doi:10.5128/ERYa10.14

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 10, 227–240

ÜHENDVERBID EESTI KEELE PINDSÜNTAKTILISES ANALÜÜSIS

Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen

Ülevaade. Artikkel käsitleb ühendverbide tuvastamist eesti keele automaatse pindsüntaktilise analüüsi käigus. Ühendverbide äratundmine on vajalik lause täpsemaks süntaktiliseks analüüsiks, sest lause osaliste süntaktilised funktsioonid, semantilised rollid ja nende keelendamine sõltub sellest, milline on lause keskmeks olev predikaatverb, sh sellest, kas predikaatverb on lihtverb või ühendverb.

Ühendverbide tuvastamiseks rakendatakse kahte strateegiat: leksikonipõhist ja reeglipõhist. Viimane tähendab seda, et osa korrapäraseid produktiivselt kombineeruvaid ühendverbe pannakse kokku reeglite abil. Artiklis kirjeldatakse kahte eksperimenti: esialgse ja täiustatud ühendverbide tuvastamise käiku. Täiustatud süsteemi tulemus on päris hea, saavutades saagise 97,4% ja täpsuse 96,6%.*

Võtmesõnad: arvutilingvistika, püsiühendite tuvastamine, eesti keel

1. Sissejuhatus

Meie igapäevane suhtlus toimub üha sagedamini elektroonsete kanalite kaudu, kogu loodav tekstiline info säilitatakse digitaalsel kujul ning seetõttu kasvab ka vajadus seda infot töödelda, parandada, tõlkida või tekstimassiivist infot otsida. Loomuliku keele töötluses mängib keskset rolli süntaktiline analüüs, olles vahelüliks morfoloogilise analüüsi ja semantilise töötluse vahel.

Eesti keele automaatse süntaksianalüüsiga kitsenduste grammatika (ka piirangute grammatika, ingl *Constraint Grammar*) abil on kõige põhjalikumalt tegelenud Kaili Müürisep ja tulemused esitanud oma doktoriväitekirjas (Müürisep 2000). Süntaksianalüsaatori sealse variandi väljundis ei analüüsitud ühend- ja väljendverbe kui kokkukuuluvaid ja tervikuna lause keskmeks olevaid ühendeid. Ühendverbi afiksaaladverbiline komponent analüüsiti adverbiaaliks ja väljendverbi substantiivne komponent vastavalt tema käändevormile kas objektiks või adverbiaaliks. Selline lahendus oli esiteks predikaadi kui lause süntaktilise keskme tuvastamise

* Artikli valmimist on toetanud Euroopa Regionaalarengufond Eesti Arvutiteaduse Tippkeskuse kaudu ning Haridus- ja Teadusministeerium institutsionaalse uurimistoetuse IUT 20-56 "Eesti keele arvutimudelid" kaudu.

seisukohalt ebatäpne ning teiseks raskendas objekti tuvastamist. Sõltub ju eesti keeles lause aspekt (perfektiivne *vs*. imperfektiivne) ja seega ka objekti kääne sageli sellest, kas predikaadiks on lihtverb või ühendverb: *luges raamatut*, *luges raamatu läbi*, **luges raamatut läbi*, **luges raamatu*. Automaatses süntaksianalüüsis sõltub seega sellest, kas predikaadiks on liht- või ühendverb, nende sõnavormide hulk, mis võiksid saada selles lauses objekti märgendi. Ühend- ja väljendverbide probleemi lahendamise vajadus on välja toodud süntaksianalüsaatori edasiste arengusuundade all (Müürisep 2000: 96).

Selle probleemi ühe osa – ühendverbide analüüsi – lahendamisega tegelebki käesolev artikkel, mis on üles ehitatud järgmiselt. Kõigepealt, osas 2, käsitletakse ühendverbi kui keelenähtust; tegeldakse tema defineerimise ja piiritlemisega. Osas 3 antakse lühike ülevaade eesti keele morfoloogilisest ühestamisest ja pindsüntaktilisest analüüsist kitsenduste grammatika formalismi abil. Ühendverbide automaatsele tuvastamisele on pühendatud osa 4, mille alaosas 4.1 selgitatakse leksikoni ja reeglite rolli ühendverbide tuvastamisel, osas 4.2 esitatakse esialgse, lihtsama tuvastamisviisi töökäik ja tulemused ning alaosas 4.3 käsitletakse täiustatud tuvastamisviisi ning selle tulemusi. Osa 5 annab lühiülevaate sarnasest uurimistööst ning artikkel lõpeb kokkuvõtte ja tulevikuplaanidega.

2. Ühendverbi määratlemisest ja piiritlemisest

Ühendverbi on eesti keeleteaduslikus kirjanduses defineeritud kui verbist ja afiksaaladverbist koosnevat, lauses liitöeldisena toimivat ühendit, mille sisuliseks tuumaks on verb, mille tähendusele afiksaaladverb lisab oma nüansi (nt EKG II: 20, Erelt 2013: 62).

Alates Huno Rätsepast (1978: 28) jagatakse eesti keele ühendverbid kahte rühma: ainukordseteks ja korrapärasteks (nt Erelt 2013: 62 jj). Ainukordsed ühendverbid on idiomaatilised ühendid, mille tähendus ei moodustu nende osade tähendustest, vaid on omane ühendile tervikuna (nt *alla vanduma*, *üles ütlema*). Korrapäraste ühendverbide kõige tüüpilisemateks esindajateks on liikumisverbide ja suunda väljendavate afiksaaladverbide kombinatsioonid, mille tähendus moodustub komponentide tähendustest kompositsionaalsel moel, ühendverbi osad on säilitanud oma tähendusliku iseseisvuse (nt *alla / ette / juurde / järele / kokku / kõrvale / laiali / ligi / läbi / mööda / otsa / peale / sisse / välja / üle / üles / vahele astuma / jalutama / jooksma / kõndima / lendama / lonkima*).

Ülalkirjeldatud rühmad on tüüpilised keeleteadusliku klassifikatsiooni näited selles mõttes, et nad esitavad iga rühma tüüpilisi esindajaid iseloomustavad tunnused. Loomulik keelekasutus on teadagi hästi kirjeldatav prototüübi mudeli abil, kus igas rühmas on kesksed eksemplarid, mis vastavad kõigile esitatud tunnustele, ja hajusatele äärealadele kuuluvad eksemplarid, mis ei vasta kõigile selle rühma tunnustele või "halvemal" juhul ühendavad endas kahe rühma tunnuseid. Keelesüsteemi kirjeldamiseks sobivad hajusate piiride ja kattuvate äärealadega kategooriad hästi, kuid automaatanalüüsil ja korpuste märgendamisel tuleb anda analüüs igale tekstisõnale; siin kirjeldataval juhul siis ära tunda kõik tekstis esinevad ühendverbid ning täpselt eristada ühendverbid verbi ja adverbi vabadest ühenditest. Nagu automaatanalüüs on ikka teatud määral lihtsustus, nii on ka piir ühendverbide ja mitte-ühendverbide vahel automaatsel analüüsil mõnevõrra suvaline.

Järgnevalt peatume lühidalt mõnel ühendverbide piiritlemise probleemil. Alustame *olema*-ühenditest osas 2.1, siis käsitleme korrapäraste ühendverbide kombineerumist osas 2.2. ning lõpuks arutleme afiksaaladverbi ja täistähendusliku adverbi piiri üle osas 2.3.

2.1. Kas verbi *olema* ja (afiksaal)adverbi ühendid on ühendverbid?

Üheks seni järjekindla lahenduseta probleemiks ühendverbide hulga piiritlemisel on *olema*-ühendid. Eesti keele grammatika (EKG II: 22) nimetab modaalsete ühendverbidena ühendeid *vaja olema* ja *tarvis olema* ning kuigi seda otse pole öeldud, jääb mulje, et muid *olema*-ühendeid ühendverbide hulka kuuluvateks ei peeta. Ka eesti keele seletussõnaraamatus (EKSS 2009) ei esitata verbi *olema* kirje juures ühtegi ühendverbi.

Seega käsitletakse tüüpiliselt ühendverbidena selliseid ühendeid nagu *alles hoidma*, *kinni panema*, *üle saama*, kuid mitte selliseid ühendeid nagu *alles olema* (*see paber on mul õnneks alles*), *kinni olema* (*pood on kinni*, *kõik on peas kinni*), *üle olema* (*meeskond on oma vastasest selgelt üle*), mis tundub *olema*-ühendite suhtes n-ö pisut ebaõiglane.

Teiselt poolt, kui ühendverb on defineeritud kui sõnaühend, mille põhitähendus tuleb verbilt, siis kas *kinni olema* põhitähendus tuleb ikka selliselt suhteliselt sisutühjalt verbilt nagu *olema*? Ning enamik muutumatu sõna ja verbi *olema* püsivaid ühendeid on sellised, kus muutumatu sõna ei klassifitseeru hästi afiksaaladverbiks kui pelgalt verbi tähendust modifitseerivaks sõnaks – nt *käsil olema*, *nõus olema*, *päri olema*, *rahul olema*, vt ka osa 2.3.

Praegu on automaatsel süntaksianalüüsil see probleem lahendatud nii, et ühendverbidena käsitletakse järgmisi *olema*-ühendeid: *alles olema*, *ära olema*, *kinni olema*, *kohal olema*, *käsil olema*, *lahti olema*, *nõus olema*, *päri olema*, *rahul olema*, *tarvis olema*, *vaja olema*, *valmis olema*, *üle olema*. Loend on ebajärjekindel ega ole kindlasti lõplik.

2.2. Korrapärased ühendverbid

Korrapäraste ühendverbide näidetena on ikka esitatud liikumisverbide kombinatsioone suunda väljendavate afiksaaladverbidega (Rätsep 1978: 28, EKG II: 21, Erelt 2013: 62–63). Nii moodustatud ühendverbide näitel saab aga ka veenduda, et teatud semantilisse välja kuuluvate verbide ja afiksaaladverbide kombineerumine korrapäraste ühendverbide moodustamiseks pole siiski täisproduktiivne, olles seotud semantilist laadi kitsendustega, mida on aga raske formaliseerida.

Kui jääda liikumisverbide näite juurde, siis need jagunevad selgelt kaheks rühmaks. Enamik neist kombineerub tõesti vabalt suunda väljendavate afiksaaladverbidega, moodustades korrapäraseid ühendverbe, nt *astus / hüppas / jalutas / kõndis / lendas / läks / marssis / roomas / sõitis / traavis / tormas / veeres / vonkles* jne *(kodust) välja / (teistest) ette / (trepist) üles* või *alla*, *(sillast) üle / (teistele) järele*. Ent väiksem osa liikumisverbe suunda näitavate partiklitega ei kombineeru, nt **hulkus / luusis / käis / seikles (kodust) välja / (teistest) ette / (trepist) üles*

või *alla / (sillast) üle / (teistele) järele* pole võimalikud kombinatsioonid. Varem mainitud verbide näol on tegemist direktsionaalidega, teise loetelu verbid ei ole aga mitte direktsionaalid, vaid liikumisverbid, mis ei spetsifitseeri liikumise algus- ega lõpp-punkti ega ka liikumisteed ja nii nad ei kombineerugi suunda väljendavate afiksaaladverbidega.

Ka afiksaaladverbide puhul eksisteerivad tüüpilised mõistemetafooriteooriale vastavad ülekanded konkreetsematest tähendusvaldkondadest (allikvaldkondadest) abstraktsematesse tähendusvaldkondadesse (sihtvaldkondadesse). Ruum ja selles liikumine on üks populaarsemaid allikvaldkondi, kohasuhteid saab kasutada ka näiteks ajast rääkides (Pajusalu 2009: 122 jj). Nii väljendab afiksaaladverb *ette* nii ruumi- kui ajasuhet, olemas on nii ühendverbid *ette astuma / hüppama / jooksma / jõudma* kui ka *ette aimama / helistama / kuulutama / teadma*. Ajasuhet väljendades võib *ette* samuti produktiivselt kombineeruda teatud semantilisse välja kuuluvate verbidega, nt teadmis- ja teatamisverbidega: *ette aimama*, *ette teadma*, *ette teatama*, *ette helistama*, *ette hoiatama*, *ette kuulutama* jne. See tähendab, et sagedased suunda väljendavad afiksaaladverbid võivad kombineeruda ka teistesse tähendusgruppidesse kuuluvate verbidega peale liikumisverbide. Sellise kombineerumise produktiivsuse kohta pole meil praegu aga piisavalt teavet.

Hetkel on süntaksianalüsaatori reeglistikus produktiivselt kombineeruvatena käsitletud lisaks liikumisverbidest (direktsionaalidest) ja suunda väljendavatest afiksaaladverbidest moodustuvatele ühendverbidele veel mitut väiksemat hulka, täpsemalt vt osa 4.2.

2.3. Afiksaaladverb *vs.* täistähenduslik adverb

Afiksaaladverbi defineeritakse kui muutumatut sõna, mis lauses verbi juurde kuuludes annab sellele uue tähendusvarjundi või konkretiseerib verbi tähendust (EKG I: 33, Erelt 2013: 21). Huno Rätsep eristab väljendverbi ühendverbist ka selle põhjal, et väljendverbi nominaalsetel komponentidel "puudub ühendverbide adverbidele (*peale*, *sisse*, *alla* jne.) omane üldistatud tähendus." (Rätsep 1978: 21)

Eksisteerib aga hulk verbi ja adverbi püsivaid ühendeid, mis ei ole "päris" ühendverbid selles mõttes, et nende adverbiline komponent ei ole redutseerunud mitteiseseisvaks ja mittetäistähenduslikuks (EKG I: 18) afiksaaladverbiks, vaid on täistähenduslik sõna, nt *hukka mõistma*, *silmitsi seisma*, *kõhuli / selili keerama* jms). Nende näidete kohta ei saa väita, et adverbiline komponent ainult täpsustaks või modifitseeriks verbiga väljendatud tähendust või et talle oleks omane üldistatud tähendus, pigem tuleneb adverbist põhiline osa ühendi kui terviku tähendusest.

Selliseid ühendeid märgendatakse süntaksianalüsaatori väljundis praegu ühendverbidena, sest kindlasti on tegemist püsiühendiga ning me ei näe hetkel põhjust postuleerida veel ühte adverbist ja verbist koosneva püsiühendi liiki lisaks ühendverbile.

3. Morfoloogiline ühestamine ja süntaksianalüüs kitsenduste grammatika abil

Eesti keele pindsüntaksi analüsaator põhineb kitsenduste grammatika formalismil (Karlsson jt 1995). Analüsaator on reeglipõhine, see tähendab, et tuvastamisreeglid on koostatud inimese poolt masinõppimist kasutamata. Analüsaator koosneb analüüsimootorist, mis on keelest sõltumatu, ja reeglite baasist. Eesti keele analüüsiks kasutatakse Lõuna-Taani Ülikoolis (*Syddansk Universitet*) väljatöötatud VISL-parserit.[1] Reeglite baas on jaotatud kaheks: morfoloogilise ühestamise reeglistik, mis valib morfoloogiliselt mitmestele sõnavormidele konteksti sobiva analüüsi, ning pindsüntaksi reeglid, mis lisavad igale sõnavormile tema süntaktilist funktsiooni näitava märgendi (vt Roosmaa jt 2003).

Analüsaator saab sisendiks morfoloogiliselt analüüsitud teksti, milles on sõnavormidele antud kõik võimalikud morfoloogilised tõlgendused (nt sõnavorm *või* võib olla sidesõna, nimisõna nominatiivis ja genitiivis või hoopis verbivorm). Morfoloogilise ühestamise reeglid eemaldavad konteksti suhtes sobimatud tõlgendused. Samal etapil toimub ka osalausepiiride esialgne määramine. Kui konteksti põhjal pole võimalik ühest analüüsi leida, jäetakse alles kõik sobivad märgendid.

Pindsüntaktilisel analüüsil lisatakse kõigepealt kõik sõnavormi grammatiliste tunnustega sobivad süntaktilised märgendid ning seejärel hakatakse iteratiivselt eemaldama neid, mis ikkagi lausesse ei sobitu (näiteks objekti märgend eemaldatakse partitiivis nimisõnalt, kui verbivorm eeldab totaalobjekti olemasolu või ei leidu osalauses üldse transitiivset verbi või on juba objekt tuvastatud ning antud sõnavorm ei ole selle tuvastatud objektiga koordinatsioonis jmt).

Analüsaatori antav märgendus on väga pindmine: tuvastatakse osalausete subjektid, objektid, predikatiivid, aga ei eristata pea- ega kõrvallauseid, objekt ning verb ei ole omavahel ühendatud, ka atribuudid ei ole seotud põhjadega, vaid lihtsalt eristatakse ees- ja järelatribuute. Samuti puudub eri tüüpi adverbiaalide eristus. Sügavam süntaktiline analüüs on järgmise, sõltuvuspuid ehitava grammatikamooduli ülesanne.

Morfoloogilise ühestamise ja pindsüntaktilise analüüsi reeglite baasis oli 2013. aasta detsembri seisuga umbes 4500 reeglit. Eksperimendid 95 000-sõnalisel käsitsi märgendatud korpusega[2] andsid kogu pindsüntaktilise analüüsi saagiseks (kõigi automaatselt korrektselt leitud märgendite arvu suhe tegelikult korrektsete märgendite arvu) oli analüsaatori sellel versioonil 92,9% ning täpsuseks (kõigi automaatselt korrektselt leitud märgendite arvu suhe kõigi leitud märgendite arvu) 69,3%; vigu oli 7,1%

See tähendab et 7% sõnadest jäävad ilma õigest märgendist ning 30–31% pakutud märgenditest on kas üleliigsed või vigased. Oluline hulk mitmesustest tekib adverbiaali ning adverbiaalse atribuudi eristamatusest, vead tekivad aga eelkõige objekti, subjekti ja predikatiivi määramisel.

Näiteks lauses *Ta on neis linnades aastaid tsaariajal elanud* on sõnavorm *linnades* jäänud mitmeseks eestäiendi (vrd *turvistes mehi*) ja adverbiaali tõlgenduse vahel ning sõnavorm *tsaariajal* on jäänud mitmeseks järeltäiendi (vrd *mehi pildil*) ja adverbiaali vahel.

[1] http://beta.visl.sdu.dk/ (30.9.2013).

[2] Testkorpus on osa käsitsi märgendatud korpusest veebiaadressil http://www.keeletehnoloogia.ee/ekt-projektid/vahendid-teksti-mitmekihiliseks-margendamiseks/soltuvussuntaktiliselt-kasitsi-analuusitud-korpus (16.2.2014).

Vigu subjekti, objekti ja predikatiivi määramisel põhjustab asjaolu, et nende süntaktiliste funktsioonide kodeerimiseks kasutatakse samu käändeid (v.a ainsuse genitiiv), lisakeerukuse toob see, et need käändevormid on paljudel sõnadel homonüümsed. Nii näiteks on lauses *Aastaid on ta korraldanud jõulukontserti.* sõnavorm *ta* saanud morfoloogilisel ühestamisel ekslikult genitiivi tõlgenduse ja süntaksianalüüsil seetõttu objekti analüüsi ning partitiivis nimisõna *jõulukontserti* on süntaktilisel analüüsil saanud subjekti tõlgenduse, st subjekt ja objekt on ära vahetatud.

Et tulemusi parandada, on vaja reeglitesse lisada rohkem leksikaalset informatsiooni ja lausemustreid, eelkõige vajab jätkuvalt täiendamist verbirektsioonide sõnastik.

4. Ühendverbide automaatne tuvastamine

Varasem lähenemine (Müürisep 2000) ei üritanudki tuvastada ühendverbe, kuid analüsaatori efektiivsuse parandamiseks ja puustruktuuri analüüsiks on nende leidmine muutunud väga oluliseks.

Selles peatükis vaadeldakse kõigepealt kahte strateegiat ühendverbide tuvastamisel: leksikoni ning kombineerimisreeglite kasutamist; kirjeldatakse lühidalt leksikoni koostamise protsessi, seejärel esitatakse kahe ühendverbide tuvastamise katse töökäik ja tulemused.

4.1. Leksikoni ning reeglite roll ühendverbide tuvastamisel

Ühendverbide tuvastamisel rakendatakse kahte strateegiat: leksikoni ning kombineerimisreegleid. Ühendverbide esialgne nimekiri loodi Paul Saagpaku eesti-inglise sõnaraamatu (Saagpakk 1992) põhjal ning selle nimekirja alusel koostati esialgne ühendverbide tuvastamise reeglistik. Kuna paljud ühendverbide koosseisu kuuluvad afiksaaladverbid võivad olla morfoloogiliselt mitmesed kaassõna või mõne nimisõnavormi suhtes (nt *juurde võtma*, kus afiksaaladverb *juurde* saab morfoloogilise analüüsi käigus ka kaassõna ning nimisõna aditiivi e lühikese sisseütleva käändevormi tõlgenduse), siis tuli esialgsete reeglite hulka lisada täiendavaid kitsendusi. Samuti tuli arvestada ühendverbide nominalisatsioonivõimalusega ning sellega, et sama afiksaaladverb võib kombineeruda erinevate verbidega, moodustades erinevaid ühendverbe.

Ühendverbide esialgset leksikoni on täiendatud eesti keele seletussõnaraamatu (EKSS 2009) andmetega ja korpuse käsitsi märgendamise käigus n-ö avastatud ühendverbidega, nt *olema*-ühenditega, täpsemalt vt osa 2.1. Leksikon kindlasti veel muutub.

Lisaks ühendverbide leksikonile on loodud ühendverbide tuvastamise grammatikas esialgu kaks suuremat gruppi reegleid korrapäraste ühendverbide tuvastamiseks: reeglid liikumisverbide ja suunda väljendavate afiksaaladverbide tuvastamiseks ning perfektiivse afiksaaladverbiga *ära* seotud verbide reeglid.

Kuigi mitte kõik liikumisverbid ja suunda väljendavad afiksaaladverbid ei saa omavahel kombineeruda, võime arvutigrammatika koostamisel eeldada, et kui nad

juba samas osalauses koos paiknevad, siis nad ilmselt ka ühendverbi moodustavad. Erandiks on juhud, kus muutumatu sõna on morfoloogiliselt mitmene afiksaaladverbi ja adpositsiooni tõlgenduste vahel ning toimib konkreetses osalauses adpositsioonina. Nagu vigade analüüsi osast näha, on need juhud ka põhiline vigade allikas.

Perfektiveeriva afiksaaladverbi *ära* puhul tuleb arvestada kontekstiga, kus *ära* võib esineda ka imperatiivi eitava vormi koosseisus (nt *ära tee*) ning sel juhul ühendverbi ei moodustu.

Ühendverbide tuvastamise protsess toimub iteratiivselt. Esimestel etappidel üritatakse afiksaaladverbikandidaate ühestada afiksaaladverbi, pre- ja postpositsiooni ning substantiivi tõlgenduste vahel, järgmisel sammul leitakse, millised selles osalauses esinevad afiksaaladverbid ja verbid omavahel kombineeruvad. Näiteks lauses *Koju minnes tuli saapal tald ära* võib afiksaaladverb *ära* moodustada ühendverbi nii verbiga *minema* kui ka *tulema*.

Ühendverbide tuvastamise reegleid kontrolliti 95 000-sõnalisel käsitsi süntaktiliselt märgendatud korpusel.[3] Korpus esindab normeeritud kirjakeelt, koosnedes tasakaalus korpuse[4] ilukirjandus-, ajakirjandus- ja teadustekstidest.

4.2. Tuvastamise esialgsed tulemused

Ühendverbide tuvastamise protsess algab morfoloogilise ühestamise etapis, kus afiksaaladverbi võimaliku tõlgendusega sõnal võib olla ka mitmeid teisi tõlgendusi: kas kaassõna (nt *alla*, *järele*, *vastu*, *üle*, *ümber*), nimisõna (nt sõnavorm *lahti* võib olla nimisõna *laht* vorm, sõnavorm *maha* võib olla nimisõna *maa* vorm), võimalikud on ka nii kaassõna kui ka nimisõna tõlgendused ühel sõnavormil (nt *ilma (ilm)*, *juurde (juur)*, *külge (külg)*, *kõrval (kõrv)*, *peale (pea)*); võimalik on ka omadussõna käändevormi tõlgendus (nt *heaks (hea)*, *üleval (ülev)*), verbivormi tõlgendus (nt *lahku (lahkuma)*, *taha (tahtma)*), nii kaassõna kui ka verbivormi tõlgendused (nt *läbi (läbima)*, *taga (tagama)*), nii nimisõna kui ka verbivormi tõlgendused (nt võib sõnavorm *välja* lisaks afiksaaladverbile olla nimisõna *väli* vorm või verbi *väljama* vorm). Mõnel sõnavormil on võimalik mitmesus afiksaaladverbi, omadussõna, nimisõna ja verbivormi tõlgenduste vahel (nt sõnavorm *täis* võib olla afiksaaladverb, omadussõna, nimisõna *täi* vorm ning verbi *täima* vorm; sõnavorm *valmis* võib olla afiksaaladverb, nimisõna *valm* ning verbi *valmima* vorm). Võimalikud mitmesused on veel afiksaaladverbi, kaassõna, nimisõna ning verbivormi tõlgenduste vahel (nt *ringi* (*ring*, *ringima*)), afiksaaladverbi, kaassõna, arvsõna, asesõna tõlgenduste vahel (nt sõnavorm *ühes* võib olla afiksaaladverb, kaassõna, arvsõna *üks* vorm või asesõna *üks* vorm) ning afiksaaladverbi, kaassõna, arvsõna, asesõna ja nimisõna tõlgenduste vahel (nt sõnavorm *ühte* võib olla afiksaaladverb, arvsõna *üks* vorm, nimisõna *üks* vorm või nimisõna *ühe* vorm).

Afiksaaladverbi ning muude tõlgenduste vahel ühestamisele pandi alus kitsenduste grammatika süntaksianalüsaatori morfoloogilise ühestamise moodulis (Puolakainen 2001), mida on käesolevas töös oluliselt täiendatud.

Kuna enamasti on ka võimalikku afiksaaladverbi tekstis ümbritsevad sõnavormid sel hetkel veel ühestamata, st kindlaks määramata võib olla kas kääne (eriti sage on mitmesus nominatiiv / genitiiv / partitiiv) või ka sõnaliik (adverb või

[3] Testkorpus on osa käsitsi märgendatud korpusest veebiaadressil http://www.keeletehnoloogia.ee/ekt-projektid/vahendid-teksti-mitmekihiliseks-margendamiseks/soltuvussuntaktiliselt-kasitsi-analuusitud-korpus (16.2.14).

[4] http://www.cl.ut.ee/korpused/grammatikakorpus/ (16.2.14).

nimisõna, verb või nimisõna), siis alati ei õnnestu afiksaaladverbi sõnaliiki õigesti määrata. Näiteks, lauses *Gripiviirus liigub kah ringi* ühestatakse sõnavorm *ringi* ekslikult nimisõna *ring* aditiivi e lühikese sisseütleva vormiks.

Ühendverbide leksikoni kasutatakse abileksikonina morfoloogilise ühestamise etapil ja põhileksikonina sellele järgneval ühendverbide osade kokkuviimisel ning süntaktiliste funktsioonide määramisel. Sellise analüüsijada tulemuse saagis ühendverbide tuvastamise osas oli 79,3%, st 20,7% ühendverbidest jäi tekstis ära tundmata. 5,2% afiksaaladverbidest tuvastati ühendverbi osana, kuid pandi seejuures kokku vale verbiga. 1,7% pakutud afiksaaladverbidest ei osutunud tegelikult ühendverbi koostisosaks, st nad ei olnud tegelikult afiksaaladverbid ja sõnaliigi määramisel oli tehtud viga.

Enamik vigu tulenes puudulikust ühendverbide leksikonist ning morfoloogilise ühestamise etapil tehtud valedest valikutest afiksaaladverbi ja adpositsiooni tõlgenduste vahel, millele liitusid veel küllaltki sagedased ühestamisvead käändsõna nominatiivi, genitiivi ja partitiivi tõlgenduste vahel.

Puuduliku leksikoni näideteks on laused .. *jäi üle ära mõõta kulminatsioonikõrguste erinevusele vastav kaare pikkus maapinnal* ning *Järjest tugevamini kiirgava Päikese kiirgus hakkab eemale puhuma vesinikust ja heeliumist koosnevat gaasi*, kus esimeses lauses tuvastati ühendverb *ära mõõtma*, kuid ei tuntud ära ühendverbi *üle jääma* ja teises lauses jäi tuvastamata ühendverb *eemale puhuma*.

4.3. Ühendverbide täiustatud tuvastamine

Tulemuse parandamiseks võeti ette järgmised sammud. Kuna enamik vigu tulenes morfoloogilise ühestamise etapil tehtud ekslikest valikutest afiksaaladverbi ja adpositsiooni tõlgenduste vahel, siis selliste olukordade minimeerimiseks toodi sisse reeglid, mis pärast morfoloogilist ühestamist tegelevad just kaassõna ja afiksaaladverbi vahekorra korrigeerimisega. Morfoloogilise ühestamise käigus on selline valik raskendatud, kuna sel ajal on üldine mitmesus väga kõrge ning sageli pole veel teada lähimas naabruses olevate noomenite käänded ega mõnikord ka sõnaliik.

Näiteks lauses *Ameerika alustas ränkadest vastasseisudest rahvuslikul pinnal ja töötas läbi vere, pisarate ja seadusloome tänase rassiliselt, usuliselt ja rahvuslikult tolerantse riigini* .. osutus võimalikuks parandada sõna *läbi* analüüsi afiksaaladverbist (enne moodustati sellega ekslikult ühendverb *töötas läbi*) kaassõnaks, mis moodustab kaassõnafraasi *läbi vere, pisarate ja seadusloome*. Samuti lauses *Erinevalt veest võib jää tagant pealpressiva jäämassi survel liikuda isegi üle kõrgustike* õnnestub kaassõna *üle* õigesti seostada kaassõnafraasiks *üle kõrgustike*, kuigi lauses esinev verb *liikuda* võimaldab teoreetiliselt moodustada ühendverbi *liikuda üle*. Ning vastupidiselt lauses .. *selgus, et osooni lagunemise reaktsioonide efektiivsus jääb tekkimise omale alla* taastati *alla* õige afiksaaladverbi tõlgendus, mis moodustab ühendverbi *jääb alla*.

Teiseks lisati võimalus reeglite abil vabalt kombineerida ühendverbe liikumisverbist ja suunda väljendavast afiksaaladverbist. Sel viisil kombineeritakse näiteks ühendverbid *ette minema/tulema/hüppama/jooksma/kargama/kiirustama/käänama/langema* jne.

Samuti loodi eraldi reegel moodustamaks produktiivselt ühendverbe perfektiveeriva afiksaaladverbiga *ära*.

Produktiivselt moodustatavate ühendverbide gruppidena esitati veel kehaasendi muutmist väljendavad ühendverbid (afiksaal)adverbidega *püsti*, *pikali*, *istuli*, *põlvili*, *selili*, *külili*, *kõhuli* ning afiksaaladverbidega *kinni* ja *lahti* kombineeruvad verbid, näiteks *ajama*, *heitma*, *hüppama*, *laskma*, *laskuma*, *jääma*, *kargama* jne.

Omaette kombineerumisgrupina lisati väga sageli ühendverbi moodustavate verbidega *jääma*, *saama* ning *minema* (mis ka üksikverbidena on väga sagedased) kombineeruvad afiksaaladverbid.

Veel ühe täiendusena lisati reeglid, mis arvestavad nominaliseerunud ühendverbidega. Näiteks lauses *Siis tuleb harjuda lühikesest basseinist pikka üle minemisega* suudetakse nende abil ära tunda ühendverbi *üle minema* nominalisatsioon.

Selle tuvastamismeetodi saagis on praegu 97,4% ehk kõikidest afiksaaladvebidest jääb tuvastamata 2,6%, seega eelmises osas kirjeldatud lihtsa meetodiga võrreldes tõusis saagis 18% võrra, mis on väga hea tulemus. 2,3% afiksaaladverbidest küll tuvastati afiksaaladverbina, kuid antud lause kontekstis pandi kokku vale verbiga. Ülegenereerimise protsent on 3,4, st lisaks igale sajale nende reeglite abil tuvastatud ühendverbile pakkus analüsaator veel 3,4 ühendverbi analüüsi, mis ei olnud tegelikult ühendverbid.

Osa vigadest on tingitud endiselt valest valikust kaassõna ja afiksaaladverbi morfoloogiliste tõlgenduste vahel. Näiteks lauses *Paneme suurde tuppa riiuli peale* .. analüüsitakse muutumatu sõna *peale* ekslikult afiksaaladverbiks ning koos verbivormiga *paneme* ühendverbiks *paneme peale*, kuigi õige on selles kontekstis kaassõnafraas *riiuli peale*. Aga tuleks tähele panna, et selles lauses on teoreetiliselt ka semantiliselt võimalik selline analüüs, kus *riiuli* on süntaktiline objekt ning *peale* on tõesti ühendverbi osa.

Lause *Nüüd lasti neid kastist välja jalutama* on näide olukorrast, kus afiksaaladverb küll tuvastatakse, kuid antud lause kontekstis pannakse kokku vale verbiga: afiksaaladverb *välja* kombineeriti ekslikult verbiga *jalutama* ühendverbiks *välja jalutama*, kuigi tegelikult esineb lauses ühendverb *lasti välja*.

Positiivse näitena on lauses *Tulin neli päeva tagasi mägedest alla* ära tuntud õige ühendverb *alla tulema*, kuigi lauses on olemas veel teise potentsiaalse produktiivselt kombineeruva ühendverbi *tagasi tulema* komponendid.

Üheks veapõhjuseks on kõrvallausega poolitatud ja üksteisest eraldatud lauseosad, kus üks osa sisaldab põhiverbi ja teine afiksaaladverbi, nt lauses .. *lõppkokkuvõttes tasub hoonete soojustamine, milleks majaomanikel kulub arvutuste järgi 15,2 miljardit eurot, ennast mitmekordselt ära* ei tuntud ära ühendverbi *ära tasuma*; analüüsi raskendava asjaoluna esineb siin pealauset poolitavas kõrvallauses verbivorm *kulub*, mis samuti saaks afiksaaladverbiga *ära* ühendverbi moodustada.

5. Sarnane uurimistöö eesti keele ja ka teiste keelte alal

2002. aastal avaldas Ivan A. Sag koos kolleegidega artikli "Multiword Expressions: Pain in the Neck for NLP", kus tõdetakse, et püsiühendid on võtmeprobleemiks tegelikke tekste lingvistilise põhjendatusega analüüsivate keeletöötlusvahendite loomisel (Sag jt 2002). Sellest teedrajavaks tunnistatud artiklist alates on püsiühendite, sh ühendverbidega seonduv probleemistik olnud arvutilingvistikas aktuaalne uurimisteema.

Eestikeelse ülevaate püsiühendite probleemistikust arvutilingvistikas annab Kadri Muischneki ja Heiki-Jaan Kaalepi artikkel "Eesti keele püsiühendid arvutilingvistikas: miks ja kuidas" (Muischnek, Kaalep 2009).

Käesolevas artiklis kirjeldatud tööle kõige sarnasem on Heiki-Jaan Kaalepi eesti keele püsiühendite märgendamise programm, mida on kirjeldatud Eesti keeletehnoloogia teisel konverentsil[5] ettekandes "Mitmesõnalised verbid" (Kaalep 2009). Selle programmi ülesandeks on tekstis märgendada kõik püsiühendite andmebaasis sisalduvad verbikesksed püsiühendid, sh ka ühendverbid. Sisendina kasutatakse morfoloogiliselt ühestatud teksti, kus pole aga märgendatud osalausepiire, st osalausepiiride määramine on püsiühendite märgendamise programmi osa, samas aga on valesti märgendatud, õigemini märkimata jäänud osalausepiir kõige tavalisem vigade põhjus selle programmi väljundis. Verbikesksete püsiühendite tuvastamise saagis on 92% ja täpsus 90%, ühendverbide tuvastamise saagist ja täpsust eraldi mõõdetud ei ole.

Eesti keele ühendverbide korpusest ekstraheerimise probleemistikku on käsitletud ka murdematerjali näitel. Kristel Uiboaed on otsinud vastust küsimusele, milline statistik sobib kõige paremini ühendverbide loendi koostamiseks murdekorpuse põhjal (Uiboaed 2010).

Üldisemalt palju käsitletud teemad on olnud tekstikorpuse baasil ühendverbide loendi koostamine (nt Baldwin, Villavicencio 2002, Ramisch jt 2008, Kaalep, Muischnek 2002), sellise loendi põhjal mitmesuguse lisainfoga varustatud leksikoni koostamine (nt Villavicencio 2003, Kaalep, Muischnek 2008) ja püsiühendite kohtlemine automaatsel süntaktilisel analüüsil.

Mis puudutab viimast allteemat, siis näiteks Aline Villavicencio ja Ann Copestake tutvustavad oma 2002. aastal avaldatud artiklis (Villavicencio, Copestake 2002) partikkelverbide käsitlemist HPSG-põhise inglise keele formaalse grammatika LinGO ERG raamistikus. Inglise keele partikkelverb (ingl *particle verb*, ka *phrasal verb*) koosneb verbist ja adverbiaalsest partiklist, *particle verb* on ka eestikeelse termini *ühendverb* ingliskeelseks vasteks (Erelt 2003: 101). LinGo ERG süsteemis esitatakse partikkelverbid leksikonis ning nad on jagatud üheteistkümneks tüübiks vastavalt sellele, kas ja milline seotud laiend neil saab olla ja milline võib olla partikli ja seotud laiendi sõnajärg. Tulevikuplaanina mainitakse teatud semantilist laadi partikkelverbide esitamist verbide ja partiklite omavahel kombineeruvate hulkadena, näitena tuuakse ikka liikumisverbide ja suunda või kohta väljendavate partiklite hulgad.

Martin Forst jt (2010) kirjeldavad võrdlevalt ühendverbide töötlemist inglise, saksa ja ungari keelt analüüsivates leksikaalfunktsionaalse grammatika formalismi

[5] http://www.keeletehnoloogia.ee/konverentsid/ekkt-teine-konverents/slaidid-2009/mitmesonaliste-verbide-aratundmine-tekstides (3.3.2014).

põhistes süntaksianalüsaatorites (ParGram LFG). Autorid väidavad esiteks seda, et nende kolme keele ühendverbe saab arvutigrammatikas esitada ühesugusel moel ja teiseks seda, et kompositsioonilistele (eesti traditsioonis nimetatud korrapärasteks) ja idiomaatilistele ühendverbidele tuleb läheneda erineval moel: idiomaatilised tuleb esitada leksikonis ja kompositsioonilised tuleb "kokku panna" süntaktilise analüüsi käigus. Kui partikli lisamine korrapärase ühendverbi puhul lisab verbi argumendistruktuuri uue argumendi, siis luuakse spetsiaalse reegli abil partikkelverbi jaoks uus formaalne kirjeldus, mis erineb vastava lihtverbi formaalsest kirjeldusest ainult selle lisatud argumendi poolest.

6. Kokkuvõte ja tulevikuplaanid

Artiklis käsitleti ühendverbide töötlemise probleeme eesti keele automaatse reeglipõhise süntaksianalüüsi raames. Ühendverbide tuvastamisel kasutatakse kahte strateegiat: leksikonipõhist ja reeglipõhist, viimasel juhul pannakse mõned korrapäraste ühendverbide hulgad kokku teatud semantilisse hulka kuuluvate verbide ja afiksaaladverbide kombineerimisel. Tuvastamine ise on iteratiivne protsess. Esimestel etappidel üritatakse ühestada afiksaaladverbina esineda võivaid sõnavorme afiksaaladverbi, pre- ja postpositsiooni ning substantiivi tõlgenduste vahel, järgmisel sammul leitakse, millised afiksaaladverbid ja millised verbid omavahel kombineeruvad. Ühendverbide tuvastamise esialgsed tulemused olid keskpärased, saagiseks saavutati 79,3%, mis jäi alla ka varasemale lihtmeetodil tehtud eksperimendile. Täiendades reeglistikku morfoloogilise järelühestamise reeglitega ning luues eraldi reeglid produktiivselt kombineeruvate korrapäraste ühendverbide tarbeks, saadi tulemuseks saagis 97,4% ning täpsus 96,6%.

Tabelis 1 on toodud erinevate eksperimentide tulemused.

Tabel 1. Ühendverbide tuvastamise eksperimentide saagised ja täpsused

	Kaalep 2009	**Esialgne**	**Täiustatud**
Saagis	92%	79,3%	97,4%
Täpsus	90%	97,9%	96,6%

Kaalepi (2009) ja siinses artiklis kirjeldatud eksperimendi tulemuste võrdlemine on natuke meelevaldne, sest tekstide hulgad, žanrid, aga ka märgendatavad ühendid on erinevad: Kaalepi tulemused on pigem üldisemalt verbikesksete püsiühendite kohta. Samuti hinnati Kaalepi eksperimendis tulemusi morfoloogilise märgenduse suhtes, kasutades sisendina perfektselt ühestatud teksti. Samas on selle katse aluseks olnud püsiühendite andmebaas väga oluliseks allikaks meie leksikoni edasiseks täiustamiseks.

Tuvastamata jäänud ühendverbide veel üheks allikaks on ebatäielik leksikon, mida on plaanis täiendada, esmajärjekorras võtta kasutusele püsiühendite andmebaasis[6] sisalduvatest ühendverbidest need, mis meie leksikonis puuduvad. Samuti tuleks täiendada korrapäraste ühendverbide kombineerimiseks kasutatavaid reegleid ning tuvastada uusi omavahel kombineeruvate verbide ja afiksaaladverbide rühmi.

[6] http://www.cl.ut.ee/ressursid/pysiyhendid/ (30.9.2013).

Järgmisena tuleb luua samasugune leksikon ja reeglid ka väljendverbide jaoks ning siis jõuda kõigi levinumate püsiühendite analüüsimiseni.

Loodud ja loodav reeglistik on väga vajalik süvasüntaktiliseks analüüsiks, ilma püsiühendite tuvastamiseta oleks analüüsipuu poolik, et mitte öelda vigane.

Viidatud kirjandus

Baldwin, Timothy; Villavicencio, Aline 2002. Extracting the unextractable: a case study on verb particles. – Proceedings of the Conference on Computaional Natural Language Learning (CoNLL 2002), Taipei, Taiwan, 31 August – 1 September 2002. Association for Computational Linguistics, 17.

EKG I = Erelt, Mati; Reet Kasik; Helle Metslang; Henno Rajandi; Kristiina Ross; Henn Saari; Kaja Tael; Silvi Vare 1995. Eesti keele grammatika I. Morfoloogia. Sõnamoodustus. [The Grammar of the Estonian Language I: Morphology. Word-formation.] Eesti Teaduste Akadeemia Eesti Keele Instituut. Tallinn.

EKG II = Erelt, Mati; Reet Kasik; Helle Metslang; Henno Rajandi; Kristiina Ross; Henn Saari; Kaja Tael; Silvi Vare 1993. Eesti keele grammatika II. Süntaks. Lisa: kiri. [The Grammar of the Estonian Language II: Syntax.] Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn.

EKSS = Langemets, Margit; Tiits, Mai; Valdre, Tiia; Veskis, Leidi; Viks, Ülle; Voll, Piret (Toim.) 2009. Eesti keele seletav sõnaraamat 1–6. [The Explanatory Dictionary of Estonian.] Tallinn: Eesti Keele Sihtasutus.

Erelt, Mati (Toim.) 2003. Estonian Language. Linguistica Uralica Supplementary Series Vol. 1. Tallinn: Estonian Academy Publishers.

Erelt, Mati 2013. Eesti keele lauseõpetus. Sissejuhatus. Öeldis. [Estonian Syntax. Introduction.] Tartu ülikooli eesti keele osakonna preprindid 4. Tartu Ülikool.

Forst, Martin; Holloway King, Tracy; Laczkó, Tibor 2010. Particle verbs in computational LFGs: Issues from English, German, and Hungarian. – Proceedings of LFG 10. CSLI Publications, 228–248.

Kaalep, Heiki-Jaan 2009. Mitmesõnalised verbid. [Multiword verbs.] – Ettekande slaidid internetiaadressil http://www.keeletehnoloogia.ee/konverentsid/ekkt-teine-konverents/slaidid-2009/mitmesonaliste-verbide-aratundmine-tekstides (30.9.2013).

Kaalep, Heiki-Jaan; Muischnek, Kadri 2002. Using the text corpus to create a comprehensive list of phrasal verbs. – Proceedings of Language Resources and Evaluations: Third International Conference on Language Resources and Evaluation, Las Palmas, May 29–31, LREC 2002. European Language Resources Association, 101–105.

Kaalep, Heiki-Jaan; Muischnek, Kadri 2008. Multi-word verbs of Estonian: a database and a corpus. – Proceedings of the LREC Workshop "Towards a Shared Task for Multiword Expressions", Marrakech, Morocco, June 1, 2008. European Language Resources Association, 23–26.

Karlsson, Fred; Voutilainen, Atro; Heikkilä, Juha; Anttila, Arto (Eds.) 1995. Constraint Grammar: A Language-Independent System for Parsing Running Text. Natural Language Processing 4. Berlin, New York: Mouton de Gruyter.

Muischnek, Kadri; Kaalep, Heiki-Jaan 2009. Eesti keele püsiühendid arvutilingvistikas: miks ja kuidas. [Estonian multiword expressions in computational linguistics.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 5, 157–172. http://dx.doi.org/10.5128/ERYa5.10

Müürisep, Kaili 2000. Eesti keele arvutigrammatika: süntaks. [Computer Grammar of Estonian: Syntax.] Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu: Tartu Ülikooli Kirjastus.

Pajusalu, Renate 2009. Sõna ja tähendus. [Word and Meaning.] Tallinn: Eesti Keele Sihtasutus.
Puolakainen, Tiina 2001. Eesti keele arvutigrammatika: morfoloogiline ühestamine. [Computer Grammar of Estonian: Morphological Disambiguation.] Dissertationes Mathematicae Universitatis Tartuensis 27. Tartu: Tartu Ülikooli Kirjastus.
Ramisch, Carlos; Villavicencio, Aline; Moura, Leonardo; Idiart, Marco 2008. Verb-particle constructions, noise and idiomaticity. – Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL 2008), Manchester, UK, August 2008. Association for Computational Linguistics, 49–56
Roosmaa, Tiit; Koit, Mare; Muischnek, Kadri; Müürisep, Kaili; Puolakainen, Tiina; Uibo, Heli 2003. Eesti keele arvutigrammatika: mis on tehtud ja kuidas edasi? [A formal grammar of Estonian: Experience and prospects.] – Keel ja Kirjandus, 46 (3), 192–209.
Rätsep, Huno 1978. Eesti keele lihtlausete tüübid. [Types of Simple Sentences in Estonian.] Eesti NSV TA Emakeele Seltsi toimetised 12. Tallinn: Valgus.
Saagpakk, Paul F. 1992. Eesti-inglise sõnaraamat. [Estonian-English Dictionary.] 2. trükk. Tallinn: Koolibri.
Sag, Ivan A.; Baldwin, Timothy; Bond, Francis; Copestake, Ann; Flickinger, Dan 2002. Multiword Expressions: A Pain in the Neck for NLP. – Alexander Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing. Third International Conference, CICLing 2002, Mexico City, Mexico, February 17-23, 2002. Proceedings. Lecture Notes in Computer Science 2276. Springer Verlag, 1–15.
Uiboaed, Kristel 2010. Statistilised meetodid murdekorpuse ühendverbide tuvastamisel. [Statistical methods for phrasal verb detection in Estonian dialects.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 6, 307–326. http://dx.doi.org/10.5128/ERYa6.19
Villavicencio, Aline 2003. Verb-particle constructions and lexical resources. – Proceedings of the Meeting of the Asociation for Computational Linguistics: 2003 workshop on Multiword Expressions, Sapporo, Japan, July 2003. Association for Computational Linguistics, 57–64
Villavicencio, Aline; Copestake, Ann 2002. Verb-particle constructions in a computational grammar of English. – Jong-Bok Kim, Stephen Wechsler (Eds.). Proceedings of the 9th International Conference on Head-Driven Phrase Structure Grammar, Kyung Hee University, Seoul 5-7 August, 2002. CSLI Publications, 357–371

Võrgumaterjalid

Eesti keele verbikesksete püsiühendite andmebaas. http://www.cl.ut.ee/ressursid/pysiyhendid/ (30.9.2013).
Sõltuvussüntaktiliselt käsitsi analüüsitud korpus. http://www.keeletehnoloogia.ee/ekt-projektid/vahendid-teksti-mitmekihiliseks-margendamiseks/soltuvussuntaktiliselt-kasitsi-analuusitud-korpus (16.2.2014).
Tasakaalus korpus. http://www.cl.ut.ee/korpused/grammatikakorpus/ (16.2.2014).
World of VISL. Visual Interactive Syntax Learning. http://beta.visl.sdu.dk/ (30.9.2013).

PARTICLE VERBS IN ESTONIAN SHALLOW SYNTACTIC ANALYSIS

Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen
University of Tartu

Particle verbs are a frequent phenomenon in Estonian and they need special attention during automatic syntactic analysis. This article deals with the computational treatment of particle verbs in the course of automatic syntactic analysis using the Constraint Grammar framework.

The recognition of particle verbs is needed for further deep syntactic analysis, as often the particle verb as a predicate introduces a different argument structure compared to the corresponding simplex verb.

The processing of particle verbs in the CG framework is mainly lexicon-based, although some groups of particle verbs are also combined by the rules, e.g. combinations of verbs of movement and directional particles.

The Constraint Grammar analyser of Estonian consists of separate modules for morphological disambiguation and syntactic analysis. The recognition of particle verbs starts during morphological disambiguation, but at that stage there is still a lot of general morphological ambiguity in text and that ambiguity hinders the recognition of particle verbs.

In the reported project a special post-disambiguating module was introduced in order to achieve more correct analysis of the uninflecting particle, which could be a part of a particle verb, function as an adposition or a declensional form of a noun or even a verb.

Recognizing particle verbs using the aforementioned extra module achieved recall 97.4% and precision 96.6%.

Keywords: computer linguistics, detection of fixed word combinations, Estonian

Kadri Muischneki (Tartu Ülikool) teaduslikud huvialad on korpuslingvistika, eesti keele süntaktiline struktuur ning automaatne süntaktiline analüüs.
Tartu Ülikool, arvutiteaduse instituut, Liivi 2, 50090 Tartu, Estonia
kadri.muischnek@ut.ee

Kaili Müürisepa (Tartu Ülikool) peamisteks uurimisuundadeks on eesti keele automaatne süntaktiline analüüs ja sellega seonduvad teemad.
Tartu Ülikool, arvutiteaduse instituut, Liivi 2, 50090 Tartu, Estonia
kaili.muurisep@ut.ee

Tiina Puolakaineni (Tartu Ülikool) põhilisteks uurimishuvideks on eesti keele automaatne süntaktiline analüüs ja morfoloogiline ühestamine.
Tartu Ülikool, arvutiteaduse instituut, Liivi 2, 50090 Tartu, Estonia
Tiina.Puolakainen@ut.ee