# EESTI KEELE SÕNAJÄRJE VEALEIDJA PROTOTÜÜBI ARENDAMINE

Erika Matsak, Pille Eslon, Jaagup Kippar

## Ülevaade

Eesti vahekeele korpuse (EVKK) materjalide põhjal on kindlaks tehtud, et eesti keele õppijale valmistab kõige sagedamini raskusi sõnajärg, nt **Majandus kiiresti arenes*. Artiklis kirjeldatakse eesti keele sõnajärjevigade tuvastamise võimalusi, meetodit ja tulemusi. Kõigepealt antakse lühiülevaade olemasolevate automaatsete vealeidjate tööpõhimõtetest ning kirjeldatakse Tallinna Ülikoolis VAKO-projekti raames välja töötatud sõnajärjevealeidja prototüüpi[1]. Sõnajärjevealeidja on statistikapõhine programm, mille tööpõhimõte sarnaneb teatud määral n-grammidega. Programmi aluseks on võetud lause sõnajärje seisukohalt üheksa olulise lauseliikme korrektsed sõnajärjemallid, mille regulaarne ilmnemine keelekasutuses moodustab sõnajärjemustreid. Programmi efektiivsuse tõstmiseks on sõnajärjemustrid paigutatud andmepuusse, kust prototüüp otsib kõigepealt õige algusmärgendiga puu ning seejärel sagedasema

[1] Arendustööd on toetanud riikliku programmi „Eesti keele keeletehnoloogiline tugi 2006–2010“ projekt „VAKO: Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine (2008–2010)“, samuti riikliku programmi „Eesti keel ja kultuurimälu 2009–2013“ projekt „REKKi käsikirjaliste materjalide digiteerimine, Eesti vahekeele korpuse alamkorpuste loomine ja korpuse kasutusvõimaluste populariseerimine (2009–2013)“.

sõnajärjemustri. Prototüübi tööd on testitud *Eesti vahekeele korpuse* tekstidel. Artiklis tuuakse välja sõnajärjevigade tuvastamisel tekkinud probleemid ja pakutakse võimalusi vealeidja efektiivsuse suurendamiseks. Sõnajärjevealeidja prototüübi näol on tegemist universaalsusele pretendeeriva programmiga, sest kasutatud analüüsimeetodit ja sõnajärjevealeidja prototüübi algoritmi saab rakendada ka teiste keelte sõnajärje uurimiseks.

**Võtmesõnad**: morfosüntaks, automaatne veatuvastus, sõnajärjevead

# 1. Vealeidjad ja meetodid

Keeletehnoloogias kasutatakse vigade tuvastamiseks erinevaid meetodeid. Üks selline tugineb spetsiaalsele sõnastikule ehk leksikonile, milles teksti iga sõnet võrreldakse õige sõnavormiga (vt nt Damerau 1964): kui sõnastikust vastet ei leita, loetakse sõnavorm valeks. Niisuguste programmidega on tavaliselt võimalik poolautomaatselt (mõningate sõnade puhul ka täisautomaatselt) parandada vigu, kusjuures aluseks võetakse algoritmid, mis leiavad lähima õige sõna.

Teine meetod põhineb n-grammidel (vt nt Beesley 1988): võrreldakse osasõna ehk näiteks kahte või kolme järjest paiknevat tähte. Ka sel juhul on õigete sõnade loendist genereeritud mallid ning on teada, millised osasõnad millise sagedusega esinevad. Väiksema sagedusega osasõnad on tihti vea tunnuseks. Lisaks on nii sõnade kui osasõnade tasemel võimalik moodustada sagedasemate vigade nimistud (vt nt Pedler 2007). Selliste loendite saamiseks on sõnad korpustes märgendatud: igale vigasele sõnale vastab õige.

Sõnajärje kontrollimiseks erinevates keeltes on n-gramme ja statistilisi meetodeid kasutatud koos (vt nt Athanaselis, Baka-

midis, Dologlou 2006). Sellisel juhul on grammiks terve sõna. Levinud on ka lingvistiline reeglipõhine lähenemine (vt nt nt Arppe 2000).

Eesti keele sõnajärge on kirjeldatud meetodeid rakendades uuritud suhteliselt vähe, kuna eesti lause sõnajärjestruktuuri analüüsiks vajalikke infotehnoloogilisi tugisüsteeme pole seni loodud. Järgnevalt tutvustame *Eesti vahekeele korpuse* sõnajärjevealeidja prototüübi tööpõhimõtteid.

## 2. Sõnajärjevealeidja prototüübi alus

Tallinna Ülikoolis on VAKO-projekti raames sõnajärje kontrollimiseks loodud prototüüp, mille aluseks on eesti kirjakeele korrektsed sõnajärjemallid. Nende väljaselgitamiseks uuriti sõnade süntaktilisi rolle lauses, misjärel töötati välja sõnajärje analüüsimeetod, mis sarnaneb teatud määral n-grammidega ning on kohandatud eesti keelele (vt Matsak, Metslang, Kippar 2010). Võimalikuks sai see tänu Kaili Müürisepa loodud parserile[2], mis on implementeeritud *Eesti vahekeele korpusse*. Parser eristab 27 erineva süntaktilise funktsiooni märgendit[3], kuid lausete sõnajärje analüüs näitas, et suurem osa neist ei mängi sõnajärjes kuigi olulist rolli. Lause sõnajärje seisukohalt on olulised verbi ja lause põhja üheksa märgendit (vt Metslang, Matsak 2010):

1. verbi märgendid
   @FMV – finiitne verb
   @IMV – infiniitne verb
   @FCV – *olema* liitaegades ning modaalverbid ahelverbides, finiitne vorm

[2] Vt: EstCG Parser 1.0a, vt http://www.cs.ut.ee/~kaili/parser/, 25.07.2010.

[3] Vt: http://www.cs.ut.ee/~kaili/parser/demo/synttags.html, 25.07.2010.

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

2. lause põhja märgendid
@SUBJ – alus ehk subjekt
@OBJ – sihitis ehk objekt
@PRD – öeldistäide ehk predikatiiv
@ADVL – määrus ehk adverbiaal, sh fraasiadverbiaal.

Sõnajärjevealeidja prototüübi loomisel on iga esimese osalause ja lihtlause piires otsitud lause põhja märgendite regulaarselt esinevaid järjendeid, mis moodustavad erineva sagedusega ilmnevaid sõnajärjemustreid. Nii näiteks annab eesti keele parser lausele (1) järgmise morfosüntakstilise väljundi:

(1) *Loomulikult ei tohi tunnistada, et sul endal ka need olemas on.*

```
"<s>"

"<Loomulikult>"
   "loomulikult" L0 D cap @ADVL #1->3
"<ei>"
   "ei" L0 V aux neg cap @NEG #2->3
"<tohi>"
   "tohti" L0 V main indic pres ps neg cap <FinV> <Intr>
   @FMV #3->3
"<tunnistada>"
   "tunnista" Lda V main inf cap <NGP-P> @IMV #4->4
"<,>"
   "," Z Com CLB #5->5
"<et>"
   "et" L0 J crd cap @J #6->12
   "et" L0 J sub cap @J #6->12
"<sul>"
      "sina" Ll P pers ps2 sg ad cap @NN> @ADVL #7->7
```

```
"<endal>"
    "ise" Ll P pos det refl sg ad cap @<NN #8->8
"<ka>"
    "ka" L0 D cap @ADVL #9->12
"<need>"
    "see" Ld P dem pl nom cap @SUBJ #10->12
"<olemas>"
    "ole" Lmas V main sup ps in cap <Intr> @ADVL #11->12
"<on>"
    "ole" L0 V main indic pres ps3 pl ps af cap <FinV>
    <Intr> @FMV #12->12
"<.>"
    "." Z Fst #13->13
"</s>"
```

Pärast sõnajärje seisukohalt ebaoluliste märgendite ning sõnade (nt *loomulikult*, *arvatavasti*, *tõenäoliselt*, *niisiis* jt) eemaldamist jätkab programm analüüsi esimese osalause piirini CLB. Toodud näites jääb alles kolme olulise süntakstilise märgendi järjend ['@NEG', '@FMV', '@IMV']. Lihtlause *Internetis* (@ADVL) *on* (@FMV) *võimalik* (@PRD) *kasutada* (@SUBJ) *mitmeid* (@NN>) *teenuseid* (@OBJ) sõnajärjemall on ['@ADVL', '@FMV', '@PRD', '@SUBJ', '@OBJ']. Mall ei arvesta sõnavormi *mitmeid*, sest nimisõnalise eestäiendina @NN> ei kuulu see sõnajärje seisukohalt oluliste märgendite hulka. Samalaadseid probleeme tekitavad sõnajärje analüüsis adverbiaali või fraasiadverbiaali rollis kasutatud sõnad. Järelikult tuleb eelnevalt hinnata, kui olulist osa mängib konkreetne sõna(vorm) lause sõnajärjes. Selleks tuleb iga sõnakasutus lauses üle kontrollida ja kindlaks teha vajalike/mittevajalike märgendite ja ebaoluliste sõnade loendid.

Lisaks tuleb eelnevalt kokku leppida ka selles, milliseid lauseid vealeidja prototüüp vaatleb ning milliseid mitte. Kuigi eesti keele sõnajärje aluseks on süntaktiliste põhimõtete asemel

pigem infostruktuurilised printsiibid (Lindström 2005: 173; EKG II: 13) ja eesti keele põhisõnajärjeks on üldiselt peetud SVX-järge (Vilkuna 1998, Koptjevskaja-Tamm, Wälchli 2001: 705) või statistiliselt pea sama sagedasti esinevat XVS-järge (Tael 1988), võib eesti keele kõrvallausetes, erituslausetes, küsisõnadega algavates küsilausetes ning hüüdlausetes öeldisverb paikneda ka lause lõpus (V3-järg). Seetõttu on prototüüp programmeeritud välja sortima kõik laused, mis algavad sõnadega *kui*, *kuna*, mistahes küsisõnade või nende käändevormidega. Samuti jäetakse välja umbisikulises kõneviisis olevad laused või osalaused. Hüüdlausete tunnuseks loeb prototüüp seni veel lauselõpulist hüüumärki.

Kuna õppijakeele puhul esineb süntaktilist valeanalüüsi, mis on tingitud õigekirjavigadest, siis on vaatluse alt välja jäetud ka kõik (osa)laused, milles esineb mõni õigekirjaviga. Nii ei analüüsi prototüüp lauseid nagu **Ema teda väga **armstas**, vanaema veel rohkem*, sest süntaksianalüsaator määrab vigaselt kirjutatud predikaadi *armstas* (= *armastas*) adverbiaaliks ning sellest tulenevalt leiaks prototüüp, et õppijakeele lausest puudub öeldis. Niisuguste lausete sõnajärge saab hakata kontrollima siis, kui *Eesti vahekeele korpusse* on implementeeritud lemmatiseerija-oletaja, mis seob vigaselt kirjutatud või moodustatud vormi korrektsega ja leiab üles lemma[4].

## 3. Sõnajärjemustrite otsimine

Korrektsete sõnajärjemallide ja -mustrite leidmiseks kasutati esialgu ilukirjandustekstide lausetest koosnevat 3000 sõnelist pilootkorpust. Valimit analüüsiti EstCG Parser 1.0a abil, parseri

[4] Eesti õppijakeele lemmatiseerija-oletaja on VAKO-projekti raames välja töötanud Kairit Sirts.

morfosüntaktiline väljund kopeeriti VBA skriptiga programmeeritud spetsiaalsesse aknasse Excelis (vt joonis 1). Järgnevalt korrastati väljund automaatselt: programm paigutas kogu morfosüntaktilise info tabelisse ja otsis iga lause algusosas sõnajärje seisukohalt oluliste märgendite järjestusi. Automaatselt leitud süntaktiliste märgendite järjendid kontrolliti käsitsi üle, korrektseks tunnistati 242 sõnajärjemalli.

| | | | Sisesta tekst | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Vea tüüp | | | | | | | | | | | |
| | Tekst korda | Paiguta | Süntaksianalüsaatori väljund | Vea tüüp | Sõna vorm | Morfosüntaktiline analüüs | 1 | 1.1 | | 1.2 | 1.2. | 1.2. |
| 4 | 1 | SLAS | | | | | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 2 | Koolist | kool+st // S_ com sg el #cap // **CLB @ADVL | | kool+st | // S_ com sg el #cap // **CLB @ADVL | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 3 | ma | mina+0 // P_ pers ps1 sg nom // @SUBJ | 7.1 | mina+0 | // P_ pers ps1 sg nom // @SUBJ | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 4 | ei | ei+0 // V_ aux neg // @NEG | | ei+0 | // V_ aux neg // @NEG | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 5 | või | või+0 // V_ mod indic pres ps neg #FinV #inf // @+FCV | | või+0 | // V_ mod indic pres ps neg #FinV #inf // @+FCV | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 6 | midagi | miski+dagi // P_ indef sg part // @ADVL | | miski+dagi | // P_ indef sg part // @ADVL | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 7 | eriti | eriti+0 // D_ // @ADVL | | eriti+0 | // D_ // @ADVL | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 8 | halba | halb+0 // A_ pos sg part // @<AN | | halb+0 | // A_ pos sg part // @<AN | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 9 | ütlelda | ütlelda+0 // S_ com sg nom #? // @PRD | | ütlelda+0 | // S_ com sg nom #? // @PRD | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 10 | . | .//_Fst // | | . | //_Fst // | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 11 | SLLS | | | | | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | sisesta | | | | | | 0 | 0 | 0 | 0 | 0 | 0 |

**Joonis 1.** Abivahend sõnajärjemallide otsimiseks

Kuna Exceli töökiirus on väike, siis sobib see abivahend idee testimiseks, mitte korpusainese automaatanalüüsiks. Seetõttu implementeeriti algoritmid Eesti vahekeele korpusse, kus lisaks korpuse tekstidele oleks võimalik analüüsida mistahes tekste ja nende keelelist korrektsust.

Prototüüpi testiti ilukirjandustekstidest võetud 20000 lausel. Valimi automaatanalüüsi tulemusena leiti vajalike süntaktiliste märgendite järjestus esimeses osalauses ja lihtlauses, määrati kindlaks ühesuguste järjendite esinemissagedus ning toodi välja regulaarselt kasutatud sõnajärjemustrid. Kinnitust leidis 600 korrektset sõnajärjemalli. Kuna uusi tekste produtseeritakse pidevalt juurde ja korpuse maht kasvabkiiresti, siis tuleb usaldusväärsete analüüsitulemuste saavutamiseks tagada programmi efektiivsus ja ka piisav töökiirus (vt Matsak, Metslang, Kippar 2010). 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. Teisalt võimaldab see kasvatada andmepuude panka, lisades näiteks sõnajärjemustrite stiilivariante jm.

# 4. Andmepuud: süntaktilised märgendid ja sõnajärjemustrid

Andmepuud genereeriti rohkem kui 10000 lause alusel, kuna algsest 20000 lauset sisaldavast valimist ei analüüsi prototüüp 8590 lauset. Sõnajärjepuid kirjeldatakse vastavalt sagedusele. Kui sõnajärjevealeidja prototüüp genereerib puid ning tuvastab õige/vigase sõnajärje süntaktiliste märgendijärjendite ja nende osade kokkulangevuse alusel ehk paradigmaatiliselt, siis eesti keele (osa)lausete regulaarselt korduvad (tüüpilised, keeleomased) sõnajärjemallid ilmnevad märgendi sageduse järgi süntagmaatiliselt. Nende interpreteerimiseks tuleb iga puu haru vaadelda lineaarselt ning märgendite sagedusest lähtudes. Järgnevalt kirjeldame lühidalt eesti keele sõnajärjepuude andmepanka ja keeleomaseid korrektseid sõnajärjemalle, mis on välja toodud algusmärgendi sageduse põhjal. Allpool ei too me iga sõnajärjepuu detailseid kirjeldusi, vaid peatume sagedasematel eesti keelele omastel sõnajärjemustritel. Sõnajärjepuude ülevaade lõpeb üldistusega, milleks on levinumate sõnajärjemustrite andmepuu. Selle puu genereerimisel on nähtavaks jäetud vaid need märgendid, mida ligi 10000-lauselises valimis on esinenud vähemalt neljakümnes lauses.

## 4.1. Verbi eitus

Verbi eitust @NEG sisaldavaid lauseid oli valimis kõige vähem – vaid 6. Seetõttu on nende alusel kõige lihtsam näidata andmepuu etappide kaupa ehitamist ja sõnajärjemustrite järk-järgulist genereerimist. Kõigepealt toodi välja (osa)lauseid iseloomustavad sõnajärjemallid (vt näitelauseid 2–7) ning seejärel võrreldi neid üksteisega.

(2) *Loomulikult ei* (@NEG) *tohi* (@FMV) *tunnistada* (@IMV) *, et sul endal ka need olemas on .*
['@NEG', '@FMV', '@IMV']

(3) *Seega ei* (@NEG) *saa* (@FMV) *me* (@SUB) *oma läkitust* (@OBJ) *kodeerida* (@IMV) *nii , et see oleks mõistetav .*
['@NEG', '@FMV', '@SUBJ', '@OBJ', '@IMV']

(4) *Siis ei (*@NEG*) lausu* (@FMV) *ma* (@SUBJ) *enam ühtki sõna* (@OBJ) *ega mõtle enam millelegi , kõik vajub enneaimamatusse õndsusemerre .*
['@NEG', '@FMV', '@SUBJ', '@OBJ']

(5) *Ei* (@NEG) *ole* (@FMV) *enam millest* (@ADVL) *rääkida* (@SUBJ) *, ta tahab maale saada , ehk ta seda küll ei ütle .*
['@NEG', '@FMV', '@ADVL', '@SUBJ']

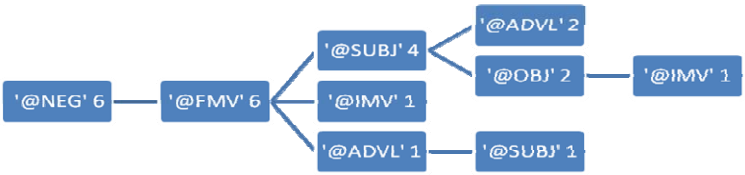
(6) *Ei* (@NEG) *ole* (@FMV) *jahu* (@SUBJ) *põskedel* (@ADVL) *ja huuled on loomulikult värsked .*
['@NEG', '@FMV', '@SUBJ', '@ADVL']

(7) *Siiski ei* (@NEG) ole (@FMV) *ma* (@SUBJ) *teie peale* (@ADVL) *väga tige , et te mu üles ajasite .*
['@NEG', '@FMV', '@SUBJ', '@ADVL']

Selgus, et 6. ja 7. näitelause sõnajärjemall ['@NEG', '@FMV', '@SUBJ', '@ADVL'] langeb kokku. Seejärel otsiti süntaktiliste märgendite järjendites osalisi kokkulangevusi, mis võimaldab esile tuua eituslause sõnajärjemustri hargnevusi. Näiteks:

| | | | | | |
|---|---|---|---|---|---|
| 1 lause | '@NEG' | '@FMV' | '@IMV' | | |
| 1 lause | '@NEG' | '@FMV' | '@SUBJ' | '@OBJ' | '@IMV' |
| 1 lause | '@NEG' | '@FMV' | '@SUBJ' | '@OBJ' | |
| 1 lause | '@NEG' | '@FMV' | '@ADVL' | '@SUBJ' | |
| 2 lauset | '@NEG' | '@FMV' | '@SUBJ' | '@ADVL' | |

Kuna algusmärgendite @NEG ja @FMV järgnevus kattub kõikides analüüsitud lausetes ja kolmandal positsioonil esineb teistest sagedamini märgend @SUBJ, siis ei pea andmepuus samu märgendeid kordama. Nende järgnevuste mugavamaks leidmiseks sorditi vastavat haru puus ülespoole (vt joonis 2).

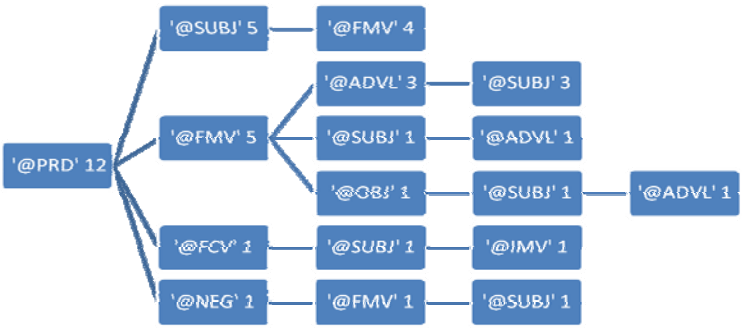



**Joonis 2.** Verbi eitusega algavate lausete sõnajärjemustrid

Joonise 2 alusel võib väita, et verbi eitusega algavatel (osa)lausetel on kaks sagedamini kasutatavat sõnajärjemustrit: ['@NEG', '@FMV', '@SUBJ', '@ADVL'] ja ['@NEG', '@FMV', '@SUBJ', '@OBJ'].

## 4.2. Predikatiiv

Predikatiiviga algavaid sõnajärjemalle leiti ilukirjandustekstide valimist 12, seega rohkem kui verbi eitusega algavaid sõnajärjemalle. Keerulisemad on ka puu hargnevused, sõnajärjemustreid on rohkem ja nad varieeruvad sagedamini, vt joonis 3).




**Joonis 3.** Predikatiiviga algavate lausete sõnajärjemustrid

Levinumad predikatiiviga algavad sõnajärjemustrid on ['@PRD', '@SUBJ', '@FMV'] – vt 8. näidet ja ['@PRD', '@FMV', '@ADVL', '@SUBJ'] – vt 9. näidet.

(8) *Ja saatanlik* (@PRD) *nagu ta* (@SUBJ) *oli* (@FMV), *ei suutnud ta jätta puusi diskreetselt nõksutamata* .

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

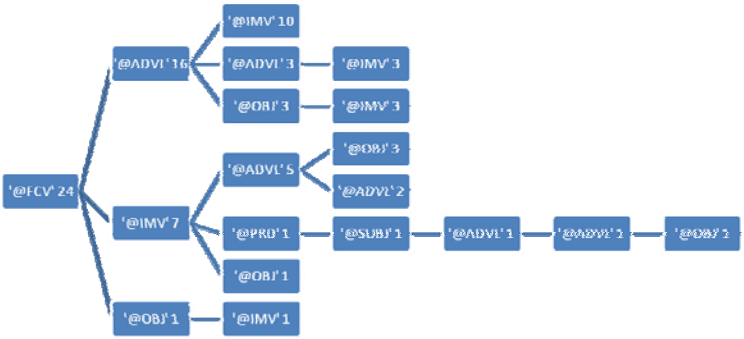
(9) *Raske* (@PRD) *on* (@FMV) *end tagasi* (@ADVL) *hoida* (@SUBJ), *kui ümberringi kõik purskab ja pritsib* .

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

## 4.3. Verb *olema* liitaegades ning modaalverbid ahelverbides (finiitvormid)

Verbi *olema* finiitse vormiga @FCV algavaid sõnajärjemalle oli valimis 24. Kõige regulaarsemalt kasutati mustrit, kus teisel positsioonil oli adverbiaal @ADVL ning poole võrra vähem

mustrit, kus verbi *olema* finiitsele vormile järgnes infiniitne verb @IMV, üksikjuhtumil objekt @OBJ (vt joonis 4).




**Joonis 4.** Sõnajärjemustrid, mille alguses on verbi *olema* või modaalverbide finiitne vorm

Enimkasutatud sõnajärjemustrites on liitaja vormi komponendid (*olema* finiitne vorm ja kesksõna) tavaliselt lahutatud teisel positsioonil asuva adverbiaaliga, vt 10. näidet.

(10) *Olin* (@FCV) *tihti* (@ADVL) *tundnud* (@IMV), *et see suvi oli mind välja valinud, kutsunud täitma prohvetlikku ülesannet.*
['@FCV', '@ADVL', '@IMV']

Järgnevad sõnajärjemustrid, milles kolmandal positsioonil on kas teine adverbiaalne lauselaiend (vt 11. näidet) või objekt (vt joonis 4).

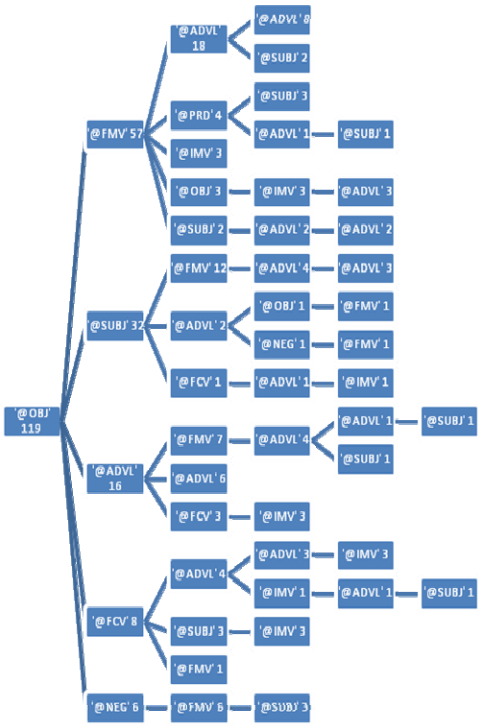
(11) *Olin* (@FCV) *juba* (@ADVL) *põlvini ürglimasse*
(@ADVL) *vajunud* (@IMV) *ja vankusin* .
['@FCV', '@ADVL', '@ADVL', '@IMV']

Harvem tuleb ette niisuguseid sõnajärjemustreid, milles liitaja vormi on kõrvuti: ['@FCV', '@IMV', '@ADVL', '@OBJ'] – (vt 12. näidet) ja ['@FCV', '@IMV', '@ADVL', '@ADVL'] (vt joonis 4).

(12) *Olen* (@FCV) *lasknud* (@IMV) *enesele* (@ADVL) *uue suve-ülikonna* (@OBJ) *valmistada , mis mu lühikest ja lihavat keha näib nägusamaks tegevat .*
['@FCV', '@IMV', '@ADVL', '@OBJ']

## 4.4. Objekt

Objektiga algavaid sõnajärjemalle oli valimis 119. Seetõttu osutus andmepuu ülesehitus eelmiste puude ülesehitusest keerukamaks ja märgendijärjendite varieeruvus mitmekesisemaks (vt joonis 5). Objektile järgneb kõige sagedamini finiitne verbivorm @FMV (57 korda), harvem subjekt @SUBJ (32) ja adverbiaal @ADVL (16), harva verb *olema* liitaegades või modaalverbid ahelverbides @FCV (8) ning verbi eitus @NEG (6). Objektiga algava sõnajärjepuu harus, kus teisel positsioonil on verbi finiitne vorm ['@OBJ', '@FMV'], ilmneb kolmandal positsioonil kõige järjekindlamalt ja seega suurema tõenäosusega adverbiaal @ADVL (18 korda) – vt 13. näidet. Võimalikud on ka predikatiiv (vt 14. näidet), verbi infiniitne vorm (vt 15. näidet) ning subjekt. Kui objektiga algava sõnajärjemustri teine komponent on subjekt, siis on üsna tõenäone, et kolmandal positsioonil seisab verbi finiitne vorm, neljandal ja viiendal adverbiaal jne (vt joonis 5). Järelikult, mida sagedasemad on keelekasutuses ühesuguste algusharudega puud, seda selgemalt tulevad esile regulaarselt kasutatavad sõnajärjemustrid. Analoogselt saab interpreteerida kõiki järgnevaid sõnajärjepuid.




**Joonis 5.** Objektiga algavate lausete sõnajärjemustrid

(13) *Nad* (@OBJ*) tunneb* (@FMV) *siin* (@ADVL) *kohe ära.*
['@OBJ', '@FMV', '@ADVL']

(14) *Luuletust* (@OBJ) *oli* (@FMV) *lihtne* (@PRD) *lugeda* (@SUBJ) *ja see jäi ruttu meelde .*
['@OBJ', '@FMV', '@PRD', '@SUBJ']

(15) *Inimesi* (@OBJ') *peab* (@FMV) *koondama* (@IMV) *ning esimesed kandidaadid on ütlematagi Internetile töötajad .*
['@OBJ', '@FMV', '@IMV']

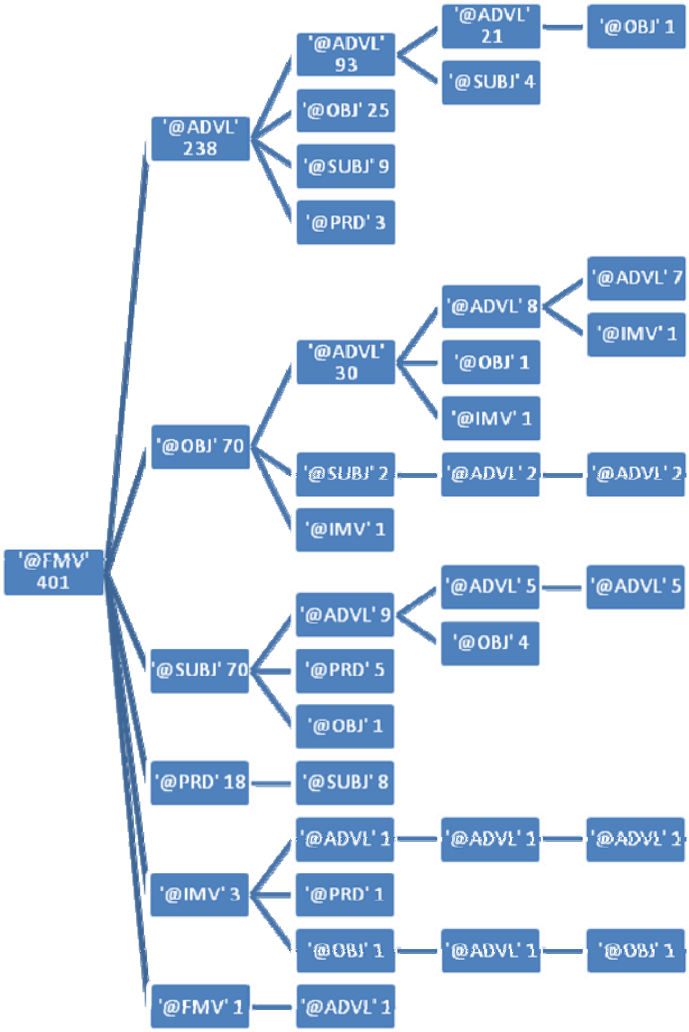
## 4.5. Finiitne verbivorm

Verbi finiitse vormiga @FMV algavaid sõnajärjemalle on valimis 401 (vt joonis 6). Sõnajärjemustreid moodustavad märgendite järjendid, milles teisel positsioonil on tavaliselt adverbiaal @ADVL (238 korda), harvem objekt @OBJ (70) või subjekt @SUBJ (70) ja predikatiiv @ PRD (18), üksikjuhtumitel ka infiniitne verb @IMV (3) ja verbi finiitne vorm @FMV (1).

Verbi finiitse vormiga algavate sõnajärjemustrite andmepuu hargnevused näitavad ühte kindlat sõnajärjeseaduspära: kui verbi finiitsele vormile järgneb adverbiaal ['@FMV', '@ADVL'], siis on kolmandal ja neljandal positsioonil suure tõenäosusega teine ning kolmaski adverbiaal (vt 16. näidet). Samalaadne märgendite järjestus hakkab silma ka siis, kui verbi finiitsele vormile järgneb objekt või subjekt: kolmandal ja neljandal positsioonil on sel juhul tavaliselt adverbiaal (vt 17. näidet); üksikjuhul kordub see seaduspärasus ka teisel positsioonil oleva infiniitse verbiga (vt tabel 6).

(16) *Jäin* (@FMV) *seejärel* (@ADVL) *kohe* (@ADVL) *magama*.
['@FMV', '@ADVL', '@ADVL']

(17) *Pane* (@FMV) *silmad* (@OBJ) *kinni* (@ADVL) *või ma tapan su ära* .
['@FMV', '@OBJ', '@ADVL']




**Joonis 6.** Verbi finiitse vormiga algavad sõnajärjemustrid

## 4.6. Adverbiaal, sh fraasiadverbiaal

Adverbiaaliga algavaid sõnajärjemalle oli valimis kokku 857. Kuna sõnajärjemustrid muutuvad järjest keerulisemaks, kirjeldame adverbiaaliga algavat andmepuud alaliikide kaupa, mis on reastatud sõnajärjemustrite kahe algusmärgendi sageduse alusel:

1) adverbiaal + finiitne verb ['@ADVL', '@FMV'] – 465 korda
2) adverbiaal + subjekt ['@ADVL', '@SUBJ'] – 128 korda
3) adverbiaal + adverbiaal ['@ADVL', '@ADVL'] – 122 korda
4) adverbiaal + verb *olema* liitaegades ning modaalverbid ahelverbides (finiitne vorm) ['@ADVL', '@FCV'] – 73 korda, adverbiaal + verbi eitus ['@ADVL', '@NEG'] – 31 korda, adverbiaal + infiniitne verb ['@ADVL','@IMV'] – 14 korda ja adverbiaal + objekt ['@ADVL', '@OBJ')] – 14 korda.

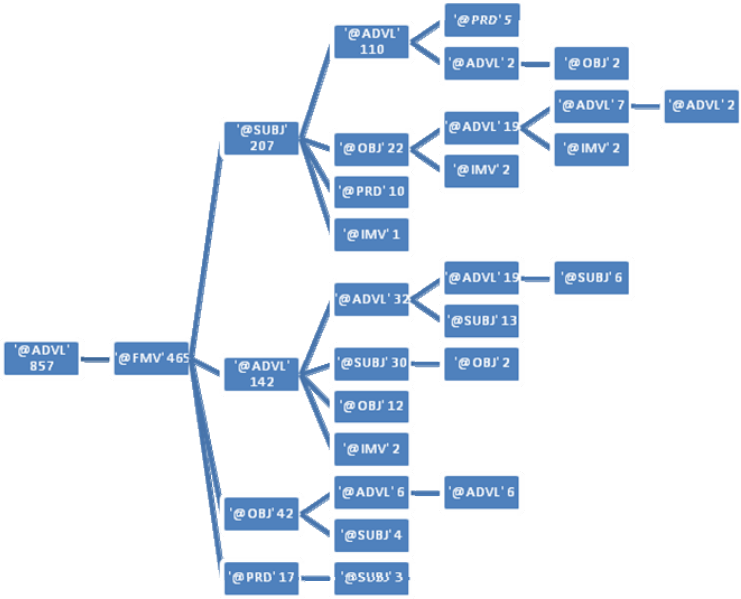
Levinum adverbiaaliga algav sõnajärjemuster on ['@ADVL', '@FMV', '@SUBJ'], kus teisel positsioonil on verbi finiitne vorm ja kolmandal subjekt (vt 18. näidet). Neljandal positsioonil järgneb subjektile tavaliselt adverbiaal, harvem objekt või predikatiiv (vt joonis 7).

(18) *Esimest korda* (@ADVL) *satub* (@FMV) *ta* (@SUBJ) *silme ette*. ['@ADVL', '@FMV', '@SUBJ']

Veidi harvem on sõnajärjemuster, kus kolmandal, neljandal ja viiendal positsioonil on adverbiaal ['@ADVL', '@FMV', '@ADVL', '@ADVL'] (vt 19. näidet ja joonis 7).

(19) *Oma ukse lävel* (@ADVL) *seisab* (@FMV) *siin* (@ADVL) *valge põllega* (@ADVL) *lihakaupleja* (@SUBJ) *ja pagari aknal on pinu sihvakaid saiu , pikki ja paksusid kui kasehalud* . ['@ADVL', '@FMV', '@ADVL', '@ADVL', '@SUBJ']

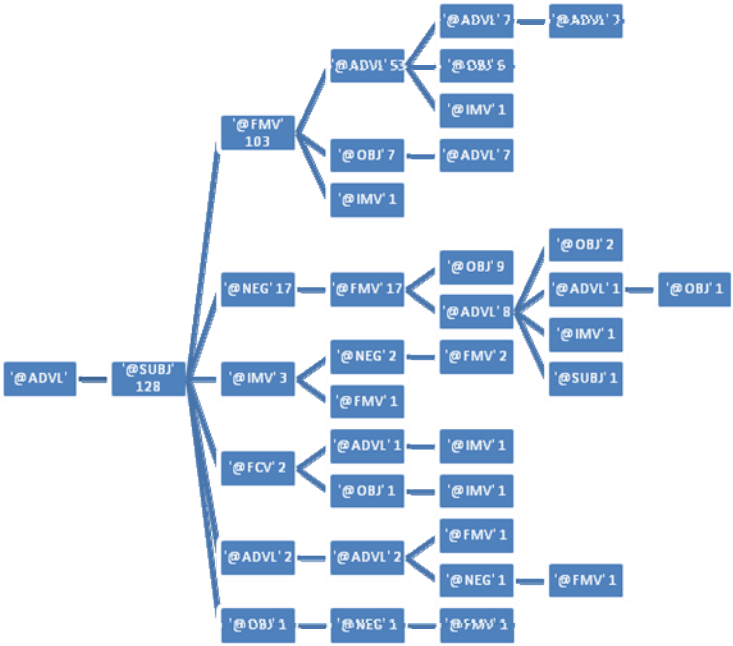
Ka selles sõnajärjemustris, kus adverbiaalile ja verbi finiitsele vormile järgneb objekt, seisab adverbiaal tavaliselt nii neljandal kui ka viiendal positsioonil (vt joonis 7).




**Joonis 7.** Adverbiaaliga algavad sõnajärjemustrid (alaliik 1)

Sageduselt teine adverbiaaliga algava andmepuu alaliik on ['@ADVL', '@SUBJ'] – 128 kasutusnäidet (vt joonis 8). Tavaliselt järgneb subjektile verbi finiitne vorm @FMV – 103 korda, tunduvalt harvemini verbi eitus @NEG (17 korda), väga harva või üksikjuhtumitel infiniitne verb, verb *olema* liitaegades ning modaalverbid ahelverbides, adverbiaal ja objekt. Sõnajärjemustri ['@ADVL', '@SUBJ', '@FMV'] neljas märgend on suure tõenäosusega teine adverbiaal (vt 20. näidet ja joonis 8). Kui

adverbiaalile ja subjektile järgneb verbi eitus, siis on võimalik, et järgmisena lisandub objekt (vt 21. näidet) või teine adverbiaal (vt joonis 8).

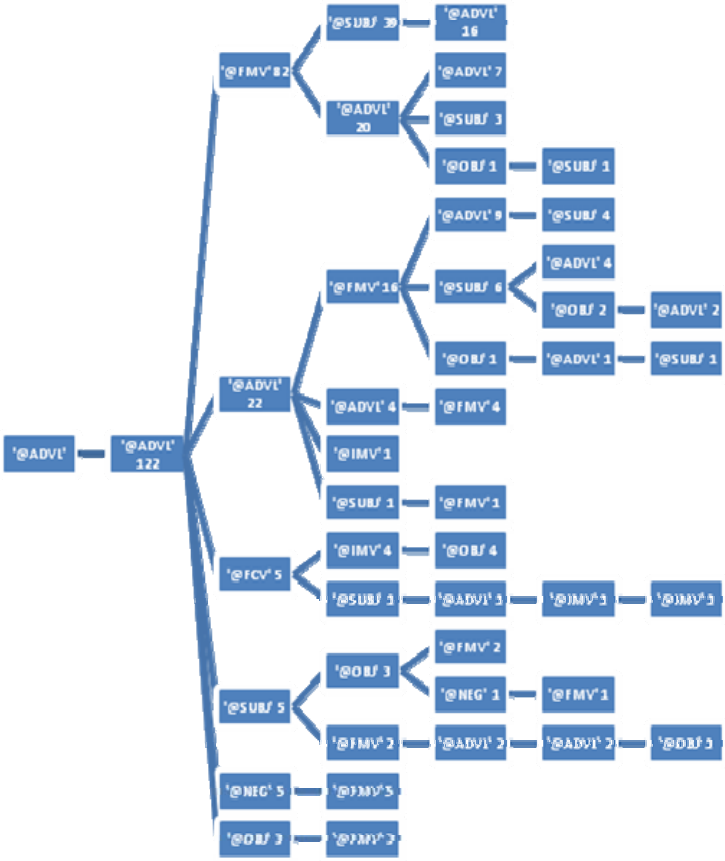



**Joonis 8.** Adverbiaaliga algavad sõnajärjemustrid (alaliik 2)

(20) *Siis* (@ADVL) *nad* (@SUBJ) *tulevad* (@FMV) *koos* (@ADVL) *alla* (@ADVL) *ja* ...

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

(21) *Siin* (@ADVL) *nad* (@SUBJ) *ei* (@NEG) *kalla* (@FMV) *õlut* (@OBJ) *maha* (@ADVL) *, ei karju , ei tuigu .*
['@ADVL', '@SUBJ', '@NEG', '@FMV', '@OBJ', '@ADVL']




**Joonis 9.** Adverbiaaliga algavad sõnajärjemustrid (alaliik 3)

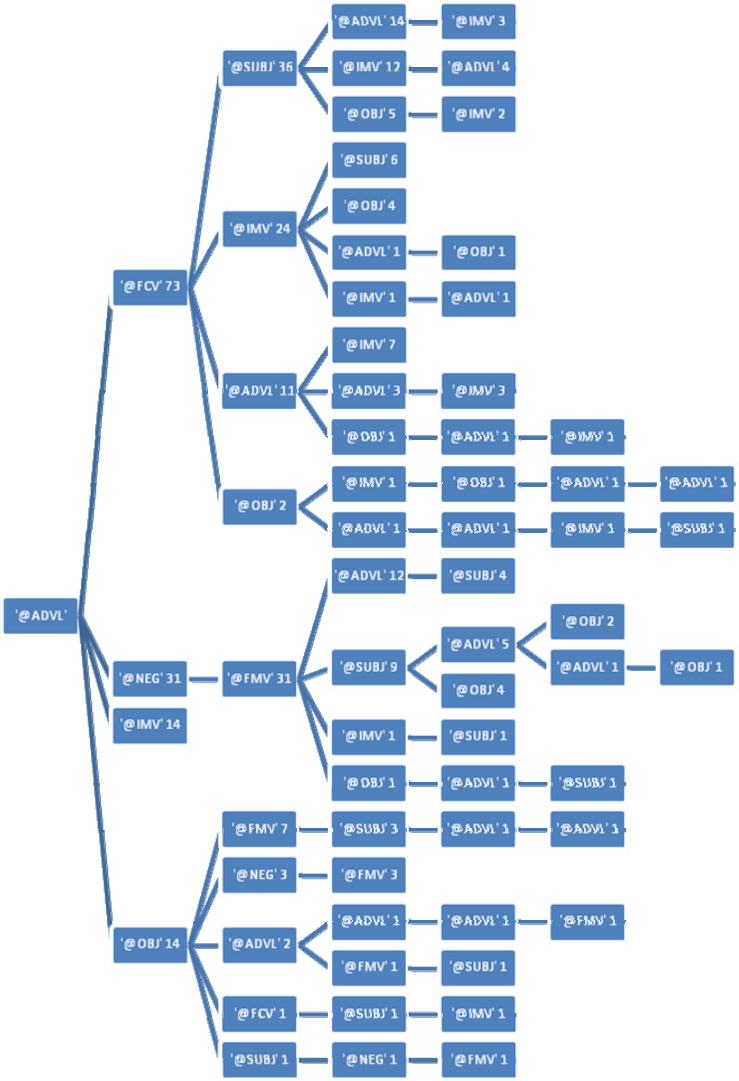
Kolmanda sõnajärjemustrite alaliigi moodustavad sõnajärjemallid, mille alguses on kaks adverbiaali ['@ADVL', '@ADVL'] (122 korda, vt joonis 9). Kõige sagedamini järgneb teisel posit-

sioonil olevale adverbiaalile verbi finiitne vorm @FMV (82 korda), harvem kolmas adverbiaal @ADVL (22 korda), väga harva verb *olema* liitaegades ning modaalverbid ahelverbides, subjekt, verbi eitus ja objekt (vt joonis 9). Sõnajärjemuster ['@ADVL', '@ADVL', '@FMV'] jätkub tavaliselt kas subjekti (vt 22. näidet) ja adverbiaaliga või siis kolmanda ja sellele järgneva neljanda adverbiaaliga (vt 23. näidet).

(22) *Sealsamas* (@ADVL) *kõrval* (@ADVL) *oli* (@FMV) *viljapõld* (@SUBJ) , *moone täis, ja põlluservas kitkus üks naine rohtu , hall seelik üles kääritud , see võis olla tuhande aasta eest .*
['@ADVL', '@ADVL', '@FMV', '@SUBJ']

(23) *Aga teel* (@ADVL) *kohvikusse* (@ADVL) *mõtled* (@FMV) *ikkagi* (@ADVL) *ümber* (@ADVL) , *pistad talle viiemargase pihku ja lased jalga .*
['@ADVL', '@ADVL', '@FMV', '@ADVL', '@ADVL']

Adverbiaaliga algava andmepuu neljanda alaliigi sõnajärjemustrid (vt joonis 10) moodustuvad nende sõnajärjemallide alusel, milles teisel positsioonil on kas verbi *olema* liitajad ja modaalverbid ahelverbides ['@ADVL', '@FCV'] (73 korda), verbi eitus ['@ADVL', '@NEG'] (31 korda), infiniitne verb ['@ADVL','@IMV'] (14 korda) või objekt ['@ADVL', '@OBJ')] (14 korda). Neljanda alaliigi sagedasemas sõnajärjemustris ['@ADVL', '@FCV'] on kolmandal positsioonil enamasti subjekt (36 korda), veidi harvem infiniitne verb (24 korda), teine adverbiaal (11 korda), üksikjuhtumitel objekt. Tavaliselt jätkab sõnajärjemustrit ['@ADVL', '@FCV', '@SUBJ'] kas teine adverbiaal (14 korda, vt 24. näidet) või infiniitne verb (12 korda, vt 25. näidet). Sõnajärjemustri ['@ADVL', '@FCV', '@IMV'] järgmine märgend on (@SUBJ (subjekt) või @OBJ (objekti) jne.



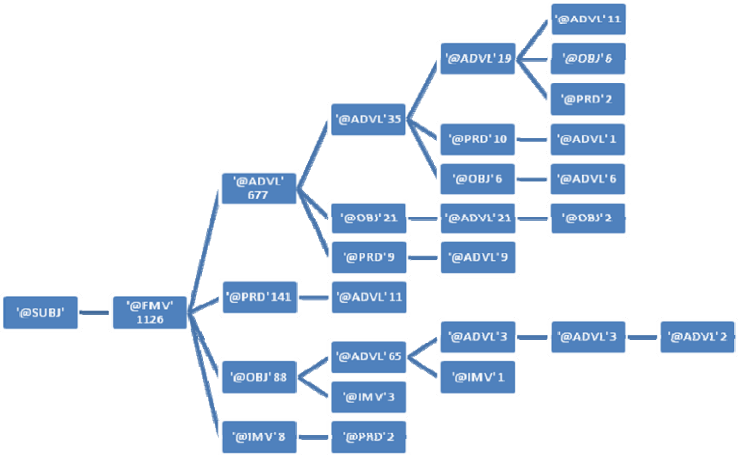

**Joonis 10.** Adverbiaaliga algavad sõnajärjemustrid (alaliik 4)

(24) *Seepärast* (@ADVL) *oli* (@FCV) *ta* (@SUBJ) *uksele* (@ADVL) *kirja* (@OBJ) *kinnitanud* (@IMV) *, et tuleb varsti tagasi ja palub toas oodata .*
['@ADVL', '@FCV', '@SUBJ', '@ADVL', '@OBJ', '@IMV']

(25) *Minu mängu* (@ADVL) *oli* (@FCV) *sugenenud* (@IMV) *pikem paus* (@SUBJ) – *polnud raha* ['@ADVL', '@FCV', '@IMV', '@SUBJ']

## 4.7. Subjekt

Subjektiga algav andmepuu on genereeritud kõige sagedamini esinenud sõnajärjemallide alusel. Regulaarsemalt kasutatud sõnajärjemustrite alguses järgneb subjektile tavaliselt verbi finiitne vorm (1236 korda, vt joonis 11). Seni vaadeldud andmepuudest erineb see puu mitme tunnuse poolest. Prototüübi genereeritud sõnajärjemallide üldise paradigma alusel on näha,




**Joonis 11.** Subjektiga algav sõnajärjemuster ['@SUBJ', '@FMV']

et subjektiga algav sõnajärjepuu pole struktuurilt ega variantide rohkuselt eelnevalt kirjeldatud puudest keerulisem. Lisaks domineerivale sõnajärjele ['@SUBJ', '@FMV'] kasutatakse sageli veel vaid kahte subjektiga algavat mustrit, milles teisel positsioonil on verb *olema* liitaegades, modaalverb ahelverbides (110 korda, vt joonis 12) või verbi eitus (110 korda, vt joonis 13). Oluliselt vähem esineb sõnajärge, mille puhul subjektile järgneb adverbiaal (46 korda, vt joonis 14). Seega sisaldub kõige sagedasema algusmärgendiga andmepuudes vähem sõnajärjemustreid kui eelnevalt kirjeldatud puudes, kuid need katavad suurema jao valimi osalausetest ja lihtlausetest.

Algusmärgendiga ['@SUBJ', '@FMV'] seisab pooltel juhtudel kolmandal positsioonil adverbiaal (677 korda 1126 kasutusest) ning sellele võib järgneda veel kolm adverbiaali (vt 26. näidet). Harvem esineb muster, kus kolmandal positsioonil olevale adverbiaalile järgneb objekt (vt 27. näidet).

(26) *Mu mõrsja* (@SUBJ) *istub* (@FMV) *salongi laua ääres* (@ADVL).
['@SUBJ', '@FMV', '@ADVL']

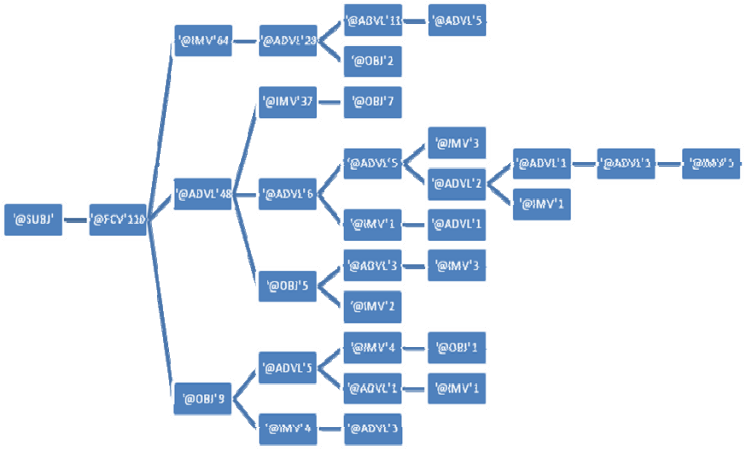
(27) *See* (@SUBJ) *tekitas* (@FMV) *alati* (@ADVL) *hirmu* (@OBJ) *, kui korrapealt ei olnud enam midagi näha läbi , vaid ainult sissepoole . Kuid asja muutis hullemaks see , et nüüd oli minu akna taga mitte tühi sein nagu lapsepõlves , vaid rõdu .*
['@SUBJ', '@FMV', '@ADVL', '@OBJ']

Algusmärgenditele ['@SUBJ', '@FMV'] järgneb veelgi harvem predikatiiv (141 korda) (vt 28. näidet) või objekt (88 korda).

(28) *Ma* (@SUBJ) *olen* (@FMV) *kindel* (@PRD) *selles* (@ADVL) *, et ta mind hakkab armastama .*
['@SUBJ', '@FMV', '@PRD', '@ADVL']

Sõnajärjepuud, mille algusmärgendid on ['@SUBJ', '@FCV'], kasutatakse teistest subjektiga algavatest puudest tunduvalt

harvem (vt joonis 12). Kolmandal positsioonil on enamasti infiniitne verbivorm (64 korda, vt 29. näidet) või adverbiaal (48 korda, vt 30. näidet).



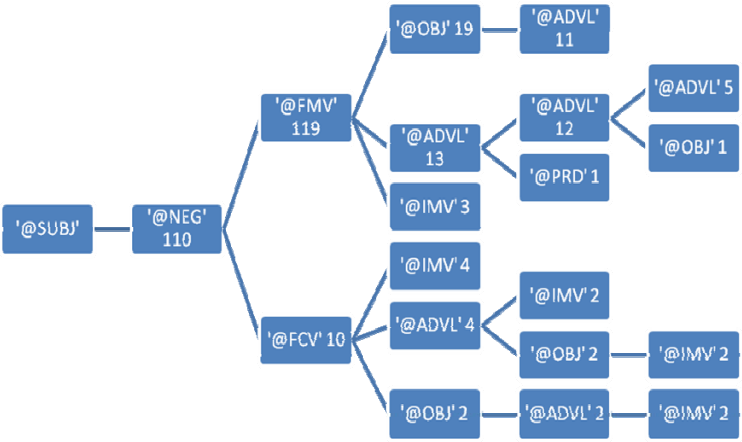

**Joonis 12.** Subjektiga algava lause sõnajärjemuster ['@SUBJ', '@FCV']

(29) *Pilved* (@SUBJ) *olid* (@FCV) *tulnud* (@IMV) *päikesele nii lähedale* (@ADVL) , *et lausa nühkisid teda , kriimustades tema kollast lakki .*
['@SUBJ', '@FCV', '@IMV', '@ADVL']

(30) *Aga tema* (@SUBJ) *oli* (@FCV) *trepi alla* (@ADVL) *läinud* (@IMV) , *kus turistid seisavad ja jõllitavad .*
['@SUBJ', '@FCV', '@ADVL', '@IMV']

Subjektiga algavas verbi eitust sisaldavas sõnajärjepuus ['@SUBJ', '@NEG'] on kolmandal positsioonil valdavalt verbi finiitne vorm @FMV (119 korda, vt 31. näidet), millele järgneb kas objekt (19 korda) või adverbiaal (13 korda).

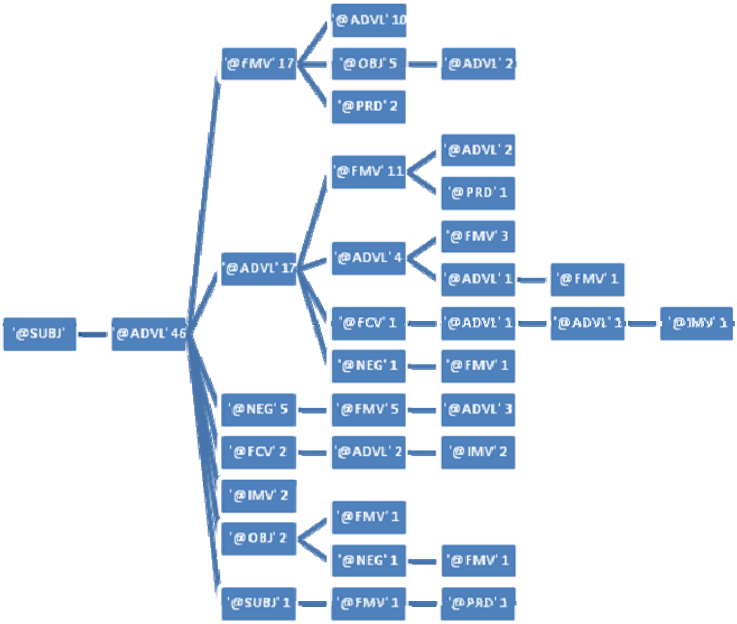
(31) *Ma* (@SUBJ) *ei* (@NEG) *tea* (@FMV) *, kas ma tahangi seda enam kirjutada , sulle .* ['@SUBJ', '@NEG', '@FMV']




**Joonis 13.** Subjektiga algav sõnajärjemuster ['@SUBJ', '@NEG']

Subjektiga algavat sõnajärjepuud ['@SUBJ', '@ADVL'], kus teisel positsioonil on adverbiaal, kasutatakse eelmistega võrreldes harva (46 korda). Kolmandal positsioonil on enamasti verbi finiitne vorm (17 korda, vt 32. näidet) või adverbiaal (17 korda, vt joonis 14).

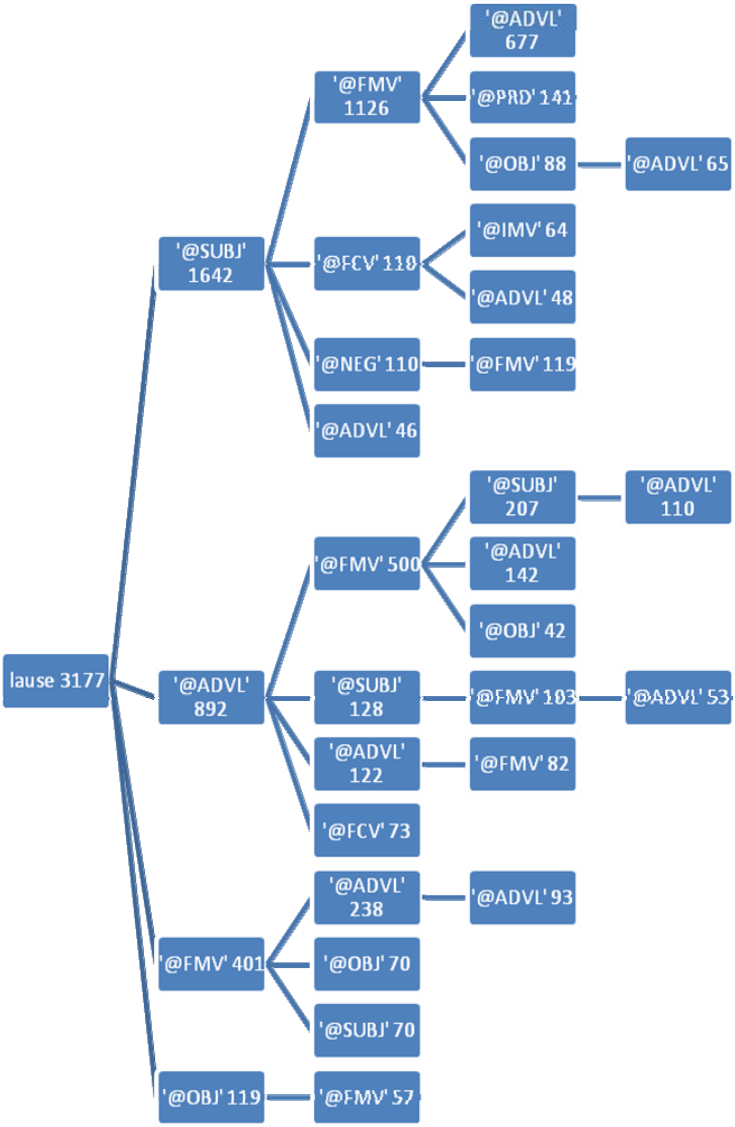
(32) *Aurutoru* (@SUBJ) *aina* (@ADVL) *puhub* (@FMV) *kaebavalt* (@ADVL) *ja kaugemalt udu seest vastavad teised laevad koledalt ja hädaohtu aimates nagu linnud , kes üksteist kiskja eest hoiatavad , kes neid kuskil luurab .*
['@SUBJ', '@ADVL', '@FMV', '@ADVL']




**Joonis 14.** Subjektiga algav sõnajärjemuster ['@SUBJ', '@ADVL']

## 4.8. Sõnajärjepuudest kokkuvõtvalt

Vaatamata sellele, et valimi alusel leitud korrektsetest sõnajärjemallidest genereeriti sõnajärjepuud, milles leidub rohkem või vähem regulaarseid sõnajärjemustreid, oleks mõttekas minna veel astme võrra kõrgemale ja tuua välja ka üks üldine levinumate sõnajärgede andmepuu (vt joonis 15). Selleks on programmi jaoks jäetud nähtavaks vaid need märgendid, mida analüüsitud valimis on esinenud vähemalt nelikümmend korda (kokku 3177 juhtumit). Selle piirangu alusel on valimis kõige levinum sõnajärjemuster ['@SUBJ', '@FMV', '@ADVL']: *Ta* (@SUBJ) *ärkab* (@FMV) *hommikul* (@ADVL), mis esines kokku 677 korral 3177-st ehk ligikaudu igal viiendal juhul.




**Joonis 15.** Valimi levinumad sõnajärjemustrid

# 5. Sõnajärjevealeidja prototüüp

Sõnajärjevealeidja prototüübi programmeerimisel võeti aluseks leitud õigete sõnajärjemallide kogum ehk paradigma. Realiseerimiseks on kasutatud Zope andmebaasi ning programmeerimiskeelt Python.

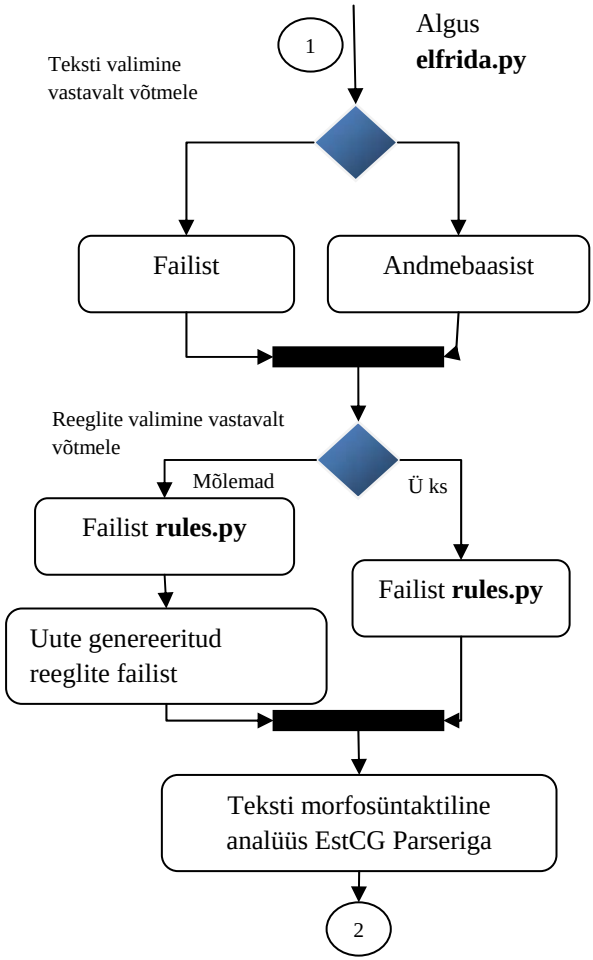
Kõigepealt tehakse kindlaks, kas lause on vaadeldav valitud sõnajärjemallide all või kuulub selliste lausete hulka, millega meie arendatav prototüüp praegu veel ei tegele. Sõna tasandil hoitakse selliseid lisaandmeid nagu sõna lemma, morfoloogiline kirje, süntaktilised märgendid, analüsaatori töö korrektsust hinnanud lingvisti parandused.

Vealeidja prototüübi põhiliseks skriptiks on *elfrida.py* (vt joonis 16). Teksti võib skripti sisestada kahel viisil: eraldi ette antud failist[5] või andmebaasist, kus paiknevad *Eesti vahekeele korpuse* õppijakeele laused. Olemasolevad korrektse sõnajärje mallid paiknevad failis *rules.py*. Samas on eraldi fail, kus hoitakse uusi sõnajärjemalle. Kui tegu on vigadeta tekstiga, siis saab selle alusel genereerida uusi korrektseid sõnajärjemalle. Kui tegu on veamärgendusega õppijakeele tekstiga[6], siis on võimalik tuvastada tüüpilisi sõnajärjevigu. Uued suure sagedusega esile tulevad sõnajärjemallid on otstarbekas lisada mallide paradigmasse. Et vealeidja oleks efektiivsem, selleks on võimalik rakendada kahepoolset kontrolli: ühelt poolt tehakse

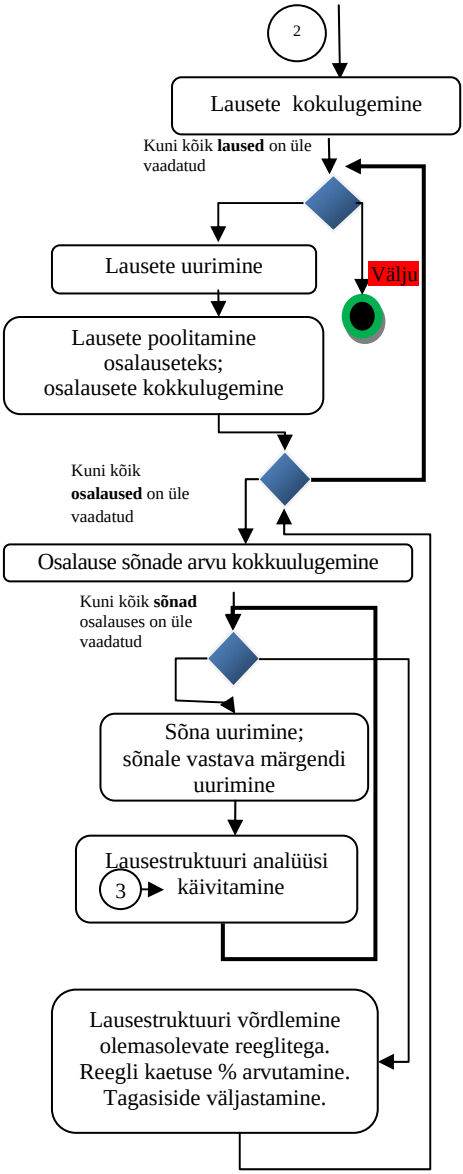
---

[5] Mõne teksti analüüs võib käia mitte läbi andmebaasi, vaid läbi faili, kus paiknevad just hetkel olulised laused, mille abil saadakse kiirvastuseid lause struktuuri ning märgendite kohta. Selline eraldi failist lugemine on oluline ka prototüübi testimiseks.

[6] Eesti vahekeele korpuses on võimalik vigu märgendada ka käsitsi. Lingivistid vaatavad õppijakeele tekstid üle ning lisavad korpuse veataksonoomia alusel vastava veamärgendi. Korpuses on kõik vead liigendatud ülem- ning alamrühmadesse, vea tüübi saab valida taksonoomiast.




**Joonis 16.** Sõnajärje vealeidja prototüübi algoritm




**Joonis 17.** Sõnajärje vealeidja prototüübi algoritm (järg)

kindlaks, kas analüüsi tulemusel korrektseks tunnistatud mall sobib konkreetse (osa)lause analüüsimiseks; teisalt vaadatakse, kas lause struktuur kuulub tüüpiliste vigaste struktuuride alla või mitte. Skriptis *elfrida.py* on arvestatud ka mõne väga sageli esineva sõnajärjeveaga nagu eksimine V2-reegli vastu. Selle reegli kohaselt peab predikaat (@FMV või @FCV) paiknema vajalike märgendite järjestuses mitte kaugemal kui kolmandal positsioonil. Teisisõnu, kui eemaldada lausest kõik sõnad, mis ei mõjuta sõnajärge, siis ülejäänud sõnade seas ei tohi öeldis paikneda kolmandast positsioonist kaugemal.

Kui on selge, kuidas teksti sõnajärg analüsaatorisse sisestatakse ning kas on vaja kasutada uute genereeritud reeglite faili, siis suunatakse tekst süntaktilisele analüüsile.

Skript *elfrida.py* suhtleb EstCG 1,0 parseriga automaatsete päringute kaudu. Parserist saadud morfosüntaktiliselt analüüsitud tekst korrastatakse, lause ning morfosüntaktiline info pannakse eraldi muutujatesse.

Järgnevalt loetakse laused kokku ning jagatakse osalauseteks, kasutades selleks osalause piiri märgendit. Iga sõna ja sellele vastavat märgendit analüüsitakse eraldi ning võrreldakse korrektseks tunnistatud sõnajärjemallidega (vt joonis 17).

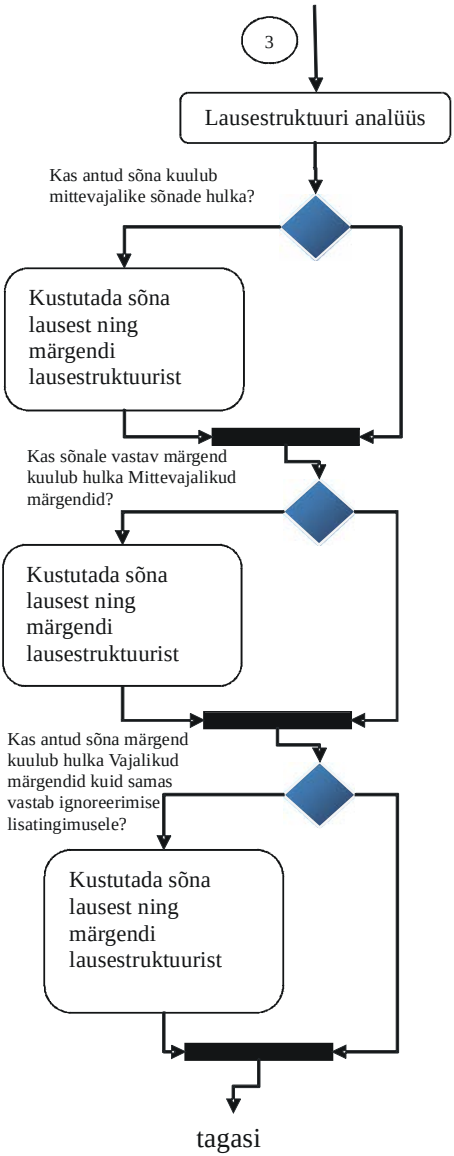
Põhimõtteliselt ei ole see algoritm piiratud esimese osalause töötlemisega, programmi ülesehitus võimaldab analüüsida ka teisi laoseosi.

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). Ka valesti kirjutatud sõna morfosüntaktiline analüüs võib vastuseks anda valed märgendid, mille

võrdlemine korrektsetega võib põhjustada vale tagasisidet ka lauseehituse kohta. Väljundi uurimisel tagastatakse Eesti vahekeele korpusse integreeritud väike lisaskript – info selle kohta, kas sõna on kirjutatud õigesti või mitte, mis annab võimaluse automaatselt välja selekteerida ainult need osalaused, mis on kirjutatud grammatiliselt korrektselt ning määrata õigesti osalausepiir. Nagu eespool seletatud, ei vaatle sõnajärjevealeidja prototüüp kirjavigu sisaldavaid lauseid.

Prototüübi järgmine samm on lausestruktuuri analüüs (vt joonis 18). Vaatluse all on üksteisele järgnevad sõnad ning nende süntaktilised märgendid. Analüüsi alustatakse esimesest sõnast ja vaadatakse, millisesse hulka see kuulub. Siin võib olla kaks võimalust, mida on eespool näidete varal demonstreeritud. Kui sõna kuulub sõnajärje seisukohalt ebaoluliste hulka, siis sõna ja sellele vastav märgend kustutatakse. Teisel juhul uuritakse sõnale vastavat märgendit ja kui selgub, et märgend kuulub mittevajalike märgendite hulka, siis nii märgend kui ka märgendile vastav sõna kustutatakse. Samas võib esineda keerulisemaid olukordi, kus märgend kuulub vajalike märgendite hulka, ent tuleb eemaldada positsiooni tõttu lauses (vt eespool).




**Joonis 18.** Sõnajärje vealeidja prototüübi algoritm (järg)

Kui kõik sõnad lauses ja neile vastavad märgendid on kirjeldatud moel analüüsitud, siis on saadud vajalike märgendite järjendeid võimalik võrrelda olemasolevate korrektseks tunnistatud sõnajärjemallidega. Samas võib tekkida olukord, kui (osa)lauses on sõnu, mida olemasolevad korrektsed sõnajärjemallid ei kata. Sel juhul tuleb välja arvutada kaetuse koefitsient. Näiteks osalausele *Kõnelemine oli lahjaks läinud* vastav sõnajärjemall on ['@SUBJ', '@FCV', '@ADVL', '@IMV'] ning kaetuse koefitsient on 1 ehk 100%. Osalause *Ta näis nüüd lambi laualt võtvat* sõnajärjemall on andmebaasis olemas, kuid üles kirjutatud lühemana: ['@OBJ', '@FMV', '@ADVL']. Sellest johtuvalt on ka kaetuse koefitsient 0,5 ehk 50% ning lauseosa *lambi laualt võtvat* ei ole struktuuri õigsuse kohta tagasisidet saanud.

Lauseid, mille puhul kaetuse koefitsient on alla 1, on võimalik välja filtreerida, et lingvist saaks jätkata nende uurimist ja leida optimaalseid formaliseermisvõimalusi.

# 6. 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.

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 töö efektiivsuse hindamiseks valiti 5880 lausest omakorda 300 juhuslikku lauset, kusjuures jälgiti, et suhtarvuliselt oleks väljajäetavate, õigeks hinnatud ja veakahtlusega lausete hulk sama.

Prototüübi hinnang lause sõnajärje korrektsuse kohta loeti õigeks, kui see langes lingvisti hinnanguga kokku[7]. Erand tehti lausete puhul, mis sisaldasid mõnda õppijakeele viga (nt valesti kasutatud kirjavahemärk või kirjavahemärgi puudumine tingib osalause piiri vale analüüsi: **Laulupidu see on väga tähtis üüritus eestlasteks, sest seal* ...). Ka need laused, mis sisaldasid pärisnime ja mida Filosofti eesti keele speller ei tundnud (nt *Kõige meelsamini vaatab **Zaura** vanu must-valgeid filme*), ei läinud prototüübi vealeidja töö efektiivsuse üle otsutamisel arvesse. Pärisnimede morfoloogilise analüüsi probleemiga seoses on Heiki-Jaan Kaalep ja Tarmo Vaino (2000) ühe lihtsaimalt rakendatava lahendusena välja pakkunud tüpograafiliste konventsioonide kasutamise: pärisnimed algavad suurtähega; nimede äratundmise teeb lihtsamaks asjaolu, et sõnastikust puuduvad sõnad kuuluvad teatud väikesesse arvu muuttüüpidesse.

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.

[7] Prototüübi tööd kontrollis lingvistika magistrant Hanna Sinijärv

Kümnel korral leidis lingvist, et lause on korrektne, kuid prototüüp pidas seda vigaseks. Nendest juhtudest kaheksal korral leidis prototüüp, et kuna öeldis paikneb lauses kaugemal kui kolmandal positsioonil (nt *Esialgu tal see **ebaõnnestub**, sest ...*), siis on see sõnajärg vigane. Kahel ülejäänul korral ei olnud predikatiiv või infiniitne verb lauses viimasel positsioonil (*Ma ise olen **olnud** Inglismaal, Soomes, Venemaal, Ukrainas, Lätis, Leedus*) ning ka selle sõnajärje tunnistas prototüüp ebakorrektseks. Põhjus on selles, et prototüüpi pole veel treenitud neid sõnajärjemalle korrektsetena vaatlema.

Kontrollitud lausetest olid 18 niisugused, mille puhul leidis programm, et lause on korrektne, kuid lingvist luges sõnajärje vigaseks. Enamasti põhjustas lahkarvamuse adverbiaali positsioon, nt **See* (@SUBJ) *on* (@FMV) *armastusest* (@ADVL) *film* (@PRD), vrd *See on armastusfilm* või *See film on armastusest*. Et prototüüp semantikaga arvestada ei oska ja sama märgendijärjend võiks olla ka täiesti korrektne (nt *Mari on hommikuti ilus*), siis on taoliste juhtumite kvalifitseerimine prototüübi edasisel arendamisel suur proovikivi.

Kuigi Eesti vahekeele korpuse vealeidja, sh sõnajärjevealeidja arendamisel on ees palju tööd, saab prototüüpi pidada küllaltki tõhusaks, mis annab soodsa stardipositsiooni edaspidiseks. Seda enam, et paljude praeguste kitsaskohtade lahendused on juba töös. Näiteks suudab automaatne sõnajärjevealeidja koos VAKO-projekti raames loodud õppijakeele lemmatiseerija-oletajaga tulevikus tõenäoliselt analüüsida ka neid lauseid, kus õppija on sõna valesti kirjutanud või eksinud vormimoodustusreeglite vastu.

# 7. 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. Edasine eesmärk on välja arendada keeleõppijale mõeldud tugisüsteem, mille abil saab nii verifitseerida lauseehituse ja sõnajärje korrektsust kui anda soovitusi õige sõnajärjemalli valimiseks ja kasutamiseks. Sellised lahendused on mõnede teiste keelte puhul juba osaliselt realiseeritud, nt inglise keele speller pakub kasutajale õige sõnajärje variandi. Senised VAKO-projekti raames saadud tulemused on olulised ka eesti keele spelleri edasiseks täiustamiseks.

Eesti keele sõnajärjevealeidja prototüüpi saab kasutada uute sõnajärjemustrite otsimiseks ning nende sarnasuse ja kasutuserinevuste võrdlemiseks näiteks sama autori eri ajal kirjutatud tekstides, erinevates allkeeltes ja žanrites, erinevate korpusaineste analüüsis. Artiklis kirjeldatud meetod lubab võrrelda eesti keele erinevaid kasutusvariante (nt õppijakeelt ja kirjakeelt), erinevate keeleoskustasemete (nt A2–B1–B2–C1) sõnajärjemustreid ja morfosüntaksit, individuaalse keelekasutuse eripära ning välja tuua ühe või teise sõnajärjemustri eelistamisega kaasnevad morfosüntaktilised piirangud. Seega on sõnajärje empiirilise uurimise käigus saadud andmepuud olulised mitte ainult keeletehnoloogilistes rakendustes, vaid ka keeleteaduses ja eesti keele õppes. Kuna tegu on universaalsusele pretendeeriva statistikal põhineva programmiga, siis saab kirjeldatud meetodit ja prototüübi algoritmi kasutada ka teiste keelte sõnajärje uurimisel. Tingimuseks on nõue, et uuri-

tava keele lause süntaktiline struktuur lubaks jaotust vajalike ja mittevajalike märgendite vahel.

Eesti keele uurimise seisukohast avab käesolev sõnajärjeuuring selle keerulise nähtuse piire, luues uusi perspektiive sõnajärjega seotud probleemide lahendamiseks.

# Kirjandus


Arppe, Antti 2000. Developing a Grammar Checker for Swedish. – Proceedings from the 12th Nordiske datalingvistikkdager, Trondheim, December 9-10, 1999 / Ed. by Torbjorn Nordgard. Department of Linguistics, Norwegian University of Science and Technology (NTNU). Trondheim: University of Trondheim, 13–27. http://www.ling.helsinki.fi/~aarppe/Publications/Nodalida-99.pdf, 28.08.2010.

Athanaselis, Theologos, Stelios Bakamidis, Ioannis Dologlou 2006. A Fast Algorithm for Words Reordering Based on Language Mode. – 16th International Conference, Athens, Greece, September 10–14, 2006. Proceedings, Part II, 943–951. http://www.springerlink.com/content/q646768285871122/fulltext.pdf, 02.01.2010.

Beesley, Kenneth R. 1988. Language identifier: a computer program for automatic natural-language of on-line text. – Language at crossroads: Proceedings of the 29th Annual Conference of the American Translators Association, Oct 12–16, 47–54.

Damerau, Fred 1964. A Technique for Computer Detection and Correction of Spelling Errors. – Communications of the ACM 7(3), 171–176.

EKG II = Erelt, Mati; Kasik, Reet; Metslang, Helle; Rajandi, Henno; Ross, Kristiina; Saari, Henn; Tael, Kaja; Vare, Silvi 1993. Eesti keele grammatika II. Süntaks. Lisa: Kiri. Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn.

Kaalep, Heiki-Jaan, Tarmo Vaino 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Arvutuslingvistikalt inimesele. Tartu: Tartu Ülikooli kirjastus, 87–99.

Koptjevskaja-Tamm, Maria, Bernhard Wälchli 2001. The Circum-Baltic languages. An areal-typological approach. – Circum-Baltic languages / Ed. by Östen Dahl, Maria Koptjevskaja-Tamm. Vol. 2. Amsterdam, Philadelphia: John Benjamins, 615–750.

Lindström, Liina 2005. Finiitverbi asend lauses. Dissertationes philologiae estonicae Universitatis Tartuensis 16. Tartu: Tartu Ülikooli kirjastus.

Matsak, Erika, Helena Metslang, Jaagup Kippar 2010. The prototype of word order assessment at the Estonian Interlanguage Corpus. – The 2010 International Conference on Artificial Intelligence ICAI 2010. Las Vegas, Nevada, USA (July 12–15, 2010). Vol. II. Las-Vegas: CSREA Press, 870–875.

Metslang, Helena, Erika Matsak 2010. Kesksete lausekomponentide järjestus õppijakeeles: arvutianalüüsi katse. – Eesti Rakenduslingvistika Ühingu aastaraamat 6 / Toim. Helle Metslang, Margit Langemets, Maria-Maren Sepper. Tallinn: Eesti Keele Sihtasutus, 175–193.

Müürsep, Kaili, Tiina Puolakainen 2007. Eesti keele formaalne grammatika: mudelist rakenduseni. Eesti Matemaatika Seltsi aastaraamat 2002–2003 / Toim. Andi Kivinukk, Gert Tamberg, Jan Willemson. http://www.matemaatika.eu/printpdf/ar2002s, 28.08.2010.

Müürisep, Kaili, Tiina Puolakainen, Kadri Muischnek, Mare Koit, Tiit Roosmaa, Heli Uibo 2003. A New Language for Constraint Grammar: Estonian. – International Conference "Recent Advances in Natural Language Processing". Proceedings. Borovets, Bulgaria, 10–12 September 2003, 304–310. http://math.ut.ee/~kaili/papers/ranlp03.pdf, 28.08.2010.

Müürisep, Kaili 2001. Parsing Estonian with Constraint Grammar. – Online proceedings of NODALIDA'01. Uppsala. http://stp.ling.uu.se/nodalida01/pdf/myyrisep.pdf, 28.08.2010.

Pedler, Jennifer 2007. Computer Correction of Real-word Spelling Errors in Dyslexic Text. PhD Thesis. Department of Computer Science and Information Systems. Birkbeck: University of London.

Tael, Kaja 1988. Sõnajärjemallid eesti keeles (võrrelduna soome keelega). Preprint KKI-56. Tallinn.

Vilkuna, Maria 1998. Word order in European Uralic. – Constituent Order in the Languages of Europe. Empirical Language Typology. EUROTYP 20–1 / Ed. by Anna Siewerska. Berlin, New York: Mouton de Gruyter, 173–233.

# The development of the prototype for an automatic word order error detector for Estonian

Erika Matsak, Pille Eslon, Jaagup Kippar

## Summary

The article presents the possibilities for recognizing word order errors in Estonian, the methods used and the current results. The article concentrates on the prototype for an automatic word order error detector for Estonian developed in Tallinn University. The statistic-based program works on a method that is similar to n-grams and the rules used are the patterns formed with 9 compulsory parts of a sentence. The set of correct word order patterns were found from the fiction sub-corpus of Tartu University's Corpus of Written Estonian.

For the statistically reliable results and the utmost efficiency and speed of the program, the rules were placed in a tree structure. The prototype starts the searches by finding a proper initial tag and continues to find a correct compatible pattern that has the highest frequency rate.

At current stage the work is focused on detecting the right/wrong position of the finite/infinite verb and the predicative (since most commonly Estonian is known as a verb second language). Prototype's efficiency was tested on Estonian learner language corpus texts. In the test described in the article 5880 sentences were analyzed with the error analyzer and 300 sentences of the output were assessed. The prototype estimated the correctitude of the word order properly in 87.82% of the cases.

Although there are a number of problems that still need to be solved including the misspelled or unknown words (i.e. proper nouns) and erringly unmarked clausal border, the method and the algorithm of the prototype for an automatic word order error detector for Estonian could also be used on other languages' word order studies as well.

The article is summarized with the survey of the problems occurred on word order detection and the possible ways to make the detector more efficient.

**Keywords**: morphosyntax, automatic error detection, word order errors

## Autorid

Erika Matsak, Tallinna Ülikooli informaatika instituudi dotsent

matsak@tlu.ee

Pille Eslon, Tallinna Ülikooli eesti keele ja kultuuri instituudi vanemteadur, dotsent

peslon@tlu.ee

Jaagup Kippar, Tallinna Ülikooli informaatika instituudi lektor

jaagup.kippar@tlu.ee