Start med rå-tekst → del teksten i setninger → tokeniser → POS-tagging → entity-detection → chunking for å finne relasjoner...
    - *Named Entity Recognition(NER)*
    - *Relasjonsekstraksjon* – avgjør:
      - Er to entiteter i en relasjon?
      - ... Hvilken?
    - *(Temporal Expression Extraction)*
    - *(Coreference resolution)*
    - *(Event extraction)*

¿ Noe mer?

- Noe som er vanskelig å forstå?
- Noe som er vanskelig å huske?
- Noe som er lett i teori, men vanskelig i praksis?
- Noe som «aldri egentlig» ble forklart?
- Noe som ikke henger på grep?
- Noe rart?
- Noe spennende?