INF1820: Introduksjon til språk-og kommunikasjonsteknologi

Niende forelesning

Lilja Øvrelid

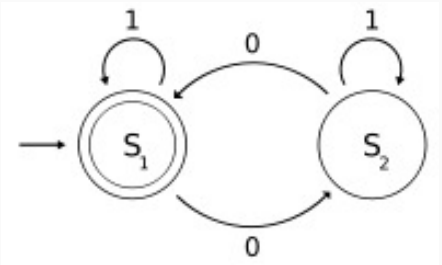
20 mars, 2017

Formelle modeller

- Kan fange inn den språklige kunnskapen v.hj.a et lite antall formelle modeller og teorier
- Hentet fra informatikk, matematikk og lingvistikk
- Disse modellene kan prosesseres ved et lite antall velkjente algoritmer

Formelle modeller

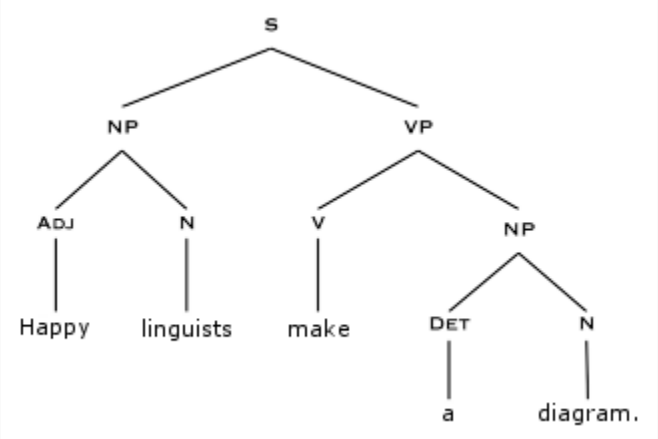
- Endelige tilstandsmaskiner ("finite state automata"):
- Består av tilstander, overganger ("transitions") og en input-representasjon
- Variasjoner: deterministiske og ikke-deterministiske, endelige tilstandsmaskiner og endelige tilstandstransdusere



Fonologi, morfologi

Formelle modeller

- Formelle regelsystemer
- feks kontekstfrie grammatikker



- S $\rightarrow$ NP VP
- NP $\rightarrow$ ADJ N
- NP $\rightarrow$ Det N
- VP $\rightarrow$ V NP

Syntaks

Kontekstfrie grammatikker

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

Derivasjoner

- En **derivasjon** av en streng fra en ikke-terminal A er resultatet av en rekke applikasjoner av reglene (fra G) til A:

NP	
Det N'	ved NP $\rightarrow$ Det N'
et N'	ved Det $\rightarrow$ *et*
et N	ved N' $\rightarrow$ N
et fly	N $\rightarrow$ *fly*

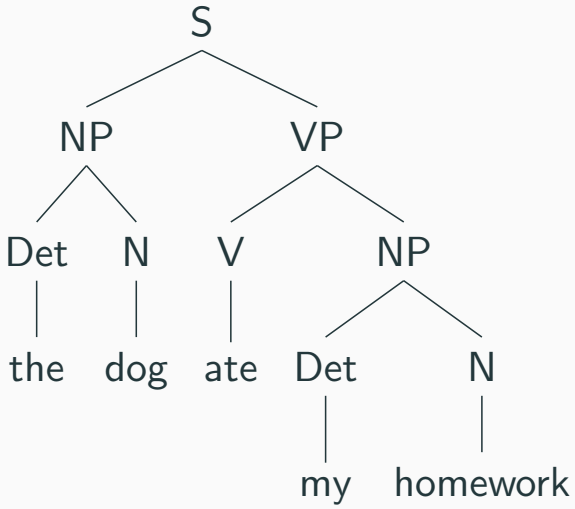
- Kan også skrive NP $\Rightarrow$ Det N' $\Rightarrow$ *et* N' $\Rightarrow$ *et* N $\Rightarrow$ *et fly*, der $\Rightarrow$ betyr "deriverer direkte" eller "gir med én regelapplikasjon"
- G genererer *et fly* (som en streng med kategori NP)

Grammatikker og språk

- CFG er en abstrakt modell for å koble strukturer med strenger;
- **Ikke** ment som en modell på hvordan mennesker produserer setninger
- En grammatikk G definerer et formelt språk L_G:
  - Språket L_G består av mengden av strenger av terminalsymboler som kan deriveres fra startsymbolet: $L_G = \{\mathrm{w} \mid \mathrm{w} \in \Sigma* \text{ og } S \text{ deriverer w } \}$
  - Setninger som kan deriveres fra G tilhører det formelle språket definert av G: de **Grammatiske Setningene** i henhold til G
  - Setninger som ikke kan deriveres fra G er de **Ugrammatiske Setningene** i henhold til G

Trær

- Derivasjoner 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]]]
```

- **Dominanse**: node x dominerer node y dersom det finnes en sammenhengende sekvens grener som strekker seg nedover ("descending") fra x til y. For eksempel:
  - NP dominerer ikke-terminalene Det, N' og N
- **Umiddelbar dominanse**: node x dominerer umiddelbart node y dersom x dominerer y og det ikke finnes noen distinkt node mellom x og y. For eksempel:
  - NP dominerer umiddelbart Det og N'.

Trær

- En node er **datter** av noden som umiddelbart dominerer den.
- Distinkte noder som umiddelbart domineres av samme node kalles **søstre**
- En node som ikke domineres av noen annen node kaller **rot**noden

CFG i motsetning til . . . ?

- Regulære uttrykk
  - Ansett som for svake til å uttrykke alle lingvistiske generaliseringer
- Kontekstsensitive grammatikker:
  - Tillater regler på formen $XAY \rightarrow X\alpha Y$, dvs hvordan A ekspanderes kan avhenge av konteksten X_Y
  - Ansett som "for sterke" - kan beskrive språk som ikke er mulige menneskelige språk
  - Regulære språk $\subset$ Kontekstfrie språk $\subset$ Kontekstsensitive språk

Rekursive strukturer

- Finnes ikke noen øvre grense på lengden av grammatiske setninger
  - derfor er mengden av setninger uendelig

  En grammatikk er en endelig beskrivelse av velformede setninger
  - for å beskrive en uendelig mengde må vi tillate iterasjon (f.eks. X^+) eller rekursjon
- **Rekursive regler**: ikke-terminalen i venstresiden i en regel dukker også opp på høyresiden

Rekursive stukturer

Direkte rekursjon:	
N' $\rightarrow$ N' PP	*fly til Bergen*
VP $\rightarrow$ VP PP	*forlot Trondheim i formiddag*
Indirekte rekursjon:	
S $\rightarrow$ NP VP	
VP $\rightarrow$ V CP	
CP $\rightarrow$ C S	*fortalte at et fly var forsinket*

Koordinering

- Enhver frasal kategori XP kan koordineres med en konstituent av samme type (XP) og danner en ny kategori av type XP
  - XP $\rightarrow$ XP *og* XP

Eksempler

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

Problemer for CFG

- Kongruens/samsvarsbøyning ("agreement")
- Subkategorisering

Samsvarsbøyning ("agreement")

- På norsk samsvarer determinativer med substantiver i NP'er:
  - *Den hunden*
  - *De hundene*
  - ** De hunden*
  - ** Den hundene*
- På engelsk samsvarer verb med subjekter:
  - *Which flight leaves in the morning?*
  - **Which flight leave in the morning?*

Samsvarsbøyning ("agreement")

- Utvide grammatikken med flere regler?
  - $NP_{sg} \rightarrow Det_{sg}\ N_{sg}$
  - $NP_{pl} \rightarrow Det_{pl}\ N_{pl}$
  - $S_{sg} \rightarrow NP_{sg}\ VP_{sg}$
  - $S_{pl} \rightarrow NP_{pl}\ VP_{pl}$
  - verre når vi legger til person og enda verre i språk med mer samsvarsbøyning, feks tre kjønn.
  - mister generaliseringer som omhandler alle verb eller substantiver

Seleksjon (subkategorisering)

- Verb selekterer visse typer konstituenter
  - *Jeg fant katta*
  - **Jeg forsvant katta*
  - *It depends* [$_{PP}$ *on the question*]
  - **It depends* [$_{PP}$ { *to/from/by*} *the question*]
- Tradisjonell subkategorisering for verb
  - transitiv (tar direkte objekt)
  - intransitiv
- I nyere teorier operer man ofte med opp mot hundre forskjellige subkategoriseringer for verb!

Seleksjon

- Flere eksempler
  - *find* selekterer en NP
  - *want* selekterer en NP eller en infinitivisk VP
  - *bet* selekterer NP NP S
- Liste av alle mulige sekvenser av komplementer kalles **subkategoriseringsrammen** ("subcategorization frame") for verbet
- Akkurat som med samsvar fører en CFG implementasjon til en regeleksplosjon
- VP $\rightarrow$ V_{intr}
- VP $\rightarrow$ V_{tr} NP
- VP $\rightarrow$ V_{ditr} NP NP

  ...

Seleksjon

Ramme	Verb	Eksempel
—	*eat, sleep*	*I like to sleep*
NP	*prefer, find, love*	*Find* [$_{NP}$ *the flight from Dallas to San Francisco*]
NP NP	*show, give*	*Show* [$_{NP}$ *me*] [$_{NP}$ *airlines with flights from Dallas*]
NP PP	*help, load*	*Can you help* [$_{NP}$ *me*] [$_{PP}$ *with a flight*]?
VP$_{inf}$	*prefer, want, need*	*I would prefer* [$_{VP_{inf}}$ *to go to Dallas*]
S	*mean*	*Does this mean* [$_{S}$ *AA has a hub in Boston*]?

Funksjonell analyse

Syntaks

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

- **Syntaktisk form** - konstituenter beskrives i form av ordklasser, fraser:
  - fraser - større konstituenter over ordnivå
  - fraser navngis etter **hodet** - det sentrale, obligatoriske medlemmet, referanse
- **Syntaktisk funksjon** - konstituenter beskrives i form av sin funksjon i setningen som helhet
  - Subjekt
  - (Direkte og indirekte) Objekt
  - Adverbial

Funksjonell analyse

- Hvilken funksjon et språklig uttrykk har i frasen eller setningen den forekommer i (ikke hva slags frase)
- **primære setningsledd**: fraser
  - Ingrid kjøpte en ny bil

Primære setningsledd

- **Predikat:** finitt verbform (++ infinitiv, partisipp)

Min	gamle	venn	kjøpte	en	bil	i	Bergen	i går
Min	gamle	venn	|**kjøpte**|	en	bil	i	Bergen	i går
			PRED					

Primære setningsledd

- **Predikat:** finitt verbform (++ infinitiv, partisipp)

Min	gamle	venn	har	kjøpt	en	bil	i	Bergen
Min	gamle	venn	| **har**	**kjøpt**|	en	bil	i	Bergen
				PRED				

Primære setningsledd

- **Predikat:** finitt verbform (++ infinitiv, partisipp)

Min	venn	liker	å	kjøre
Min	venn	|**liker**	**å**	**kjøre**|
		PRED		

Primære setningsledd

- **Subjekt:** hvem eller hva er/var det som PRED?

Min	gamle	venn	| kjøpte|	en	bil	i	Bergen	i går
			PRED					
| **Min**	**gamle**	**venn**|	| kjøpte|	en	bil	i	Bergen	i går
SUBJ			PRED					

Primære setningsledd

- **Subjekt:** hvem eller hva er/var det som PRED?

Min	gamle	venn	| har	kjøpt|	en	bil	i	Bergen
			PRED					
|**Min**	**gamle**	**venn**|	| har	kjøpt|	en	bil	i	Bergen
SUBJ			PRED					

Primære setningsledd

- **Direkte objekt:** hvem eller hva er/var det som SUBJ PRED?

| Min	gamle	venn|	| har	kjøpt|	en	bil	i	Bergen
SUBJ			PRED						
| Min	gamle	venn|	| har	kjøpt|	|**en**	**bil**|	i	Bergen
SUBJ			PRED		D.OBJ				

Primære setningsledd

- **Direkte objekt:** hvem eller hva er/var det som SUBJ PRED?

|Ida| |ønsker å bli| |en berømt sanger|
SUBJ PRED
|Ida| |ønsker å bli| |**en berømt sanger**|
SUBJ PRED

- **Predikativ**: *være, bli*
- utypiske objekter

Primære setningsledd

- **Indirekte objekt:** hvem eller hva er/var det som SUBJ PRED D.OBJ?

|Politimannen| |ga| dem |en bot|
SUBJ PRED D.OBJ
|Politimannen| |ga| |**dem**| |en bot|
SUBJ PRED I.OBJ D.OBJ

Primære setningsledd

- **Adverbial (rom):** hvor eller fra hvor er/var det at SUBJ PRED (D.OBJ)?

| Min | gamle | venn| | har | kjøpt| | en | bil| | i | Bergen
SUBJ PRED D.OBJ
| Min | gamle | venn| | har | kjøpt| | en | bil| | **i** | **Bergen**|
SUBJ PRED D.OBJ RAL

Primære setningsledd

- **Adverbial (tid):** når, hvor lenge eller hvor ofte er/var det at SUBJ PRED (D.OBJ) (RAL)?

| Min | gamle | venn| | kjøpte| | en | bil| |i | Bergen| | i går
SUBJ PRED D.OBJ RAL
| Min | gamle | venn| | kjøpte| | en | bil| |i | Bergen| |**i går**|
SUBJ PRED D.OBJ RAL TAL

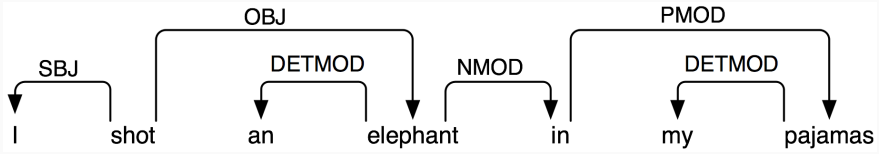
Primære setningsledd

- **Setningsadverbial** omskrivning: "Det er SetnAL slik at S - SetnAL"
  - *Harry vil* **sjelden** *lage blomkålsuppe*
  - *Jeg går* **ikke** *alene i havneområdet*

Dependensgrammatikk

- Alternativ grammatikkformalisme med økende popularitet innenfor språkteknologi
- Konstituenter og frasestruktur er ikke eksplisitt representert
- Syntaktisk struktur er gitt ved **ord** og **binære relasjoner mellom ord.**
- Relasjoner stort sett funksjonelle
- Fordel for språk med friere leddstilling enn engelsk (feks norsk, tysk, tsjekkisk)

Dependensgrammatikk



- Dependensgrammatikk representerer:
  - hode-dependent forhold (rettede kanter)
  - funksjonelle kategorier (kantetiketter)
  - muligens noen strukturelle kategorier (ordklasser)
- Frasestrukturgrammatikk representerer:
  - fraser (ikke-terminale noder)
  - strukturelle kategorier (ikke-terminale etiketter)
  - muligens noen funksjonelle kategorier

Syntaks og språkteknologi

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

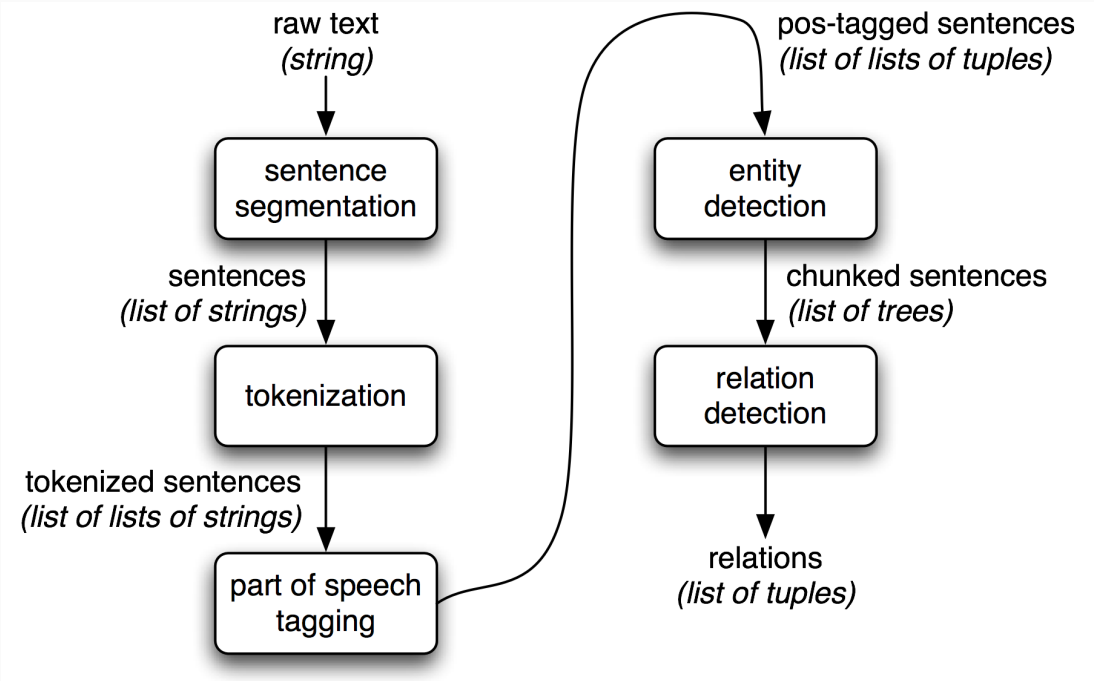
- Dele setningen inn i en sekvens **"chunks"**: ikke-overlappende sekvenser med tekst
  - [when I read] [a sentence], [I read it] [a chunk] [at a time]
- En chunk inneholder et **hode**, muligens med noen funksjonsord/modifikatorer først
  [walk] [straight past] [the lake]
- Ikke-rekursive: en chunk 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

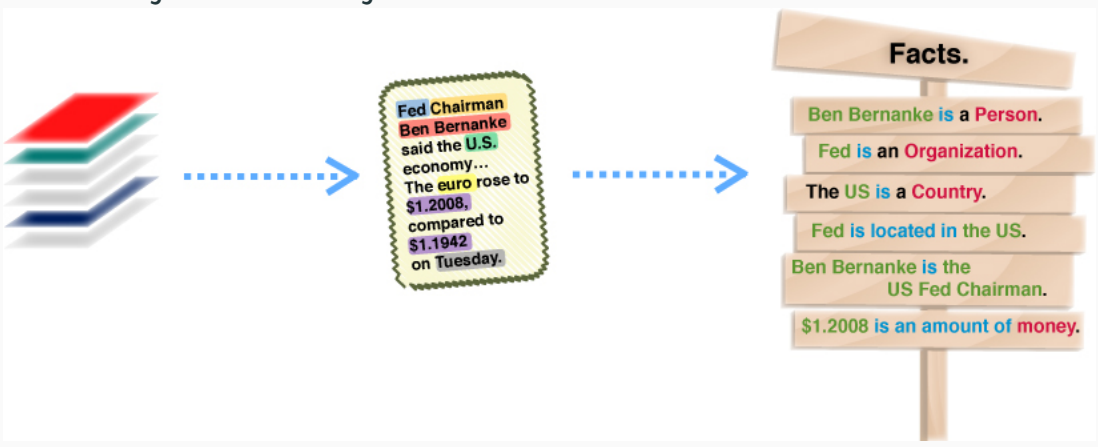
Named Entity Recognition

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell-Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:
LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

Bruk av chunking

Informasjonsekstraksjon



Bruk av chunking

- Studere syntaktiske mønstre, feks mulige subkategoriseringsrammer for verb
  - `gave NP`
  - `gave NP up`
  - `gave NP help`
  - `gave NP to NP`
  - etc.
- bruke feks i grammatikkutvikling

Syntaktisk parsing

- Automatisk tildele 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						**Attachment ambiguities**
		VB				
	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!!

Statistisk parsing

- Klassisk NLP parsing:
  - grammatikk og leksikon
- Flertydighet er et stort problem!
  - minimal grammatikk for tidligere setning (*Fed raises etc.*): 36 ulike analyser
  - stor, realistisk grammatikk: millioner av analyser
- Bruker sannsynlighet for å velge mest sannsynlige analysen
- Trebanker sentrale

Trebanker

- Korpus manuelt annotert med syntaktisk struktur:
  $\Rightarrow$ en **trebank**
- Penn Treebank: trebanker fra Brown, Switchboard, ATIS og *Wall Street Journal* korpusene
- Trebanker for andre språk:
  - Prague Dependency Treebank (tsjekkisk)
  - Negra (tysk)
  - Penn (kinesisk)
  - Norwegian Dependency Treebank (norsk)

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

- Kontekstfri grammatikk (CFG)
  - formelt regelsystem: konstituenter, hierarkisk gruppering, lineær rekkefølge
  - ved **derivasjon** definerer vi de de grammatiske setningene i henhold til en grammatikk
  - tillater **rekursjon** (direkte og indirekte)
  - problem: regeleksplosjon ved eksempelvis samsvarbøyning og seleksjon

Oppsummering

- Funksjonell analyse
  - annet aspekt ved syntaks: **funksjon**
    - analyse av primære setningsledd som predikat, subjekt og objekt
- Syntaks i språkteknologi
  - viktig skritt mot semantisk tolkning
  - **Chunking**: "fattigmannssyntaks"
    - analyserer ikke-rekursive fraser
    - nyttig for eksempelvis informasjonsekstraksjon
  - Syntaktisk parsing: automatisk syntaktisk analyse