

Syntax och parsning

Syntaktiska teorier och syntaxformalismen
Fullständig syntaxanalys (klassisk parsning)
Ytsyntaktisk analys
Grammatisk analys med dependenser

Mål med grammatik

- Att definiera ett språk med ett regelsystem
- Att tilldela varje mening i språket en eller flera grammatiska strukturer
 - Om en mening kan härledas med grammatikens regler sägs grammatiken generera meningen.

Översikt

Teorier: Government-Binding Theory
X-bar, Minimalism, Cognitive grammar
Formalism: DCG, PATR, CG, Finite-state-grammar
Teori med integrerad formalism:
GPSG och LFG
System: Maskinöversättning,
grammatikkontroll m.fl.

Chomsky

Noam Chomsky (1928-)
Grammatik – ett mentalt organ, ett färdigutvecklat språkanlag
Studierna av strukturella egenskaper
+formell beskrivning ger
karaktärisering av denna förmåga
Grammatik – relationen mellan en talares språklig kompetens och performans.

Generativ grammatik

Mål: Hitta de grammatiska regler som genererar ett språk

- Produktion och tolkning av språkliga satser involverar successiva tillämpningar av omskrivningsregler.
- En talare av ett språk måste känna språkets grammatik.
- Grammatiken definierar vilka strängar som tillhör språket eller inte.

Grammatikalitet och acceptabilitet

- Grammatisk och acceptabel: *Det regnar.*
 - Grammatisk men oacceptabel: *Mannen gav stenarna bumlingarna.*
 - Ogrammatisk men acceptabel: *Jag såg han.*
 - Ogrammatisk och oacceptabel: *Mannen ser sten bumlingarna.*
- Problem:** att skilja ut oacceptabla meningar från ogrammatiska meningar.

Svensk ordföljd – rätt eller fel bland 24 teoretiskt möjliga satsor

- a. Rune köpte sin nya klocka i lördags.
- b. Rune köpte i lördags sin nya klocka.
- c. I lördags köpte Rune sin nya klocka.
- d. Sin nya klocka köpte Rune i lördags.
- e. Sin nya klocka köpte i lördags Rune.
- f. Köpte Rune sin nya klocka i lördags?
- g. Köpte Rune i lördags sin nya klocka?
- h. Köpte i lördags Rune sin nya klocka?
- i. Rune sin nya klocka köpte i lördags.
- j. Rune i lördags köpte sin nya klocka.
- k. Rune sin nya klocka köpte i lördags.
- l. Rune i lördags sin nya klocka köpte.
- m. I lördags Rune köpte sin nya klocka.
- n. I lördags Rune sin nya klocka köpte.
- o. I lördags sin nya klocka Rune köpte.
- p. I lördags sin nya klocka köpte Rune.
- q. I lördags köpte sin nya klocka.
- r. Sin nya klocka Rune köpte i lördags.
- s. Sin nya klocka Rune i lördags köpte.
- t. Sin nya klocka i lördags Rune köpte.
- u. Sin nya klocka i lördags köpte Rune.
- v. Köpte sin nya klocka i lördags Rune?
- x. Köpte i lördags sin nya klocka Rune?
- y. Köpte sin nya klocka Rune i lördags?

Syntax och semantik

Semantiska faktorer påverkar den syntaktiska beskrivningen –för att analysera något syntaktiskt måste man veta något om omvärlden.

Syntax ett steg på vägen mot semantik
Skall beskrivningarna skiljas åt? Ja, syntax – relationer mellan ord och fraser, semantik – ordens och frasernas relationer till omvärlden

Problem: flertydighet

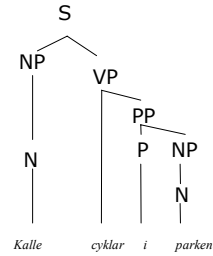
Ordens flertydighet:

Svenska: 60 %

Engelska: 45 %

Finska: 15 %

+ andra flertydigheter: konstituenten, huvuden



Konstituentstruktur

- Konstituenten: grupper av ord kring ett huvudord
den röda hundens tofflor
besegrades vid Waterloo
deras gröngulblåa egendomliga egenskaper
en hund
- Huvudord: det centrala elementet i en fras
tofflor, besegrades, egenskaper

N (Nomen), V (Verb), (P) prepositioner, (A) Adjektiv

Nominalfraser (NP), Verbfraser (VP)

Prepositionsfraser (PP), Adjektivfraser (AP)

Argument för konstituentstruktur

- uppträder i liknande kontexter:
en hund sågs i parken
i parken sågs en hund
sågs en hund i parken
- enskilda ord kan inte ersätta konstituenten
*hund sågs i parken
*i parken sågs en
- konstituenten kan flyttas men inte brytas upp
i parken sågs en hund
sågs en hund i parken
*i hund en parken sågs
*hund sågs i parken en

Argument mot konstituentanalys

Flertydighet mellan konstituenterna är stor och när de sätts ihop ökar flertydigheten igen
Ganska mycket talar för att ordbaserade system är både enklare att konstruera och effektivare att parse.

En generativ grammatik:

S	→ NP VP	NP	→ kalle
S	→ VP	NP	→ regnet
VP	→ V	V	→ springer
VP	→ V PP	P	→ i
PP	→ P NP		

Kalle springer i regnet.

Springer i regnet.

?Regnet springer i Kalle

Vad menas med syntaxformalism?

- Ett språk för att formalisera grammatiska regler.
- Ganska ofta med fokus på:
 - en given syntaktisk teori (ofta knepiga lingvistiska relationer), några kärnmeningar, sällan något omfattande textmaterial.
- Representera lingvistisk kunskap i ett system
- För att tilldela indata en strukturell analys

Syntaktiska teorier/formalismer

- Transformationsgrammatik (Chomsky et al)
- HPSG (Pollard & Sag)
- Tree Adjoining Grammar
- Dependency Grammar och Constraint Grammar
- Finite-state grammars

Parsning

Att utföra bästa möjliga (av grammatiken tillåten) analys av ett språkligt påstående.

Problem:

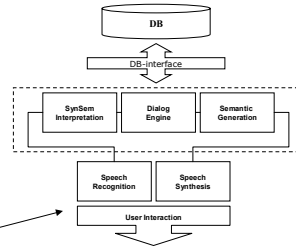
- Språkets tvetydighet

Generering

- Genereringsproblemet är parsningsproblemet motsats: Att från en logisk representation skapa ett påstående i naturligt språk.
- Påståendet måste givetvis ha samma innebörd som den logiska formen.
- "tvåvägsgrammatikor" = grammatikor som kan användas både för analys och generering.

Parsning användningsområden

- Maskinell syntaktisk analys av språk
- Mellansteg i många tillämpningar
- Semantisk analys
 - Maskinöversättning
 - Grammatikkontroll
 - Frågebesvarande system
 - Dialogsystem
 - Informationsextraktion
 - Textsummering



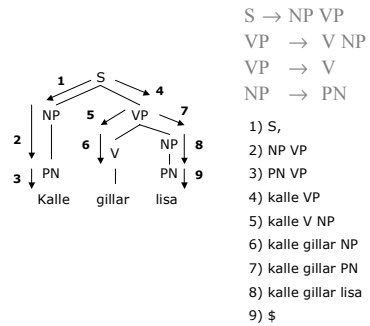
Top-down parsing

- Mål-driven (goal-directed) sökning
- Utgår från toppnoden
 - Toppen-ner, djupet-först
 - Toppen-ner, bredden-först

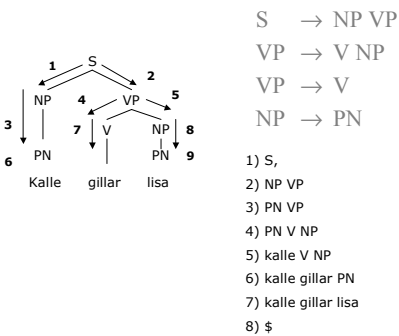
Bottom-up parsing

Data-driven sökning
Utgår från strängen (data)
Bottén-upp, bredden-först

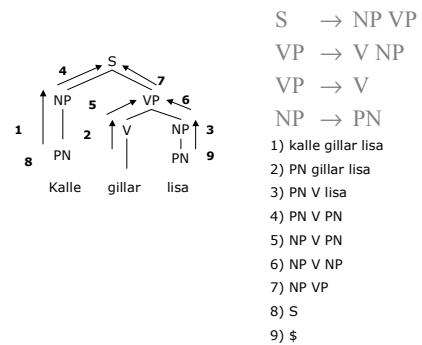
Toppen-ner, djupet-först



Toppen-ner, bredden-först



Botten-upp, bredden-först



Flertydighet

Olika typer av flertydigheter:

Flertydig lexikal kategori: Ordklasstagning med disambiguering.

Strukturell flertydighet: vi får flera parsetråd till en sats.

PP-attachment: *Jag ser mannen med kikaren.*

Han äter pizza med gaffel. Jämför:

Han äter pizza med ansjovis.

Coordination: *Hunden eller katten äter fisken och köttet. Jämför: Hunden äter köttet och katten äter fisken.*

Autentiskt exempel

Ett förhållande som stark försvårar personvalet till riksdagen för kandidater som är bosatta inom en valkrets är när partier, utöver en lista som upptar deras namn, går fram med ytterligare en lista med samma namn i samtliga valkretsar i riket.

Två relativsätser ...

Multipla prepositionsfraser

Ineffektiv parsning av delträd

Toppen-ner parsern bygger upp giltiga träd för delar av indata som sedan förkastas, för att sedan byggas upp igen när parsern försöker bygga upp ett nytt träd ...

En lista med samma namn i samtliga valkretsar i riket

Fras:	Antal analyser:
En lista	4
med samma namn	3
i samtliga valkretsar	2
i riket	1
En lista med samma namn	3
En lista med samma namn i samtliga valkretsar	2
En lista med samma namn i samtliga valkretsar i riket	1

Parsningssystem för svenska

- SLE- Swedish Language Engine (Gambäck)
- Uppsala Chart Parser (UCP) (Sågvall-Hein)

Del II Ytsyntaktisk analys

- Taggmönster
- Ordklasstagning
- Finite-state parsning
- Constraint grammar, functional dependency grammar
- Parsning av ogrammatisk indata

Problem med fullständig parsning

Flertydigheten gör att vi får många syntaxträd
Hur skall man välja ut det bästa trädet?
Många meningar får ingen analys alls?
Det tar för mycket tid och resurser?
Behöver vi en fullständig syntaktisk analys?

Robust parsing

En robust parser skall ge en korrekt eller användbar analys av 90 % av meningarna (Briscoe). Åtminstone följande tre problem måste lösas:

1. Chunking, vad är en mening och en sats?
2. Disambiguation, att välja rätt analys bland alla syntaktiskt möjliga.
3. Undergeneration, eller att hantera fall som ligger utanför systemets lexikala och syntaktiska täckning.

Val: teoretisk korrekthet eller effektiv tillämpbarhet?

Taggmönster som en enkel syntaktisk struktur

- Min bror väntade på flygplatsen
PS + NN + VB + PP + NN
- Min flygplatsen väntade på bror
PS + NN + VB + PP + NN

Taggmönster med särdrag

PS + NN.IND + VB + PP + NN.DEF

Godkänner: *Min bror väntade på flygplatsen.*
Godkänner inte: * *Min flygplatsen väntade på bror.*

Ordklasstaggning

Statistisk (Viggos föreläsning) eller regelbaserad (Morp, Constraint grammar)
Maskininläring: Brills taggare

Det var en man som stod där.
PN VB DT NN HP VB AB
DT NN NN PN HP VB AB
Löser ytsyntaktisk flertydighet
Bra som förprocess

```

"<min>"
"min" <*> N UTR INDEF SG NOM
"min" <*> <POSS-SG1> <MD> DET UTR DEF SG NOM (@DN>)
"min" <*> <POSS-SG1> PRON UTR DEF SG NOM
"min" <*> <MEASURE> ABBR UTR/NEU INDEF SG NOM

"<glada>"
"glad" A UTR/NEU DEF SG NOM
"glad" A UTR/NEU DEF/INDEF PL NOM
"glada" N UTR INDEF SG NOM

"<glada>"
"glad" A UTR/NEU DEF SG NOM
"glad" A UTR/NEU DEF/INDEF PL NOM
"glada" N UTR INDEF SG NOM

"<satt>"
"sitta" V ACT PAST
"sitta" V ACT SUPINE
"sitta" V PCP2 UTR/NEU INDEF SG NOM
"satt" A UTR/NEU INDEF SG NOM

"<på>"
"på" ADV "på" PREP

"<en>"
"en" <NUM/ART> <ID> DET UTR INDEF SG NOM (@DN>)
"en" <NUM> PRON UTR INDEF SG NOM
"en" N UTR INDEF SG NOM "en" ADV (@AD>)

"<för>"
"föra" V ACT PRES
"föra" V ACT IMP
"för" N UTR INDEF SG NOM
"för" <CLB> CC
"för" ADV (@AD> @ADV)
"för" PREP

```

Språklig ytanlys med regler

Några olika system för:

- Ordklasstagning
- Dependensgrammatik

Constraint Grammar presenteras i detalj

Ordklasstagning med funktionsordslexikon

Gunnel Källgrens MorP parser

Den ytliga informationen i en text underskattad

Minimalt lexikon med funktionsord

Mönstermatchning av morfologiska och ytsyntaktiska ytmönster

Märker orden med ordklass, känner igen vissa NP och PP.

Jabberwocky (Källgren, 1992)

De iggla skviggarna trassade vombigt i den harliga gopen.

- Hurk, najdade den ena skviggen, maffar pem en bunne?

- Snå, bebbade den umre skviggen aldrigt. Snafs på nerfen.

- Lytrik snafs det?

- Pej gemmer det fraskar för kloxigt.

Så vatrik jeggade skviggen snafen bunne.

Brills tagger

Maskininlärning

Det behövs en taggad korpus

Ord-taggar-frekvenser

Korrigeringsregler

Finite-state parsing I

Många språkssystem behöver inte "fullständig" syntaktisk analys.

Partiell parsning räcker ofta ganska långt ((partial | shallow | light) parsing)

Några vanliga tillämpningsområden är informationsextrahering och även grammatikkontroll (mer senare).

Informationsextrahering behöver ta fram begränsat med information för att fylla i olika typer av informationsmallar, t.ex. företagsnamn, personer och tid och plats, kanske också vem som gjorde vad mot vem.

PG → Gyll

P G Gyllenhammar lämnar sin post som VD för Volvo. Ny VD blir Sören Gyll.

Position: VD

Company: Volvo

In-Person: Sören Gyll

Out-Person: P G Gyllenhammar

Finite-state parsing III

- Nomengrupper: en glad gäst, en hund, han, hunden

NG --> PN | (DT) (JJ) NN

- Verbgrupper: har spelat, att spela
- VG --> VB | AUX VB | IE VB

Dependensgrammatik

- Tesnière (1959), Mel'cuk (1988)
- Grammatiska relationer: Subjekt, finit verb (main), och objekt viktiga, huvud och modifierare
- Mycket generella semantiska relationer

Constraint grammar

Ett språkoberoende system för att analysera löpande text

Fred Karlsson et al Helsingfors Universitet

All indata skall ges någon analys

- Parseern skall kunna analysera obegränsad text (unrestricted text) – den skall vara robust.
- Tar inga beslut om vad som är grammatiskt – ogrammatiskt
- I praktiken måste man ändå satsa på att hantera de ord och meningar som kan anses tillhöra språket (Master lexicon, Core Lexicon)
- Heuristiska metoder

Grammatikformalismen skall vara språkoberoende

- Grammatikformalismen skall inte utgå från något speciellt språk.
- Formalismen skall kunna rymma många olika språk från olika språkfamiljer.
- Grammatikformalismen måste vara separerad från programkoden för att vara språkoberoende.

Anpassning till olika texttyper

- Hur stort lexikon man än har så kommer det alltid att saknas ord
 - Egennamn, Datatermer
- Vissa konstruktioner finns endast i vissa texter
- Tillåta att ha möjlighet att anpassa systemet till texttypen
 - Domänlexikon
- Gissare för okända ord

Förprocessning är ett viktigt steg

Tokenisering
Idiom-igenkänning
Stavningskontroll

Utdata skall vara läsbart och inte
nersölat med ett ohanterbart antal
analysalternativ.

Många system reducerar inte
tolkningsmöjligheter utan
introducerar istället nya möjliga
tolkningar.

En kärna av regelbaserade restriktioner

- Men det skall finnas möjlighet till probabilistiska (alt. heuristiska) villkor "på toppen" på ett lingvistiskt sätt
- Korpusstudier viktiga för att formulera regler – basera grammatiken på hur det verkligen ser ut.
- Det är en styrka om så många regler som möjligt är icke-heuristiska
- Det är upp till grammatikern att använda heuristik och ta risker (kan vara nödvändigt för vissa språk/tillämpningar).

Morfologisk och lexikal analys är grunden

Full lexikal täckning: alla böjningsformer,
basformsreduktion, hantering av sammansättningar
(ej lexikaliserade)

TWOL

Minst 30 000 lexem (kärnvokabulären)

ENGTWOL: 56 000 SWETWOL: 48 000

Morfologisk heuristik

Flertydighet är kärnproblemet

Gazdar-Mellish (1989): Flertydighet är det
enskilt största problemet inom NLP.

CG vill lyfta fram beskrivningen av
flertydighet.

CG är i grunden ett system för att skriva
disambigeringsregler

CG bryter upp parsningsproblemet i tre delar

- Morfologisk disambiguering
- Bestäm satsgränserna inom en mening
- Ytsyntaktisk disambiguering

Disambiguering skall ske på alla tre delproblemen

- Utgå från värsta fallet: Tilldelning av alla möjliga tolkningar.
- Målet med varje restriktion är att förkasta så många alternativ som möjligt.
- En lyckosam parsning gör meningen morfologiskt och syntaktiskt otvetydig

Den syntaktiska beskrivningen

- Varje ord skall tilldelas en ytsyntaktisk funktion
- Beskrivningen ger också grundläggande dependensrelationer inom satsen och meningen.
- En platt representation (inget träd).

Constraint grammar parsing Förinställda steg:

1. Satsgränsmappning
2. Kontextkänslig morfologisk disambiguering
3. Satsgränsmappning (ny kan ha uppstått)
4. Kontextkänslig morfologisk disambiguering
5. Satsgränsmappning
6. Morfosyntaktisk mappning
7. Syntaktisk analys

Regelsyntax

Morfosyntaktisk mappning:
<matchningsankare contextvillkor
syntaktiska_taggar>
Ex: ((N) ((-1C PREP) (1 <<<)) (@<P))

N tilldelas den syntaktiska funktionen
prepositionskomplement om N föregås av
en preposition och följs av meningsslut.

Syntaktiska regler

- =s! väljer en tolkning och förkastar övriga
 - =s0 förkastar en tolkning sparar övriga
- Ej matchning av ordformer
@w är standard – vilket ord som helst
Jag spelar fotboll
(@w =s! (@SUBJ) (1 VFIN) (1 ACTIVE)
(NOT *-1 @SUBJ) (NOT *-1 @SUBJ))

1. Förprocessning

*Dessa entreprenöriella faktorer hade än så länge
dämpat explosionen*

*dessa
entreprenöriella
faktorer
hade
än_så_länge
dämpat
explosionen
\$.

2. Morfologisk analys

"<*dessa>" "denna" <**c> <DEM> <MD> DET UTR/NEU
DEF PL NOM @DN>
"denna" <**c> <DEM> PRON
UTR/NEU DEF PL NOM
"<entreprenöriella>"
"<faktorer>" "faktor" N UTR INDEF PL NOM
"<hade>" "ha" <AUX> V ACT PAST
"<än_så_länge>" "än_så_länge" <COLLOCATION> ADV
"<dämpat>" "dämpa" V ACT SUPINE
"dämpa" <PCP2> A NEU INDEF SG NOM
"<explosionen>" "explosion" N UTR DEF SG NOM ”
<\$.>" "\$." CLB <PUNCT>

3. Morfologisk heuristik

- Analys av ord som inte analyseras av SWETWOL
- 60 regler baserade på ordformer
- "<entreprenöriella>" "entreprenöriella" <NON-SWETWOL> A UTR/NEU DEF SG NOM
"entreprenöriella" <NON-SWETWOL> A
UTR/NEU DEF/INDEF PL NOM
- Saknas regel tilldelas ordet substantiv-taggar

4. Morfologisk disambiguering

- Över 50 % av orden i svenskan är flertydiga (SVD, 1.9 milj. ord. SWETWOL-analys)
 - 1.82 tolkningar/ord i medeltal
- Testresultat:** 4631 ord (4304 full disambiguering)
Täckning: 99.57 %
Precision: 95.36 %
- Utvärderingsproblem:** för lite testdata, vad som är en korrekt tolkning avgjordes av regelskrivaren
- Förbättringar:** Allt som går att göra är inte gjort, framförallt behövs det en större utvecklingskorpus i nuläget 120 000 ord.

"<*dessa>" "denna" <**c> <DEM> <MD> DET UTR/NEU
DEF PL NOM @DN>
"<entreprenöriella>" "entreprenöriella" <NON-SWETWOL>
A UTR/NEU DEF/INDEF PL NOM
"<faktorer>" "faktor" N UTR INDEF PL NOM
"<hade>" "ha" <AUX> V ACT PAST
"<än_så_länge>" "än_så_länge" <COLLOCATION>
ADV "<dämpat>" "dämpa" V ACT SUPINE "<explosionen>"
"explosion" N UTR DEF SG NOM ”
<\$.>" "\$." CLB <PUNCT>

5. Morfosyntaktisk mappning

Huvudsyfte: Lägga på den syntaktiska flertydigheten.

Syntaktiska huvudtyper:

- verbfunktion – huvudverb, verbkedjeverb, finita/infinita verb
- subjekt, objekt och adverbial (ord med huvudfunktion i satsen)
- modifierare – bestämmningar till huvudorden

"<*dessa>" "denna" <**c> <DEM> <MD> DET
UTR/NEU DEF PL NOM @DN>

"<entreprenöriella>" "entreprenöriella"
<NON-SWETWOL> A UTR/NEU DEF/INDEF PL
NOM @AN> @SCOMP @OCOMP @SUBJ @OBJ @IOBJ
@AOBJ @ADVL @AD> @P>> @<P @<N @NPHR

"<faktorer>" "faktor" N UTR INDEF PL
NOM @SUBJ @OBJ @IOBJ @AOBJ @SCOMP
@OCOMP @ADVL @NN> @P>> @<P @<NN @NPHR

Syntaktisk disambiguerining

- inte fullt utvecklad
- 400 regler finns i nuläget
- Det behövs 3-4 * 400 regler för att lösa upp alla tvetydigheter
- Inga resultat finns att tillgå eftersom denna del inte är fullt utvecklad
- Denna del används inte i grammatikkontrollen i Word 2000

Functional Dependency Grammar

Functional Dependency Grammar
(Tapanainen & Järvinen, 1997)

Syntaxträd, länkar mellan ord, inte konstituentier

Ca 35 dependensrelationer

Grammatiska funktioner

Huvud – modifierare

Enkla semantiska relationer

Functional dependency parser för svenska:

www.conexoroy.com/products.htm

0				
1	The	the	det:>2	@DN> DET SG/PL
2	dog	dog	subj:>3	@SUBJ N NOM SG
3	sniffs	sniff	main:>0	@+FMAINV V PRES SG3
4	at	at	loc:>3	@ADVL PREP
5	the	the	det:>6	@DN> DET SG/PL
6	meat	meat	pcomp:>4	@<P N NOM SG
7	and	and	cc:>3	@CC CC
8	the	the	det:>9	@DN> DET SG/PL
9	cat	cat	subj:>10	@SUBJ N NOM SG
10	sniffs	sniff	cc:>3	@+FMAINV V PRES SG3
11	at	at	loc:>10	@ADVL PREP
12	the	the	det:>13	@DN> DET SG/PL
13	fish	fish	pcomp:>11	@<P N NOM SG/PL