doi:10.5128/ERYa8.04

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 8, 55–68

OSALAUSETE TUVASTAMINE EESTIKEELSES TEKSTIS KUI ISESEISEV ÜLESANNE

Heiki-Jaan Kaalep, Kadri Muischnek

Ülevaade. Artiklis esitatakse üks viis, kuidas eestikeelses tekstis automaatselt tuvastada osalauseid ja nendega võrdsustatud osa infiniittarinditest, kusjuures omaette üksusena eristatakse kiilud kui sellised üksused, mis katkestavad endast mõlemal pool asuvat sama osalauset.

Kirjeldatav süsteem toetub kirjavahemärkidele, osalause piiril olevatele üksiksõnadele ja verbi finiitsetele vormidele, olles põhijoontes kooskõlas EKK ja EKG II osalause-käsitlusega. Sõnade morfoloogiline analüüs ja ühestamine peavad olema tehtud, kuid süntaktilist analüüsi ei eeldata. Ehkki kirjeldatav algoritm on üsna lihtne, võimaldab ta osalausepiiri tuvastada küllalt hästi: kõigist võimalikest osalause ja kiilu piiridest tuvastas süsteem 95% (saak) ja kõigist väljapakutud piiridest olid õiged 96% (täpsus). Programmi abil on märgendatud osalaused Tartu Ülikooli koondkorpuse veebiversioonis.*

Võtmesõnad: arvutilingvistika, süntaks, lause, osalause, infiniittarind, eesti keel

1. Sissejuhatus

Kirjaliku keelekasutuse pikkade ortograafiliste lausete tükeldamine osalauseteks (ingl *clause splitting*) on oluline samm paljudes loomuliku keele automaattöötluse valdkondades, mh automaatses süntaktilises analüüsis, püsiühendite tuvastamises, anafooride lahendamises, mitmekeelsete tekstide paralleelistamises, masintõlkes jne.

Näiteks eestikeelses tekstis ühendverbide tuvastamisel on info osalausepiiride kohta oluline, sest just osalause on optimaalne kontekst kandidaatpaaride moodustamiseks. Nii ei moodusta lauses *Vaatasin tehtu üle ja jalutasin minema* sõnavormid *üle* ja *minema* ühendverbi, kuigi nende vahel on tekstis ainult kaks sõna, erinevalt lausest *Eesti läheb kevadel samaaegselt oma lähiümbruse riikidega üle suveajale*, kus üksteisest viie tekstisõnaga eraldatud sõnavormid *läheb* ja *üle* moodustavad ühe

* Artikli valmimist on toetanud Euroopa Regionaalarengute Fond Eesti Arvutiteaduse Tippkeskuse kaudu ning Haridus- ja Teadusministeerium (sihtfinantseeritav teema sF0180078s08 "Loomulike keelte arvutitöötluse formalismide ja efektiivsete algoritmide väljatöötamine ning eesti keelele rakendamine"). Täname anonüümseid retsensente asjatundliku kriitika ning soovituste eest.

leksikaalse üksuse. Osalause piiri tuvastamine sellistes lausetes võimaldab kitsendada sõnavormidega *üle* ja *minema* potentsiaalselt kokkukuuluvate sõnade hulka.

Ka Kristel Uiboaed (2010) nendib eesti murrete korpusest ühendverbide automaatse tuvastamise katseid kokku võttes, et üsna määrav oli lausestaja töö efektiivsus: mida korrektsemalt olid osalaused eraldatud, seda adekvaatsem tuli kandidaatandmestik.

Käesolevas artiklis tutvustataksegi selle probleemi ühte võimalikku lahendust: osalausestamise programmi, mille sisendiks on morfoloogiliselt ühestatud tekst ja mis lahendab osalausestamise ülesande ilma täieliku süntaktilise analüüsita.

Selle artiklis kirjeldatava osalausestamise programmiga on osalauseteks jagatud kõik Keeleveebi[1] kaudu kasutatavad tänapäeva eesti keele korpused, mis võimaldab kasutajal korpusest mitme keelendi koosesinemisi otsides valida kontekstiks osalause.[2] Nimetatud korpustest pärinevad ka selles artiklis esitatud näited, mida vahel on küll lihtsustatud ja lühendatud, kontsentreerimaks tähelepanu käsitletavatele nähtustele.

Artikkel on üles ehitatud nii, et selle järgmises osas vaadeldakse osalause mõiste ja piiride käsitlemist eesti keele teaduslikes kirjeldustes, osas 3 määratletakse, mida kirjeldatava osalausestamise süsteemi seisukohalt käsitletakse osalausena, osas 4 esitatakse eestikeelse teksti osalausepiiride leidmise algoritm, 5. osas analüüsitakse selle osalausestaja töö tulemusi ja tehtud vigu ning 6. osas võrreldakse kirjeldatavat programmi teiste samalaadsete süsteemidega.

2. Lause, osalause, kõrvallause, sekundaartarind, klaus

Selles osas vaadeldakse lühidalt lause, osalause, kõrvallause, infiniittarindi ning vähemkäibiva klausi mõiste käsitlemist eesti keele grammatikakirjelduses, loomaks mõistelist tausta artikli järgmistele osadele.

"Eesti keele käsiraamat" kirjeldab lauset kui keelelise suhtluse põhiüksust (EKK: 429), mis tüüpiliselt sisaldab finiitset verbivormi ja ühte või mitut muud moodustajat, milleks võib olla ka osalause (EKK: 434). Osalause mõiste ühendab endas nii pea- kui kõrvallause: liitlause jaguneb osalauseteks. Tüüpilise osalause öeldisverb on finiitne, erandina võib *da*-infinitiiv samuti toimida iseseisva öeldisverbina teatud tüüpi lausetes; ka iseseisvas lauses võib teatud tingimustel esineda *da*-infinitiivne öeldisverb, täpsemalt vt EKG II: 244–246.

"Eesti keele grammatika" (EKG II: 276) kirjeldab osalauset, eriti just põimlause osalauset lause ja sekundaartarindi vahepealse nähtusena. Eriti *da*-infinitiivse öeldisega kõrvallaused sarnanevad infiniittarinditega mitme tunnuse poolest, millest olulisematena olgu nimetatud finiitse öeldisverbi ja subjekti esinemise võimaluse puudumine.

Osalause kui ühte sündmust väljendava süntaktilise tervikuga piirneb sekundaartarind – sekundaarne sündmuse väljendamise vahend, millel on lausega teatud ühisjooni (EKG II: 232). Infiniitsete sekundaartarindite kohta kasutatakse eesti keeleteaduses ka nimetust lauselühend, mis "Eesti keele käsiraamatu" (EKK: 436) järgi on verbifraasi ja osalause vahepealne moodustaja, ja mis, hoolimata sellest, et

[1] www.keeleveeb.ee (01.03.2012).

[2] Korpusepäringu vastuses on osalausepiirid nähtavad roheliste kandiliste sulgudena.

tema peasõnaks on käändeline verbivorm ja temas alust esineda ei saa, sarnaneb rohkem lause kui fraasiga.

Kõrvallaused jagunevad EKG II järgi komplement-, adverbiaal- ja relatiivlauseteks, viimane laiendab tüüpiliselt nimisõna, erandjuhul ka pealauset kui tervikut. Kuna relatiivlause järgneb enamasti vahetult oma peasõnale, paikneb ta sageli oma pealause sees, tükeldades kirjas selle kaheks üksteisest eraldatud katkeks – nt lauses S*ee mees, kes meile vastu tuli, oli meie direktor* on üksteisest eraldatud lauseosad *see mees* ja *oli meie direktor*.

Lauses üldlaiendina esinev, ülejäänud lausega grammatiliselt seostamata lause on kiillause, nt *Uus Peetri katlamaja (asub Peetri pargi lähedal) läheb käiku jaanuaris*. Kiiluga lauset eristab põimlausest see, et kiildumisega ei kaasne grammatilisi muutusi ei kiilduvas lauses ega teda hõlmavas lauses, mõlemad säilitavad oma süntaktilise sõltumatuse ja terviklikkuse. (EKG II: 102)

Uuemates käsitlustes koondab Mati Erelt kõik sündmust väljendavad struktuurid – nii osalaused kui sekundaartarindid – klausi nimetuse alla (Erelt 2004, Erelt 2011). Selles artiklis kirjeldatav programm tuvastab teatud osa klausidest, nimelt osalaused ja väikese osa sekundaartarinditest. Nii on klaus nende tuvastatavate üksuste jaoks liiga lai termin ja osalause jälle liiga kitsas. Edaspidi kasutatakse selles artiklis käsitletavate üksuste kohta siiski koondnimetust osalause, mõeldes selle all osalauseid ja allpool (osas 3) täpsemalt kirjeldatud tunnustele vastavaid infiniitseid klause.

3. Mida kirjeldatav süsteem käsitleb osalause ja kiiluna?

Kirjeldatav süsteem mõistab osalause või osalausega võrdsustatud klausina keeleüksust, millel on lihtlause või lausesarnase süntaktiliselt tervikliku üksuse struktuur. Osalausestamise eesmärgiks on, nagu öeldud, kirjalike tekstide pikkade ortograafiliste lausete tükeldamine "mõistlikeks juppideks" – väiksemateks süntaktilisteks terviküksusteks, mille süntaktiliseks keskmeks on tüüpjuhul finiitne verb, ent võib olla ka infiniitne – seda nii *da*-infinitiivse öeldisverbiga osalauses kui teatud tüüpi infiniittarindites.

Osalausetena märgendab siinkirjeldatav süsteem osa *des*-, *mata*- ja *maks*-klausidest. Esiteks võrdsustatakse osalausega sellised põhilause järel paiknevad komaga eraldatud *des*-, *mata*- ja *maks*-klausid, mille peasõna on tarindi algul, ning teiseks lausealgulised ja komaga eraldatud *des*- ja *maks*-klausid, mille peasõna asub tarindi (ja ka kogu lause) algul. Põhjenduseks esiteks EKG II sedastus (lk 272), et sekundaartarindi eraldamine komadega oleneb tema lauselisuse astmest ja teiseks – komaga eraldatud tarindit on lihtsam automaatselt tuvastada.

Keelenorm suhtub erinevalt *des*- ja *mata*-tarinditesse ühelt poolt ning *maks*-tarindisse teiselt poolt. Kohustuslik on eraldada komaga põhilause järel paiknevad *des*- ja *mata*-tarindid, kui infiniitne verbivorm asub tarindi algul, ning soovitav on eraldada komaga põhilause algul paiknevad pikemad *des*- ja *mata*-tarindid, mille peasõna on tarindi algul. *maks*-tarindite komakasutus ei ole normeeritud, soovi korral võib järgida *des*- ja *mata*-tarindite puhul kehtivaid reegleid.

Eelnevast järeldub, et eeldusel, et kirjakeeles järgitakse täpselt sekundaartarindite komastamise reegleid, ei sõltu osalausestamise tulemus mitte ainult

sekundaartarindi lauselisuse astmest, vaid ka kirjutaja suvast eraldada *maks*-klaus ning põhilause algul paiknev *des*- ja *mata*-klaus komaga või mitte.

Nagu on näidanud Ellen Uuspõld (2001), on tekstides sagedased grammatiseeruvad *des*- ja *mata*-tarindid, mis on muutumas infiniitsest klausist kaassõnafraasiks ja mille komakasutus kõigub. Seejuures on keelekasutaja suhtumine kaassõnastuvatesse *des*- vs. *mata*-tarinditesse erinev: tarindeid, mille peasõnaks on kaassõnastuvad *des*-vormid *alates* ja *võrreldes*, tüüpiliselt komadega ei eraldata, kuid samuti kaassõnafraasiks muutumise teel olevaid tarindeid *mata*-vormidega *hoolimata* ja *vaatamata* kirjavahemärgistatakse suhteliselt tihti (Uuspõld 2001: 320). Sellest tulenevalt ei käsitleta meie osalausestaja väljundis osalausena lausealgulist *mata*-tarindit, mille põhisõna on tarindi algul, sest selles positsioonis kipuvad sageli esinema kaassõnastuvad *mata*-vormid *hoolimata*, *vaatamata* ja *rääkimata* ning komakasutus ei erista kaassõnafraasi infiniittarindist.

Osalausestaja praegune versioon ei erista osalauseid nende süntaktiliste omaduste alusel, s.t ei erista nt relatiiv-, adverbiaal- või komplementlauseid, kuid see võiks olla üks töö edasiarendusi, seda enam, et relatiivlausete äratundmine vormitunnuste põhjal on tegelikult suhteliselt lihtne, erinevalt komplement- ja adverbiaallause eristamisest (vt ka Erelt 2004: 402–404). Samuti ei seata osalauseid omavahel hierarhiasse (pealause, kõrvallause, kõrvallause funktsioon pealause suhtes).

Ent kirjeldatav süsteem märgendab eraldi nn kiilud, mida käesolevas artiklis (erinevalt EKG II ja EKK terminikasutusest) mõistetakse laiemalt kui kiillauset. Kiil on siinses käsitluses tervikuna teise süntaktilise üksuse sees paiknev süntaktiline üksus, s.t. sama süntaktiline struktuur algab enne kiilu ja jätkub pärast kiilu lõppu. Just see, et kiilu taga võib jätkuda eespool alanud osalause, eristab kiilu tavalisest osalausest, mille taga algab tingimata juba uus osalause. Kiiluks võib olla relatiivlause (1) või kiillause (2), aga ka fraas või üksiksõna (3).

(1) Tänased noored eestlased, kel pole huvi kaasaegse maailma vastu kustunud, saavad tekstilise elamuse pigem ..

(2) Oli väga ilus tüdruk, aga nüüd on (tasandab häält) vanaeit!!

(3) See stiil tähendab ülisubjektiivset teksti üldjuhul räpastel (narkootikumid, seks, joomarlus, poliitika) teemadel.

Kiile püütakse eristada muudest osalausetest esmajärjekorras seetõttu, et kiil lõhub lause terviklikkuse, kui ta aga ära tunda ja tinglikult eemaldada, muutub alles jääv lause palju kergemini analüüsitavaks. Kiilu alguse või lõpu märkimine pelgalt osalausepiirina jätaks programmi ilma teadmisest, et samas osalauses on kuskil (s.t teisel pool kiilu) veel sõnu, mis vaadeldavatega kokku kuuluvad, ja seega teeks programmi töö raskemaks.

Kiile on kerge eristada kirjavahemärgistusele tuginedes, kuna nad paigutatakse kirjas (üla)komade, mõttekriipsude või sulgude vahele.

Kiilude eristamine tavalistest osalausetest, mis võimaldab kiilu ümbritsevat struktuuri kohelda tervikuna, on autorite arvates üks kirjeldatava süsteemi eeliseid teiste omasarnaste ees.

4. Osalausestamise algoritmi kirjeldus

Algoritmi luues oletati, et lauseid konstrueerides kasutatakse teatud võtteid – kirjavahemärke, sidesõnu, sõnade järjekorda jms – süsteemi huvitavate klauside üksteisest eraldamiseks.

Liitlause loomisel kasutatakse teatud kindlaid malle või lausekonstruktsioone, mille tuvastamisel võib olla kindel, et lause on liitlauselise struktuuriga, olenemata sellest, kas mõned lausele muidu kohustuslikuks peetavad elemendid on igas lauseosas olemas või mitte. Näiteks lauses *Kui osta külmkapp, siis ikka selleks, et toitu säilitada* ei ole ühtegi finiitset verbi, aga on selgelt kolm iseseisvat osa ja neid signaliseerivad sõnad *kui*, *siis* ja *et*.

Oletati ka seda, et kirjavahemärkide panemise tava vastab küll suurel määral eesti keele grammatikates esitatud ortograafiareeglitele, kuid ei pruugi nendega päris täpselt kattuda.

Osalausete leidmise puhul seisab programm tavaliselt vastamisi pika kirjaliku keele lausega, milles on rohkem kui üks sidesõna, sidend või osalause piiri tähistav kirjavahemärk ja seejuures pole teada, millised neist märkidest või sõnadest eristavad osalauseid, millised aga rinnastatud sõnu või fraase. Nii ongi algoritmi üldiseks ideeks, et osalauseid ja kiile tuvastatakse järk-järgult, alustades lihtsamatest ja kindlamatest juhtumitest. Loodetavasti muutub lause struktuur pärast iga etappi programmi jaoks lihtsamaks. Iga lause vaadatakse läbi mitu korda.

Sisendiks on morfoloogiliselt ühestatud tekst, s.t iga tekstisõna on varustatud informatsiooniga tema lemma ja grammatiliste kategooriate kohta. Morfoloogiline analüüs (s.h sõnastikust puuduvate sõnade analüüs) ja statistiline ühestamine tehti OÜ Filosofti programmiga *etmrf*.

Kõigepealt märgitakse kindlad kiilud, siis tähistatakse osalause keskmeks sobivad verbid, seejärel märgitakse teatud kirjavahemärgid ja sõnad võimalike osalausepiiridena ja siis püütakse otsustada, millised neist võimalikest on kindlad osalausepiirid ja millised omakorda kindlasti osalauseid ei eralda.

Kõige lõpuks tähistatakse kiiludena veel mõned kindlalt tuvastatud osalaused, et hilisemad tekstitöötlusprogrammid saaksid paremini töötada; nt (4).

(4) Avastus, <kiilu algus> et asi on halb <kiilu lõpp>, tekitas hirmu.

Järgneb algoritmi detailsem kirjeldus.

1. Sulgudes olev tekst, ka ühesõnaline, märgitakse kui kiil. Sulud on kindel indikaator selle kohta, et nendes olev tekst ei ole süntaktiliselt ümbritseva teksti osa.
2. Tähistatakse osalause keskmeks sobivad verbid: verbi finiitsed vormid; *nud*-partitsiibid, kui nad on kindlalt määratletavad kui mineviku (*on teinud*) või eituse (*ei teinud*) vormid ja seega ei esine atribuudi rollis (*kadunud auto*), *tud*-partitsiibid eituse vormidena (*ei tehtud*) ja *des-*, *mata-* ning *maks*-vormid, kui nad on vahetult koma järel ehk esimene sõna lauselühendis, mis on järelasendis. (Ka korrelaatsidendite *hoolimata sellest et* ja *vaatamata sellele et* puhul, kus ainuke koma on terve sidendi ees, ei tee *hoolimata* ja *vaatamata* ekslik tähistamine osalause keskmeks sobivana otseselt kahju, sest hiljem kasutatakse potentsiaalsete osalausepiiride säilitamiseks/eemaldamiseks ainult keskmeks sobiva verbi olemasolu

informatsiooni, ja antud juhul on kindel, et selline keskmeks sobiv verb samas osalauses ka leidub.) Ka lausealgulisi *des*- ja *maks*-vorme käsitletakse osalause keskmeks sobivatena.

Kõiki liitöeldise koosseisu kuuluvaid sõnavorme ei püütagi tuvastada, sest tähtis on ainult teada, kas potentsiaalses osalauses leidub tema keskmeks sobiv verbivorm või mitte, verbi liitvormi või perifrastilise verbivormi täpne koosseis pole selle ülesande seisukohalt oluline.

On selge, et vead verbi määratlemisel tingivad hiljem vigu ka osalausestamisel. Praeguse algoritmi puuduseks on asjaolu, et liitajavormide puhul õnnestub verbi *olema* vorm ja *nud*-partitsiip ühtekuuluvaks määrata ainult siis, kui nad asuvad lauses kõrvuti. Kui *olema*-vormi ja *nud*-partitsiibi vahel on sõnu, nt *oli Mart uskunud*, siis partitsiibivormi ei märgita öeldise osaks, aga kuna *olema* vorm on nagunii märgitud osalause keskmeks sobiva finiitse verbivormina, siis osalausestamiseks sellest piisab. Osalausestamises tekivad vead juhul, kui lauses on hiljem veel kord kasutatud täismineviku liitaega, kuid selliselt, et *olema*-verb on väljajätteline ning reaalselt esineb ainult *nud*-partitsiibi vorm. Siis on vaja otsustada, kas tegemist on atribuudina kasutatud partitsiibi või väljajättelise öeldisega, selle otsustamiseks tuleb aga omakorda kontrollida, kas *nud*-vormi on samas lauses juba varem kasutatud koos verbi *olema* vormiga (s.t öeldisena). Lauses (5) määratakse *lõhkunud* võimalikuks öeldiseks, sest talle eelnevas lauseosas on *avanud* märgitud öeldise osana, lauses (6) jätab aga algoritm sõnavormi *avanud* öeldise osa märgendita, seetõttu jääb selleta ka *lõhkunud* ning osalause piir jääbki antud lauses lisamata. Lauses (5) on korrektselt tuvastatud osalausepiir märgitud püstkriipsuga.

(5) Ta oli avanud akna ja | lõhkunud ukse.
(6) Ta oli kiiresti avanud akna ja lõhkunud ukse.

Et selliseid *nud*-partitsiipide öeldisena mitte-äratundmisest tingitud vigu vähendada, on edaspidi kavas uurida, kas väljajättelise öeldise kasutamine järgib mingeid sagedasi ja kindlaid malle, nt hüpoteetiline reegel "kui *nud*-partitsiip asub vahetult sidendi järel ja ta ei kuulu mõne omadussõnaga samasse loendisse (nt *õnnelik ja väsinud*), siis on ta tõenäoliselt öeldis".

tud-partitsiibile ei rakendata *nud*-partitsiibi reegleid, põhjuseks asjaolu, et automaatselt ei suudeta ilma süntaksianalüüsita nagunii eristada impersonaali liitmineviku vorme seisundipassiivi vormidest. EKG II (lk 30–31) põhjal on lauses (7) *tud*-partitsiibi vormid impersonaali liitajavormi osad ja järelikult on tegu kahe osalausega, lauses (8) aga on tegu seisundipassiiviga – rinnastatud predikatiividega ühes osalauses. Ka EKK (lk 456–457) möönab, et passiivi korral on partitsiip pigem adjektiivi ja verbi piirijuhtum kui üks neist. On veel teinegi põhjus *tud*-partitsiibi käsitlemiseks *nud*-partitsiibist erinevana: nimelt tundus esialgse korpusuuringu põhjal, et *tud*-partitsiipi kasutatakse atribuudina *nud*-partitsiibist sagedamini, s.t liiga palju on vigu põhjustavaid lauseid nagu (9).

(7) Siin on magatud ja õpitud.
(8) Praod olid vatti täis topitud ja paberiga kinni kleebitud.
(9) Palju on kollaseid värvitud pindu ja poleeritud alumiiniumi.

3. Märgitakse võimalikud osalausete piiride kohad, milleks on kirjavahemärgid *: , ; – . ? !* ja sidesõnad *ja*, *ning*, *ega*, *või*.
4. Otsese kõne algus (koolon, millele järgneb jutumärk) ja lõpp (punkt, hüüumärk, küsimärk või koma, millele järgneb jutumärk) märgendatakse kui kindlad osalausepiirid.
5. Kui jutumärkide paarsus on kindlaks tehtav ja osalause keskmeks sobiv verb asub nii jutumärkidest seespool kui ka väljas, siis on jutumärkide vahel tõenäoliselt tsitaat (s.t kiil) ja see märgitaksegi sellisena; muud jutumärgid ei tähista ei osalause ega kiilu piiri (nt *"Sõda ja rahu"*).
6. Kirjavahemärgid *:* ja *;* märgitakse kui kindlad osalause piirid.
7. Koma või mõttekriips, millele järgneb vahetult kas *ja*, *ning*, *ega*, *või*, *et*, *kui*, *kus*, *kuhu*, *kust*, *sest*, *kuid*, *nagu*, *ehkki*, *siis*, *kuni*, *otsekui*, *justkui*, *kuna*, *kuidas* või *kas*, märgitakse kui kindel osalause piir.

   Koma või mõttekriips, millele järgneb sõna *mis*, *kes*, *missugune*, *milline* või *see* mistahes käändes, märgitakse samuti kui kindel osalause piir.

   Kui komale või mõttekriipsule järgneb mingi sõna, millele omakorda vahetult järgneb sõna *et* (*ilma et*, *nii et*, *ainult et* jms), siis ka seal on kindel osalause piir.

   Nagu näha, peetakse mõningaid kirjavahemärgi ja sõna kombinatsioone niivõrd usaldusväärseteks osalausepiiride signaliseerijateks, et ei kontrollita isegi osalause keskmesse sobiva verbi olemasolu. Põhjenduseks kaalutlus, et selline kindlal moel tähistatav koht lauses tähendab seda, et järgnevad sõnad selles lauseosas kindlasti ei kuulu eelnevatega sama tihedalt kokku kui need, mis asuvad enne piiri.

   Seevastu koma või mõttekriips, millele järgneb vahetult sõna *aga* või *kuigi*, märgitakse kui kindel osalause piir ainult juhul, kui neile sõnadele omakorda enne järgmist võimalikku osalausepiiri järgneb osalause keskmeks sobiv verb. Sellega välditakse liigse piiri panemist näiteks kontekstides *aitab lastel, aga ka täiskasvanutel head tuju säilitada*; *jooksis vaikselt, aga kiiresti* ja *sisaldab vett, kuigi väheses koguses*.

   Tuleb tunnistada, et asjaolu, et sõnade *kuid* ja *ehkki* puhul osalause keskmesse sobiva verbi olemasolu ei kontrollita, *aga* ja *kuigi* puhul aga kontrollitakse, näitab programmi ebajärjekindlust. Kas ja kuidas seda muuta, nii et osalausestaja muutuks paremaks, vajab edasist uurimist.
8. Vaadeldakse allesjäänud oletuslikke piire. Kui mõlemal pool piiri on osalause keskmeks sobiv verb, siis lihtsalt tähistatakse see piir kindlana.

   Aga mida teha juhul, kui oletuslike piiride vahel on sellist verbi mittesisaldav fraas, aga mõlemal pool fraasi on osalause keskmeks sobiv verb, nt fraas *endine üliõpilane* lauses (10)? Kas osalause piiriks tuleks valida sellise fraasi ees või taga olev piir? Kõige üldisem reegel sellistel puhkudel on, et kui *ja*-le, *ning*-ile või komale järgneb vahetult osalause keskmeks sobiv verb, siis see ongi osalause piir, antud näite puhul seega jääb osalauseks *Mari on tegelikult Maiu, endine üliõpilane*; korrektselt tuvastatud osalausepiir on näites (10) märgitud püstkriipsuga.

(10) Mari on tegelikult Maiu, endine üliõpilane, | elab Tammsaare teel ja | armastab rahvalaule laulda.

9. Eemaldatakse oletuslikud osalause piirid, kui need paistavad olevat rinnastatud elementide vahel. S.t kui mõlemal pool oletuslikku piiri on ühesuguses käändes sõna (nt *rohelistes, punastes ja sinistes pükstes*). Mõnikord saabub sellise eemaldamise tulemusena olukord, kus veel mõnede oletuslike osalausepiiride puhul on näha, et neist mõlemale poole jäävas osalauses leidub selle keskmeks sobiv verb ja seega saabki antud piiri kindlana tähistada.

   Lause (11) on kahe *ja* ja ühe *ning*-iga jagatud neljaks osaks, millest esimeses ja viimases on osalause keskmeks sobiv verb, kahes keskmises aga mitte. Kus peaks olema osalausepiir? Märganud, et *pikkade* ja *pingeliste* on samas käändes, võib nendevahelise *ja* oletuslike osalausepiiride hulgast kustutada; samal moel võib käituda *tujukuse* ja *isepäisuse* vahelise *ning*-iga. Järele jääb ainult üks oletuslik piir, millest mõlemal pool on osalause keskmeks sobiv verb, seega võib selle tähistada kindlana (näites püstkriipsuga).

   (11) Ma ei nurisenud pikkade ja pingeliste tööpäevade üle ja | Vabariigi Presidendi paljukirutud tujukus ning isepäisus ei häirinud mind.

10. Mõned osalaused nimetatakse ümber kiiludeks; seda saab teha ainult sellistel juhtudel, kui on kindel, et osalause järel jätkub sama lause, mis algas osalause ees, nt (12). S.t ühel pool oletatavat kiilu peab osalause keskmeks sobiv verb kindlasti puuduma. Kiiluks sobivad kõrvallaused, mis algavad sõnaga *kes*, *mis*, *missugune*, *milline* (mistahes käändes), *kus*, *kuhu*, *kust*, *millal* või *et*, s.t relatiivlaused.

    (12) Mees, <kiilu algus> kes tuli vastu <kiilu lõpp>, kandis musta kaabut.

5. Tulemused

Osalausestaja töö hindamiseks kontrolliti käsitsi u 16 000 sõna mahus osalausestatud tekste, mis jagunesid enam-vähem võrdselt ajakirjanduse, ilukirjanduse ja populaarteaduse (ajakiri Horisont) tekstiklasside vahel. Tulemuste hindamiseks arvutati osalausepiiride märgendamise täpsus (süsteemi poolt õigesti tuvastatud osalausepiiride arv jagatud kõigi süsteemi poolt tuvastatud osalausepiiride arvuga) ja saak (süsteemi poolt õigesti tuvastatud osalausepiiride arv jagatud tegelike osalausepiiride arvuga). Seejuures loeti veaks ka sellised juhtumid, kui kiilu alguse või lõpu asemel oli pandud lihtsalt osalausepiir. Üldiselt võib tulemusi nimetada korralikeks: üldine täpsus oli 94% ja saak 96%. Tulemused tekstiklassiti on esitatud tabelis 1.

Tabel 1. Osalausepiiride (sh kiilude) tuvastamise kvaliteet

Tekstiklass	Sõnu	Lauseid	Süsteemi poolt märgendatud osalausepiire	Neist õigeid	Süsteemi poolt märgendamata jäänud osalausepiire	Saak	Täpsus
ajakirjandus	5205	328	308	294	15	95%	95%
populaarteadus	5944	439	333	318	18	95%	95%
ilukirjandus	4926	286	440	427	21	95%	97%
KOKKU	16075	1053	1081	1039	54	95%	96%

Kui vaadata eraldi tavalise osalausepiiri ja kiilu tuvastamist, siis tavalise osalausepiiri tuvastamine on lihtsam ülesanne (saak 98% ja täpsus 96%) kui kiilu piiride märgendamine (saak 79% ja täpsus 100%). Kõige sagedasem viga osalausepiiride märgendamisel seisneski selles, et üks või mõlemad kiilu piirid märgendati tavaliste osalausepiiridena, näiteks on süsteem märgendanud lauses (13) esimese relatiivlauset eraldava koma kindla osalausepiirina (näites märgitud püstkriipsuga), kuid teisele komale pole lisanud mingit piiri märgendit. Lauses (14) on mõlemad kiillauset *kui see kedagi huvitab* eraldavad komad märgendatud lihtsalt osalausepiiridena. Kuna kiilu märgendid on asendunud tavaliste osalausepiiride märgenditega, jääb kiilu poolt jagatud lause *Meie neljaliikmelise meeskonna taust oli selline* tervikuks ühendamata.

(13) Seega oli samm, | mille astus Eesti, palju pikem ja otsustavam.

(14) Meie neljaliikmelise meeskonna taust, | kui see kedagi huvitab, | oli selline ..

Vigu põhjustab ka vale morfoloogiline analüüs, näiteks lauses (15) oli sõnavorm *kaaluvad* analüüsitud ekslikult verbi *kaaluma* oleviku mitmuse kolmanda isiku vormiks, mitte oleviku partitsiibi mitmuse nimetavaliseks vormiks, ja see morfoloogilise analüüsi viga põhjustas vale osalausepiiri lisamise viimase koma juurde. Lauses (16) on liigne osalausepiir lisatud selletõttu, et sõnavorm *minevat* on ekslikult morfoloogiliselt tõlgendatud kui kaudse kõneviisi vorm, mitte kui *vat*-infinitiiv (nad on nimelt alati homonüümsed) ja selle tulemusena on meil lause, kus mõlemal pool sidesõna *ja* on verbi finiitne vorm, ehk siis kindel tunnus, et tegemist on kahe osalausega.

(15) Seda arvestavad ärimehed, investorid, | krediidiliine kaaluvad pangad.
(16) Tüdruk paistis ikka rohkem ja | rohkem närvi minevat.

Muidugi põhjustavad vigu ortograafiareeglitele mittevastavad kirjavahemärgid, näiteks lauses (17).

(17) .. mida lähiaastatel saab kujundama Eesti, | kui kindla EL-iga ühineja maine.

Märgendaja jaoks on problemaatilised rinnastatud osalaused, mille öeldisverbid on liitaegades ja teises osalauses on *olema*-verb välja jäetud (vt ka osa 4). Kuna lause (18) teises osalauses *luule triviaalsüsteem teisenenud* on *olema*-verbi finiitne vorm välja jäetud ja osalauses on eksplitsiitselt olemas ainult *nud*-partitsiip *teisenenud*, jääb tuvastamata ka osalausepiir.

(18) Aeg oli temast mööda läinud, luule triviaalsüsteem teisenenud ..

Osalausepiiride automaatse määramise süsteemi töö hindamisel loeti veaks ainult sellised süsteemi poolt märkimata jäänud osalausepiirid, mida süsteem oma aluspõhimõtetest lähtuvalt oleks pidanud tuvastama. Osalausestaja väljundis on ka selliseid kohti, kuhu järjekindluse huvides oleks tulnud osalausepiir panna, aga kuna tekstis polnud selgeid pidepunkte, millele toetudes seda teha, siis, et vältida paratamatuid vigu, ei üritatudki neid piire märgendada. Järgnevalt kirjeldatakse näidetena mõningaid selliseid konstruktsioone.

Automaatselt on raske eristada rinnastatud öeldisverbe ja rinnastatud osalauseid (mis võivad ju samuti koosneda ainult öeldisverbist). EKG II (lk 214) järgi pole korduvate verbifraasidega lauseid enamasti võimalik üheselt rindlauseteks ja rinnastatud öeldistega koondlauseteks jagada; öeldiste endi rinnastusega on kindlalt tegemist ainult siis, kui rinnastuse funktsiooniks on eelkõige verbi tähenduse intensiivistamine, nt *Ma aina ootan ja ootan* või *Ta nuttis ja naeris ühekorraga*. Tähenduse intensiivistamist automaatselt tuvastada muidugi ei õnnestu ja nii pannakse osalausete automaatsel tuvastamisel kahe finiitse verbivormi vahele alati osalausepiir, nt (19-20).

(19) Ühe treeninguga kadus paar kilo, | pärast sõi ja | jõi need aga jälle tagasi.
(20) Ainult ootan ja | ootan.

Osa infiniitseid klause, mh ka teatud tingimustele vastavad gerundiivitarindid peaksid olema märgendatud omaette osalausetena (vt osa 3). *des*-vormi minevikulisteks vasteteks on *olema*-verbist ja *nud*- või *tud*-partitsiibist koosnevad analüütilised verbivormid (nt *olles teinud*, *olles tehtud*). Tekstis jäetakse sõnavorm *olles* tavaliselt ära ja gerundiivi minevikuvorm esinebki *nud*- või *tud*-partitsiibi kujul. Mineviku partitsiipidel on aga ka palju teisi kasutusi, mh täiendina, nii et iga *nud*- või *tud*-vormi kuulutamine potentsiaalseks osalausesarnase süntaktilise struktuuri verbiliseks keskmeks põhjustaks palju vigu. (Siinkohal tuleb jällegi rõhutada, et liitajavormide puhul piisab üksi *olema*-verbi finiitsest vormist selleks, et osalausel oleks finiitne kese.) Nii jäävad gerundiivi minevikuvormis tarindid, kus sõnavorm *olles* on väljajätteline, omaette osalausena märgendamata, nt (21).

(21) Tajunud oma võõrdumist tegelikust elust, asub ta ..

6. Võrdlus eesti ning teiste keelte jaoks loodud sarnaste süsteemidega

Osalausepiiride tuvastamiseks kasutatakse nii reeglipõhiseid (nt Müürisep 2000, Puolakainen 2001), masinõppel põhinevaid (nt Mitkov jt 1999) kui ka hübriidseid (nt Orasan 2000, Puscasu 2004) süsteeme. Osalauseteks jagamise ülesande lahendamine algab alati osalausena mõistetava üksuse täpse piiritlemisega; selle kohta täpsemalt nt (Orasan 2000: 130–131).

Käesolevas artiklis kirjeldatud süsteemile sarnaneb näiteks Georgiana Puscasu (2004) algselt rumeeniakeelsete tekstide jaoks loodud ja hiljem inglise keele jaoks kohandatud osalausestamise süsteem, mille eesmärgiks on küll ainult finiitverbi sisaldavate osalausete tuvastamine. Puscasu süsteem saab samuti sisendiks morfoloogiliselt ühestatud teksti ning kasutab osalausepiiride määramiseks põhiliselt kirjavahemärke, sidesõnu ning infot verbi finiitsuse/infiniitsuse kohta. Puscasu süsteemi *F-mõõt* osalause alguse määramisel oli 95% rumeenia ja 92% inglise keele jaoks.

F-mõõt on täpsuse ja saagi harmooniline keskmine, mida arvutatakse järgmise valemi järgi:

$$F = \frac{2 \cdot saak \cdot täpsus}{saak + täpsus}$$

Käesolevas artiklis kirjeldatud eesti keele osalausestaja *F-mõõt* on järelikult 95%.

Sarnasest tööst võiks nimetada veel saksa keele korpuse DEREKO (DEutsches REferenzKOrpus) märgendamise projekti, mille lõpparuandes (Dipper jt 2002) käsitletakse eraldi osalausete märgendamise probleemi, kusjuures osalause on defineeritud kui üksus, millel on üks põhiverb, mis võib olla nii finiitne kui infiniitne. DEREKO lingvistiliste märgendite süsteemis eristatakse kolme tüüpi osalauseid: üldisest osalause klassist eristatakse relatiivlauseid ja eraldi märgendatakse infiniitse verbiga osalauseid (originaalis: *clauses*). Osalausete omavahelisi suhteid eksplitsiitselt ei esitata (s.t lausepuus pole näidatud, millist sõna/fraasi/osalauset mingi osalause laiendab).

Eesti keele automaattöötluse vallas on osalausestamisega varem tegeldud reeglipõhise morfoloogilise ühestamise (Puolakainen 2001) ning reeglipõhise süntaktilise analüüsi (Müürisep 2000) raames. Sarnaselt siin artiklis kirjeldatud osalausestajale kasutavad mõlemad nimetatud reeglipõhised süsteemid osalausepiiride määramisel pidepunktidena sidesõnu, kirjavahemärke ja finiitseid verbivorme, kuid määratlevad osalauset täpselt EKG II järgi, s.t ühtegi infiniittarindit ei käsitleta osalausena. Nii morfoloogilise ühestaja kui ka süntaksianalüsaatori koosseisus olevate osalausepiiride leidmise reeglite eesmärgiks ei ole kiilude (kiillausete, relatiivlausete, sulgudes lausete või lauseosade jms) eristamine tavalisest osalausest. Selle tulemusena lõhub kiil alati oma pealause, s.t lauses *Mees, kes seal seisab, ostis minu naabermaja ära* jääb lausealguline moodustaja *mees* ühendamata ülejäänud lauseosaga *ostis minu naabermaja ära*.

Müürisep ega Puolakainen ei ole eraldi hinnanud osalausepiiride määramise tulemuslikkust, mistõttu ei ole võimalik võrrelda siinkirjeldatud süsteemi tulemusi nende omadega.

Müürisepa väljatöötatud eesti keele kitsenduste grammatika süntaksianalüsaatorit (Müürisep 2000) on kohandatud ka suulise keele (Müürisep jt 2006) ja murdetekstide (Lindström, Müürisep 2009) analüüsiks. Osalausepiiride määramisel suulise kõne tanskriptsioonides ei saanud lähtuda kirjavahemärkidest, sest neid ei kasutata transkribeeritud tekstides normeeritud kirjakeele reeglite järgi, vaid lähtuda tuli transkriptsiooniks kasutatud intonatsioonimärkidest ja partiklitest. Suulise kõne automaatse süntaktilise analüüsi kvaliteedi hindamisel väidavadki autorid, et osalausepiiride valesti määramine põhjustas enim vigu (Müürisep jt 2006: 79). Ka murdetekstide – mis esindavad samuti suulist keelekasutust – analüüsil on tuvastamata osalausepiir kõige sagedasem veatüüp (Lindström, Müürisep 2009: 27).

7. Kokkuvõte

Paljude lingvistiliste probleemide (nii teoreetiliste kui praktiliste) lahendamiseks oleks hea, kui pikad laused oleks jagatud mõistlikeks üksusteks, mille piiridesse antud probleemi seisukohalt olulised lingvistilised nähtused teadaolevalt jäävad. Üheks selliseks mõistlikuks üksuseks on osalause ja osalausele lähedasem osa infiniitsetest klausidest. Asjaolu, et alati ei ole võimalik üheselt otsustada, kas ja mitmeks mõistlikuks osaks tuleks lause jagada, ei tähenda, et sellist jagamist üldse ei saaks teha või et automaatne jagamine oleks väga vearohke.

Selles artiklis esitati üks viis, kuidas automaatselt tuvastada osalauseid ja nendega samastatud osa infiniittarinditest, nimelt teatud osa *des*-, *mata*- ning *maks*-tarinditest. Tuvastatavate konstruktsioonide hulga piiritlemine on paratamatult seotud ka tuvastamise viisiga: milleks defineerida kategooriat, mille esinemisi ei suudeta teistest eritleda?

Kirjeldatud süsteem toetub kirjavahemärkidele, osalause piiril olevatele üksiksõnadele ja verbi finiitsetele vormidele, olles põhijoontes kooskõlas EKK ja EKG II osalausekäsitlusega. Sõnade morfoloogiline analüüs ja ühestamine peavad olema tehtud, kuid süntaktilist analüüsi ei eeldata. Omaette üksusena eristatakse kiilud kui sellised üksused, mis katkestavad endast mõlemal pool asuvat sama osalauset.

Ehkki kirjeldatav algoritm on üsna lihtne, võimaldab ta osalausepiire tuvastada küllalt hästi. Käsitsi kontrollimise tulemusel selgus, et kõigist võimalikest osalause ja kiilu piiridest tuvastas süsteem 95% (saak) ja et kõigist väljapakutud piiridest olid õiged 96% (täpsus). Programmi abil on märgendatud osalaused TÜ koondkorpuse veebiversioonis.

Viidatud kirjandus

Dipper, Stefanie; Kermes, Hannah; König-Baumer, Esther; Lezius, Wolfhang; Müller, Frank H.; Tylman, Ule 2002. DEREKO. German Reference Corpus. Final Report (Part I). Internetidokument aadressil www.sfs.uni-tuebingen.de/dereko/DEREKOReport.pdf (20.09.2011).

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

EKK = Erelt, Mati; Erelt, Tiiu; Ross, Kristiina 2007. Eesti keele käsiraamat. Tallinn: Eesti Keele Sihtasutus.

Erelt, Mati 2004. Märkmeid eesti keele komplekslause kohta. – Keel ja Kirjandus, 6, 401–413.

Erelt, Mati 2011. Tänapäeva eesti kirjakeele morfosüntaksi ja süntaksi uurimisest Tartu ülikoolis. – Emakeele Seltsi aastaraamat, 56 (2010), 37–62.

Lindström, Liina; Müürisep, Kaili 2009. Parsing corpus of Estonian dialects. – E. Bick, K. Hagen, K. Müürisep, T. Trosterud (Eds.). Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing, Odense, Denmark, 14.05.2009, NEALT Proceedings Series, 8. Tartu: Tartu University Library.

Mitkov, Ruslan; Orasan, Constantin; Evans, Richard 1999. The importance of annotated corpora for NLP: The cases of anaphora resolution and clause splitting . – Proceeding of "Corpora and NLP: Reflecting on Methodology Workshop", TALN'99. Veebidokument aadressil http://clg.wlv.ac.uk/papers/show_paper.php?ID=10 (16.09.2011).

Müürisep, Kaili 2000. Eesti keele arvutigrammatika: süntaks. Dissertationes mathematicae Universitatis Tartuensis, 22. Tartu: TÜ kirjastus.

Müürisep, Kaili; Nigol, Helen; Uibo, Heli 2006. Eesti suulise keele korpuse automaatne pindsüntaktiline analüüs. – M. Koit, R. Pajusalu, H. Õim (Toim.). Keel ja arvuti. Tartu: TÜ Kirjastus, 72–84.

Orasan, Constantin 2000. A hybrid method for clause splitting in unrestricted English texts. – Proceedings of ACIDCA 2000, Corpora and Natural Language Processing, March 22–24, Monastir, Tunisia, 129–134.

Puolakainen, Tiina 2001. Eesti keele arvutigrammatika: morfoloogiline ühestamine. Dissertationes mathematicae Universitatis Tartuensis, 27. Tartu: TÜ kirjastus.

Puscasu, Georgiana 2004. A Multilingual method for clause splitting. – Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics,

Birmingham, UK. Internetidokument aadressil http://clg.wlv.ac.uk/papers/puscasu-04a.pdf (04.09.2011).
Uiboaed, Kristel 2010. Statistilised meetodid murdekorpuse ühendverbide tuvastamisel. – Eesti Rakenduslingvistika Ühingu aastaraamat, 6, 307–326. http://dx.doi.org/10.5128/ERYa6.19
Uuspõld, Ellen 2001. *des*- ja *mata*-vormide kaassõnastumine ja eesti komareeglid. – Reet Kasik (Toim.). Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks. TÜ eesti keele õppetooli toimetised, 17. Tartu: TÜ kirjastus, 306–321.

Heiki-Jaan Kaalepi (Tartu Ülikool) põhilised uurimisvaldkonnad on korpuslingvistika ja eesti keele morfoloogia.
heiki-jaan.kaalep@ut.ee

Kadri Muischneki (Tartu Ülikool) teaduslikeks huvialadeks on korpuslingvistika ning eesti keele (automaatne) morfosüntaktiline ja süntaktiline analüüs.
kadri.muischnek@ut.ee

CLAUSE SPLITTING AS A SEPARATE TASK (IN THE ANALYSIS OF ESTONIAN TEXTS)

Heiki-Jaan Kaalep, Kadri Muischnek
University of Tartu

Clause splitting is usually included in syntactic analysis, but it can be also regarded as a task in itself. For example, for a corpus query system it would be convenient to have some knowledge about clause boundaries in order to enable the user to retrieve co-occurrences of words or grammatical categories not in the usual "window" of three or four words, but in the whole clause. This paper presents a rule-based clause splitting system for written Estonian. The input text has to be morphologically disambiguated, but no parsing is required. So the presented system offers a convenient option for tasks that need information about clause boundaries. Two types of clauses are recognized: parentheses are treated separately from other clauses. By parenthesis we mean a clause that is situated in another clause dividing the latter into two separate strings. Recognizing parentheses enables us to treat the divided clause as a whole. The clause splitting system achieves 96% precision and 95% recall in unrestricted Estonian texts.

Keywords: computational linguistics, syntax, sentence, clause, non-finite clause, Estonian