Eesti keele morfoloogilised ühestajad

Tiina Puolakainen
Eesti Keele Instituut

Morfoloogiline ühestamine – mis see on?

Morfoloogiline analüsaator annab sõnavormile tihti mitu morfoloogilist tõlgendust. Alternatiivsete analüüside tekkimise peamiseks põhjuseks on vormide homonüümia, näiteks sõnavorm *ilma* võib olla nimisõna omastavas, osastavas või lühikeses sisseütlevas käändes, kaassõna või määrsõna. Morfoloogiline analüsaator vaatab antud üksikut sõnavormi ja muidugi ei saa nende tõlgenduste hulgast valikut teha. Enamikus rakendustes oleks aga hea teada mitte ainult seda, millised võimalikud tõlgendused on sõnal, vaid milline on tema tõlgendus konkreetses kontekstis. Näiteks kasvõi sagedussõnastiku koostamisel on vaja teada, kas sõnavormi *peeti* lugeda nimisõnaks *peet* või verbiks *pidama*. Selliste sõnavormide näiteid võiks tuua veelgi: *viis* (nimisõna *viis*, arvsõna *viis*, verb *viima*), *sai* (nimisõna *sai*, verb *saama*), *või* (nimisõna *või*, sidesõna *või*, verb *võima*), *tee* (nimisõna *tee*, verb *tegema*), *tuli* (nimisõna *tuli*, verb *tulema*). Ilma iga sõnavormi morfoloogilist tõlgendust teadmata ei ole mõeldav ka teksti edaspidine süntaktiline, semantiline, pragmaatiline analüüs, ei saa hästi funktsioneerida infootsisüsteemid, lingvistilised rakendused, grammatikakorrektorid, masintõlke- ja dialoogsüsteemid.

Morfoloogiline ühestamine on loomuliku keele automaatse analüüsi etapp, mis järgneb morfoloogilisele ja eelneb süntaktilisele analüüsile. Morfoloogilist ühestamist teostava programmi e ühestaja sisendiks on morfoloogilise analüsaatori väljund – morfoloogiliselt analüüsitud tekst – ning morfoloogilise

ühestaja väljund on omakorda süntaktilise analüsaatori sisendiks. Morfoloogilise ühestaja ülesanne on eemaldada konkreetsesse konteksti mittesobivad morfoloogilised tõlgendused.

Erinevatel autoritel võib kohata erinevaid käsitlusi morfoloogilise ühestamise kohast loomuliku keele analüüsi etappides: kord on ühestamine eraldi ülesanne, kord on morfoloogilise analüüsi osa või lisand, kord süntaktilise analüüsi osa. Igal seisukohal on ka oma põhjendus. Kui me vaatame ühestajat kui nn musta kasti – ei vaata, kuidas ta töötab, vaid ainult tema väljundit – siis selleks väljundiks on morfoloogiliste tõlgendustega varustatud tekst, seega võib morfoloogilist ühestamist pidada morfoloogilise analüüsi osaks või lisaks. Kui aga vaadata ühestaja sisulist tegevust (seestpoolt), siis ebakorrektsed tõlgendused eemaldatakse vaadates konteksti, mitte üksikut sõnavormi – ja seega kuulub ühestamine juba mitte morfoloogia, vaid pigem süntaksi juurde. Kuid mõne ülesande puhul, näiteks tekstist sõnastiku tegemisel, on küll tarvis morfoloogilist ühestamist, aga süntaksianalüüsi vaja ei lähe. Võttes arvesse need põhjendused, võib kokkuvõtteks väita, et morfoloogiline ühestamine on suhteliselt iseseisev loomuliku keele analüüsi etapp.

Morfoloogilise analüsaatori poolt pakutud võimalike morfoloogiliste tõlgenduste hulgast on tavaliselt konkreetses lauses või fraasis grammatiliselt korrektne ainult üks tõlgendus. Näiteks lauses *Peeter sai piletid kätte viis minutit enne etendust* on sõnavorm *sai* loomulikult verbi *saama* vorm, mitte nimisõna *sai*. Kuid mõnikord saab ka konkreetses kontekstis olla "õige" enam kui üks tõlgendus – sellisel juhul saab erinevate tõlgenduste valimisel ka semantiliselt erinevaid fraase. Näiteks lausest *Naine tuli talle järele* on võimalik aru saada kahel viisil: *naine tuli temale järele* või *naine tuli talle (tall) järele*. Sellisel juhul peab ka ühestaja jätma alles kõik antud kontekstis grammatiliselt korrektsed morfoloogilised tõlgendused.

Erinevates keeltes on mitmeti tõlgendatavate sõnade osakaal erinev: näiteks inglise keeles on mitme tõlgendusega 40% sõnadest, rootsi keeles üle 60% sõnadest, aga soome keele

sõnadest võib ainult 11 protsendil leida mitu grammatilist tõlgendust (Karlsson jt 1995). Katsed näitavad, et eesti keeles on üle 45% sõnadest ilukirjandustekstides morfoloogiliselt mitmeti tõlgendatavad.

Erinevaid lähenemisi ühestamisele

Loomuliku keele morfoloogilisele ühestamisele on käesoleval ajal kolm põhilähenemist: reeglipõhine (nt ENGCG (Karlsson jt 1995)), tõenäosuslik (nt CLAWS (Garside jt 1987)) ning segalähenemine (nt Brilli N-Best ühestaja (Brill 1994, 1995, 1997)). On ka veel selline võimalus, et ühestamise etappi kui niisugust ei eristata, vaid õiged morfoloogilised tõlgendused leitakse süntaktilise struktuuri määramise käigus (nt PEG (Jensen, Heidorn 1993)).

Tõenäosuslik ühestaja töötab üldjuhul järgmiselt: lause analüüsimisel arvutab ta iga võimaliku tõlgenduste kombinatsiooni korral selle tõenäosuse ja valib sellise kombinatsiooni, mille puhul tõenäosus on kõige suurem. Tõlgenduste kombinatsiooni tõenäosuse arvutamisel kasutab ta tavaliselt käsitsi ühestatud tekstil treenimise käigus arvutatud tõlgenduste paaride järjest esinemise tõenäosusi. Reeglipõhised ühestajad ei vaata kõikvõimalike tõlgenduste jadasid, vaid eemaldavad (või teisendavad, nagu näiteks Brilli N-Best ühestajas) mittesobivaid tõlgendusi järk-järgult, juhindudes sõna kontekstist. Seejuures see, millist infot saavad reeglid kontekstis kontrollida, sõltub juba konkreetsest formalismist.

Reeglipõhised ühestajad jätavad tüüpiliselt mõned mitmesused otsustamata ja tänu sellele teevad väga vähe vigu; seevastu tõenäosuslikud ühestajad annavad tavaliselt täiesti ühese väljundi, seda aga suurema veaprotsendi arvelt. Reeglipõhises formalismis kasutatakse nn grammatikareegleid, kus iga reegel esitab mõnda keelereeglilaadset fakti. Selliseid ühestajaid iseloomustavad samuti parem arusaadavus, ülevaatlikkus ja kohaldatavus — inimesel on võimalik ühestamise käiku

(ja järelikult ka tulemusi) otseselt mõjutada, kuna ühestamise eeskirjad on esitatud üheselt tõlgendatavate reeglite abil ja iga üksiku reegli mõju on võimalik hinnata. Tõenäosuslike ühestajate töö käiku on raske kõrvalt juhtida: ei ole võimalik täpselt prognoosida, kuidas mõjub tulemusele üksikute tõenäosuste käsitsi muutmine.

Eesti keele jaoks on praegu olemas mõlema suuna esindajate esimesed ühestajate versioonid, mis mõlemad valmisid Tartu Ülikoolis. Statistiline morfoloogiline ühestaja (Kaalep, Vaino 1998 ja 2000) põhineb Markovi varjatud mudelil (ingl *Hidden Markov Model*) (Dermatas, Kokkinakis 1995; Weischedel jt 1993). Reeglipõhine morfoloogiline ühestaja (Puolakainen 2000, 2001) põhineb kitsenduste grammatika formalismil (ingl *Constraint Grammar*) (Karlsson 1990).

Reeglipõhine eesti keele ühestaja

Kitsenduste grammatika (KG) ühestaja sisendiks on kasutatud eesti keele morfoloogilise analüsaatori ESTMORF (Kaalep 1999) väljundit. Kuid analüsaatori poolt väljastatavad märgendid ei sobinud hästi kitsenduste grammatika reeglite formuleerimiseks ja said kohati teisendatud. Morfoloogilise analüsaatori väljundis koosneb iga sõna tõlgendus sõnaliigi märgendist, lisaks käändsõna puhul arvu ja käände märgendist ning pöördsõna puhul vormitähisest, mida esindab üks formatiivivariantidest. Seejuures sõnaliigi märgend sisaldab endas ka muud informatsiooni, nt algvõrdes, keskvõrdes ja ülivõrdes omadussõna sõnaliigimärgendid on erinevad. Selgem oleks selline märgendus, kus sõltumata sellest, kas omadussõna on alg-, kesk- või ülivõrdes, on ikkagi üks ja sama märgend, mis ütleb, et tegu on adjektiiviga, ja lisaks veel märgend, mis näitab võrdlusastet. Teiseks, noomeni ja verbi grammatiline info peaksid olema tähistatud samade põhimõtete järgi. Morfoloogilise analüsaatori väljundis väljendavad noomenivormi märgendid vastavaid grammatilisi tähendusi (nt

pluural partitiiv), verbivormi märgendiks on üks tema formatiivivariantidest (nt *sid*), süntaksianalüsaator aga vajab grammatilist infot kajastavaid märgendeid. Lisaks sellele on mõnel morfoloogilise analüsaatori märgendil mitu grammatilist tähendust:

1) Märgend *sid* tähistab lihtminevikus nii *singulari 2. pööret* kui *pluurali 3. pööret*:
   luge+sid - *//_V_ sid //* – indikatiiv imperfekt singular 2. pööre,
   luge+sid - *//_V_ sid //* – indikatiiv imperfekt pluural 3. pööre.
2) Märgend *n* tähistab nii *käändsõna* (nimi-, omadus-, asja arvsõna) *nominatiivi* kui ka *verbi indikatiivi preesensi singulari 1. pöörde jaatava kõne* vormi.
3) Märgend *_V_ o* (nt sõnavormil *loe*) võib grammatiliselt tähendada *verb imperatiiv preesens singular 2. pööre jaatav kõne*, kuid see võib olla ka eitusvormi osa (*ära loe* ja *ei loe*), kus grammatilisteks tõlgendusteks on vastavalt *verb imperatiiv preesens personaal (eituse osa)* ja *verb indikatiiv preesens personaal (eituse osa)*. Seejuures viimases grammatilises tähenduses muutub kõneviis: imperatiivi asemel on seal indikatiiv ja see peaks olema kajastatud kasutatavates märgendites.

Ühestaja märgendisüsteemis on oluliselt teisendatud verbi märgendeid, lisatud on märgendeid asesõnade, kaassõnade, sidesõnade ja kirjavahemärkide tõlgendustele, samuti on lisatud verbi rektsiooni näitavad märgendid. Teisendused toovad mõnes kohas mitmesust sisse – mõni ühene sõnavorm saab mitmeseks, näiteks tõlgendus *luge+sid //_V_ sid //* teiseneb kaheks tõlgenduseks:

luge+sid //_V_ main indic impf ps2 sg ps af #FinV #NGP-P //
luge+sid //_V_ main indic impf ps3 pl ps af #FinV #NGP-P //

Kitsenduste grammatika ühestaja koosneb keelest sõltumatust programmist ja eesti keele arvutigrammatikast. See grammatika on käsitsi koostatud, tuginedes ligi 50000-sõnalisele

eelnevalt käsitsi morfoloogiliselt ühestatud tekstikorpusele ja sellelt saadud statistilistele andmetele. Eesti keele morfoloogilise ühestamise grammatika sisaldab 40 osalause piiride määramise reeglit ja üle 1200 ühestamise kitsenduse.

Kitsendustes on võimalik püstitada tingimusi ümbritsevate sõnavormide algvormide, morfoloogiliste tõlgenduste ning sõnajärje kohta. Tingimusi saab esitada terve lause ulatuses, kuid vajadusel saab piirduda ka osalausega, mis on võimalik tänu osalause piiride määramise reeglitele. Osalausepiiride määramine toimub sidesõnade, kirjavahemärkide ja verbide põhjal ning sisuline reegel on järgmine: kui sõnale eelneb vastav kirjavahemärk ja/või sõna ise on sidesõna ning vasakul ja paremal pool seda sõna leidub verbi pöördeline vorm, siis see sõna on osalause esimene sõna. Kuna osalausete määramise järgus pole kõigil sõnavormidel ühest tõlgendust, sh ka sidesõnadel ja verbidel, siis eristatakse kindlaid ja oletatavaid osalausepiire.

KG ühestaja tööpõhimõte on järgmine: korraga vaadatakse ühte lauset. Esiteks püütakse määrata osalausete piirid. Järgmise sammuna rakendatakse lause igale mitmesele sõnavormile ühestamise reegleid, vaadeldes sõnu järjest lause algusest kuni lõpuni. Reeglid püüavad eemaldada konteksti mittesobivad tõlgendused, kuid viimast tõlgendust ei eemaldata kunagi. Et tänu sellisele esimesele ühestamise ringile on mõned mitmesed sõnavormid saanud ühese tõlgenduse, siis on lootust, et uuesti lause algusest alustades on võimalik veel mõningaid reegleid rakendada: paljud reeglid nõuavad, et kontrollitav sõna oleks ühene – ainult siis võib teise sõna kohta järeldusi teha. Sellepärast korrataksegi osalausete piiride määramist ja ühestamist veel kaks korda.

Ühestamise kitsenduste näiteid:

- valida sõnavormi *tuli* puhul verbi tõlgendus õigeks, kui osalauses eespool leidub ühene nimisõna nimetavas käändes ning osalauses ei leidu rohkem sõnu verbi tõlgendusega (selline reegel rakendub näiteks lauses *Hääl tuli tuhmi peegli moodi metallplaadist*);

- kustutada omastavat käänet nõudva postpositsiooni tõlgendus, kui eelmisel sõnal pole nimi- või asesõna omastava käände tõlgendust (*Kas ta ei tule tuttav ette?*);
- kustutada omadussõna ablatiivi tõlgendus, kui sellest sõnast osalause lõpuni ei leidu nimi- või asesõna, millel vähemalt üheks tõlgenduseks on ablatiivi tõlgendus (selline reegel aitab tihti määrsõna ja omadussõna valikul, näiteks *eriliselt – eriline, tähelepanelikult – tähelepanelik, hooletult – hooletu*);
- valida õigeks nimisõna nimetava käände tõlgendus, kui osalauses pole nimi-, ase- ega arvsõnu nimetavas, lauses paremal leidub predikatiivi lubav verb (nagu *olema, näima, saama*) ja ei leidu muid finiitseid verbe, lauses paremal leidub omadussõna ainsuse nimetavas ja omakorda sellest paremal ei ole nimisõnu nimetavas (*Mu ajataju pole eriti tugev*).

Ühestaja ei anna ideaalseid tulemusi: ühelt poolt mõnikord ei õnnestu saada sõna üheseks ning teiselt poolt eemaldatakse mõnikord just õige tõlgendus ja alles jääb vale. Mineviku partitsiibid on üks selline riskirühm, mis on morfoloogilisel ühestamisel põhiliseks raskuseks. Partitsiipidele antakse neli tõlgendust: verbi vorm, omadussõna ainsuse ja mitmuse nimetavas ja ka nimisõna mitmuse nimetavas ning nende vahel valiku tegemiseks pole senini head lahendust leitud. Kasutatakse vaid kõige lihtsamaid reegleid, mis eemaldavad verbi tõlgenduse juhul, kui liitaja või eituse moodustamine ei ole võimalik, või eemaldavad adjektiivi tõlgenduse juhul, kui see sõna ei saa olla öeldistäiteks (lauses puudub verb *olema*) ning ei saa olla ka atribuudiks (lauses ei leidu nimisõna, mida laiendada). Kuid tihti on lauses formaalselt võimalik moodustada nii liitaega kui laiendada nimisõna (tänu sellele, et mineviku partitsiibid ei käändu, võivad nad laiendada suvalises käändes nimisõna). Näiteks järgmised laused on struktuuri poolest väga sarnased, kuid valida tuleb vastupidised tõlgendused:

Ma olen väsinud (omadussõna). – *Ma olen väsinud* (verb) *inimestega vaidlemisest.*

Ma olen saabunud (verb). – *Ma olen saabunud* (omadussõna) *inimeste nimekirjas.*

Teiseks suuremaks riskirühmaks on esimese kolme käände ühestamine: nimetava ja omastava käände tõlgendused jäävad tihti mõlemad alles. Formaalselt pole selge, miks näiteks ei või järgmises lauses *tema* olla subjekt: *Tema põiklevatest vastustest võis fantaasiat appi võttes välja lugeda, et abahakiiridel kehtis midagi ürgkondliku kommuuni või patriarhaadi taolist.*

Raske on otsustada, et esimeses lauses on *raha* osastavas käändes, kui teises analoogilises lauses on *poiss* nimetavas:

Katsugu raha (osastav kääne) *paremini hoida.*

Katsugu poiss (nimetav kääne) *paremini õppida.*

Järgmises lauses on sõnavormil *süüd* huvitaval kombel kolm tõlgendust – *süü* mitmuse nimetavas ning ainsuse osastavas ja *süüd* ainsuse nimetavas: *Süüd oli siin muidugi ka naisel.* Ühestaja eelistab sellises lauses aluse positsioonis ainsuse nimetava käände tõlgendust ning seega eksib.

Vead võivad tekkida ka liitlausetes, kus üks osalause asub teise sees, ning need on tingitud kitsenduste grammatika formalismi eripärast. Näiteks kuna lauses *Nagu oleks enamik sellest, mida päevast päeva tarvis, liikunud läbi elukaaslase kõhetu meele, läbi jagamiste ja jagelemiste, mille eest Pärtel enese arvates kogu aeg põgenenud oli, aga mitte nii varmasti, et naine teda peaaegu iga hetk tabanud poleks* ei osata kokku panna osalause kaks osa, siis ka öeldise koostisosad *oleks* ja *liikunud* kaotavad oma õiged tõlgendused.

Ühestaja töö tulemusena saavad ühese tõlgenduse 85–90% sõnadest ja vigaselt analüüsitakse kuni 2% sõnadest. Eesti keele reeglipõhine morfoloogiline ühestaja on arendatud sellisel viisil, et teda on võimalik kasutada automaatses süntaksianalüüsis, mis on välja töötatud TÜ-s (Müürisep 2000) ning baseerub samuti kitsenduste grammatikal. Ühestaja arendamisel oli algselt kasutatud Tartu Ülikoolis väljatöötatud (Kaalep 1999) morfoloogilist analüsaatorit. Käesoleval ajal ühestajat täiendatakse Eesti Keele Instituudis eesmärgiga võimaldada ühilduvust arendatava reeglipõhise morfoloogilise analüsaatoriga.

Statistiline eesti keele ühestaja

Markovi varjatud mudelil (MVM) põhineva ühestaja tööpõhimõte on järgmine: igale morfoloogilisele tõlgendusele vastab üks märgend, erinevad tõlgendused võivad olla kokku võetud üheks märgendiks. Treenimise etapil, kasutades suuremahulist tekstikorpust, arvutatakse iga märgendi (või märgendite) kohta tõenäosusi: teatud märgend alustab lauset; teatud märgendile lauses järgneb teatud märgend; kui sõnavormil on teatud märgendite komplekt, siis milline märgend tuleb nende hulgast valida. Ühestamise etapil leitakse, kasutades arvutatud tõenäosuste tabeleid, lause jaoks antud märgendite võimalike jadade tõenäosused ning valitakse maksimaalse tõenäosusega märgendite jada, seega valitakse ka märgenditele vastavad morfoloogilised tõlgendused.

Statistilise ühestaja kohta avaldatud testimise tulemuste kohaselt umbes 3% sõnadest saab ühestamise tagajärjel vale analüüsi (Kaalep, Vaino 2000). Tekstis allesjääva mitmesuse määra ei saa otseselt võrrelda, sest ühestajate kasutatavad märgendite süsteemid erinevad oma detailsuse astmelt. Nimelt on statistilise ühestaja märgendite süsteem üldisem kui reeglipõhisel ühestajal. See on täiesti piisav mõne konkreetse ülesande jaoks, kus morfoloogiline ühestamine pole eesmärk omaette, vaid mõne muu ülesande hädavajalik koostisosa, nt sagedussõnastiku koostamine või masintõlge. Nii näiteks ei püüagi statistiline ühestaja eristada mõnesid väga raskesti automaatselt (aga mõnes olukorras ka haritud lingvisti poolt) määratavaid kategooriaid, mis kokku iseloomustavad 13,5% sisendsõnadest (Kaalep, Vaino 2000). Need on: *nud-*, *tud-* partitsiipide puhul tegusõna või omadussõna valik; sõna *ta*, *tema*

nimetava või omastava käände valik; verbivormi *on* ning asesõnade *kes* ja *mis* puhul ainsuse või mitmuse valik; sõnade *kui* ja *nagu* puhul määrsõna või sidesõna valik; sõnade *üks* ja *teine* puhul arvsõna või asesõna valik.

Väike võrdlus

Reeglipõhise ühestaja märgendite süsteem on detailsem, sest ta on orienteeritud süntaksianalüüsile, ning eelnimetatud mitmesusi püütakse hoolimata otsustamise raskusest siiski määrata ja paratamatult nendes kohtades mõnikord ka eksitakse. Suurem detailsus, mida küll alati ei suudeta vigadeta garanteerida, on vajalik süntaksianalüsaatoris, mille koosseisu reeglipõhine ühestaja kuulub, ja muidugi ka kõikides tema rakendustes. Samas saab detailse analüüsi märgendust alati üldisemaks muuta ning sel teel ühestaja numbrilisi näitajaid parandada. Statistilise ühestaja eeliseks on oluliselt väiksem pingutus selle töölepanekuks, võrreldes reeglite käsitsi formuleerimisega. Käsitsi kirjutatud reeglitega ühestaja eeliseks on aga asjaolu, et süstemaatiliste vigade leidmise korral on võimalik need kohe parandada.

Kirjeldatud eesti keele morfoloogiliste ühestajate võrdluseks viidi läbi väike test. Valiti üks kahetuhandesõnaline ilukirjandustekst, mida kasutati ühtlasi MVM ühestaja treenimisel ja KG ühestaja testimisel. Testi tulemused on toodud tabelis 1. Võrdlemist segab veidi asjaolu, et ühestajad kasutavad küll sama morfoloogilist analüsaatorit ja samades märgendites on ka MVM väljund, kuid KG ühestaja märgendid on mõnevõrra detailsemad. See asjaolu kajastub kahe ühestaja algseisu veergudes (2 ja 4): sama tekst MVM ühestaja puhul sisaldab 34,5% mitme tõlgendusega sõnu ning 48,8% KG ühestaja puhul. Ühestamise järel jäi teksti MVM ühestaja puhul 13,4% mitme tõlgendusega sõnu ning KG ühestaja puhul – 7,8%, seejuures õige tõlgenduse kaotas vastavalt 3% ja 1,4% sõnadest.

	MVM algseis	MVM lõppseis	KG algseis	KG lõppseis
Mitmesuse protsent	34,5	13,4	48,8	7,8
Tõlgendusi sõna kohta	1,62	1,26	2,17	1,10
Vigade protsent		3,0		1,4

Tabel 1. Statistilise ja reeglipõhise morfoloogilise ühestaja võrdluse testi tulemused.

MVM ühestaja tegi 57 viga, nendest 8 olid KG ühestajaga ühised vead. Põhilised veagrupid olid: valik nimetava, omastava, osastava või lühikese sisseütleva käände vahel (30 viga), nimisõna või määrsõna või kaassõna valik (17 viga), nimisõna või verbi valik (10 viga). Näiteks lauses *Pärast seda kui naine suri, maha maetud sai ja esimene hämmeldus uuest olukorrast vaibuma hakkas, asus Pärtli hinge kõle tühjus* on koguni 3 viga: sõnavormil *sai* nimisõna nimetavas käändes verbi *saama* vormi asemel; sõnavormil *hämmeldus* verbi *hämmelduma* vormi tõlgendus nimetavas käändes nimisõna asemel ning sõnavormil *hinge* nimisõna omastav kääne lühikese sisseütleva käände asemel.

KG ühestaja tegi 37 viga, nendest 17 kuulub selliste juhtumite hulka, mida statistiline ühestaja ei püüagi lahendada. Põhilised veagrupid olid: partitsiibid (verb või omadussõna) ja "olema" (vastavalt kas abi- või põhiverb) (14 viga), valik nimetava, omastava, osastava või lühikese sisseütleva käände vahel (12 viga), nimisõna või määrsõna või kaassõna valik (5 viga), nimisõna või verbi valik (4 viga). Näiteks lauses *Pärtel oleks võinud otsida ka sõprade seltsi* otsustas ühestaja sõnavormi *seltsi* puhul kaassõna tõlgenduse kasuks, aga õige on nimisõna osastava käände tõlgendus. Ühise vea näiteks võib tuua lause: *Oli ju iga toiduaine vaevaga toodetud, looduselt tänuga vastu võetud.* Siin valisid mõlemad ühestajad sõnavormide *iga* ja *toiduaine* puhul omastava käände, kuigi tegelikult on õige nimetav kääne.

Kokkuvõttes võib öelda, et praegusel hetkel on mõlemad ühestajad eestikeelsetele tekstidele rakendatavad. Iga konkreetse ülesande puhul tuleb aga eraldi kaaluda, kumma ühestaja kasutamist eelistada, sõltuvalt nõutavast märgendite detailsusest ja ka vajalike muudatuste tegemise raskusest rahuldava lahendi saamisel. Huvitavaid tulemusi võiks anda ka nende ühestajate kombineerimine. Näiteks, kui eesmärgiks on täielikult ühene väljund, siis on mõttekas rakendada alguses reeglipõhist ühestajat, seejärel kasutada alles jäänud mitmesuste lahendamiseks statistilist ühestajat või vastupidi. Või kui eesmärgiks on teha nii vähe vigu kui võimalik (seega kompromissina lubada väljundisse ka rohkem mitme tõlgendusega sõnu), siis lasta mõlemal ühestajal pakkuda oma variandid ning juhul, kui nad ei lange kokku, võtta arvesse mõlemad.

Viited

Brill, E. 1994. Some Advances in Transformation-Based Part of Speech Tagging. – Proceedings of AAAI94, http://www.cs.jhu.edu/~brill/acadpubs.html.

Brill, E. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. – Computational Linguistics, 21 (4), lk 543–565.

Brill, E. 1997. Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging. Ilmumas kogumikus "Natural Language Processing Using Very Large Corpora". Kluwer Academic Press.

Dermatas, E., Kokkinakis, G. 1995. Automatic Stochastic Tagging of Natural Language Texts. – Computational Linguistics, 21 (2), lk 137–163.

Garside R., Leech G., Sampson G. 1987. The Computational Analysis of English: a corpus-based approach. London, New York: Longman.

Jensen, K. and Heidorn, G. 1993. Natural Langauage Processing: The PLNLP Approach. Boston, Dordrecht, London: Kluwer Academic Press.

Kaalep H.-J, Vaino T. 1998. Kas vale meetodiga õiged tulemused? Statistikale tuginev eesti keele morfoloogiline ühestamine. – Keel ja Kirjandus, nr 1, lk 30–38.

Kaalep, H.-J. 1999. Eesti keele ressursside loomine ja kasutamine keeletehnoloogilises arendustöös. Dissertationes Philologiae Estonicae Universitatis Tartuensis 7. Tartu.

Kaalep H.-J, Vaino T. 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Arvutuslingvistikalt inimesele. Toim T. Hennoste. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu. Lk 87–99.

Karlsson F. 1990. Constraint Grammar as a Framework for Parsing Running Text. – The 13th International Conference on Computational Linguistics, Vol. 3, Helsinki. Lk 168–173.

Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A. 1995. Constraint Grammar: a Language Independent System for Parsing Unrestricted Text. Berlin and New York: Mouton de Gruyter.

Müürisep, K. 2000. Eesti keele arvutigrammatika: süntaks. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu.

Puolakainen T. 2000. Eesti keele reeglipõhise morfoloogilise ühestamise probleemseid kohti. – Arvutuslingvistikalt inimesele. Toim. T. Hennoste. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu. Lk 73–86.

Puolakainen, T. 2001. Eesti keele arvutigrammatika: morfoloogiline ühestamine. Dissertationes Mathematicae Universitatis Tartuensis 27. Tartu.

Weischedel, R., Meteer, M., Schwartz, R., Ramshow, L., Palmucci, J. 1993. Coping with Ambiguity and Unknown Words through Probabilistic Models. – Computational Linguistics, 19 (2), lk 359–382.