

IN1140: Introduksjon til språkteknologi

Forelesning #7

Lilja Øvrelid

Universitetet i Oslo

28 september 2020

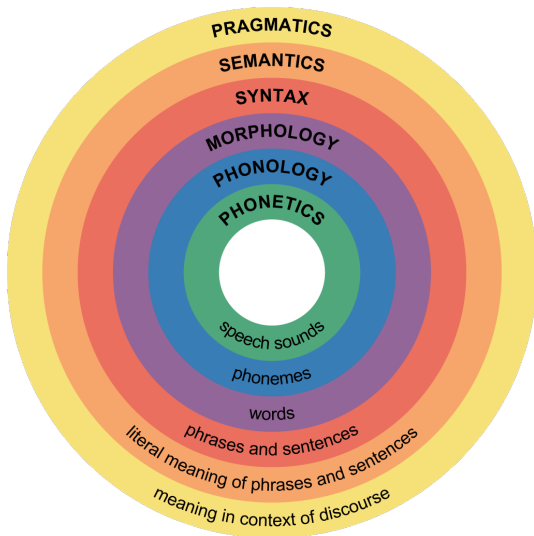


Forrige uke

- ▶ Ordklasser
- ▶ Ordklassetagging
- ▶ Oblig1b: regulære uttrykk (frist: i morgen)
- ▶ Oblig2a: språkmodeller og ordklassetagging (frist: 16/10)

I dag

- ▶ Syntaks
- ▶ Kontekstfrie grammatikker
- ▶ Midtveisevaluering



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

ORD – FRASER – SETNINGER

Fraser

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

Setninger

- ▶ Inneholder et verb og (som regel) et subjekt
- ▶ *Spis!*
- ▶ *Bea lukket vinduet fort.*

- ▶ Flere mulige grupperinger av ord
 - ▶ Ulik gruppering gir ulik betydning
-
- ▶ *Gamle menn og kvinner kan forlate skipet*
 - ▶ [*Gamle menn*] og [*kvinner*]
 - ▶ [*Gamle* [*menn og kvinner*]]

- ▶ **Konstituenten** – grupperinger av ord i en setning, fungerer som en enhet
 - ▶ The dog ate my homework
 - ▶ The dog ate my homework
- ▶ Hvordan kan vi avgjøre konstituentstatus?
 - ▶ Lingvistiske tester

The dog ate my homework

“stå alene”-testen:

- ▶ *What did the dog eat?*
- ▶ *my homework*
- ▶ **ate my*

“erstattes med pronomen”-testen:

- ▶ *The dog ate it*
- ▶ *The dog ate my homework and the cat did too*

“Flyttes som enhet”-testen:

- ▶ *It was my homework that the dog ate.*
- ▶ *My homework was eaten by the dog.*

- ▶ 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)

- ▶ Et enkelt ord kan bygges ut til en gruppe ord, slik at den nye gruppen har samme funksjon i setningen (en konstituent)
 - ▶ Kan substitueres for hverandre
- ▶ *The dog ate the cake*
 - ▶ *The dog ate the birthday cake*
 - ▶ *The dog ate the delicious birthday cake*
 - ▶ *The dog ate the delicious birthday cake that was meant for Bea*
- ▶ Fraser:
 - ▶ (adledd) **hode** (adledd: utfylling)
 - ▶ (*the delicious birthday*) **cake** (*that was meant for Bea*)

- ▶ Hodet er et substantiv
- ▶ Fungerer typisk som subjekt eller objekt i setningen

Eksempler:

- ▶ determinativ + substantiv: *the dog, en hund*
- ▶ egennavn: *Barack Obama, Japan*
- ▶ pronomen: *he, they, han, henne*
- ▶ Hodet bestemmer **kongruens** – feks tall, kjønn, bestemthet

- ▶ Hodet er en preposisjon
- ▶ Etterfølges av en NP-utfylling (preposisjonsobjekt)

Eksempler:

- ▶ prep + NP *in the garden, over the rooftops*
- ▶ foranstilt adledd (Adj) *dypest ned i skuffen*

- ▶ Hodet er et adjektiv
- ▶ Kan ha foranstilt adledd: adverb eller andre adjektiv
- ▶ Noen adjektiv tar etterstilt utfylling (som verb)

Eksempler:

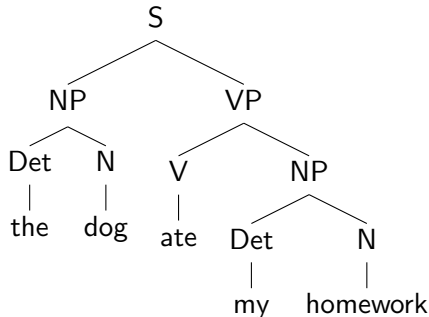
- ▶ Adj: *crazy, red, brilliant*
- ▶ foranstilt adledd: *almost crazy, pretty big*
- ▶ etterstilt adledd: *crazy about dogs, lik sin far*

- ▶ Hodet er et verb i finitt eller infinitt form
- ▶ Kan fullføre: *Jeg/Barnet/Den rare mannen ...*

Eksempler:

- ▶ verb *sover, danset*
- ▶ verb + NP: *spiste kaken*
- ▶ verb + NP + NP: *ga ham kaken*
- ▶ verb + NP + PP: *la alle papirene i skuffen*

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



- ▶ *The dog ate my homework*
 - ▶ leksikale kategorier
 - ▶ finne umiddelbare konstituenten
 - ▶ finne hodet
 - ▶ ikke-leksikale kategorier (fraser)

- ▶ Informasjonen i et frasestrukturtre kan også representeres som frasestrukturregler
- ▶ Generaliserer over vår syntaktiske kunnskap
- ▶ Spesifiserer de velformede strukturene i et språk
- ▶ Så langt har vi sett:
 1. $S \rightarrow NP VP$
 2. $NP \rightarrow D N$
 3. $VP \rightarrow V NP$

- ▶ Noen flere regler: intransitive verb

1. *The cat purred*
2. *The woman laughed*

- ▶ $VP \rightarrow V$

- ▶ Noen flere regler: PP i VP

1. *The dog played in the garden*
2. *The cat ate the cake on the terrace*

- ▶ $VP \rightarrow VP PP$
- ▶ $PP \rightarrow P NP$

- ▶ Noen flere regler: leddsetninger (innledes av subjunksjon “complementizer” (C)):

1. *My brother said **that the dog purred***
2. *We wondered **whether the cat ate the cake***

- ▶ $VP \rightarrow V\ CP$
- ▶ $CP \rightarrow C\ S$

Kontekstfrie grammatikker (CFGer)

- ▶ Formell modell som fanger inn konstituentstatus og rekkefølge
- ▶ Brukes mye innenfor lingvistikk
- ▶ Fungerer best for språk som engelsk, med nogenlunde fast leddstilling
- ▶ De fleste moderne lingvistiske teorier inneholder en form for kontekstfri grammatikk
- ▶ En CFG består av:
 - ▶ Fraser og ordklasser (ikke-terminale kategorier)
 - ▶ Ord (terminale kategorier)
 - ▶ Regler

Kontekstfrie grammatikker (CFGer)

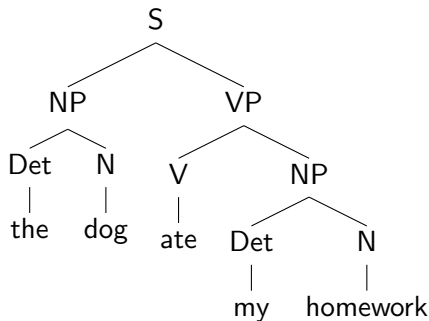
- ▶ Formelt: en CFG er en 4-tupple $\langle N, \Sigma, R, S \rangle$, der
 - ▶ N er en mengde **ikke-terminale** symboler (syntaktiske kategorier)
 - ▶ Σ er en mengde **terminale** symboler (ord)
 - ▶ R er en mengde **regler** på formen $A \rightarrow \alpha$, der
 - ▶ A er en ikke-terminal
 - ▶ α er en streng av symboler hentet fra mengden $(\Sigma \cup N)^*$, dvs både terminaler og ikke-terminaler
 - ▶ S er et særskilt startsymbol

Kontekstfrie grammatikker (CFGer)

Eksempel CFG

- ▶ La $G = \langle N, \Sigma, R, S \rangle$ der
 - ▶ $N = \{S, NP, VP, DT, N', V, N\}$
 - ▶ $\Sigma = \{et, fly, ankom\}$
 - ▶ $R = \{S \rightarrow NP VP,$
 $NP \rightarrow Det N',$
 $N' \rightarrow N,$
 $VP \rightarrow V,$
 $Det \rightarrow et,$
 $N \rightarrow fly,$
 $V \rightarrow ankom,$
 $\}$
 - ▶ $S = S$

- Applikasjoner av regler kan også visualiseres som **trær**



- Trær uttrykker:
 - Hierarkisk gruppering av konstituenten
 - Syntaktisk kategori for konstituenten
 - Lineær rekkefølge av konstituenten

- ▶ Trær kan også skrives som klammer med etiketter (“labelled bracketings”)

```
[NP  
  [Det a]  
  [N' [N flight]]]
```


- ▶ Mengden av setninger i et naturlig språk antas å være uendelig
 - ▶ Språkets **kreativitet**
- ▶ 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 du mener at jeg tror at jeg skriver akkurat nå
 - ▶ osv.

- ▶ Trenger mekanisme som kan generere (i prinsipp) uendelige strukturer
- ▶ **Rekursive** strukturer: inneholder en delstruktur av samme type som helheten
 - ▶ programmeringsspråk: feks Python
 - ▶ frasestrukturgrammatikk: en trestruktur er rekursiv dersom den inneholder en node som dominerer en annen node av samme type

- ▶ Rekursive regler gjør at grammatikken kan generere et uendelig antall strukturer
- ▶ *The dog played in the garden on Monday*
- ▶ *The dog played in the garden on Monday for an hour*
- ▶ *The dog played in the garden on Monday for an hour with a stick*
- ▶ $VP \rightarrow V PP ??$
- ▶ $VP \rightarrow VP PP$

- ▶ Andre rekursive regler

- ▶ *The dog with the collar barked*
 - ▶ *The dog with the collar around its neck barked*
 - ▶ *The dog with the collar around its neck on the sofa barked*
 - ▶ ...
- ▶ $\text{NP} \rightarrow \text{NP PP}$

- ▶ Fraser av samme type kan **koordineres** og danne en ny kategori av samme type
 - ▶ $XP \rightarrow XP \text{ og } XP$
- ▶ I need to know $[_{NP}[_{NP}\text{the aircraft}] \text{ and } [_{NP}\text{the flight number}]]$
- ▶ What flights do you have $[_{VP}[_{VP}\text{leaving Denver}] \text{ and } [_{VP}\text{arriving in San Francisco}]]$

- ▶ Annet eksempel på rekursjon: adjektiver
 - ▶ The kindhearted intelligent handsome boy
- ▶ Introduserer et nivå mellom NP og N: **N'**
- ▶ $NP \rightarrow \text{Det } N'$
- ▶ $N' \rightarrow \text{Adj } N'$
- ▶ $N' \rightarrow N$

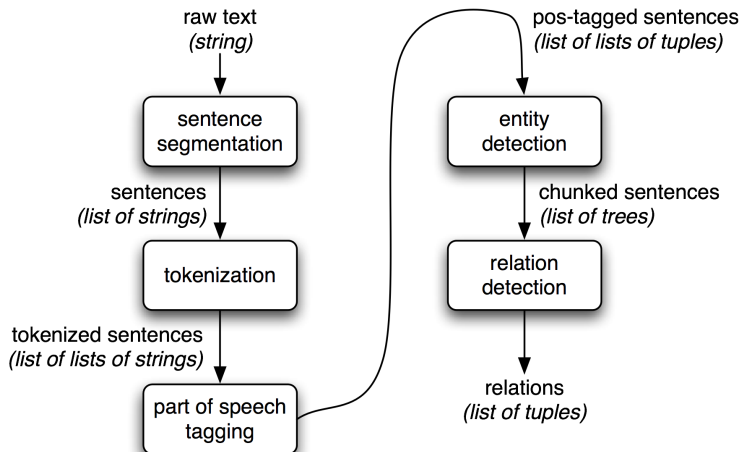
- ▶ Syntaktisk analyse brukes i en rekke språkteknologiske applikasjoner:
 - ▶ Grammatikkontroll
 - ▶ Spørsmål-Svar systemer
 - ▶ Informasjonsekstraksjon
 - ▶ Tekstgenerering
 - ▶ Maskinoversettelse
 - ▶ Opinion Mining
 - ▶ osv.
- ▶ Trenger syntaktisk analyse for å få tilgang til semantisk tolkning

- ▶ Dele setningen inn i en sekvens “**chunks**”: ikke-overlappende sekvenser med tekst
- ▶ En chunk inneholder et **hode**, muligens med noen funksjonsord/modifikatorer først
[walk] [straight past] [the lake]
- ▶ Ikke-rekursive: en chunk kan ikke inneholde en chunk av samme kategori

- ▶ Forenklede fraser (“fram til hodet”)
- ▶ Ikke komplett syntaktisk beskrivelse, men tilstrekkelig for mange applikasjoner
- ▶ NP-utfyllinger (PP’er, relativsetninger) er ofte rekursive og/eller flertydige: **ikke** inkludert i NP-chunker

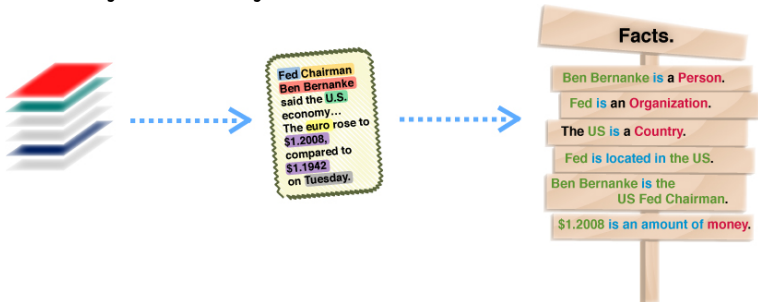
[G.K. Chesterton],
[author] of
[The Man] who was
[Thursday]

Bruk av chunking



Bruk av chunking

Informasjonsekstraksjon



- ▶ Automatisk tildele full syntaktisk struktur til en gitt setning
- ▶ Tradisjonelt (for CFG'er):
 - ▶ søk gjennom alle mulige trær for en setning
 - ▶ “bottom-up” vs “top-down” algoritmer

- ▶ Mer enn én mulig struktur for en setning
- ▶ Veldig vanlig

PoS-ambiguities

		VB			
	VBZ	VBP	VBZ		
NNP	NNS	NN	NNS	CD	NN
Fed	raises	interest	rates	0.5	%

Attachment ambiguities

in effort
to control
inflation

Back in the days (90-tallet)

- ▶ Grammatikk-drevet parsing: mulige trær definert av grammatikk
- ▶ Problemer med **dekningsgrad**
 - ▶ bare rundt 70% av alle setninger ble tildelt en analyse
- ▶ de fleste setninger ble tildelt mer enn én analyse av grammatikken
 - ▶ hvordan velge?

Data-drevet (statistisk) parsing

- ▶ I dag finnes det data-drevne/statistiske parsere for en rekke språk og syntaktiske representasjoner
- ▶ Data-drevet parsing: mulige trær er definert av en trebank (noen ganger også en grammatikk)
- ▶ Tildeler én analyse per setning
- ▶ Og får flesteparten rett
- ▶ Fortsatt et aktivt forskningsfelt, forbedringer mulig!!

- ▶ Korpus manuelt annotert med syntaktisk struktur:
⇒ en **trebank**
- ▶ Penn Treebank: mye brukt engelsk trebank
- ▶ Trebanker for andre språk:
 - ▶ Prague Dependency Treebank (tsjekkisk)
 - ▶ Negra (tysk)
 - ▶ Penn (kinesisk)
 - ▶ Norwegian Dependency Treebank (norsk)
 - ▶ **Universal Dependencies** (70 språk!)

Trebanker

Eksempel fra Penn Treebank (WSJ)

```
( (S
  (PP-LOC (IN In)
    (NP
      (NP (NNP Thursday) (POS 's) )
      (NN edition) ))
    (, ,)
    (NP-SBJ (PRP it) )
    (VP (VBD was)
      (VP
        (ADVP-MNR (RB incorrectly) )
        (VBN indicated)
        (SBAR (IN that)
          (S
            (NP-SBJ (DT the) (NN union) )
            (VP (VBD had)
              (VP (VBN paid)
                (NP (DT a) (NN fee) )
                (PP-DTV (TO to)
                  (NP
                    (NML (JJ former) (NNP House) (NNP Speaker) )
                    (NNP Jim) (NNP Wright) ))))))))
          (. .) ))
```

- ▶ Syntaks: hvordan setninger bygges opp av ord og ordkombinasjoner, såkalte **konstituent**er
- ▶ Konstituentstatus avgjøres ved tester
- ▶ **Fraser** - bygger opp setningen eller andre fraser (hierarkisk) og navngis etter hodet (NP, VP, PP, etc.)
- ▶ **Flertydighet**:
 - ▶ Flertydighet grunnet flere mulige strukturer for en setning
 - ▶ Forklarer hvordan gruppering av ord relaterer til betydning

- ▶ Syntaktiske regler:
 - ▶ Beskriver frasestrukturtrær
 - ▶ Kombinerer ord til fraser og fraser til setninger
- ▶ Kontekstfri grammatikk (CFG)
 - ▶ Formelt regelsystem: konstituenten, hierarkisk gruppering, lineær rekkefølge
 - ▶ Tillater rekursjon

- ▶ Syntaks i språkteknologi
 - ▶ viktig skritt mot semantisk tolkning
 - ▶ **Chunking**: “fattigmannssyntaks”
 - ▶ analyserer ikke-rekursive fraser
 - ▶ nyttig for eksempelvis informasjonsekstraksjon
 - ▶ Syntaktisk parsing: automatisk syntaktisk analyse
 - ▶ Bruk av trebanker for statistisk parsing