Süntaksiteooriad ja -mudelid
MTAT.06.031

4 AP

Kaili Müürisep
kaili.muurisep@ut.ee

2007/08 kevad

Tänane loeng

Contents

1 Ülevaade kursusest

Eesmärgid

- Eesmärk: Saada ülevaade tähtsamatest süntaksiteooriatest. Õppida kasutama ja testima loomuliku keele süntaksianalüsaatorit nii reeglipõhiseid kui statistikapõhiseid meetodeid kasutades.
- Kirjeldus: Põhilised süntaksiteooriad - fraasistruktuurigrammatika ja transformatsioonigrammatika. Chomsky grammatikate hierarhia. Loomuliku keele süntaksianalüüs reeglipõhiste ja statistiliste meetodite abil. Peajuhitav fraasistruktuurigrammatika (HPSG). Leksikaalfunktsionaalne grammatika (LFG). Sõltuvusgrammatika. Kitsenduste grammatika. Reeglipõhise parseri kirjutamine. Süntaktiliselt märgendatud tekstikorpused. Puude pangad (treebank). Puude pankade loomise ja kasutamise tarkvara. Statistikapõhise parseri treenimine süntaktiliselt märgendatud korpusel.

Ligikaudne ajakava

Veebruar Põhimõisted. Teooriate liigitus. Kontekstivabad grammatikad. Definite Clause Grammar. GB.

Märts Unifikatsioonigrammatikad. HPSG.

Aprill LFG. CG. TAG

Mai Sõltuvusgrammatikad. Lõplikud automaadid. Puudepangad. Mudelid.

Juuni Eksam.

Lõpp

- Eksam 50 p

  #### Näiteküsimused

  1. Koostage fraasistruktuurigrammatika, mille abil saaks analüüsida lauset Ilus pikk poiss läks üle silla.
  2. Tooge eestikeelne oma näide lihtlausest, milles kaassõnafraasi kuuluvus on mitmene, joonistage vastavad puud.

3. Joonistage suluehituse põhjal puu või vastupidi.
4. Joonistage Lause Peeter kõditab Mari süntaksipuu kasutades SUBCAT tunnust.
5. Kirjutage fraasi sild üle jõe tunnusstruktuur
6. Kirjutage lause Peeter sõi jäätist f-struktuur (LFG)
7. Kirjeldage keelt, mida tunneb antud automaat.
8. Leidke väljundsõne, kui on antud muundur ja sisendsõne.
9. jne jne

- Referaat 10 p

  Referaadi teemad (vanad)

  1. HPSG rakendused praktikas.
  2. LFG rakendused praktikas
  3. Sõltuvusgrammatikate rakendused praktikas
  4. Saksa keele süntaktiline analüüs.
  5. Skandinaavia keelte süntaktiline analüüs.
  6. Ungari keele süntaktiline analüüs.
  7. Süntaksianalüsaatori hindamise kriteeriumid.
  8. Ülevaade fraasistruktuuripuude pankadest
  9. Ülevaade sõltuvuspuude pankadest.
  10. Suulise keele süntaktiliseks analüüsiks kasutatavad meetodid.

- Praktikumide ülesanded 40 p

Koduleht
http://math.ut.ee/~kaili/Loengud/Mudelid08/

Kirjandus

1. Fred Karlsson. Üldine keeleteadus. ptk 1.3.4.; 5.
2. Kaisa Häkkinen. Keeleteaduse alused. ptk 14-16.

2 Sissejuhatus

Põhimõisted

Süntaksi ehk lauseõpetuse uurimisobjekt on lausete ehitus: see, millisest väiksematest osadest laused koosnevad, missugused on osade ülesanded ja omavahelised suhted, kuidas lauseosi saab omavahel ühendada. (Fred Karlsson. Üldkeeleteadus)

Süntaksiteooria - kooskõlaline kogum põhimõtteid ja printsiipe, mis kehtivad keele lausete moodustamisel.

Süntaksimudel - mingi süntaksiteooria formaliseering matemaatiliste struktuuride abil.

Lause

- Lause on grammatika ehk keele struktuurilise organiseerituse suurim üksus.
- Minimaalne täielik lause koosneb finiitverbist ja selle juurde kuuluvatest nominaalsetest liikmetest, mida tavaliselt on vähemalt üks.
- Traditsioonilises grammatikas nimetatakse nominaalseid lauseliikmeid nende funktsiooni järgi: alus, sihitis, määrus jne

Lauseliigid

- Lihtlause

  Poiss jookseb. Rabin kohtus Arafatiga. Tüdruk vajutas nuppu. Ruum oli tühi. Eestis süüakse tonnide kaupa viinereid.

- Rindlause

  Poiss jooksis ja tema sammude müdin kostis kaugele.

- Põimlause

  Epideemia algas siis, kui külas lõhkes kanalisatsioonitoru.

Kirjalik või suuline keel

- Kirjalikus keeles peetakse lauseks ühest või mitmest osalausest koosnevat üksust, mis lõpeb punkti, hüüumärgi või küsimärgiga.
- Suulise keele põhiüksuseks peetakse lausungit, mis on intonatsiooniline tervik.

Dialoog
A: see oli kõik ee ausatel eesmärkidel et perele auto
B: ei no loomulikult
A: saada. aga lissalt mai saand seda autot

Grammatika

- keele ehituse süsteemipärane esitus
- õpik ehk reeglite kogu
- keele süsteem

Kogu keelesüsteem koosneb fonoloogia, morfoloogia, lekskikoni, süntaksi ja semantika allsüsteemidest.

Grammatika - reeglite süsteem, mis määrab keele struktuuri.

Jaguneb:

- kollektiivne
- indiviidi poolt esimese keele omandamise käigus alateadlikult vastuvõetav (grammatiline pädevus)

Noam Chomsky on süntaksi eesmärgiks pidanud sellise kirjelduse andmist, mis genereeriks kõik keele grammatikale vastavad laused ega genereeriks ühtegi mittegrammatilist lauset.

2.1 Fraasistruktuur

Lause moodustajastruktuur

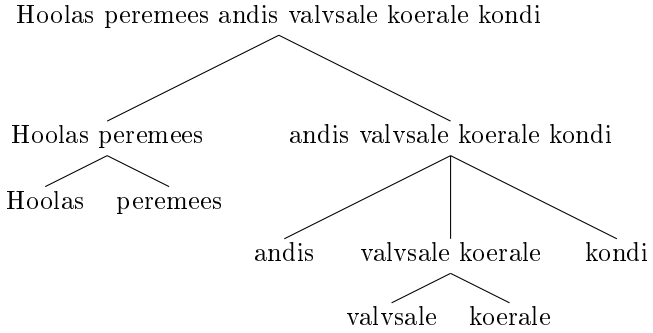
- Lausetel on hierarhiline struktuur.
- Moodustajate leidmiseks kasutatakse substitutsiooni ja permutatsiooniteste.

Moodustajate analüüs põhineb näitelausete osadeks jagamisel.

Hoolas peremees andis valvsale koerale suure kondi.
Hoolas / peremees andis valvsale koerale suure kondi.
Hoolas peremees / andis valvsale koerale suure kondi.
Hoolas peremees andis / valvsale koerale suure kondi.

Moodustajastruktuur

Lause hierarhiline struktuur avaldub igal tasandil **vahetute moodustajatena**.



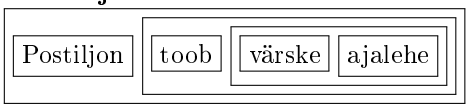
Konstruktsioonid ehk tarindid

- endotsentriline konstruktsioon: üks moodustajatest on põhi ja käitub lauses nagu kogu konstruktsioon.
  - Õues haukus väga tige peni.
  - Õues haukus tige peni.
  - Õues haukus peni.
- eksotsentriline konstruktsioon: osad on terviku võrdsed komponendid. nt kaassõnafraas, hulgafraas, alus-öeldis.

– Ta hüppas üle aia
– *Ta hüppas aia
– Ta hüppas üle

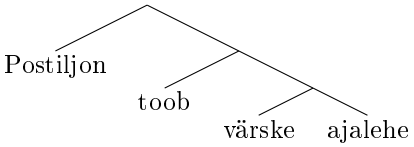
– Toas on kolm lauda
– *Toas on kolm
– *Toas on lauda

Moodustajate hierarhia esitusviisid



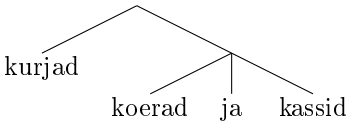
[[Postiljon] [[toob] [[värske][ajalehe]]]]

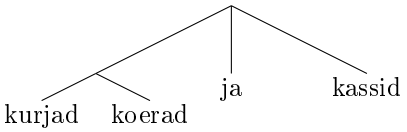
Aga
Did you go?



Struktuurilised mitmetitõlgendatavused

Millised veel?





Fraasid

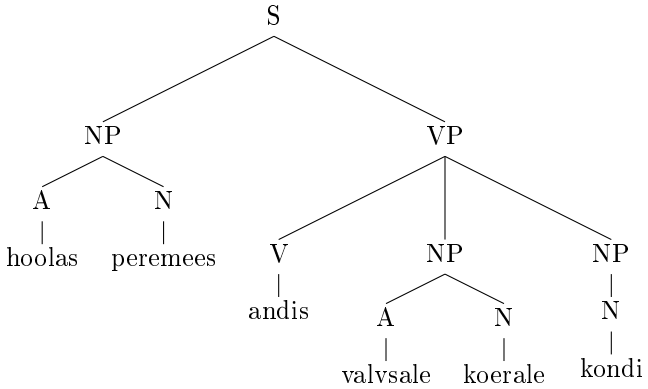
Puudes on nelja liiki tippe:

- sõnad lehtedes
- sõnaliik e kategooria
- fraasid
- lauset tähistav S juurtipus

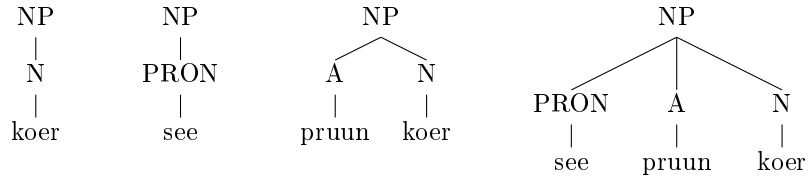
Fraasitüübid

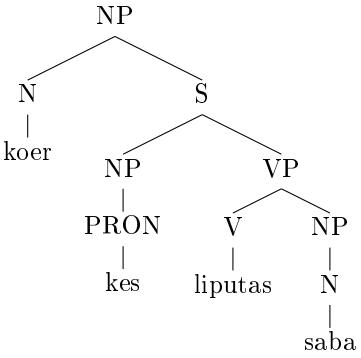
N	NP	nimisõnafraas	noun phrase
A	AP	omadussõnafraas	adjective phrase
ADV	ADVP	määrsõnafraas	adverb phrase
P	PP	kaasssõnafraas	prepositional phrase
V	VP	tegusõnafraas	verb phrase

Fraasistruktuuripuu

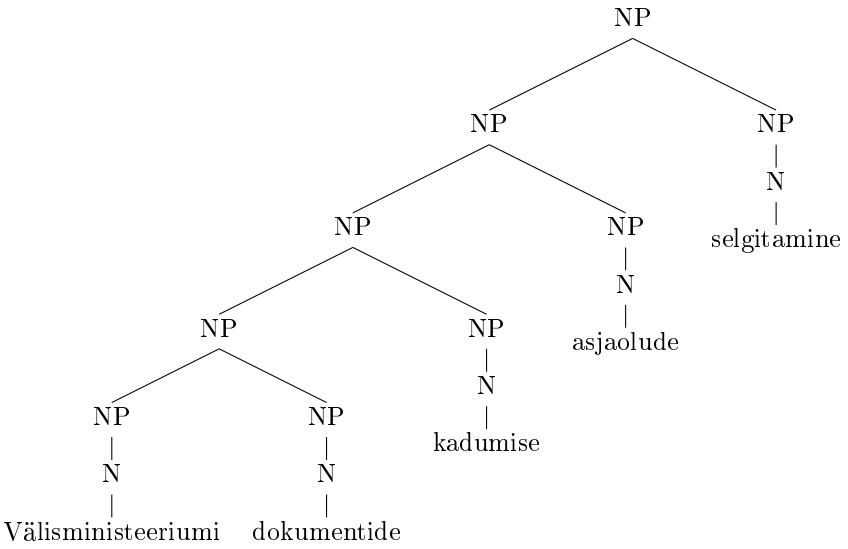


Nimisõnafraas





Vasakule hargnev puu



Teised fraasid

AP üsna vana koer

PP üle minu laiba

PP vana maja taga

ADVP väga hilja

QP palju koeri

VP on tahtnud

VP oleks võinud tahta ujuma minna

Fraasitüüpide omadused

- asendatavus - nt fraasi asendamine asesõnaga
- nihutatavus - kogu fraasi asukoha vahetus lause sees
- osade lahutamatus - võimaluse puudumine viia fraasi sisse struktuuri mittekuuluvaid moodustajaid
- grammatilised protsessid - käändemärgistus ja rektsiooninähud

NP

- Valvas koer liputas saba - Ta liputas saba
- Saba liputas valvas koer
- See koer liputas eile saba - *See eile koer liputas saba
- eeslaiendi käände ja arvuühildumine

Mittesidusad fraasid

Aga
Ladina keeles paiknevad adjektiivsed laiendid lauses põhisõnast eraldi, eriti kui nad on rõhutatud.

hoc	in	loco
see+ABL	sees	koht+ABL
'selles kohas'		

Eesti keele VP omadused

- asendatav
- fraasisisesed grammatilised protsessid töötavad: sihitise ja osade määruste kääne sõltub verbist
- ei liigu lauses tervikuna:

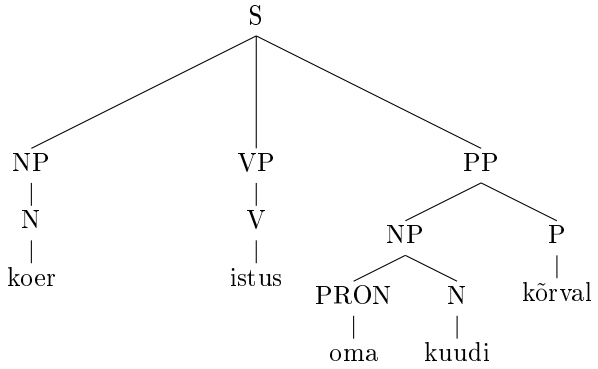
  Postiljon toob iga päev värske ajalehe. Iga päev toob postiljon värske ajalehe.
- eesti VPd on võimalik lõhkuda:

  Ohver jõi ilmselt palju õlut

Eesti keele grammatikas käsitletakse verbifraasidena ainult:

- ahelverbe ehk verbi ühendeid infiniitsete verbivormidega (*hakkas sööma*)
- ühendverbe (*jooksis ära*)
- väljendverbe (*laskis jalga*)

Verbifraasiharuta puu



Fraasid kui sõnaliikide projektsioonid

- Fraasid jaotatud:
  1. alumine tasand - teatud sõnaliiki kuuluv sõna: N, A, PostP
  2. kõrgem tasand - NP, AP, PP

- Kas vahepealsed npd erinevad kuidagi alumise ja kõrgeima tasandi npdest?
- Äkki on mõistlik neid märgistusega eristada? N, N', N"
- vahepealsed npd erinevad maksimaalsest oma omaduste poolest, näiteks ei ole lubatud järeltäiendid - tähistatakse N'

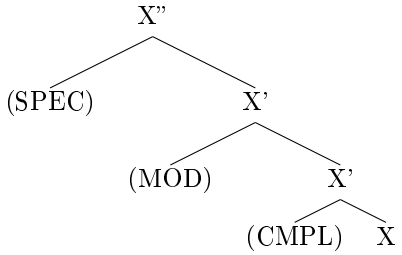
  Täiendi järeltäiend

  - *kilo võid hind
  - *kavatsuse, et võiksime minna kohe, otstarbekus

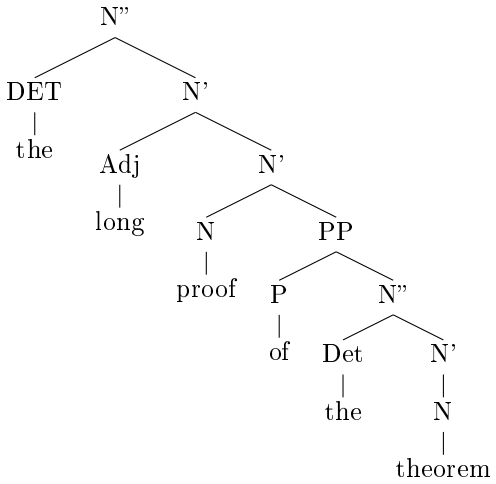
- Maksimaalne projektsioon on N"

X-prim-teooria

Sama põhimõte kehtib ka teiste sõnaliikide kohta.
X viitab mis tahes sõnaliigile (N, A, V, P), primm struktuuritasandile.



X' näide



Fraasistruktuurigrammatika

Asendusreeglid
$S \rightarrow NP\ VP$
$NP \rightarrow AP\ NP$
$NP \rightarrow N$
$AP \rightarrow A$
$VP \rightarrow V\ NP$

$VP \rightarrow V\ NP\ NP$
Leksikon
$N \rightarrow$ *{peremees, koerale, kondi}*
$A \rightarrow$ *{hoolas, valvsale}*
$V \rightarrow$ *{andis}*

Kontekstivaba grammatika

- noolest vasakul on ainult üks sümbol
- paremal pool võib olla sümboleid üks või rohkem
- Ainus operatsioon, mida kasutatakse sõne moodustamiseks alamsõnedest, on konkatenatsioon.
- Süntaktilistest seostest kirjeldatakse ära ainult fraasistruktuur
- Terminaalsetel sümbolitel pole omadusi
- Mitteterminaalid on atomaarsed, s.t neil puudub sisemine struktuur
- Puudub otsene võimalus siduda süntaksit ja semantikat

Need teised grammatikad

- 0-tüüpi grammatika - piiranguid reeglitele ei ole.
- 1-tüüpi grammatika e kontekstitundlik grammatika: $\alpha\ A\ \beta \rightarrow \alpha\ \psi\ \beta$
- 2-tüüpi grammatika e kontekstivaba: $A \rightarrow \psi$
- 3-tüüpi grammatika: $A \rightarrow xB$, $A \rightarrow x$ või $A \rightarrow Bx$, $A \rightarrow x$

ID/LP reeeglid
Asendusreeglid kannavad kahte liiki informatsiooni:

- Immediate dominance ehk vahetud moodustajad $PP \rightarrow NP\ P$ postpositsioonifraas koosneb nimisõnafraasist ja tagasõnast
- Linear precedence ehk lineaarne järjestus nimisõnafraas paikneb tagasõna ees
- ID/LP reeglites hoitakse see info lahus

ID/LP reeglid

- $NP \rightarrow N'$,*(DET)*
- $N' \rightarrow N'$,PP
- $N' \rightarrow N'$, QP
- $N' \rightarrow N'$, $AdjP$
- $N' \rightarrow N$, *(PP)*
- $[DET, QP, AdjP] > Head > PP$

Kas KVG sobib?

- ei suuda katta kõiki loomulike keelte nähtusi: ühildumine, verbi subkategorisatsioon jne
- ei ole leksikaliseeritav
- vaja midagi muud

2.2 Sõltuvusstruktuur

Traditsiooniline lauseliigendus

Traditsioonilises grammatikas kirjeldatakse lausestruktuuri lauseliikme mõistet kasutades. Lauses esinevad sõnad jagunevad kõigepealt oma staatuse alusel põhisõnadeks ja laienditeks.

Alusrühm		Öeldisrühm		
Atribuut e täiend	Subjekt e alus	Predikaat e öeldis	Adverbiaal e määrus	Objekt e sihitis
tähelepanelik adj	tudeng n	uuris v	pingsalt adv	konspekti n

Eesti keele lauseliikmed

Predikaat e öeldis - tegevus

Subjekt e alus - tegija

Objekt e sihitis - millele tegevus on suunatud, tulemus

Predikatiiv e öeldistäide - omadus

Adverbiaal e määrus - välised detailid

Atribuut e täiend - omadus

Süntaktilised funktsioonid tuleb määrata tekstilausete morfosüntaktiliste omaduste abil.

- grammatiliste morfeemide, eriti käändetunnuste ning partiklite ja adpositsioonide kasutus
- kongruents ja rektsioon
- sõnajärg

Eesti keele süntaktiliste funktsioonide eristamise kriteeriumid

1. Kas nominatiiv on x-i tüüpiline kääne?
2. Kas partitiiv on x-i tüüpiline kääne?
3. Kas x ühildub öeldisega?
4. Kas x asub tüüpiliselt enne verbi?

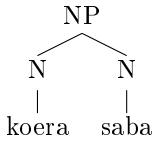
5. Kas x-ga lause finiitverb on peaaegu alati verb olema?
6. Kas x ühildub subjektiga isikus ja arvus?
7. Kas x esineb kohakäänetes?

Ebamugavad näited

- Eestlased vahetasid välja läti rahuvalvajad.
- Ta lõi teibaga meest.
- Orav viis selle käbi peidukohta.
- Lauale ilmus üha uusi klaase.

Sõltuvus

Fraasistruktuurist alati ei selgu, mis laadi ja kui tihe kokkukuuluvus osade vahel valitseb, näiteks mis on konstruktsiooni põhisõna ja mis tema laiend.



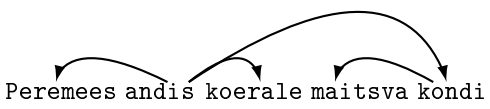
3 Lühiülevaade levinumatest grammatikaformalismidest

Levinumad grammatikaformalismid

- Dependency Grammar (DG)
- TreeAdjoining Grammar (TAG)
- Combinatory Categorial Grammar (CCG)
- Lexical Functional Grammar (LFG)
- Head Driven Phrase Structure Grammar (HPSG)

Sõltuvusgrammatikad

- Vaatab sõnadevahelisi seoseid.
- Analüüsi keskmeks on verb ja tema rektsioonid.
- Ei kasutata grupeerimist



Puuühendamisgrammatika (TAG)

- Aravind Joshi loodud teooria
- Lihtstruktuurideks on puud.
- Tuletusreeglid genereerivad puu tippe.
- Puud on leksikaliseeritud
- Puid kombineerides saadakse analüüsipuu

Leidub kahte tüüpi puid:

- algsed puud (tähistatakse α), mis kirjeldavad valentsi. Leidub nn *foot node*, mis on seotud konkreetse sõnaga.
- abipuud (tähistatakse β), mis võimaldavad rekursiooni. Juurtipp ja alustipp on sama sümboliga.

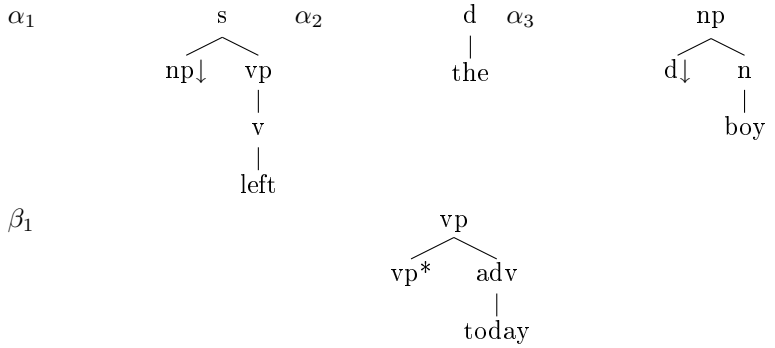
Derivatsioon koosneb kas

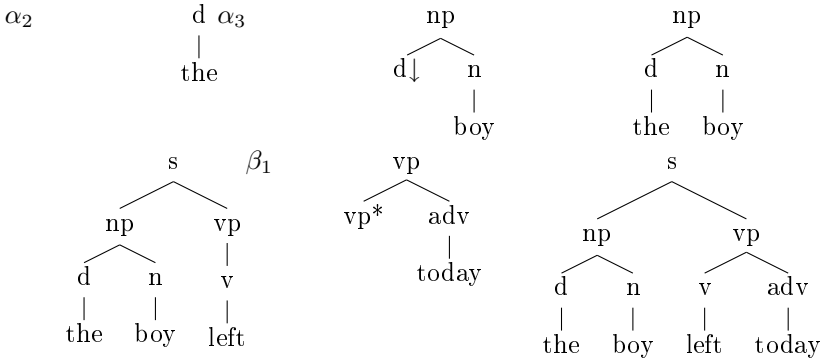
- substitutsioonist või
- lisamisest

Substitutsioon
Substitutsioon asendab tipu sellise puuga, mille juurtipul on sama sümbol.

Lisamine
Lisamise käigus lisatakse puu teise puu sisse. Juur- ja alustipp peavad olema samad selle sisemise tipuga.

- Peetakse kergelt kontekstisõltuvaks (võimsamad kui kv-grammatikad, nõrgemad kui kt-grammatikad)
- Loetakse piisavalt väljendusrikkaks, et töödelda loomulikke keeli.
- Samas enamikel juhtudel efektiivsed analüüsi keerukuse suhtes.





Leksikaal-Funktsionaalne Grammatika

- kombineerib fraasistruktuuri ja sõltuvuste infot

$$\begin{bmatrix} \text{TOPIC} & \begin{bmatrix} \text{PRED} & \text{'who'} \end{bmatrix} \\ \text{PRED} & \text{'wash} \left\langle \left(\uparrow \text{SUBJ} \right) \left(\downarrow \text{OBJ} \right) \right\rangle\text{'} \\ \text{TENSE} & \text{past} \\ \text{SUBJ} & \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'car'} \\ \text{DET} & + \end{bmatrix} \end{bmatrix}$$

Peajuhitav fraasistruktuurigrammatika (HPSG)

- Kontekstivaba grammatika
- Tunnuste struktuurid
- Tüübihierarhia
- Unifikatsioon
- Leksikon oluline
- Konstruktsioonid

$$_{type}\begin{bmatrix} \text{PHON} & \langle \textit{sleeps} \rangle \\ \text{SYNSEM} \mid \text{LOC} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{SUBCAT} & \langle \text{NP}\boxed{1} \rangle \end{bmatrix} \\ \text{CONTENT} & \begin{bmatrix} \text{AGENT} & \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Kategoriaalne grammatika

- adjektiivid ja verbid kirjeldavad funktsioone, mis määravad, kuidas neid võib kombineerida teiste kategooriatega.

$nice :\equiv NP/N$
$dogs :\equiv N$
$fly :\equiv S \setminus NP$
$like :\equiv (S \setminus NP)/NP$
$NP/N \; N \Rightarrow NP$
$NP \; S \setminus NP \Rightarrow S$