Sõnajärjevealeidja prototüübi kasutus süntaktiliste struktuuride uurimisel

Erika Matsak, PhD

Associate Professor of Tallinn University, Institute of Informatics

The present work was supported by the National Program for Estonian Language Technology (2006-2010), project "VAKO – Developing the language software and the language technology resources of Estonian Interlanguage Corpus (2008-2010)

Sisukord

- Vealeidjad ja metoodikad
- Sõnajärje vealeidja prototüübi alused
- Sõnajärjemallide otsimine ning analüüs
- Andmepuud: süntaktilised märgendid ja sõnajärjemustrid
- Sõnajärjevealeidja prototüüp
- Prototüübi testimistulemused
- Kokkuvõte

Vealeidjad ja metoodikad

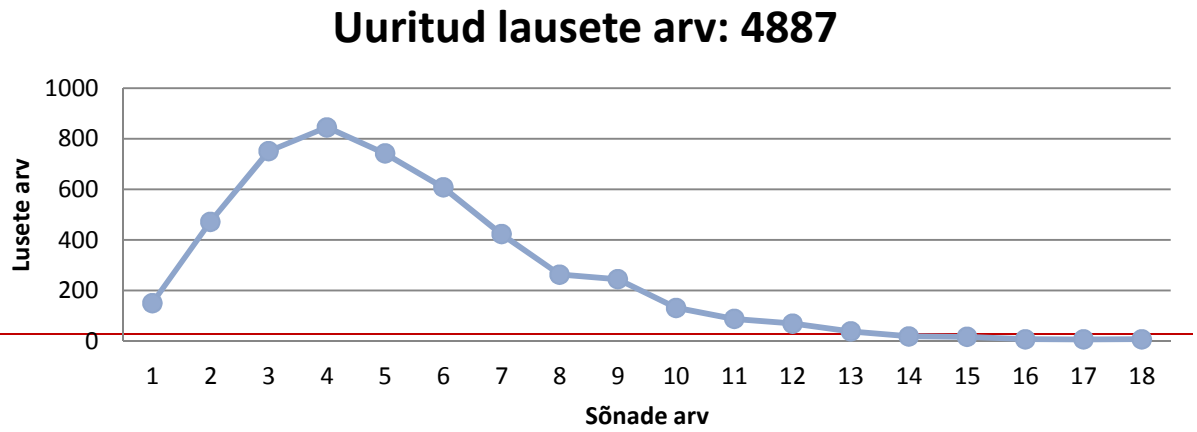
- Üks meetod tugineb spetsiaalsele sõnastikule, milles võrreldakse iga sõnet õige sõnavormiga (Damerau 1964)
- Teine meetod põhineb n-grammidel (Beesley 1998), kui võrreldakse osasõna ehk näiteks kahte või kolme järjest paiknevat tähte.
- Lisaks on nii sõnade kui osasõnade tasemel võimalik moodustada sagedasemate vigade nimistud (Pedler 2007). Selliste loendite moodustamiseks on sõnad korpustes märgendatud: igale vigasele sõnale vastab õige sõna.
- Sõnajärje õigsuse kontrollimiseks on erinevate keelte puhul kasutatud n-gramme ja statistilisi meetodeid (Athanaselis, Bakamidis, Dologlou 2006). Grammiks on terve sõna.
- Levinud on ka lingvistiline reeglipõhine lähenemine (vt näiteks Arppe 2000).

Sõnajärje vealeidja prototüübi alused

- Tallinna Ülikoolis on VAKO-projekti raames sõnajärje kontrollimiseks loodud prototüüp, mis on integreeritud EVKK õppijakeele korpusesse.
- Vealeidja kasutab oma töös reeglipõhist lähenemist, kus reegliteks on õigete sõnajärjemallide kogu.

Sõnajärje vealeidja prototüübi alused

Sõnade arv (grammi pikkus)	Erinevate grammide arv (27 märgendit)	Erinevate grammide arv (9 märgendit)
2	$27^2=729$	$9^2=81$
3	$27^3=19683$	$9^3=729$
4	$27^4=531441$	$9^4=6561$
5	$27^5=14348907$	$9^5=59049$
6	$27^6=387420489$	
Total	402321249	66420



Sõnajärje vealeidja prototüübi alused

- Lausete süntaktilist struktuuri uurides selgus, et suurem osa märgenditest ei mängi sõnajärjes olulist rolli ning arvestada võib ainult üheksa märgendiga (Metslang, Matsak 2010):

  Öeldise märgendid

  @FMV – finiitne verb

  @IMV – infiniitne verb

  @FCV – *olema* liitaegades ning modaalverbid ahelverbides, finiitne vorm

  @ICV – *olema* liitaegades ning modaalverbid ahelverbides, infiniitne vorm

  @NEG – verbi eitus

  Lause põhja märgendid

  @SUBJ — alus ehk subjekt

  @OBJ — sihitis ehk objekt

  @PRD — öeldistäide ehk predikatiiv

  @ADVL — määrus ehk adverbiaal, ka fraasiadverbiaal.

Sõnajärje vealeidja prototüübi alused

- Osalause (erijuhul lihtlause) süntaktiliseks malliks nimetame öeldise või lause põhja märgendite järjestatust. Näiteks, kui tegu on lihtlausega *Internetis on võimalik kasutada mitmeid teenuseid*, siis selle malliks on **@ADVL** (*Internetis*), **@+FMV** (*on*), **@PRD** (*võimalik*), **@SUBJ** (*kasutada*), **@OBJ** (*teenuseid*). Sõnale *mitmeid* vastab nimisõnalise eestäiendi märgend **@NN>**, mis ei kuulu vajalike märgendite hulka ning seega seda märgendit sõnajärjemalli ei lisata.
- Sõnajärjemall on

  ['@ADVL', '@FMV', '@PRD', '@SUBJ', '@OBJ']

Sõnajärje vealeidja prototüübi alused

- Siiski ei ole osalausele vastava süntaktilise malli tuvastamine eesti keeles alati nii lihtne. Näiteks lauses *Juba paar nädalat valitses põud* on sõna *juba,* mis ei mängi sõnajärjes olulist rolli, esitatud vajaliku märgendiga **@ADVL**. Et oleks võimalik iga sõna kohta ütelda, kas see sõna on sõnajärje määramiseks oluline või mitte, on vaja moodustada kolm hulka, mille abil kontrollitakse sõna ja sellele vastavale märgendi olulisust:
  - vajalikud märgendid,
  - mittevajalikud märgendid,
  - mittevajalikud sõnad

Sõnajärje vealeidja prototüübi alused

- Prototüüp on programmeeritud välja sortima kõik laused, mis algavad sõnadega *kui*, *kuna*, mistahes küsisõnade või nende käändevormidega.
- Samuti jäetakse välja umbisikulises kõneviisis olevad laused või osalaused ning hüüumärgiga lõppevaid lauseid .
- Kuna õppijakeele puhul esineb süntaktilist valeanalüüsi, mis on tingitud õigekirjavigadest, siis on vaatluse alt välja jäetud ka kõik (osa)laused, milles esineb mõni õigekirjaviga.

Sõnajärjemallide otsimine ning analüüs

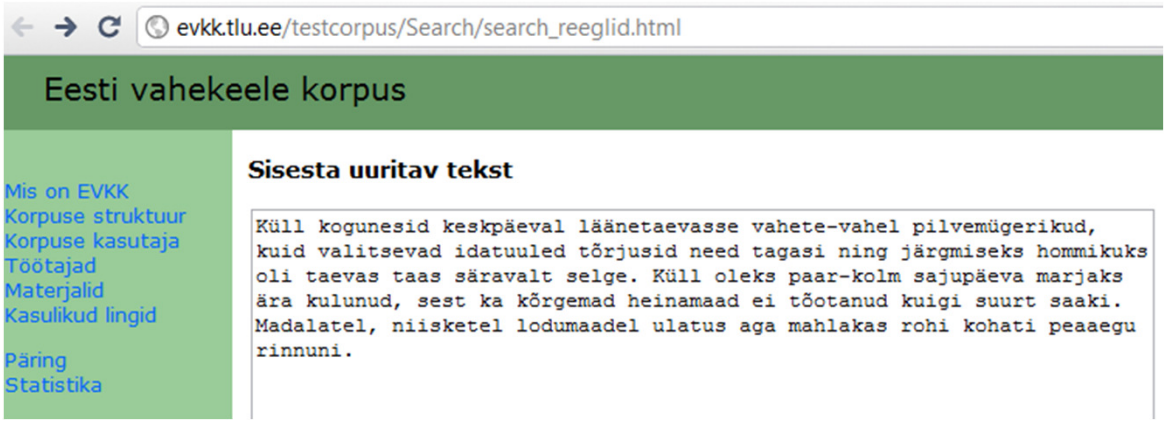
- Õigete lausestruktuuride selgitamiseks kasutasime lauseid, mis olid võetud Tartu Ülikooli eesti kirjakeele korpuse ilukirjanduse tekstidest.
- Kinnitust leidis 600 korrektset sõnajärjemalli.
- Prototüüpi testiti ilukirjandustekstidest võetud 20000 lausel.
- Kuna uusi tekste produtseeritakse pidevalt juurde ja tekstid pole ei tekstiliigilt, stiililt ega mahult samalaadsed, siis tuleb usaldusväärsete analüüsitulemuste saavutamiseks tagada programmi efektiivsus ja ka piisav töökiirus (vt Matsak, Metslang, Kippar 2010).

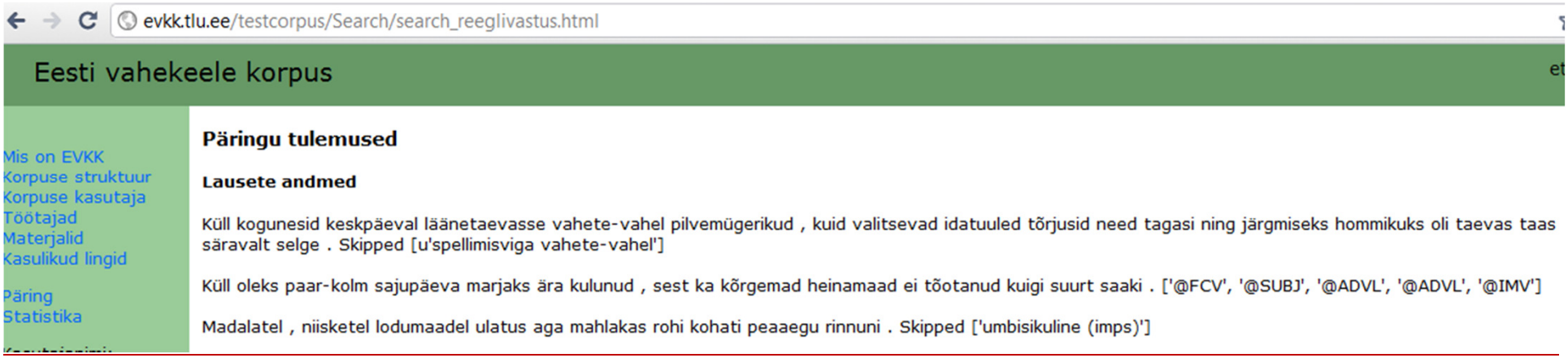
Sõnajärjemallide otsimine ning analüüs

- Üheks võimaluseks on paigutada süntaktiliste märgendite järjendid sõnajärjemustritesse ja ühesuguse algusmärgendiga mustrid andmepuudesse.
- Programm leiab kõigepealt üles vastava algusmärgendiga puu ja otsib puu ladvast allapoole liikudes sagedasemaid ja regulaarselt ilmnevaid sõnajärjemustreid.
- Kuna suure tõenäosusega moodustavad mustreid just sagedased sõnajärjemallid, siis tõhustab andmepuude kasutamine programmi tööd üsna oluliselt

Andmepuud: süntaktilised märgendid ja sõnajärjemustrid

- Andmepuud genereeriti rohkem kui 10000 lause alusel, kuna algsest 20000 lauset sisaldavast valimist ei analüüsi prototüüp 8590 lauset.





Andmepuud: süntaktilised märgendid ja sõnajärjemustrid

- *Loomulikult ei (@NEG) tohi (@FMV) tunnistada (@IMV) , et sul endal ka need olemas on .*
- *Seega ei (@NEG) saa (@FMV) me (@SUB) oma läkitust (@OBJ) kodeerida (@IMV) nii , et see oleks mõistetav .*
- *Siis ei (@NEG) lausu (@FMV) ma (@SUBJ) enam ühtki sõna (@OBJ) ega mõtle enam millelegi , kõik vajub enneaimamatusse õndsusemerre .*
- *Ei (@NEG) ole (@FMV) enam millest (@ADVL) rääkida (@SUBJ) , ta tahab maale saada , ehk ta seda küll ei ütle .*
- *Ei (@NEG) ole (@FMV) jahu (@SUBJ) põskedel (@ADVL) ja huuled on loomulikult värsked .*
- *Siiski ei (@NEG) ole (@FMV) ma (@SUBJ) teie peale (@ADVL) väga tige , et te mu üles ajasite .*

Andmepuud: süntaktilised märgendid ja sõnajärjemustrid

['@NEG', '@FMV', '@IMV']

['@NEG', '@FMV', '@SUBJ', '@OBJ', '@IMV‘]

['@NEG', '@FMV', '@SUBJ', '@OBJ']

['@NEG', '@FMV', '@ADVL', '@SUBJ‘]

2

['@NEG', '@FMV', '@SUBJ', '@ADVL‘]

['@NEG', '@FMV', '@SUBJ', '@ADVL‘]

Andmepuud: süntaktilised märgendid ja sõnajärjemustrid

2	['@NEG', '@FMV', '@SUBJ', '@ADVL‘]
	['@NEG', '@FMV', '@SUBJ', '@ADVL‘]
	['@NEG', '@FMV', '@IMV']
	['@NEG', '@FMV', '@SUBJ', '@OBJ', '@IMV‘]
	['@NEG', '@FMV', '@SUBJ', '@OBJ']
	['@NEG', '@FMV', '@ADVL', '@SUBJ‘]

Andmepuud: süntaktilised märgendid ja sõnajärjemustrid

2

['@NEG', '@FMV', '@SUBJ', '@ADVL‘]

['@NEG', '@FMV', '@SUBJ', '@ADVL‘]

['@NEG', '@FMV', '@IMV']

['@NEG', '@FMV', '@SUBJ', '@OBJ', '@IMV‘]

['@NEG', '@FMV', '@SUBJ', '@OBJ']

['@NEG', '@FMV', '@ADVL', '@SUBJ‘]

Andmepuud: süntaktilised märgendid ja sõnajärjemustrid

2

['@NEG', '@FMV', '@SUBJ', '@ADVL‘]

['@NEG', '@FMV', '@SUBJ', '@ADVL‘]

['@NEG', '@FMV', '@SUBJ', '@OBJ', '@IMV‘]

['@NEG', '@FMV', '@SUBJ', '@OBJ']

['@NEG', '@FMV', '@IMV']

['@NEG', '@FMV', '@ADVL', '@SUBJ‘]

Andmepuud: süntaktilised märgendid ja sõnajärjemustrid

2

['@NEG', '@FMV', '@SUBJ', '@ADVL']

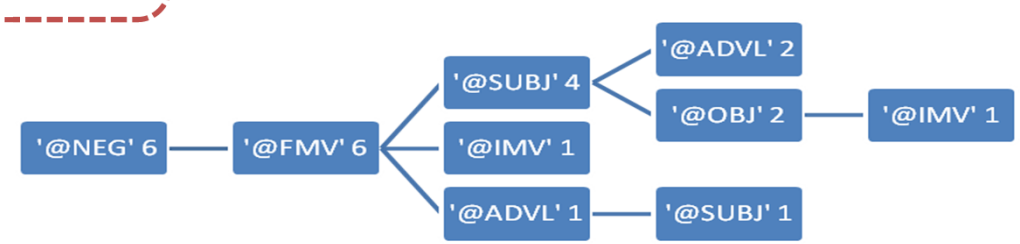
['@NEG', '@FMV', '@SUBJ', '@ADVL']

['@NEG', '@FMV', '@SUBJ', '@OBJ', '@IMV']

['@NEG', '@FMV', '@SUBJ', '@OBJ']

['@NEG', '@FMV', '@IMV']

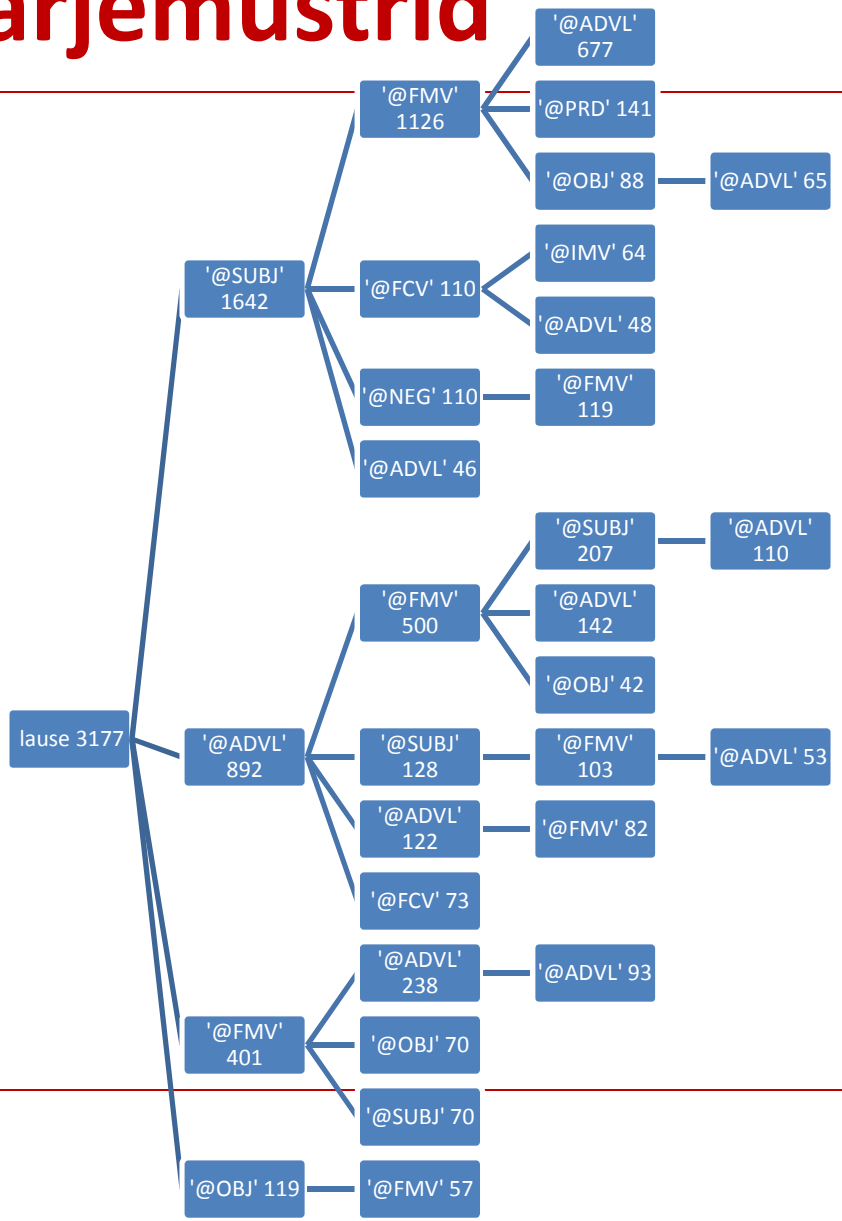
['@NEG', '@FMV', '@ADVL', '@SUBJ']



Andmepuud: süntaktilised märgendid ja sõnajärjemustrid

Antud slaidil on programmi jaoks jäetud nähtavaks vaid need märgendid, mida analüüsitud valimis on esinenud vähemalt nelikümmend korda

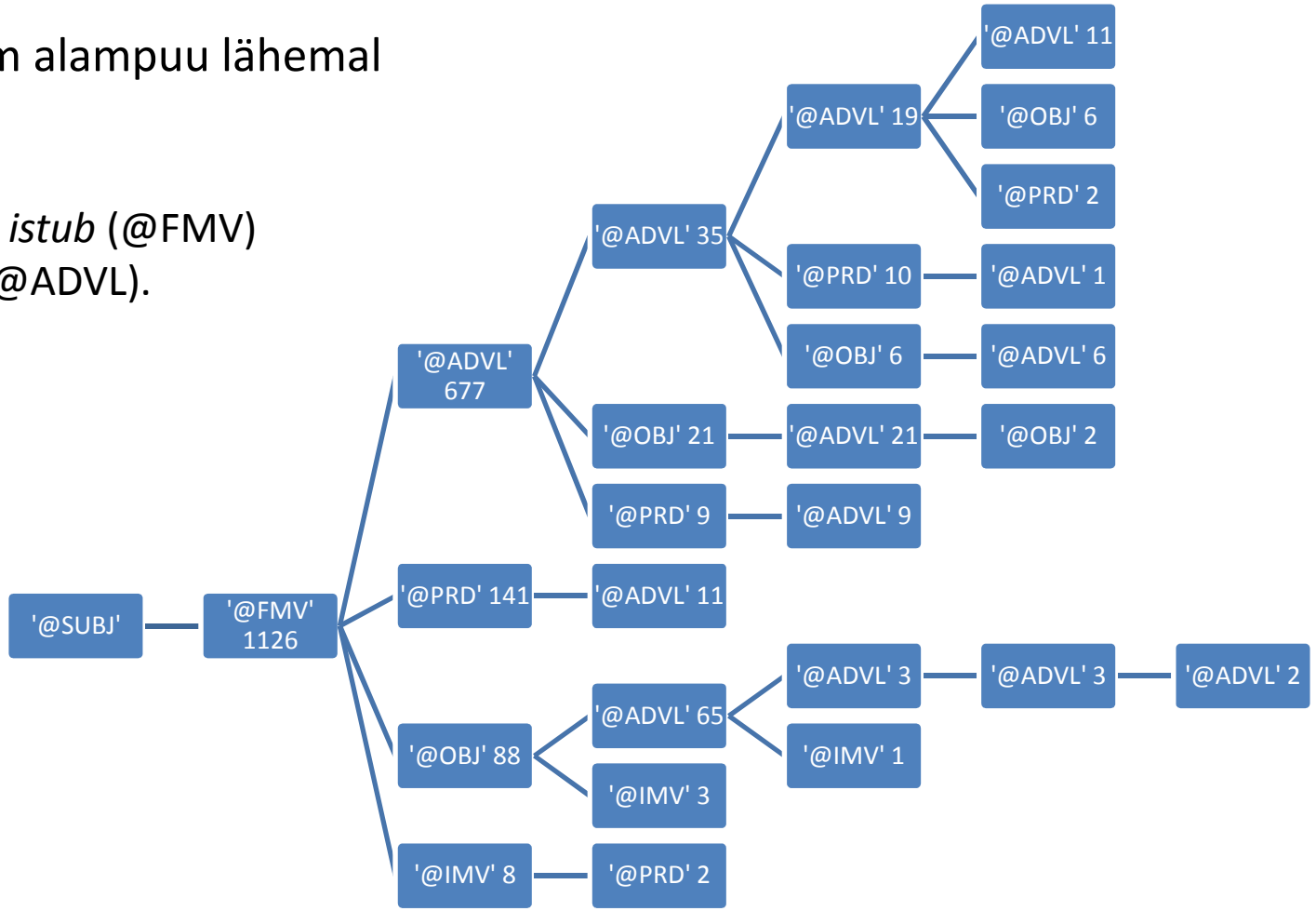
2/3 lausetest (lausete arv ≈10000) ei kasuta sagedasi malle.
Praegu on kinnitamata ≈2500 malle, mis on esinenud 1 või 2 korda.



Andmepuud: süntaktilised märgendid ja sõnajärjemustrid

- Populaarsem alampuu lähemal

Mu mõrsja (@SUBJ) *istub* (@FMV) *salongi laua ääres* (@ADVL).



Sõnajärjevealeidja prototüüp

- Sõnajärjevealeidja prototüübi programmeerimisel võeti aluseks leitud õigete sõnajärjemallide kogum. Realiseerimiseks on kasutatud Zope andmebaasi ning programmeerimiskeelt Python.
- Kõigepealt tehakse kindlaks, kas lause on vaadeldav valitud sõnajärjemallide all või kuulub selliste lausete hulka, millega meie arendatav prototüüp praegu veel ei tegele. Sõna tasandil hoitakse selliseid lisaandmeid nagu sõna lemma, morfoloogiline kirje, süntaktilised märgendid, analüsaatori töö korrektsust hinnanud lingvisti parandused.
- Olemasolevad korrektse sõnajärje mallid paiknevad failis
- Samas on eraldi fail, kus hoitakse uusi (kinnitamata) sõnajärjemalle.
- Kui tegu on vigadeta tekstiga, siis saab selle alusel genereerida uusi korrektseid sõnajärjemalle.
- Kui tegu on veamärgendusega õppijakeele tekstiga, siis on võimalik tuvastada tüüpilisi sõnajärjevigu

Sõnajärjevealeidja prototüüp

- Et vealeidja oleks efektiivsem, selleks on võimalik rakendada kahepoolset kontrolli: ühelt poolt tehakse kindlaks, kas analüüsi tulemusel korrektseks tunnistatud mall sobib konkreetse (osa)lause analüüsimiseks; teisalt vaadatakse, kas lause struktuur kuulub tüüpiliste vigaste struktuuride alla või mitte.
- Osalause leidmisel on suurimaks probleemiks õppijakeeles tehtud grammatilised ja interpunktuatsiooni vead. Kui lauses on koma puudu või sõna valesti kirjutatud, on morfoloogiliselt valesti analüüsitud ka sellised sõnaliigid nagu side- ja tegusõna ning osalausepiiri leidmine võib nurjuda (vt Müürisep, Puolakainen 2007).

Prototüübi testimistulemused

- Prototüübi testimiseks kasutati Eesti vahekeele korpuse B-taseme tekste, mille hulgast võeti juhuvaliku alusel 5880 lauset. Nagu eespool selgitatud, lasti prototüübil välja sorteerida laused, mis algavad sõnaga *kui*, *kuna* või ükskõik millise ja millises käändes küsisõnaga. Sõnajärjevigu ei otsitud ka õigekirjavigu sisaldavatest, hüüumärgiga lõppevatest või umbisikulises kõneviisis olevatest lausetest.

Prototüübi testimistulemused

- Kõiki ülejäänud lauseid kontrollis prototüüp järgmiste vigade osas:

  1) märgend @FMV paikneb kaugemal kui teisel positsioonil;

  2) enne osalausepiiri (CLB) ei ole märgendit @FMV ega @FCV;

  3) märgend @PRD ei ole lauses viimasel positsioonil;

  4) märgend @IMV ei ole lauses viimasel positsioonil.

- Kui prototüüp ei leidnud ühtegi loetletud veakirjeldust, siis võrreldi lausete sõnajärge korrektsete sõnajärjemustritega, mida oli kokku 600. Prototüüp luges laused õigeks, kui lause oli kaetud sobiva õige sõnajärjemalliga.

Prototüübi testimistulemused

- Kokkuvõttes pidas prototüüp juhuvaliku alusel saadud 300 lausest vigaseks 143, korrektseks 72 ja väljajätmisele kvalifitseeruvaks 85 lauset. Lingvisti hinnangu alusel olid samad näitajad vastavalt 146, 75 ja 79. Seega langesid vaadeldud lausete puhul prototüübi töö ja lingvisti hinnangud kokku 87,82% ulatuses.

Kokkuvõte

- VAKO-projekti *Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine (2008–2010)* raames on uuritud eesti keele sõnajärjemalle ja -mustreid, genereeritud sõnajärje andmepuud ning loodud sõnajärje vealeidja prototüüp, mis võimaldab kontrollida (osa)lause süntaktilist struktuuri. Prototüübi graafiline liides on valmimisjärgus
- Sõnajärje empiirilise uurimise käigus saadud andmepuud on olulised mitte ainult keeletehnoloogilistes rakendustes, vaid ka keeleteaduses ja eesti keele õppes.

THANK YOU FOR YOUR ATTENTION!