Süntaksiteooriad ja -mudelid
MTAT.06.031

2. loeng

Kaili Müürisep
kaili.muurisep@ut.ee

21. veebruar 2008

Tänane loeng

Contents

Kasutatud kirjandus

- Fred Karlsson. Üldkeeletadus. ptk 6.2.2.
- Kaisa Häkkinen. Keeleteaduse alused.ptk 13.
- Tiina Puolakainen. Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu 2001.
- Syntactic Wordclass Tagging. H. van Halteren (ed.) Kluwer 1999.
- Roche, Schabes. Deterministic Part-of-Speech Tagging with Finite-State Transducers.

1 Sõnaliikidest

Sõnaliikide ajaloost

1. Pānini (350 e.m.a, India) eristas noomeneid ja verbe. Noomeniteks luges ka muutumatud sõnad.
2. Sībawaihi (760-793, Araabia): noomenid (nimisõnad, omadussõnad, asesõnad, partitsiibid), verbid ja partiklid.
3. Platon (424-348 e.m.a, Kreeka): noomen ehk nimi - mille kohta midagi öeldakse, verb - mis ütleb midagi noomenite kohta.
4. Aristoteles(384-322 e.m.a, Kreeka): sõnad, millel on iseseisev tähendus (verb, noomen), ja sõnad, mis väljendavad tähenduslike sõnade vahel loogilisi suhteid.
5. Dionysios Thrax (100 e.m.a, Kreeka): nimisõna, verb, partitsiip, artikkel, asesõna, eessõna, määrsõna, sidesõna.
6. Marcus Terentius Varro (116-27 e.m.a, Rooma) jagas neljaks selle põhjal, kas leidub kääne või aeg: nimisõna = +k,-a; verb = -k,+a; partitsiip = +k, +a; määrsõna = -k, -a;

Moodne sõnaliikideks jaotus

Kriteeriumid:

- muutmine
- sõnade omatähendus
- süntaktiline funktsioon

Jaotus

- avatud klass
- suletud klass

Eesti keele sõnaliigid

- Verbid
  - pöörduvad isikus,
  - esinevad eri aegades ja kõneviisides,
  - võivad olla öeldiseks.
- Nimisõnad
  - muutuvad käändes ja arvus,
  - pole võimalik moodustada võrdlusastmeid ega sti- ja lt-tuletisi,
  - eeslaiendiks on atribuut, kuid ei saa olla intensifikaatorit.
- Omadussõnad
  - moodustavad sti- ja lt-tuletisi,
  - esinevad kas nimisõna eesatribuudina või öeldistäitena.
- Kaassõnad
  - laiendiks on nimisõnafraas,
  - rektsioon.
- Asesõnad
- Arvsõnad
- Määrsõnad
- Sidesõnad
- Hüüdsõnad

Teiste keelte sõnaliigid

Schahter (1985) väidab, et igal keelel on olemas mingisugune sõnaliigisüsteem, kuid see on keeliti väga erinev.

Näited

- Nigeeria ibo keeles vähem kui 10 omadussõna.
- Lõuna-Ameerika ketšua keeles pole ühtki.
- Ida-Aasia keeltes on mõõdusõnad eraldi sõnaliigina.
- Eesti keeles pole artiklit.
- Jaapani ja korea keeles on viisakussõnad.

Sõnaliikide määramine

- They can fish.
- Täpselt **neid** asju muidugi ei tea, kuid tasuks uurida.
- Võib-olla oleks siis juba mõttekas pakkida sigaretid näiteks äravisatud kooretopsikutesse vms. ja müüa **neid** pruugitud prügikonteineritest.

Eesti keele sõnaliikide mitmesus

1. määrsõna ja sidesõna: *aga*, *kui*, *siis*, *nagu*
2. määrsõna ja "*ei*"
3. nud-partitsiibid: *surnud*, *alustanud* ...
4. määrsõna ja alaltütlevas käändes omadussõna: *kindlalt*, *tugevalt*
5. määrsõna ja kaassõna: *mööda*, *üle*, *vastu*
6. määrsõna ja nimisõna: *siin*, *kord*, *siia*, *lõpuni*
7. "*oma*"
8. tud-partitsiibid: *liigutatud*, *kaotatud*
9. nimisõna nimetavas ja verb 3. pöördes: *kõrgus*, *muutus*, *tuli*, *käis*, *teadis*
10. nimisõna seesütlevas ja verb 3. pöördes: *ajas*, *piisas*, *mälus*
11. nimisõna alalütlevas ja määrsõna: *veel*
12. nimisõna, määrsõna, sidesõna, verb: *või*
13. nimisõna sisseütlevas, määrsõna: *ikka*, *maha*, *kokku*

Sõnaliikide määramine ja süntaksiteooriad

- inimesel on sisemine leksikon, määrab ise;
- vaadatakse, millistest variantidest saab strukture moodustada ja neist moodustataksegi. Päris elus kustutatakse vähetõenäolised puud analüüsi käigus/lõpus ära.
- reeglipõhised meetodid morfosüntaktiliseks ühestamiseks;
- statistilised meetodid morfosüntaktiliseks ühestamiseks.

1.1 Reeglipõhine meetod

Kitsenduste grammatika

Fred Karlsson, 1995. a. morfoloogiliseks ühestamiseks ja pindmiseks süntaktiliseks analüüsiks.

Morfoloogiaanalüsaator lisab kõik märgendid, ühestaja eemaldab need, mis konteksti ei sobi.

Kahte liiki reegleid:

- vali üks tõlgendus ja kustuta teised

  SELECT

  SELECT (V) (-1 (“<ei>”));

  ei pea, ei vea, ei kohta.

- kustuta üks tõlgendus

  REMOVE

  REMOVE (K prep) (NOT *1 (S) OR (P) OR (N) BARRIER (V));

  *Ta sõitis õigest tänavaotsast **mööda**.*

Jätab alles mitu tõlgendust, kui ei oska või kahtleb.

Alati jääb alles vähemalt üks tõlgendus.

Reeglid on käsitsi koostatud, regulaaravaldise sarnased, kontrollivad konteksti, sõnajärge ja sõna enda morfoloogilist informatsiooni

Näide

```
Peeter          peeter+0 //_S_ prop sg nom //
ei              ei+0 //_V_ aux neg //
pea             pida+0 //_V_ main pres neg //
                pea+0 //_S_ com sg nom //
                pea+0 //_S_ com sg gen //
SELECT(_V_) (-1 (“ei”)); seda            see+0 //_P_ dem sg part //
tänavat         tänav+0 //_S_ com sg part //
REMOVE (_K_ prep) (NOT *1 (_S_) BARRIER (_V_))mööda          mööda+0 //_D_ //
                mööda+0 //_K_ prep //
                mööda+0 //_K_ post //
minema          mine+ma //_V_ sup ill //
.               . //_Z_ Fst //
words 8, removed 3, kept 9.
```

1.2 Statistiline meetod

Inglise keele morfoloogilise ühestamise probleemist

The	can	will	rust
det	modal	**modal**	noun
	noun	noun	**verb**
	verb	verb	

Inglise keeles 30-150 POS-märgendit.

Baseline
Kui on 300000-sõnaline märgendatud korpus. Leida iga sõna jaoks selle kõige sagedasem märgend. Võtta uus tekst ja omistada igale sõnale see märgend. Tundmatud sõnad saavad nime märgendi. Mis võiks olla sellise märgendaja korrektsus?

Vastus
90%

Brilli ühestaja ehk tranformatsioonipõhine õppimine

- Leksikaalne märgendaja - lisab kõige tõenäolisema märgendi.
- Tundmatute sõnade mõistataja.
- Kontekstipõhine märgendaja.

Brilli märgendaja näide

Algsed laused

1. Chapman /np killed /vbn John /np Lennon /np
2. John /np Lennon /np was /bedz shot /vbd by /by Chapman /np
3. He /pps witnessed /vbd Lennon /np killed /vbn by /by Chapman /np

vbn - verb past participle;
vbd - verb past tense.

Reeglid

1. vbn vbd PREVTAG np
2. vbd vbn NEXTTAG by

Analüüsitud laused - reegel 1

1. Chapman/np killed/*vbd* John/np Lennon/np
2. John/np Lennon/np was/bedz shot/vbd by/by Chapman/np
3. He/pps witnessed/vbd Lennon/np killed/*vbd* by/by Chapman/np

Analüüsitud laused - reegel 2

1. Chapman/np killed/vbd John/np Lennon/np
2. John/np Lennon/np was/bedz shot/*vbn* by/by Chapman/np
3. He/pps witnessed/vbd Lennon/np killed/*vbn* by/by Chapman/np

Reeglite liigid

- A B PREVTAG C
- A B PREV1OR2OR3TAG C
- A B PREV1OR2TAG C
- A B NEXTTAG C
- A B NEXT1OR2TAG C
- A B SURROUNDTAG C D
- A B NEXTBIGRAM C D
- A B PREVBIGRAM C D

t → t' in context C

Keerukus RKn, kus R - reeglite arv, K - konteksti pikkus, n -sõnade arv Reeglite teisendamisel automaadiks saavutati keerukus, mis sõltub ainult n-st.

TBL algoritm

1. Märgenda iga sõna tema kõige sagedama POS-märgendiga
2. for k= 1, 2, ...
   (a) Leia kõik võimalikud transformatsioonid, mis võivad korpuses aset leida;
   (b) omista t_k -le reegel, mis põhjustab suurimat vigade vähenemist;
   (c) rakenda t_k korpusel;
   (d) lõpeta, kui lõpetamiskriteerium on täidetud;
3. Väljasta t_1, t_2, ... t_k

Algoritmi rakendamine

Muster: X → Y kui järgmine sõna on Z

TRAIN: The/Det race/Noun ends/Verb in/Prep a/Det defeat/Noun
GUESS: The/*Det* race/*Verb* ends/Noun in/Prep a/Det defeat/Verb

Det → Y kui järgmine sõna on 'race' ⇒ -=1

Verb → Y kui järgmine sõna on 'ends'
Y=Noun ⇒ +=1
Y≠Noun ⇒ +=0

Algoritmi probleemid

- ahne algortm - leitakse hetkel parim lahendus
- transformatsioone rakendatakse vasakult paremale
  - vasakult paremale AAAA → ABAB
  - paremalt vasakule AAAA → ABBB

Sellegipoolest on korrektsus 96-98%, kui pole tundmatuid sõnu.

Kui suurendada treeningkorpust 64.000 sõnalt 640.000 sõnale, kasvab korrektsus 0.4%

5 popimat reeglit

1. NN → VB , kui eelmine märgend on TO
2. VBP → VB , kui üks 3 eelmisest märgendist on MD
3. VBP → VB , kui üks 2 eelmisest märgendist on MD
4. VB → NN , kui üks 2 eelmisest märgendist on DT
5. VBD → VBN , kui üks 3 eelmisest märgendist on VBZ

N parimat märgendit

n-best wordclass tagging

Vahel on vaja, et märgendus oleks pigem korrektne kui ühene.

Kui erinevused tõenäosuste vahel on väga väikesed, väljastada kõik ligilähedase tõenäosusega märgendid.

Kuidas see võiks välja näha Brilli ühestajas?

Märgendite lisamise reeglid

Pärast seda, kui ühene märgendaja on oma töö teinud, panna tööle reeglid, mis lubavad lisada mitu märgendit.

'*yen*' esineb ingliskeelsetes tekstides kord ainsuses nimisõnana, kord mitmuses.

Märgendaja arvab, et see on the järel mitmuses.

Lisa märgend SingularNoun kui vaadeldav sõna on '*yen*' ja eelnev sõna on '*the*'.

$Skoor = \frac{Märgenduse\ korrektsuse\ suurenemine}{Lisatud\ märgendite\ arv}$

Juhendamata õppimine

Paljud (pooled) sõnad on ühesed - õpi nende põhjal.

The/AT can/NN is/BEZ open

can on NN, kui AT _ BEZ kontekstis on enamik üheseid sõnu NN-märgendiga

Kasutatakse teistsugust lähenemist:

- Algselt lisatakse kõik leksikonis olevad märgendid

- Reeglid on kujul:

  $X_1 X_2 ... X_n \rightarrow X_i$ kui

  - eelmine sõna on W
  - järgmine sõna on W
  - eelmine märgend on T
  - järgmine märgend on T

Noun_Verb → Noun, kui eelmine on Adj

Korrektsus 95-96%

2 Sissejuhatus Government & Binding teooriasse

Government & Binding teooria

Liliane Haegeman. Introduction to Government & Binding Theory. Blackwell 1994.

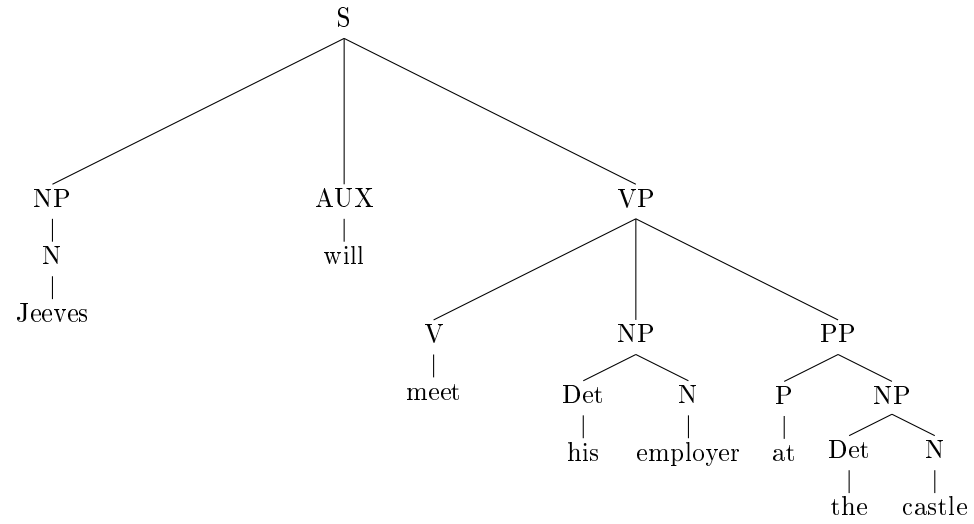
Süntaksianalüüsi ühikud

(1) Jeeves will meet his employer at the castle.

(2)
1. $S \rightarrow NP\ AUX\ VP$
2. $NP \rightarrow (Det)\ N$
3. $VP \rightarrow V\ NP\ PP$
4. $PP \rightarrow P\ NP$
5. $N \rightarrow Jeeves,\ employer, castle$
6. $V \rightarrow meet$
7. $AUX \rightarrow will$
8. $P \rightarrow at$
9. $Det \rightarrow the,\ his$

(3) $[_S[_{NP}[_N Jeeves]][_{AUX} will][_{VP}[_V meet][_{NP}[_{Det} his][_N employer]]$
$[_{PP}[_P at][_{NP}[_{Det} the][_N castle]]]]]$

(4)



Permutatsioonid ja küsimused

(5) *Permutatsioonid*

1. At the castle, Jeeves will meet his employer.
2. His employer, Jeeves will meet at the castle.
3. Meet his employer at the castle, Jeeves will (indeed).
4. *Employer at the, Jeeves will meet his castle.

(6) *Küsimused*

1. Will Jeeves meet his employer at the castle?
2. Who will Jeeves meet at the castle?
3. Where will Jeeves meet his employer?
4. What will Jeeves do?
5. Who will meet his employer at the castle?

Sõnad ja fraasid

Sõnad kuuluvad erinevatesse süntaktilistesse kategooriatesse (sõnaliikidesse) ja see kategooria määrab sõna distributsiooni (konteksti, kus ta asub).

(7) * Jeeves will appointment his employer at the castle.

Mentaalne leksikon

1. *meet*: verb
2. *employer*: noun
3. *castle*: noun

4. *at*: preposition
5. *his*: determiner
6. *appointment*: noun

Sõnad ja fraasid - 2

(8) *Grammatiline korrektsus*

1. Jeeves will meet his employer at the castle
2. ?Jeeves will meet his castle at the meeting.
3. ?Jeeves will meet his castle at the employer.

(9) *Olulisus*

1. Jeeves will meet his employer at the castle.
2. Jeeves will meet his employer.
3. *Jeeves will meet at the castle.

Predikaadid ja argumendid

(10)

1. Maigret will [$_{VP}$imitate [$_{NP}$Poirot] [$_{PP}$with enthusiasm]]
2. Bertie will [$_{VP}$abandon [$_{NP}$the race] [$_{PP}$after the first lap]]
3. Miss Marple will [$_{VP}$ reconstruct [$_{NP}$the crime] [$_{PP}$in the kitchen]]

(11)

1. Maigret will imitate Poirot.
2. Bertie will abandon the race.
3. Miss Marple will reconstruct the crime.

PP ei ole obligatoorne, vaid kirjeldab viisi, aega, kohta. Selliseid moodustajaid nimetatakse **adjunktideks**.

(12)

1. *Maigret will imitate.
2. *Bertie will abandon.
3. *Miss Marple will reconstruct.
4. *Jeeves will meet.

Midagi on puudu.

Alamkategooriad

(13)

1. Hercule is dithering.
2. Wooster gave Jeeves the money.

(14)

1. *Wooster gave Jeeves.
2. *Hercule is dithering the crime.

Verbide klassifitseerimine nende kohustuslike argumentide järgi.

1. *meet*: verb, transitive
2. *dither*: verb, intrasitive
3. *give*: verb, ditransitive

1. *meet*: V,[-NP];
2. *dither*: V,[–];
3. *give*: V,[-NP,NP] või V,[$_N$P, PP]

Valents

Sõna **valentsi** all mõeldakse selle kohtade või argumentide ehk kohustuslike laiendite hulka.

Valentsi all olevad komplemendid on **aktandid**.

Verbide valents võib olla nullist kolmeni.

Valents

1. Sajab.
2. Kõneleja aevastas. Laps magab.
3. Nad vaatasid jalgpallivõistlust. Oskar Luts on kirjanik.
4. The boss offered her a better salay.

Hägusus

1. Kalev suitsetab (piipu).
2. Elli ootab (sõpra).
3. Mart istus (diivanile).

Tähenduse muutus

1. Peeter loeb raamatut.
2. Peeter loeb Priitu oma sõbraks.

Teiste sõnaliikide valents

- Kaassõnad
    1. Eesõnad: Prep+NP.
    2. Tagasõnad:NP+Post
- Omadussõnad: NP + *pikkune*, *suurune* jne

Argumendistrukuur loogikas

Transitiivne verb
Maigret imitates Poirot.
A(mp)
imitate(Maigret, Poirot)

Intransitiivne verb
Maigret stumbled.
S(m)
stumble(Maigret)

Argumendistruktuuri esitus

Valentsi esitus

meet:	verb;	1	2	
		NP	NP	
imitate:	verb;	1	2	
		NP	NP	
dither:	verb;	1		
		NP		
give:	verb;	1	2	3
		NP	NP	NP
		NP	NP	PP

Keerulisem verb

1. Hercule bought Jane a detective story.
2. Hercule bought a detective story.

buy:	verb;	1	(2)	3
		NP	NP	NP

Adjektiivide valents

1. Jeeves is envious of Bertie.
2. Jeeves envies Bertie.
3. *Jeeves is envious Bertie.
4. *Poirot envies.

5. Poirot is envious.

envious:	adjective;	1	(2)
		NP	PP
restless:	adjective;	1	
		NP	
concious$_1$:	adjective;	1	2
		NP	PP
concious$_2$:	adjective;	1	
		NP	

Nimisõnade valents

1. Poirot's analysis of data was superfluous.
2. The analysis of data was superfluous.
3. The analysis was superfluous.

analyse	verb;	1	2
		NP	NP
analysis	noun;	(1)	(2)
		NP	PP

Kaassõnade valents

1. John is in London.
2. Florence is between Milan and Rome.

in:	preposition;	1	2	
		NP	NP	
between:	preposition;	1	2	3
		NP	NP	NP

θ-teooria

- Chapman killed Lennon.
kill:	verb;	1	2
		NP	NP

  Chapman on **AGENT**; Lennon on **PATIENT**.

θ-rollid

AGENT - keegi, kes alustab predikaadiga tähistatud tegevust.

PATSIENT - predikaadiga tähistatud tegevuse objekt.

TEEMA - asi või olend, keda liigutati tegevuse käigus.

KOGEJA - olend, kes koges tegevuse käigus.

BENEFITSIENT - olend, kes sai kasu tgevuse käigus.

SIHT - asi või olend või koht, mille suunas käib tegevus.

ALLIKAS - asi või koht, kust midagi liigutati.

KOHT - koht, kus tegevus toimub.

Näited

1. Constance rolled the ball towards Poirot.
2. The ball rolled towards the pigsty.
3. Madame Maigret had been cold all day.
4. Maigret likes love stories.

θ-grid

- Chapman$_i$ killed Lennon$_j$.

Grid

kill: verb

AGENT NP	PATIENT NP
i	j

Kriteerium 2.1. *θ-kriteerium*

- *Igal argumendil on ainult üks θ-roll.*
- *Igale θ-rollile vastab ainult üks argument.*

Projektsiooniprintsiip

Printsiip 2.1. ***Projektsioni printsiip***
Leksikaalne informatsioon on süntaktilisel tasandil esitatud.

Osalaused θ-rollis

- Maigret$_i$ believes [this story]$_j$.
- Maigret$_i$ believes [that the taxi driver is innocent]$_j$.
- Maigret$_i$ believes [the taxi driver to be innocent]$_j$.
- Maigret$_i$ believes [the taxi driver innocent]$_j$.

believe

believe: verb

1 NP	2 NP/S
i	j

It ja extrapositsioon

surprise

- The burglary surprised Jeeves.
- That the pig had been stolen surprised Jeeves.
- It surpised Jeeves that the pig had been stolen.

surprise

surprise: verb

1	2
NP/S	NP
i	j

There
There are three pigs escaping.

Abiverbid

accuse

1. Poirot accuses Maigret.
2. Poirot has accused Maigret.
3. Poirot is accusing Maigret.
4. Poirot does not accuse Maigret.

Abiverbid ja öeldistäitena esinev *olema* (koopula) ei omista temaatilisi rolle.

Laiendatud projektsiooniprintsiip

Printsiip 2.2. *Laiendatud projektsiooniprintsiip*
$S \Rightarrow NP$ - AUX - VP

Igas lauses peab olema alus.