Sõltuvusgrammatika

MTAT.06.031 Süntaksiteooriad ja -mudelid

Kaarel Kaljurand
(kaljurand@gmail.com)
Tartu Ülikool

2006

Loengute sisu

- Sissejuhatus, näited, ajalugu, võrdlus fraasistruktuuri grammatikaga
- Terminoloogia: pea, alluv, valents, jne
- Erinevad formalismid. Probleemid
- Sõltuvusgrammatika ja semantika
- Implementatsioonid
- Rakendused

Lugeda

- Joakim Nivre (2005). Dependency Grammar and Dependency Parsing
  - ülevaade fookusega parsimisalgoritmidel
- Timo Järvinen, Pasi Tapanainen (1997). A Dependency Parser for English
  - palju näiteid
- Anne Abeille, ed (2003). Building and using Parsed Corpora
  - palju erinevaid artikleid, paar tükki sõltuvusgrammatikast

Lugeda

- Ralph Debusmann and Denys Duchier (2003). A Meta-Grammatical Framework for Dependency Grammar
  - Topological Dependency Grammar
- Michael Collins (1996). A New Statistical Parser Based on Bigram Lexical Dependencies
  - eduka parseri kirjeldus
- Huno Rätsep (1978). Eesti keele lihtlausete tüübid
  - valentsist eesti keeles

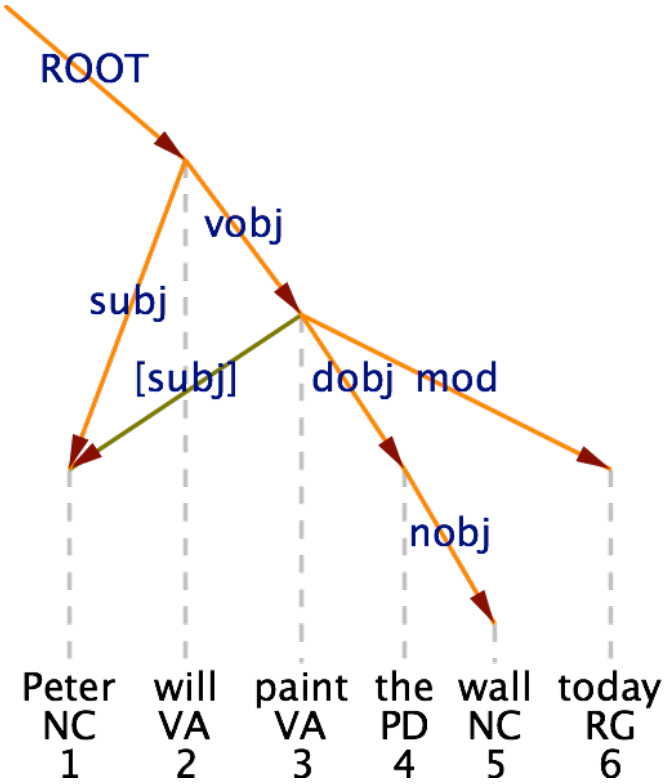
Põhiline

- Lause struktuur on esitatud binaarsete seoste hulgana
  - S = { r(x, y), r(z, y), r(y, w) }
- Seosed on lekseemide vahel
  - S = { r(a, man), r(rich, man), r(man, waits) }
- Seostel on nimed ja suund
  - S = { det(a, man), mod(rich, man), subj(waits, man) }

Põhiline

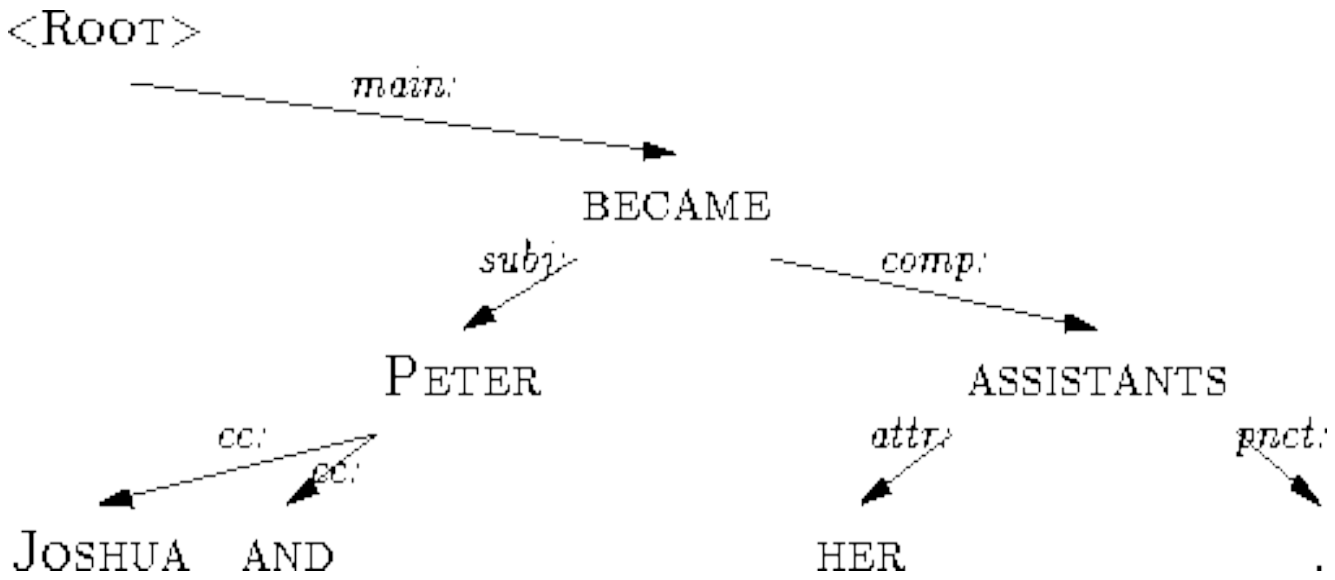
- Põhiline on formaliseerida
  - milliste üksuste vahel on seosed võimalikud
  - millised on kitsendused seoste struktuurile
    - kas kõik sõnad peavad olema seotud?
    - kas tsüklid tohivad olla?
    - ...
- Ei ole üht selget formalismi, on palju erinevaid formalisme, mis põhijoontes kattuvad aga detailides erinevad

Graafiline esitus



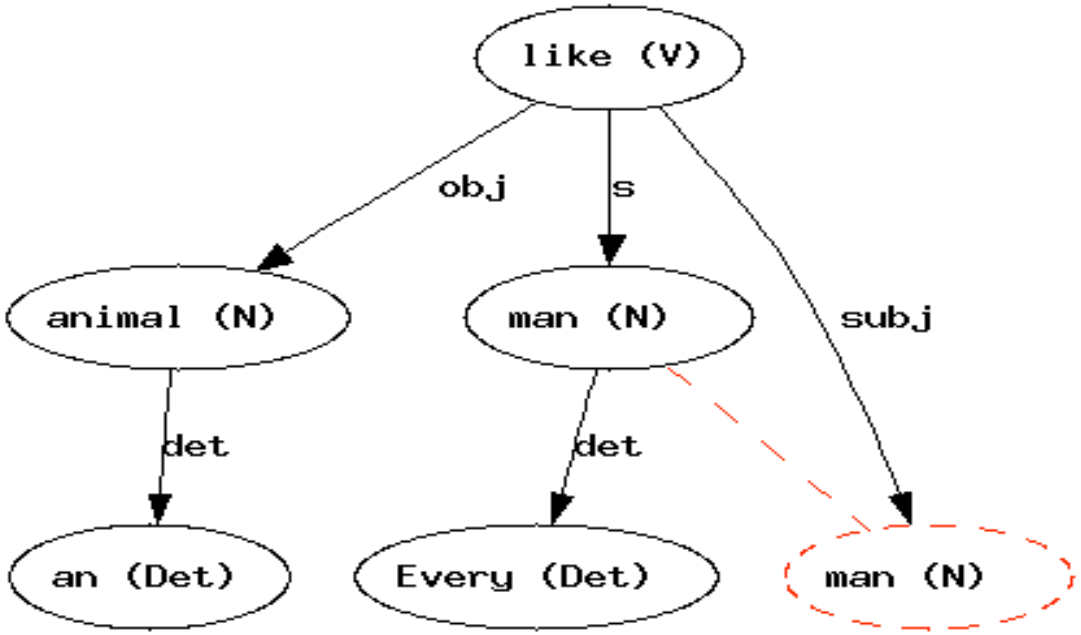
- Peter will paint the wall today.

Graafiline esitus



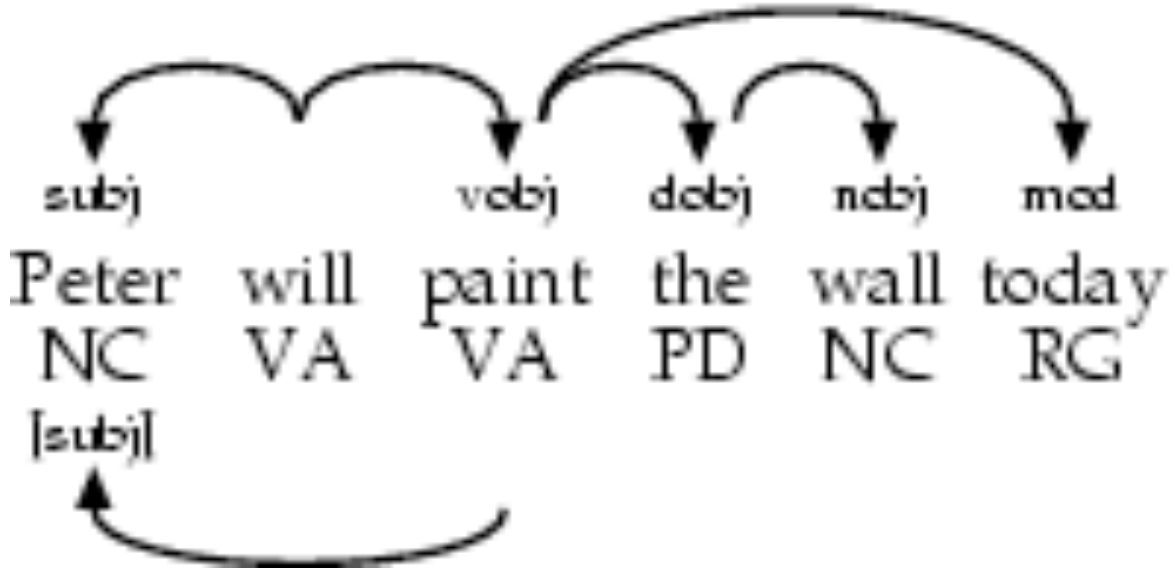
- Joshua and Peter became her assistants.

Graafiline esitus



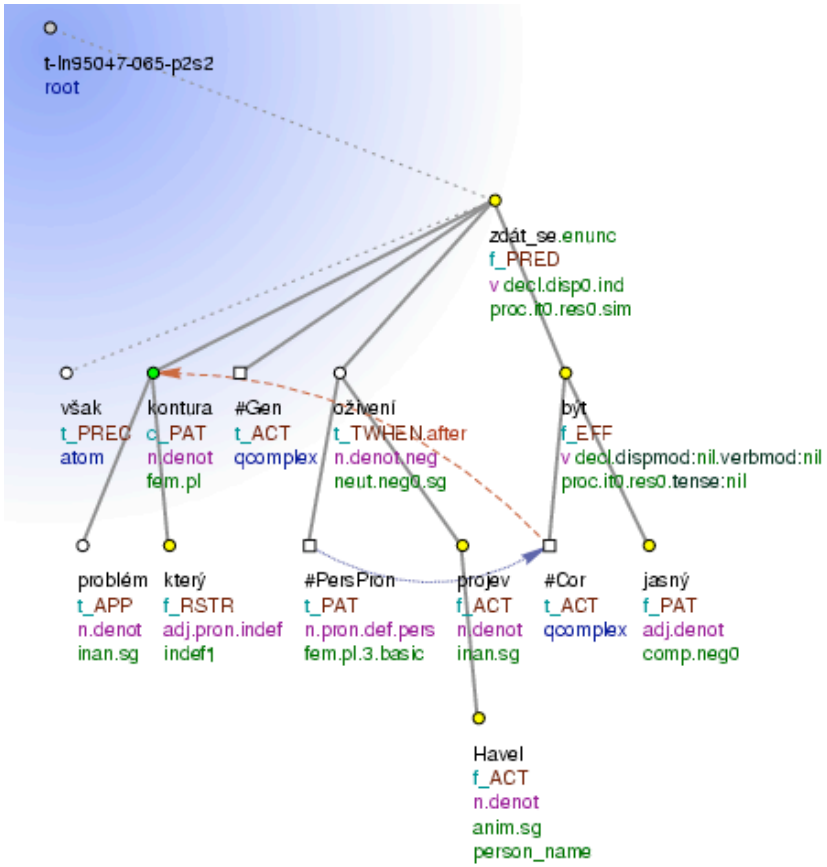
- Every man likes an animal.

Graafiline esitus



- Peter will paint the wall today.

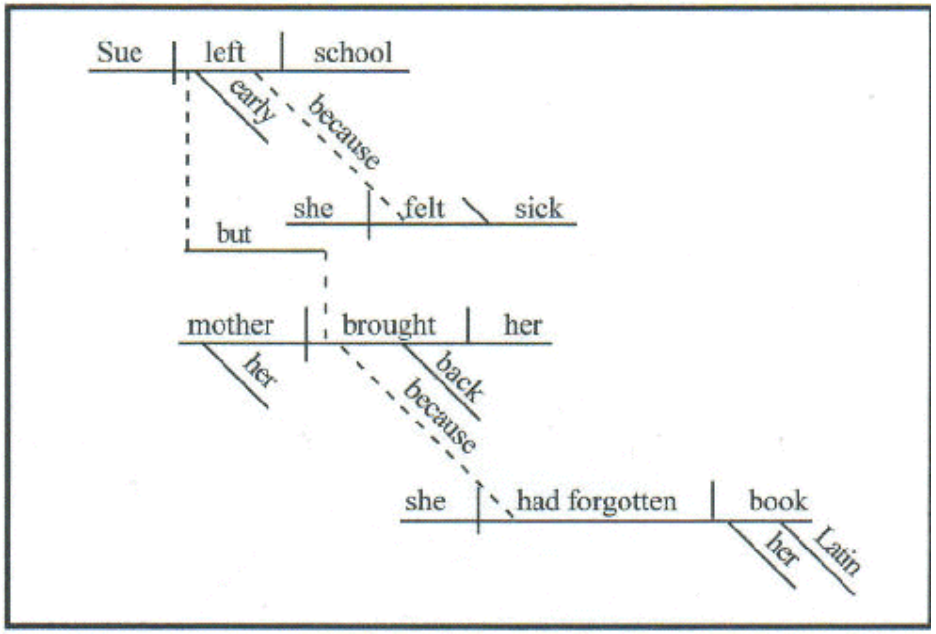
Graafiline esitus



Ajalugu

- Juba keskajal…
- Fraasistruktuurigrammatika: Chomsky 1957
- Sõltuvusgrammatika: Tesniere 1959
- Sõltuvusgrammatika juba keskkoolist tuttav
  - alus, öeldis, sihitis; mitte nimisõnafraas jne
  - USA koolides: sentence diagramming

Sentence diagramming



- Sue left school early because she felt sick, but her mother brought her back because she had forgotten her Latin book.

DG vs PSG

- Seosed sõnade vahel vs fraaside hierarhia
- Võrdle: grupeerimine fraasistruktuuri grammatikas
  - N $\subset$ NP, NP $\subset$ VP, V $\subset$ VP, VP $\subset$ S
- DG on konseptuaalselt lihtsam
  - subjekt, objekt
  - vs: NP, VP, gap/trace

Formalisimd

- Tesniere (1959)
- Word Grammar (Hudson)
- Functional Generative Description (Sgall)
- Meaning-Text Theory (Melcuk)
- Functional Dependency Grammar (Järvinen, Tapanainen)
- Topological Dependency Grammar, Extensible Dependency Grammar (Duchier, Debusmann)

Terminoloogia

- Pea ja alluv
- Sõltuvusseos
- Funktsionaalsus. Case role, thematic role (theta-role)
- Valents. Komplement. Adjunkt
- Projektiivsus
- Monostratal vs multistratal

Atomaarne üksus sõltuvuspuus

- Mis on atomaarne üksus mille vahele sõltuvussuhteid leida püütakse?
  - sõna?
  - morfeem?
  - sõnade hulk?
  - lemma?
- Nt ingl. k. *to* on iseseisev sõna, eesti k. on sõnalõpp
- Erinevad teooriad erinevad siinkohal

Pea

- Inglise keeles: *head*, *governor*, *parent*
- Pea mõiste esineb paljudes süntaksiformalismides (nt HPSG)
- Otsustamaks kumb sõna on seoses pea, arvestatakse morfoloogilisi, süntaktilisi ja semantilisi tunnuseid

Mis on pea?

- Pea on (süntaktiliselt) kohustuslik, alluva võib ära jätta
  - A rich man waits. → A man waits.
- (Di)transitiivne verb on pea, selle objekt on alluv
  - A man sees Mary.
  - John gives an apple to Mary.
- Pea kannab (käände)lõppu
  - red books
  - ??? punased raamatud
  - a piros könyvek
  - ??? die rote Bücher

Mis on pea?

- A on B pea kui A+B on A hüponüüm
  - African elephant
- Mõjutab teiste sõnade käändelõppu (ühildumine)
  - red books read well
  - red book reads well
- A ja A+B on sama distributsiooniga
  - I like red books/Mary/books
- A otsustab fraasi tüübi (N → NP)

Mis on pea?

- Vahel on raske otsustada mis on pea
- Vt nt Taani korpuse stiiljuhti
- Sõltuvalt analüüsi tasemest (morfoloogiline, süntaktiline, semantiline) võib pea olla erinev

Alluv

- Inglise keeles: *dependent*, *modifier*, *daughter*
- Alluv on see, mis pole pea :)

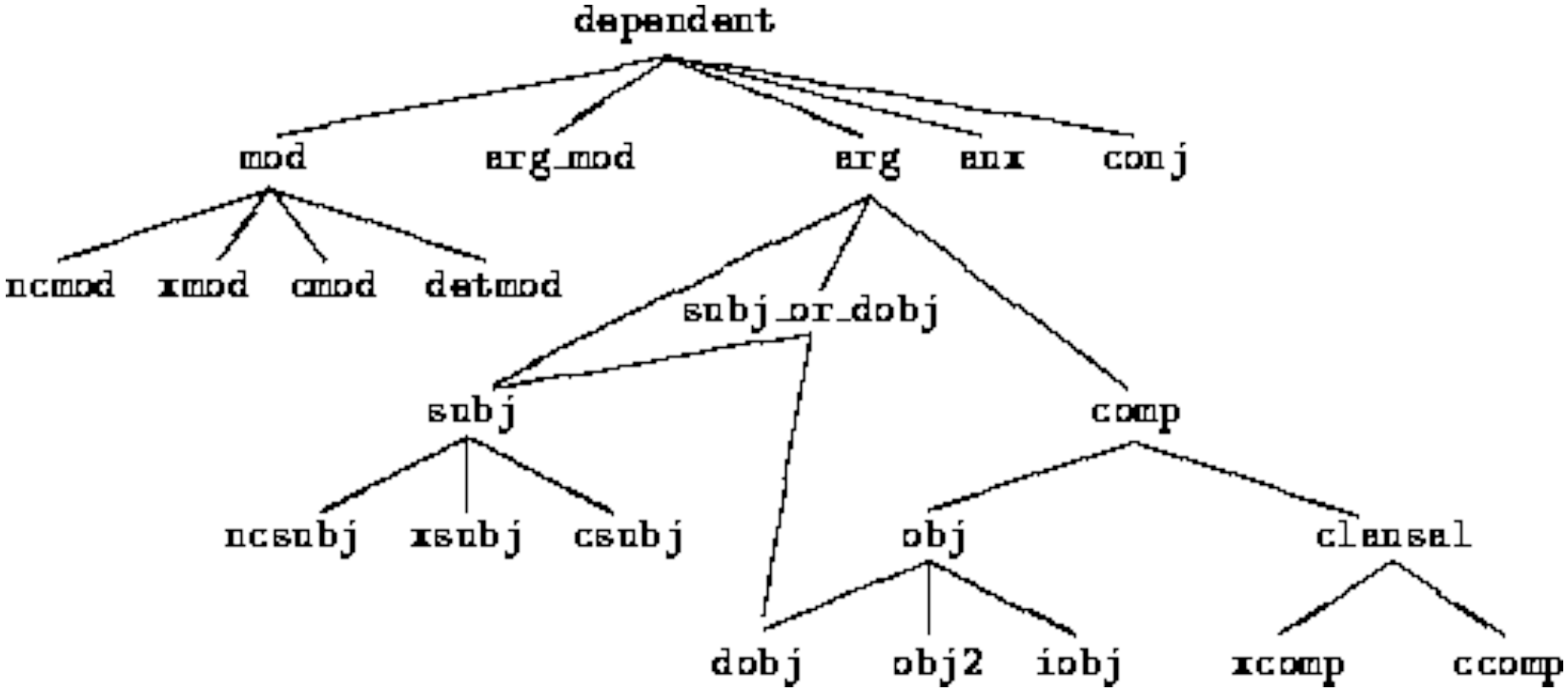
Sõltuvusseosed

- Seose nimi näitab alluva funktsiooni
- Süntaktiline
  - subjekt
  - objekt
  - adverbiaal
- Semantiline (temaatiline roll)
  - agent
  - patsient

Põhilised seosed

- subjekt (A *dog* barks.)
- objekt (John sees *Mary*.)
- kaudobjekt (John gives Mary an apple.)
- predikatiiv (John is rich.)
- attribuut (A *red* cat sleeps.)
- artikkel (*A* dog sees *a* cat.)
- eessõna (A man waits *in* a park.)
- verbi laiend (A man waits in a *park*.)
- nimisõna laiend (A man with glasses waits.)
- adverbiaal (A man waits *silently*.)
- …

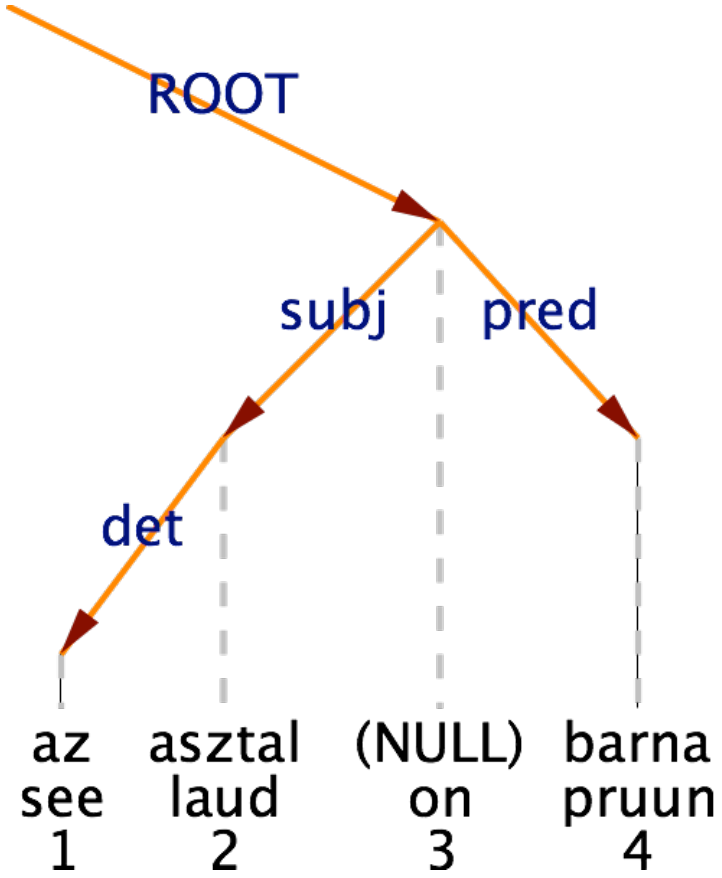
Seoste hierarhia näide: English Parser Evaluation Corpus



Sõltuvuspuu omadused

- A→C & B→C => A=B (ainult üks pea)
- A→B => NOT B→A (antisümmeetria)
- NOT A→A (antirefleksiivsus)
- A→B & B→C => NOT A→C (antitransitiivsus)
- Kõikidel seostel on nimi
- Kõik sõnad on seotud (saari pole)
- Kõik sõnad alluvad millelegi (põhiverb allub juurele *ROOT*)
- Nulllekseemid vahel lubatud (*Az asztal barna*)

Nulllekseemid



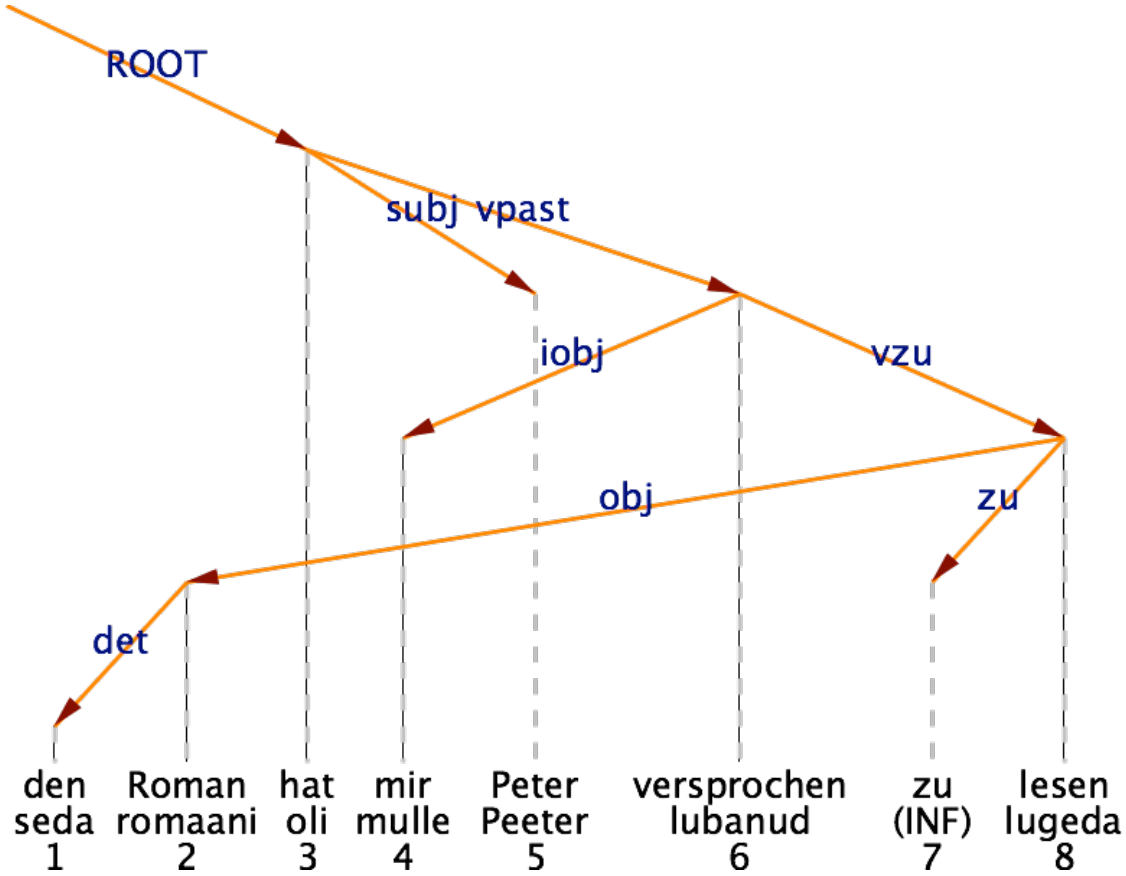
Sõnade järjekord

- Vaba sõnajärjega keeltes on lihtne lause kanoonilist sõnajärge muuta, ilma et sellest midagi oluliselt (semantiliselt) muutuks
- Reeglina ei pööra sõltuvusanalüüs sõnade järjekorrale erilist tähelepanu
- Samas saab sõnajärge erinevate kitsenduse näol arvesse võtta

Projektiivsus

- Projektiivsus
  - Kitsendus, mis võtab arvesse sõnade järjekorra lauses
  - Def: pea ja alluva “vahelised” sõnad peavad alluma peale või selle alluvale
  - Ei sobi vaba sõnajärjega keeltele

Näide: mitte-projektiivne analüüs



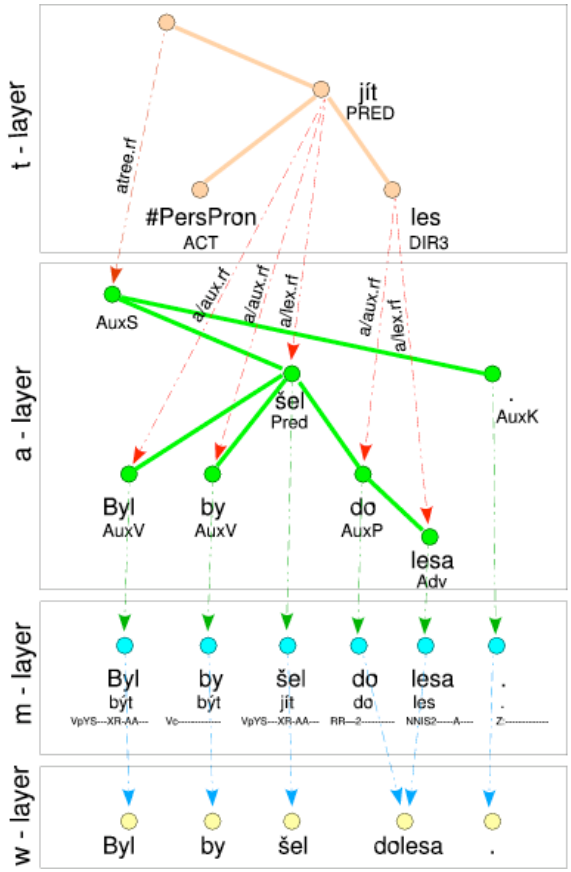
Näide: TDG

- Topological DG eristab Linear Precedence (LP) ja Immediate Dominance (ID) reegleid
- Lineral Precedence
  - *Topological fields* teooria: Vorfeld, Mittelfeld, Nachfeld
- Immediate Dominance
  - Tavaline sõltuvusanalüüs: subj, obj, …
- LP ja ID dimensioonid kitsendavad teineteist
  - LP-puu on ID-puu lamendatud versioon (nn *climbing* printsiip)

Monostratal vs multistratal

- Monostratal
  - Analüüsi tulemuseks on ainult üks sõltuvuspuu
    - Tesniere
- Multistratal
  - Erinevad puud morfoloogilise, süntaktiline ja semantilise analüüsi jaoks
    - FGD (semantilisemates puudes mõned lekseemid kaovad, nt funktsioonisõnad)
    - XDG (lekseemid säilivad)

Linking the Layers



Valents

- Kitsendus mis on defineeritud läbi sõna
- Kirjeldab sõnade tüüpe ja nende omadust funktsioneerida lauses peana
- Reeglina kirjeldab verbe, aga sageli ka adjektiive ja nimisõnu (nt kui need tulenevad verbist)
- Seotud mõisted:
  - predikaat-argument struktuur
  - verbi subkategorisatsioon (IV, TV, DTV)

Valentsifreim

- Valentsifreim kirjeldab reeglina kohustuslikke alluvaid (aktante) ja nende tüüpe
  - Agent
  - Patsient
  - Instrument
- Alluv on reeglina nimisõna, aga võib olla ka lause või koordinatsioon
- Erinev valentsifreim vastab tihti erinevale sõnatähendusele
- Valentsi eiramine muudab lause süntaktiliselt vigaseks
- Vt nt PDT-VALLEX

Valentsinäiteid

- Verbid (nt *run*)
  - John runs. (freim: run Agent/nom)
  - John runs a firm. (freim: run Agent Patient/acc)
- Verbid (nt *disappear*)
  - The problem disappeared.
  - *The teacher disappeared the problem.
- Adjektiivid (nt *fond of*)
  - *John is fond of.
  - John is fond of Mary.
- Verbid (nt *otsustama*)
  - Kohtunik otsustas saatuse.
  - Kohtunik otsustas matkale minna.

Komplement vs adjunkt

- Komplement (complement, laiend?)
  - osa valentsifreimist
  - lauses kohustuslik
  - lauses ei tohi korduda
- Adjunkt (adjunct, vaba laiend?)
  - valentsifreimis reeglina ei sisaldu
  - võib ära jätta (John runs quickly.)
  - võib korduda (John waits in a bank in the morning.)
  - peab järgnema komplementidele

Komplement vs adjunkt

- Näide:
  - John gives Mary an apple in the park.
  - * John gives Mary.
  - * John gives an apple.
  - ? John gives quickly Mary an apple.
  - John quickly gives Mary an apple.
  - John gives an apple to Mary to eat. (Compl, Compl, Adj)
- Näide:
  - John waits on the table.
  - freim 1: wait + 0
  - freim 2: wait + PP-complement
  - John waits. (→ freim 1)
  - * John waits. (→ freim 2)

Täna loengus

- Probleemid sõltuvusgrammatikaga
- Sügav süntaks ja semantika
- Parserid, nende hindamine
- Sõltuvuspuude pangad
- Sõltuvusgrammatika rakendamine

Lugeda

- Tuomo Kakkonen (2005). Dependency treebanks: methods, annotation schemes and tools
- John Carroll, Ted Briscoe, Antonio Sanfilippo. Parser Evaluation: a Survey and a New Proposal

Probleemid

- Alati tuleb otsustada, mis on pea ja mis on alluv
- Vahel on selline eristamine ebaloomulik
- Nt koordinatsioon
  - John, Mary and Bill wait.
- Nt mitmesõnalised üksused
  - Rahvas valis New Yorki linnapea ära.
- Fraasi mõiste puudub: sõna ei saa modifitseerida fraasi
  - John runs quickly.
  - John runs probably. (lauseadverbiaal)

NEGRA projekti tõdemus

- We used Dependency Grammar as the starting point for the development of our annotation scheme. It soon turned out that the sharp distinction between heads and modifiers stipulated by the formalism causes difficulties in practice. In particular, all kinds of constructions without a clear syntactic head are difficult to analyse: ellipses, sentences without a verb (e.g., copula-less predicatives), and coordinations.

Probleemid

- Koordinatsioon
  - They operate ships and banks.
  - 3 võimalikku analüüsi:
    - ‘and’ on pea alluvatega ‘ships’ ja ‘banks’
    - → ships → and → banks
    - ‘ships and banks’ on üks üksus (fraas)
- Verbiahelad
  - Mis on pea?
    - John has not been walking in the park.

Sügav süntaks (*deep syntax*)

- Probleem: infinitiivlause subjekt või objekt on lauses ilmutamata
- Control
  - John wants to run. (run → John)
  - John promises Mary to run. (run → John)
  - John forces Mary to run. (run → Mary)
- Raising
  - John seems to run. (run → John, aga *John seems)
  - John expects Mary to run. (run → Mary, *John expects, *John expects Mary)

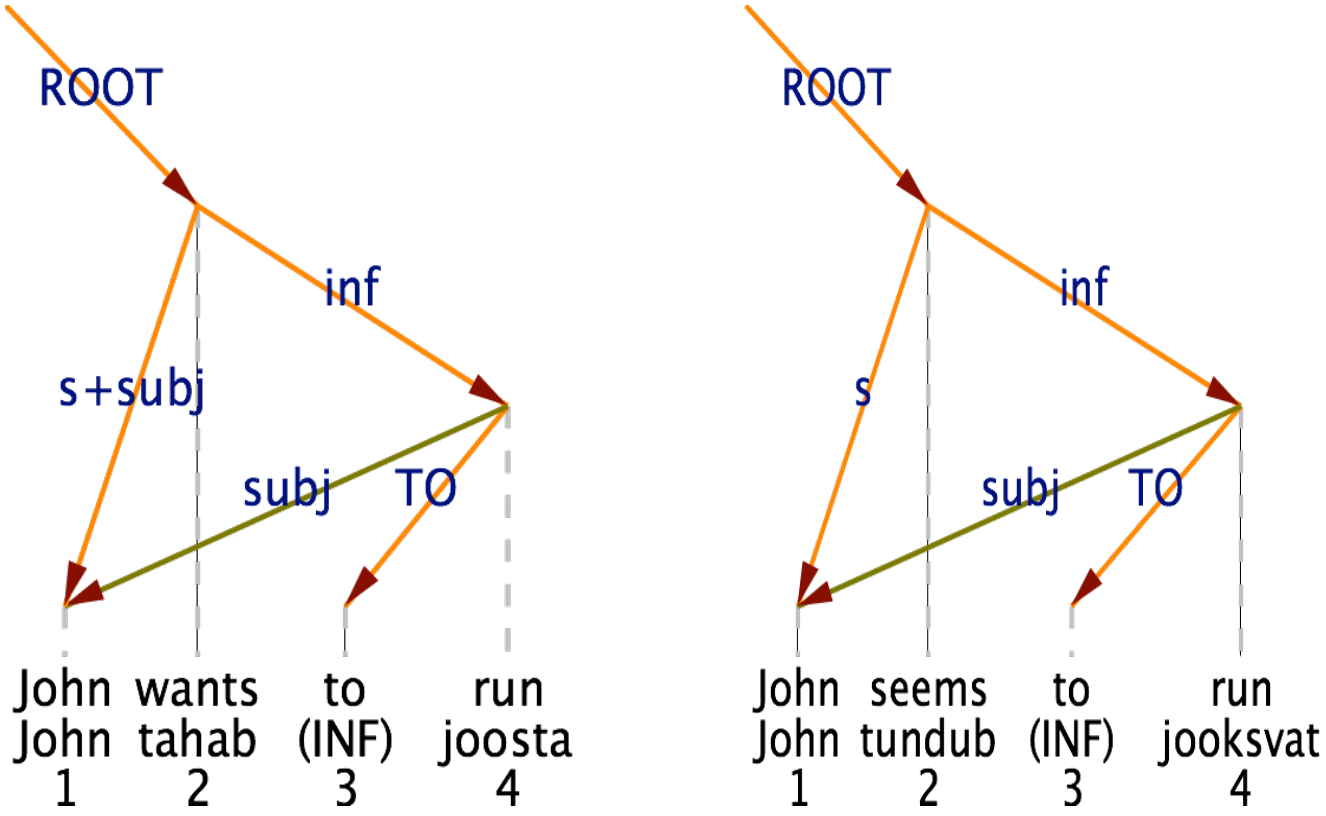
Sügav süntaks (*deep syntax*)

- Infinitiivlause tõeväärtus pole muidugi selge
  - John wants to run. → true
  - John wants. → true
  - John runs. → ???
- … aga mingit infot selline analüüs ikkagi annab:
  - Tee hakkab jahtuma. (VEDELIK-TEE)
  - Tee hakkab lagunema. (EHITIS-TEE)

Sügav süntaks (*deep syntax*)

- Nt XDG analüüsis pole sügava süntaksi dimensiooni poolt omistatud struktuur mitte puu vaid tsükliteta orienteeritud graaf (*directed acyclic graph*), sest kontrolli loomulik esitus tekitab subjektile või objektile 2 pead

Sügav süntaks



Semantika

- Ron Kaplan: in the end we don’t need the trees but to know WHO did WHAT to WHOM, WHEN and WHERE
- John saw Mary in the park in the morning.
- Sõltuvusanalüüsi väljund on “semantiline”
  - Teisendus Subjekt→Agent on lihtne
  - Teisendus NP→Agent ei ole nii lihtne

Semantika

- (Lause)semantika all mõistetakse arvutilingistikas tavaliselt lause teisendust predikaatarvutuse valemiks
- Sõltuvusanalüüs on juba päris ilusal kujul
  - subj(see, John) AND obj(see, dog)
  - subj → agent, obj → patient
  - agent(see, John) AND patient(see, dog)

Semantika

- Aga kvantorid, disjunktsioon, eitus, jne vajavad veel tööd
- Anafooride lahendamine?

Näide: semantika

- Semantiliselt erinevad laused
  - Every man is a human.
  - No man is a human.
- Sõltuvusanalüüsilt sarnased
  - det(man, every) AND subj(be, man) AND obj(be, human)
  - det(man, no) AND subj(be, man) AND obj(be, human)
- FOL analüüsilt erinevad
  - $\forall$ x (man(x) $\rightarrow$ human(x))
  - $\forall$ x (man(x) $\rightarrow$ NOT human(x))

Underspecification

- *PP-attachment*
  - John eats a soup with a spoon.
  - Luba analüüsiks katkendlikke puid
- PP tüüp
  - Täpsed tüübid: location, time, comitative, etc
  - Ebatäpne tüüp: modifier/adjunct
- Koordinatsioon, argumentide distributsioon
  - John and Mary meet. (* John meets.)
  - John and Mary eat. (John eats.)

Sõltuvusparser ja kitsendused

- Parser leiab lahenduse(d) arvestades järgmiste kitsendustega:
  - Valentsi sõnastik
  - Sõnade järjekord (nt projektiivsus)
  - Seostepuu matemaatilised omadused
  - …

Leksikon

- Leksikonil on tähtis roll
- Mitte pelgalt sõnade loend vaid valentsileksikon
- Aravind Joshi: Complicate locally, simplify globally
  - nagu pusles: mida keerulisemad on tükid (sõnad), seda lihtsam on neid kokku panna (lauseteks)

Parsimisalgoritmid

- Kompromiss:
  - Mida rohkem on kitsendusi seda kiirem on algoritm
  - Mida rohkem on kitsendusi seda vähem ekspressiivsem on formalism
- Töötavad keerukusega $O(n^2)$ ja $O(n^3)$

Parsimisalgoritmid

- XDG kasutab *Constraint Programming* lähenemist (konseptuaalselt lihtne, deklaratiivne)
- Lihtsaim imperatiivne algoritm
  - *Left-to-right without backtracking*
  - Igal sammul otsusta, kas saab ühega eelnevaist suhte tekitada

Näide. sisemine kuju

- Kuidas esitada binaarseid seoseid arvutis:
  - (subj see John), …
  - rel(subj, see, John), …
- Kui me tahame säilitada kogu info sõnade kohta:
  - rel(w1, subj, w2), type(w1, lemma, see), type(w1, morph, V), ...

Implementatsioonid

- Machinese Syntax (Connexor)
- Pro3Gres (Schneider)
- Extensible Dependency Grammar (Debusmann, Duchier)
- MINIPAR (Lin)
- Vt ka RASP (Carroll), Stanford Parser, Collins parser (Collins)

Machinese Syntax

- Autor: Connexor Soomes
- Erinevad keeled
- Kommerts
- Seostetüüpide arv: ~40
- Mõned tüübid üpris semantilised: nt PP-tüübid: location, time, etc
- Reeglipõhine
- Ehitatud kitsenduste grammatika peale
- Vt ka: Machinese Semantics

Pro3Gres

- PRObability-based, PROlog-implemented Parser for RObust Grammatical relation extraction System
- Gerold Schneider, Zürichi ülikool
- Inglise keel
- Tasuta
- Seoseid: ~20

Pro3Gres grammatika

- Käsitsi kirjutatud grammatika
  - verbile eelneb subjekt
    - * Here sits the king.
  - *utterance* verbidele võib ka järgneda
    - “I saw Mary”, said John.
  - transitiivsele verbile järgneb vahetult objekt
    - * John saw in the park Mary.

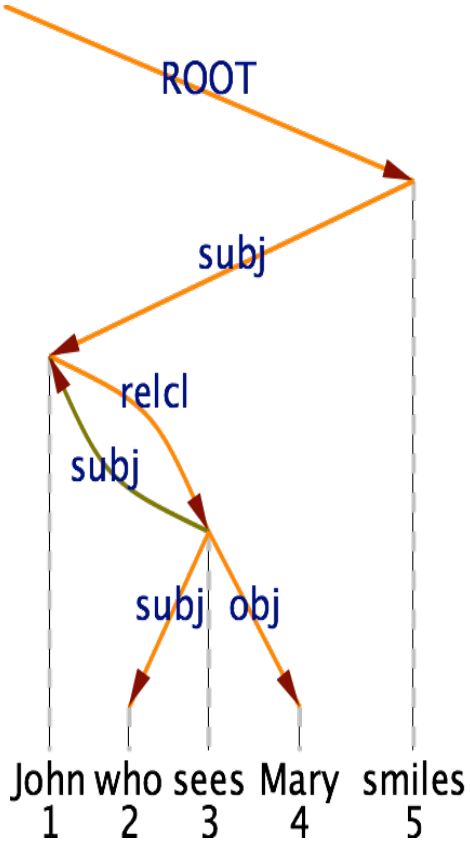
Pro3Gres: *chunking*

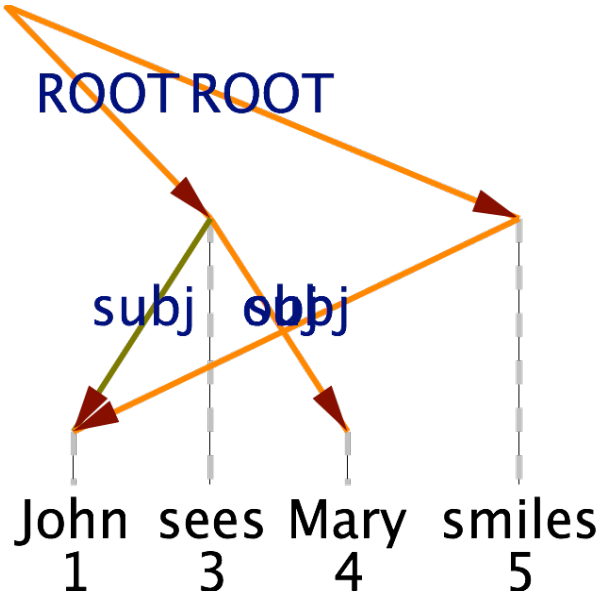
- Eeltöötlus: NP ja VP *chunking*
- *chunk* = *nucleus*
- Iga *chunk*i esindab selle pea, seosed leitakse peade vahel
  - a rich [man], the [year] 2000, does not [wait]
- Statistiline komponent automaatselt Penn puudepangast
  - lekseemide vahelise suhte tõenäosus, arvestades ka lekseemide vahelist kaugust lauses
  - tulemuseks järjestatud analüüside loend

Pro3Gres: järeltöötlus

- Järeltöötlusena analüüsitakse lauset semantilisemalt:
  - *Passive*
  - *Control/raising*
  - Relatiivlause pronoomenid

Näide: relatiivlause





XDG

- Debusmann, Duchier
- Vabavara. Mozart/Oz
- Kitsenduste levitamine (*Constraint Propagation*)
- Palju erinevaid kitsenduste tüüpe nn dimensioone
- Teoreetiline raamistik, mida testitakse mitmel keelel

XDG

- Extensible dependency grammar
- Topological dependecy grammar edasiarendus
- Dim: *Immediate dominance* kitsendused
  - subkatfreim
  - ühildumine
  - valents (millised pead/alluvad konkreetsel sõnal olla võivad)
- Dim: *Linear precedence* kitsendused
  - *Vorfeld*, *Mittefeld*, *Nachfeld*
- Dim: *Deep Syntax* kitsendused
  - agent, patient, raising & control
- Kitsendused dimensioonide vahel (*climbing*, *linking*)

MINIPAR

- Lin
- Inglise keel
- Vabavara
- Reeglid + statistika
- Sõnastik WordNetist

Sarnased formalismid

- Collins (1997, 1999)
  - viimase aja parimaid/tuntumaid parsereid
  - statistika Penn puudepanga põhjal
  - väljastab PSG esituse
  - sisemiselt kasutab sõltuvusesitust
- Constraint Grammar (Karlsson)
  - *underspecified* sõltuvusgrammatika
  - Connexori FDG ehitub sellele
- Link Grammar
  - tohutu hulk ähmaseid seoseid
  - seosed pole suunaga

Parserite hindamine

- *Precision* ja *recall*
- Leitud seoste kattumine tegelikega, mitte fraaside võrdlemine
- Subjektide ja objektide tuvastamise täpsus hetkel 90%

Korpused

- Praha korpus
- NEGRA ja TIGER korpused
  - segu sõltuvus- ja fraasistruktuurianalüüsist
- Susanne korpus
- Taani korpus

Praha korpus

- Umbes miljon sõna
- 3 taset
  - morfoloogiline
  - analüütiline (süntaktiline) (23 seosetüüpi)
  - tektogrammatiline (semantiline)
- Kasutati Collinsi parserit, mis töötas 80% täpsusega
- Seega muutus töö pool-automaatseks, st Collinsi parseri väljundi kontrollimiseks

TIGER korpus

- Umbes pool miljonit sõna
- Sisaldab nii fraasistruktuuri- kui ka sõltuvusannotatsiooni
  - Sõnad on grupeeritud fraasideks
  - Iga sõna roll fraasis on annoteeritud

Susanne 500

- 500 lauset Susanne korpusest
- Kasutatakse tihti parserite hindamisel

Taani korpus

- 100 000 sõna
- Hea stiilijuht

Sõltuvusgrammatika ja eesti keel

- Töö valentsi alal: Rätsepa lausemallid
- Eesti keele kitsenduste grammatika
  - Parser
  - Korpus
- Soome keele kitsenduste grammatikale on edukalt ehitatud sõltuvusgrammatika

Rakendused eesti keelel

- DepDict
  - sõna tähendus esitatakse süntaktiliselt seotud sõnade loendi kaudu
- Sõnatähenduste ühestamine
  - süntaktiliselt seotud sõnad lauses kitsendavad sõna tähendust (nn distributiivsuse printsiip)
    - Tee tolmas.
    - John valas tee tassi.