IN1140: Introduksjon til språkteknologi

Forelesning #7

Sondre Wold

Universitetet i Oslo

14 oktober 2024



Tema for i dag

Forrige gang

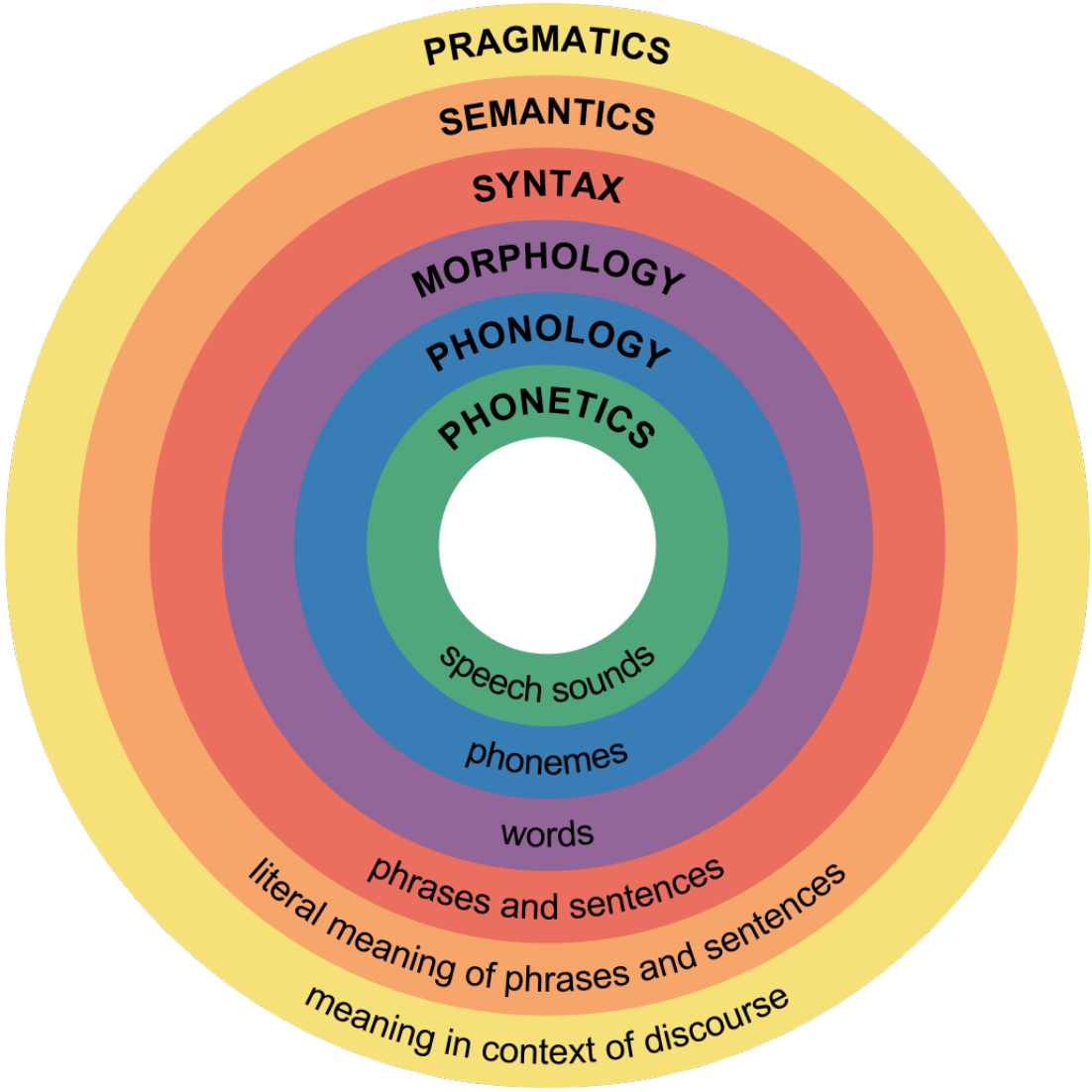
- Ordklasser
- Ordklassetagging
- Oblig2a: språkmodeller og ordklassetagging (frist ons. 18/10)

I dag

- Syntaks
- Kontekstfrie grammatikker
- Midtveisevaluering

Jeg liker epler som er rødere enn rødfargen jeg så i fargekartet til malingsbutikken som ligger til venstre for bensinstasjonen øverst i veien i bydelen med flest innbyggere i hovedstaden i landet som ligger vest for landet med gule landslagsdrakter.

Tema for i dag – Syntaks



Syntaks

"Studiet av hvordan setninger bygges opp av ord og ordkombinasjoner"

ORD – FRASER – SETNINGER

Syntaks

"Studiet av hvordan setninger bygges opp av ord og ordkombinasjoner"

ORD – FRASER – SETNINGER

Fraser

- Bygger opp setningen eller andre fraser
- Substantivfraser, f.eks. *det pene huset*
- Verbfraser, f.eks. *liker fotball*
- Preposisjonsfraser, f.eks. *i skogen*

  etc.

Setninger

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

Flertydighet

- En underliggende hypotese; vi mennesker grupperer ord også mentalt.
- Et av flere 'bevis' på dette:
- Ulik gruppering kan gi ulik betydning:

- *Gamle menn og kvinner kan forlate skipet*

Flertydighet

- En underliggende hypotese; vi mennesker grupperer ord også mentalt.
- Et av flere 'bevis' på dette:
- Ulik gruppering kan gi ulik betydning:

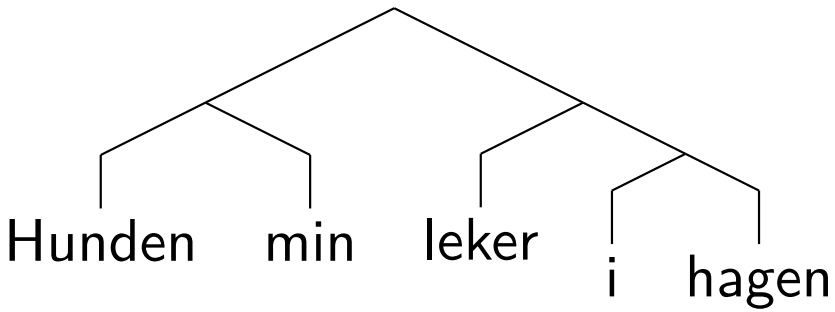
- *Gamle menn og kvinner kan forlate skipet*
- [*Gamle menn*] *og* [*kvinner*]
- [*Gamle* [*menn og kvinner*]]

Konstituenter

- Grupperinger av ord i en setning som fungerer som en enhet
- Hunden min leker i hagen

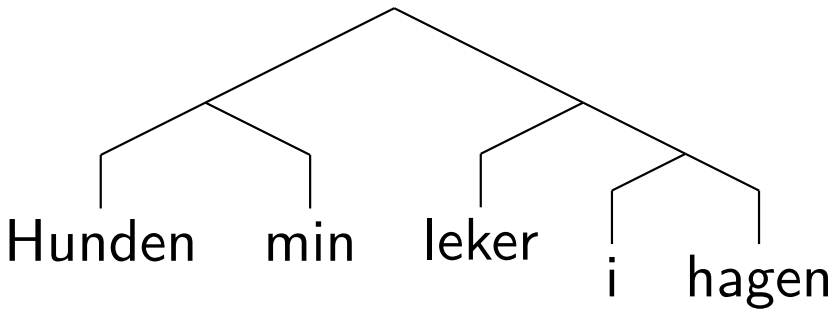
Konstituenter

- Grupperinger av ord i en setning som fungerer som en enhet
- Hunden min leker i hagen
- Kan nøstes; danner en hierarkisk struktur som kan visualiseres som en tre-graf:



Konstituenter

- Grupperinger av ord i en setning som fungerer som en enhet
- Hunden min leker i hagen
- Kan nøstes; danner en hierarkisk struktur som kan visualiseres som en tre-graf:



- Kan også representeres mer kompakt med «klammenotasjon»:
  [Huden min [leker [i hagen]]]
- Hvordan kan vi avgjøre konstituentstatus?
  - Lingvistiske tester

Konstituenttester

“stå alene”-testen:

Hunden min leker i hagen

- *Hvor leker hunden min?*

Konstituenttester

“stå alene”-testen:

Hunden min leker i hagen

- *Hvor leker hunden min?*
- *i hagen*
- **leker i*

Konstituenttester

"stå alene"-testen:

Hunden min leker i hagen

- *Hvor leker hunden min?*
- *i hagen*
- **leker i*

"erstattes med pronomen"-testen:

Hunden ødela blomsterbedet

- *Hunden ødela det*
- *Hunden ødela blomsterbedet og det gjorde katten også*

Konstituenttester

"stå alene"-testen:

Hunden min leker i hagen

- *Hvor leker hunden min?*
- *i hagen*
- **leker i*

"erstattes med pronomen"-testen:

Hunden ødela blomsterbedet

- *Hunden ødela det*
- *Hunden ødela blomsterbedet og det gjorde katten også*

"Flyttes som enhet"-testen:

Hunden min leker i hagen

- *I hagen leker hunden min*

Konstituenttester

- Engelsk eksempel: *The dog ate my homework*

Konstituenttester

- Engelsk eksempel: *The dog ate my homework*
  - *(What did the dog eat?) my homework* (stå alene)
  - *The dog ate* **it** (erstattes med pronomen)
  - *My homework was eaten by the dog* (flytter som en enhet)

Fraser og syntaktiske kategorier

- Et enkelt ord kan bygges ut til en gruppe ord med samme syntaktiske kategori, slik at den nye gruppen har samme funksjon i setningen (en konstituent)
- Kan erstattes med hverandre

- *Vi spiste kake*
- *Vi spiste en kake*
- *Vi spiste en god kake*
- *Vi spiste en god kake som egentlig skulle spares til i morgen*

Fraser og syntaktiske kategorier

- Et enkelt ord kan bygges ut til en gruppe ord med samme syntaktiske kategori, slik at den nye gruppen har samme funksjon i setningen (en konstituent)
- Kan erstattes med hverandre

- *Vi spiste kake*
- *Vi spiste en kake*
- *Vi spiste en god kake*
- *Vi spiste en god kake som egentlig skulle spares til i morgen*

- Oppbygging av fraser:
  - (adledd) hode (adledd: utfylling)
  - (*en god*) *kake* (*som egentlig skulle spares til i morgen*)
- Vi skal se på noen eksempler på ulike type fraser: NP, PP, VP, osv.

Substantivfraser (NP)

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

Eksempler:

- determinativ + substantiv: *the dog, en hund*
- egennavn: *Noam Chomsky, Japan*
- pronomen: *he, they, han, henne*

- Hodet bestemmer kongruens – feks tall, kjønn, bestemthet

Preposisjonsfraser (PP)

- 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*

Adjektivfraser (AdjP)

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

Eksempler:

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

Verbfraser (VP)

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

Eksempler:

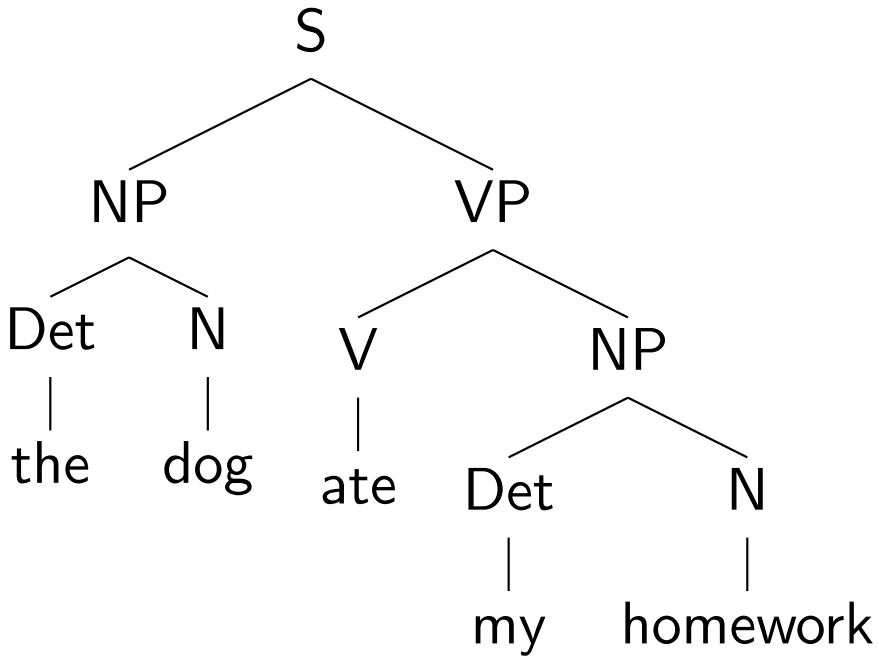
- verb: *sover, danset*
- verb + NP: *spiste kaken*
- verb + NP + NP: *ga ham kaken*
- verb + NP + PP: *la alle papirene i skuffen*

Syntaktiske kategorier

- Frasale kategorier: NP, VP, AdjP, PP
- Leksikale kategorier (ordklasser): N, V, P, Adj, Adv

Syntaktiske kategorier

- Frasale kategorier: NP, VP, AdjP, PP
- Leksikale kategorier (ordklasser): N, V, P, Adj, Adv
- Frasestrukturtre (Phrase Structure tree) – annotert med syntaktiske kategorier på de interne nodene:



- Eller med annoterte klammer:
  [S [NP [Det the] [N dog]] [VP [V ate] [NP [Det my] [N homework]]]]

Frasestrukturregler

- 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

Frasestrukturregler

- Noen flere regler: intransitive verb

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

- VP $\rightarrow$ V

Frasestrukturregler

- 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

Frasestrukturregler

- Noen flere regler: leddsetninger (innledes av subjunksjon "complementizer" (C)):

1. *My brother said that the cat 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 i programmeringsspråk

Hvordan vet datamaskinen hva som er et gyldig utsagn i f.eks Python?

Kontekstfrie grammatikker i programmeringsspråk

Hvordan vet datamaskinen hva som er et gyldig utsagn i f.eks Python?

```
>>> print(ast.dump(ast.parse('x = 1; y = 2', mode='single'), indent=4))
Interactive(
    body=[
        Assign(
            targets=[
                Name(id='x', ctx=Store())],
            value=Constant(value=1)),
        Assign(
            targets=[
                Name(id='y', ctx=Store())],
            value=Constant(value=2))])
```

Kontekstfrie grammatikker (CFGer)

- Formelt: en CFG er en 4-tuppel $\langle N, \Sigma, R, S \rangle$, der

1: N er en mengde **ikke-terminale** symboler (syntaktiske kategorier)

2: Σ er en mengde **terminale** symboler (ord)

3: 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

4: S er et særskilt startsymbol

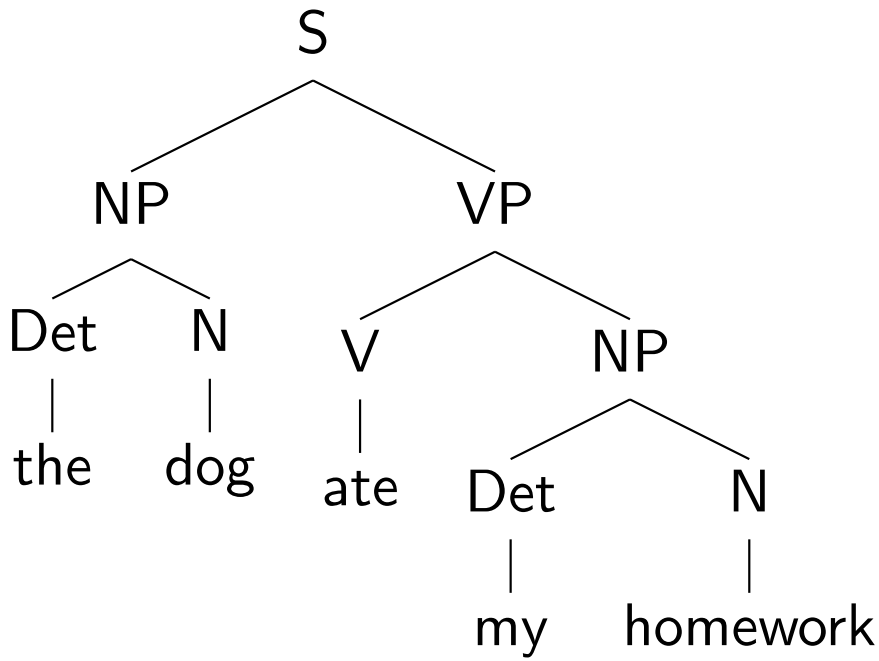
Kontekstfrie grammatikker (CFGer)

Eksempel CFG

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

Trær og regler

- Kan tenke på et frasestrukturtre som en visualisering av applikasjoner av regler i grammatikken:



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

Rekursjon

- 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.

Rekursjon

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

Rekursjon

- Rekursive regler gjør at grammatikken kan generere et uendelig antall strukturer

- *The dog played*
- *The dog played in the garden*
- *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*
- ...

Rekursjon

- Rekursive regler gjør at grammatikken kan generere et uendelig antall strukturer

- *The dog played*
- *The dog played in the garden*
- *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
- VP $\rightarrow$ VP PP

Rekursjon

- Andre rekursive regler

- *The dog barked*
- *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*
- ...

- NP $\rightarrow$ Det N
- NP $\rightarrow$ NP PP

Rekursjon

- Andre rekursive regler

- *The dog barked*
- *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*
- ...

- NP $\rightarrow$ Det N
- NP $\rightarrow$ NP PP
- PP $\rightarrow$ P NP

Rekursive strukturer

- Fraser av samme type kan koordineres og danne en ny kategori av samme type
  - XP $\rightarrow$ XP *og* XP

- I need to know $[_{NP}[_{NP}$the aircraft] and $[_{NP}$the flight number]]
- What flights do you have $[_{VP}[_{VP}$leaving Denver] and $[_{VP}$arriving in San Francisco]]

Syntaks i språkteknologi

- Syntaktisk analyse brukes i en rekke språkteknologiske applikasjoner:
  - Grammatikkontroll
  - Tekstgenerering
  - Maskinoversettelse
  - Spørsmål–svar systemer
  - Informasjonsekstraksjon
  - Sentimentanalyse
  - osv.
- Trenger syntaktisk analyse for å få tilgang til semantisk tolkning

Syntaktisk parsing

- Automatisk tildele full syntaktisk struktur til en gitt setning
- Ulike tilnærminger til parsing; regel-basert, data-drevet, eller hybrid
- Parsere blir gjerne utviklet mot en såkalt trebank.

Trebanker

- Korpus manuelt annotert med syntaktisk struktur
- Penn Treebank: mye brukt engelsk trebank
- Trebanker for andre språk:
  - Prague Dependency Treebank (tsjekkisk)
  - Negra (tysk)
  - Penn (kinesisk)
  - Norwegian Dependency Treebank (norsk)
  - Universal Dependencies (over 100 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) ))))))))
    (. .) ))
```

Regel-basert parsing

- Den tradisjonelle tilnærmingen
- Grammatikk-drevet parsing: mulige trær definert av grammatikk
- Tillærmet som et søkeproblem: 'bottom-up' vs 'top-down' algoritmer for generering av trær
- Problemer med *dekningsgrad* og *robushet*:
  - bare rundt 70% av alle setninger ble tildelt en analyse
- de fleste setninger ble tildelt mer enn én analyse av grammatikken.
  - Hvordan velge?
  - Ofte løst ved å legge til en statistisk komponent, f.eks med sannsynlighet for ulike trær estimert fra en trebank (hybrid metode):
  - Probabilistic Context-Free Grammar (PCFG)

Flertydighet

- Svært ofte mer enn én mulig struktur for en setning.

PoS-ambiguities						**Attachment ambiguities**
		VB				
	VBZ	VBP	VBZ			
NNP	NNS	NN	NNS	CD	NN	
Fed	raises	interest	rates	0.5	%	in effort to control inflation

Data-drevet parsing

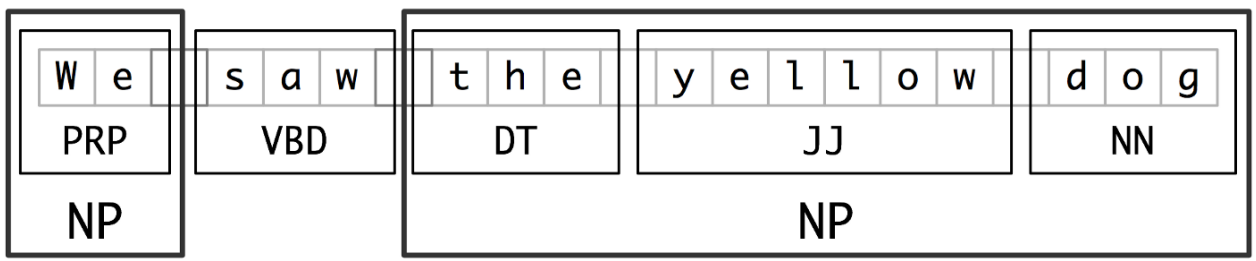
- I dag finnes det data-drevne parsere basert på maskinlæring for en rekke språk og syntaktiske representasjoner
- Mulige trær er definert av en trebank
- Trebanken brukes som treningsdata for en ML-modell for å predikere trestrukturer
- Tildeler én analyse per setning
- Og får flesteparten rett
- Fortsatt et aktivt forskningsfelt, forbedringer mulig!!

Chunking

- Dele setningen inn i 'chunks': ikke-overlappende sekvenser med tekst
- En chunk inneholder et hode, muligens med noen funksjonsord / modifikatorer først: [walk] [straight past] [the lake]
- Ingen intern struktur, ikke hierarkiske

Chunking

- Dele setningen inn i 'chunks': ikke-overlappende sekvenser med tekst
- En chunk inneholder et hode, muligens med noen funksjonsord / modifikatorer først: [walk] [straight past] [the lake]
- Ingen intern struktur, ikke hierarkiske
- En chunker tar typisk PoS-tagget tekst som input.
- Hver chunk tildeles typisk også en syntaktisk kategori (NP-chunks, VP-chunks, osv.).
- Eksempel på NP-chunker (fra NLTK):



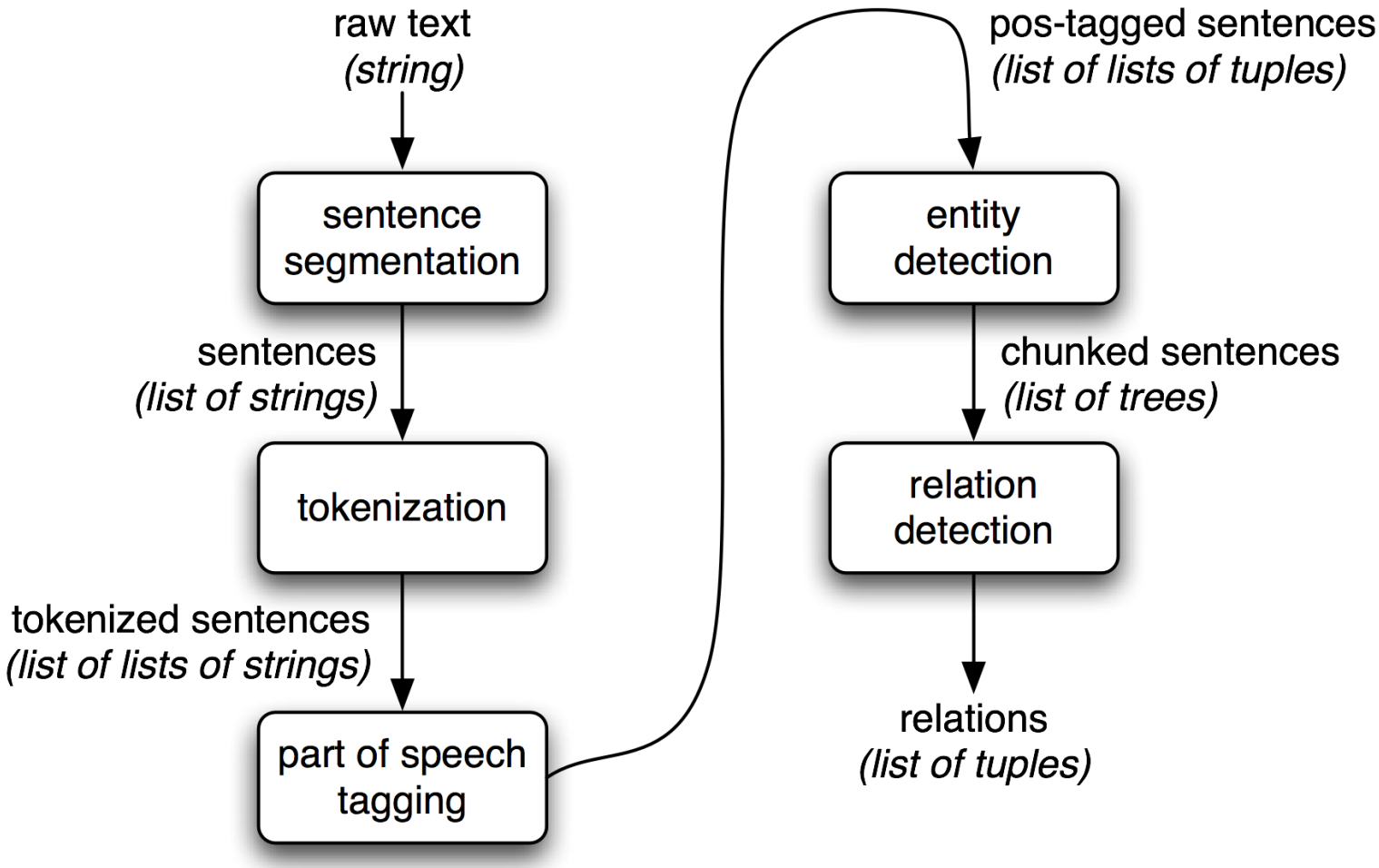
Chunking

- Forenklede fraser ("fram til hodet")
- Ikke-rekursive: kan ikke inneholde en chunk av samme kategori
- 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 ]
```

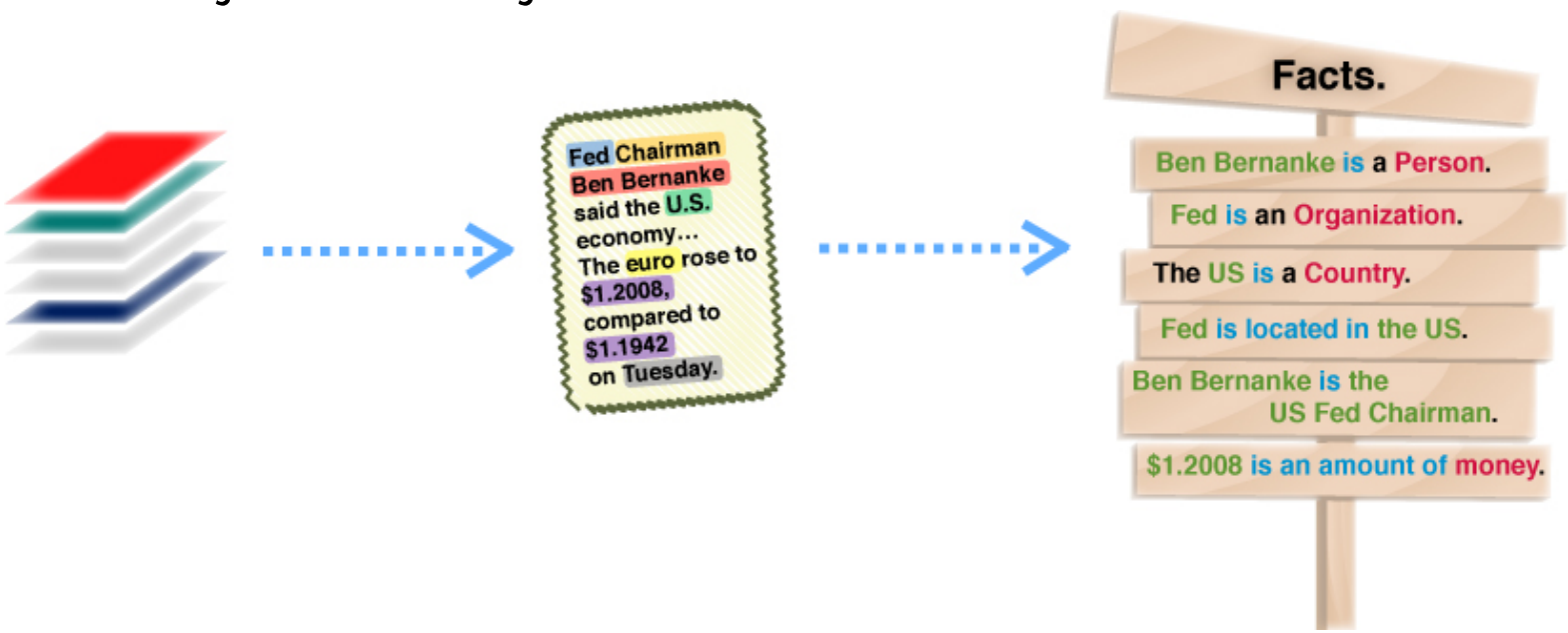
- Ikke komplett syntaktisk beskrivelse, men tilstrekkelig for mange applikasjoner

Bruk av chunking



Bruk av chunking

Informasjonsekstraksjon



Oppsummering

- Syntaks: hvordan setninger bygges opp av ord og ordkombinasjoner, såkalte konstituenter
- 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

Oppsummering

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

Oppsummering

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