IN1140: Introduksjon til språkteknologi

Forelesning #7

Lilja Øvrelid

Universitetet i Oslo

10 oktober 2019

Tema for i dag

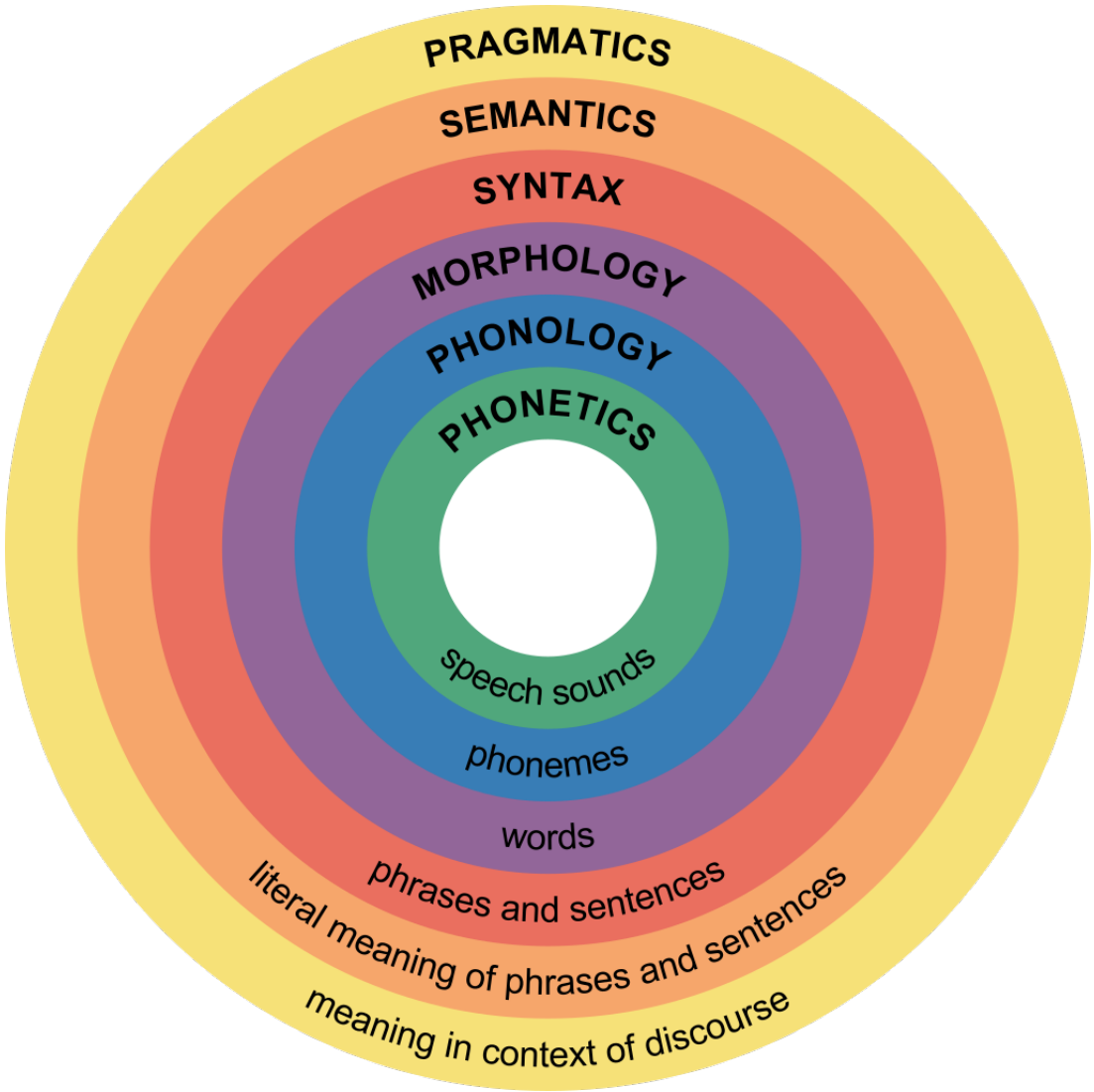
Forrige uke

- Ordklasser
- Ordklassetagging
- Oblig2: språkmodeller og ordklassetagging (frist: 23/9)

I dag

- Syntaks
- Kontekstfrie grammatikker
- Midtveisevaluering

Tema for i dag – Syntaks



Syntaks

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

ORD – FRASER – SETNINGER

- Setninger - inneholder et verb og (som regel) et subjekt

  - *Spis!*
  - *Bea lukket vinduet fort.*

- 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.

Syntaktiske regler

- Kombinerer ord til fraser og fraser til setninger
- Beskriver forholdet mellom grupper av ord (ordklasser) og plassering i setningen
  - artikler (determinativ) liker å komme foran et substantiv
- Andre begrensninger som påvirker grammatikalitet

- **Store huset*
- **Det huset store*
- *Det store huset*

Flertydighet

- Beskriver alle mulige grupperinger av ord

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

Flertydighet

Strukturell flertydighet

- *For sale: an antique desk suitable for lady with thick legs and large drawers*

- Flertydighet grunnet flere mulige strukturer for en setning
- Forklarer hvordan gruppering av ord relaterer til betydning

Konstituenter

- Konstituenter – 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

Konstituenttester

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.*

Konstituenttester

- 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: syntaktiske kategorier

- 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*)

Substantivfraser (NP)

- 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

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*, *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 . . .*

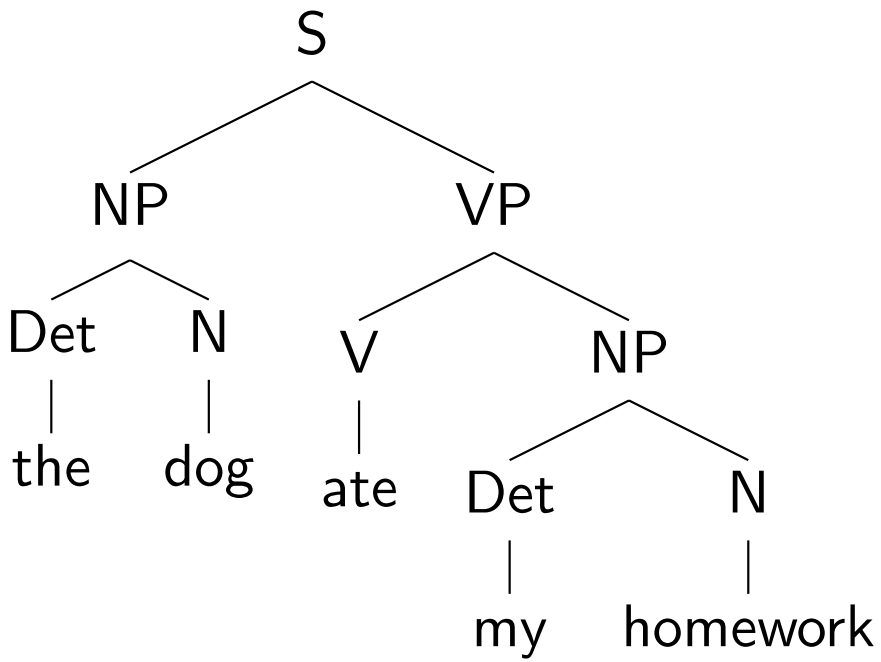
Eksempler:

- 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

- Frasestrukturtre (Phrase Structure (PS) tree)



Frasestrukturtrær

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

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 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

Kontekstfrie grammatikker (CFGer)

- Formelt: en CFG er en 4-tuppel $< N, \Sigma, R, S >$, 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

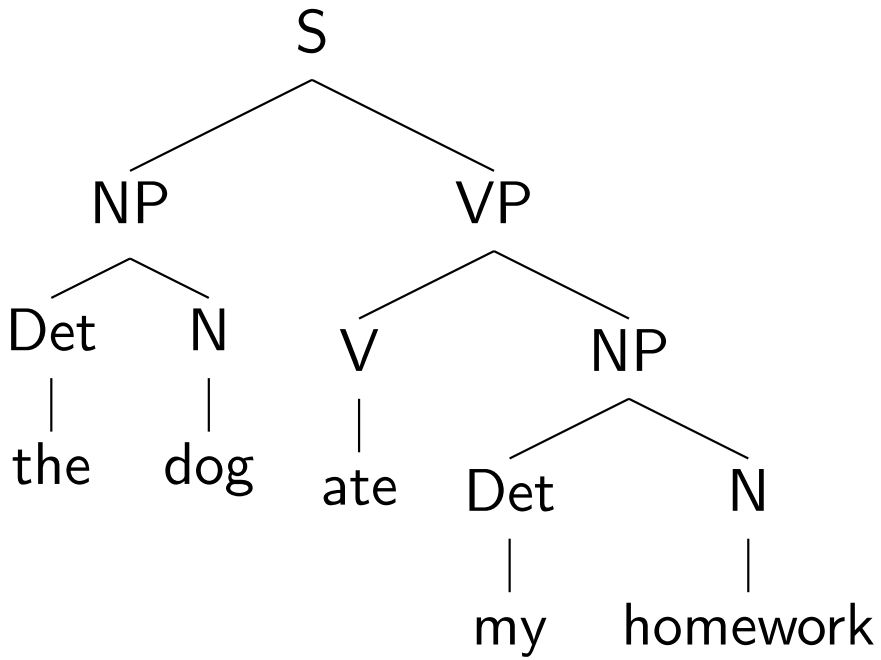
Kontekstfrie grammatikker (CFGer)

Eksempel CFG

- La $G =< N, \Sigma, R, S >$ 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

Trær

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



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

Trær

- Trær kan også skrives som klammer med etiketter ("labelled bracketings")

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

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 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

Rekursjon

- 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*
- ...

- NP $\rightarrow$ NP PP

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]]

Bar-kategorier

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

Syntaks i språkteknologi

- 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

Chunking

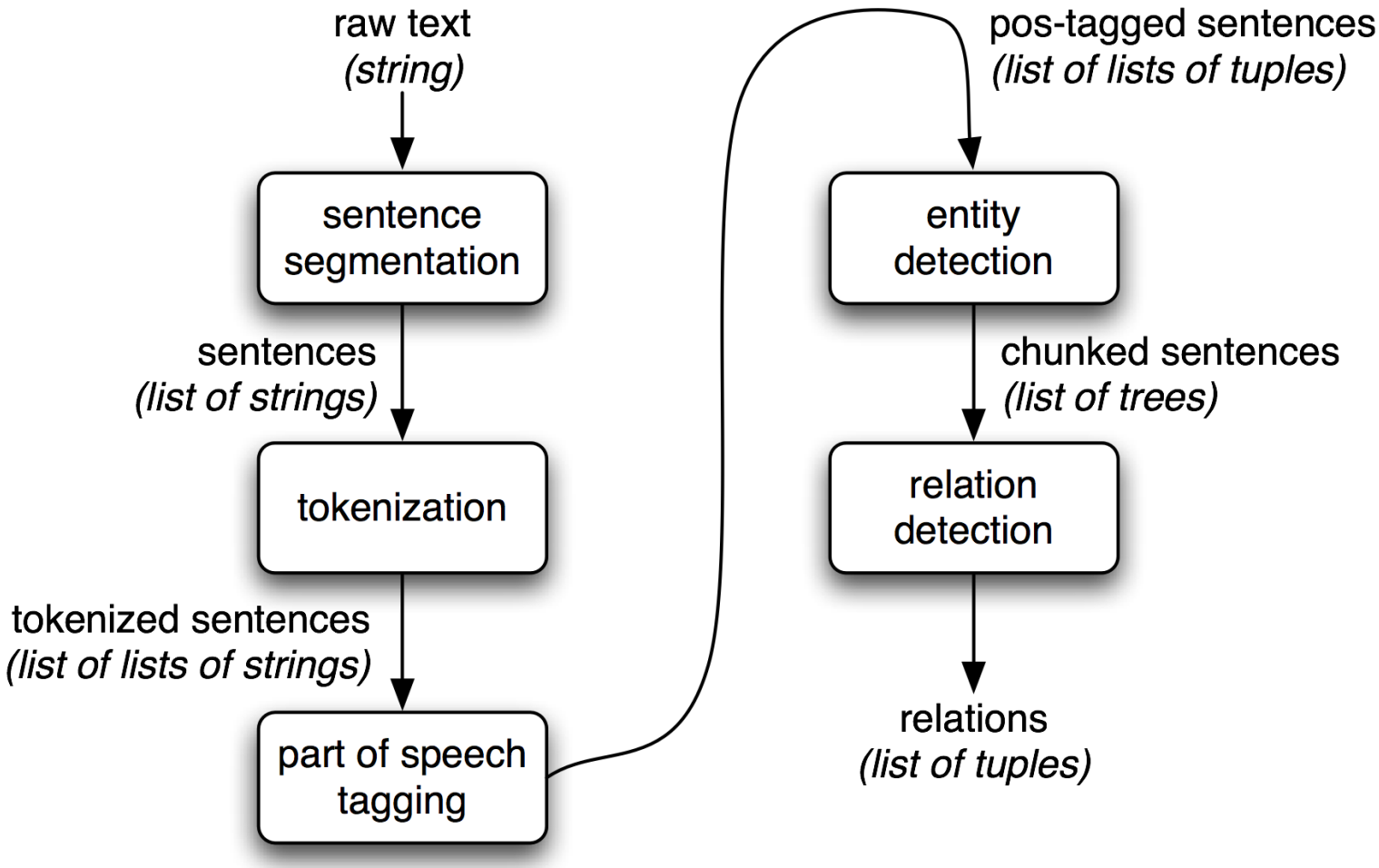
- Dele setningen inn i en sekvens **“chunks”**: ikke-overlappende sekvenser med tekst
- En chunk inneholder et **hode**, muligens med noen funksjonsord/modifikatorer først
  [walk] [straight past] [the lake]
- Ikke-rekursive: en chunk kan ikke inneholde en chunk av samme kategori

Chunking

- 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

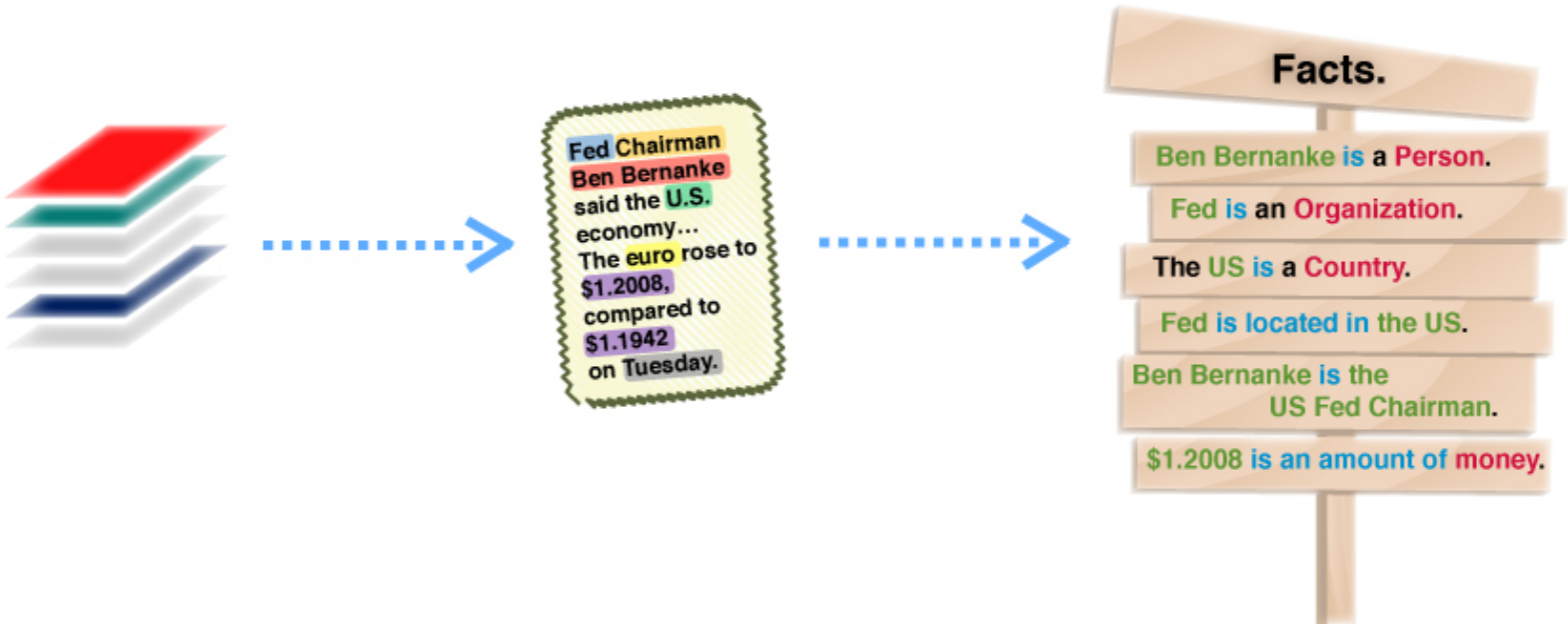
```
[ G.K. Chesterton ],
[ author ] of
[ The Man ] who was
[ Thursday ]
```

Bruk av chunking



Bruk av chunking

Informasjonsekstraksjon



Syntaktisk parsing

- 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

Flertydighet

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

PoS-ambiguities

		VB				**Attachment ambiguities**
	VBZ	VBP	VBZ			
NNP	NNS	NN	NNS	CD	NN	
Fed	raises	interest	rates	0.5	%	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!!

Trebanker

- Korpus manuelt annotert med syntaktisk struktur:
  $\Rightarrow$ 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) ))))))))
    (. .) ))
```

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
  - Chunking: "fattigmannssyntaks"
    - analyserer ikke-rekursive fraser
    - nyttig for eksempelvis informasjonsekstraksjon
  - Syntaktisk parsing: automatisk syntaktisk analyse
  - Bruk av trebanker for statistisk parsing