SÕNALIIGITUSE KITSASKOHAD EESTI KEELE ARVUTIANALÜÜSIS

Kadri Muischnek, Kadri Vider

Tartu Ülikool, arvutilingvistika uurimisrühm

Ülevaade. Artiklis vaadeldakse tekstikorpuste morfoloogilise ja semantilise ühestamise käigus kerkinud lingvistilisi probleeme, keskendudes tekstisõna sõnaliigilise kuuluvuse probleemile.

Võtmesõnad: tekstikorpused, sõnaliigid, morfoloogiline ühestamine, sõnatähenduste ühestamine.

1. Sissejuhatus

Tartu Ülikooli arvutuslingvistika uurimisrühmas tehti riikliku sihtprogrammi „Eesti keel ja rahvuskultuur“ toel valmis oluline keeleressurss – ühestati morfoloogiliselt 500 000 tekstisõna mahus eestikeelseid tekste, täpsemalt:

- 100 000 sõna ajalehetekste tänapäeva eesti keele korpusest;
- 100 000 sõna ilukirjandustekste tänapäeva eesti keele korpusest;
- 100 000 sõna seadustekste;
- 100 000 sõna populaarteaduslikke tekste ajakirjast Horisont;
- 100 000 sõna suulist kõnet TÜ suulise kõne korpusest.

Kirjalikud tekstid on ühestatud kujul saadaval koduleheküljel www.cl.ut.ee ja loodetavasti on selle artikli ilmumise ajaks valmis saanud ka morfoloogiliselt ühestatud tekstide kasutajaliides, mis võimaldab esitada päringuid sõna algvormi ja morfoloogilise kategooria ning nende kombinatsioonide alusel. Suulise kõne ühestatud tekstidega on asi keerulisem. Vastavalt isikuandmete kaitse seadusele ei saa neid sellisel kujul interneti kaudu kättesaadavaks teha ja neid saab kasutamiseks TÜ suulise kõne uurimisrühmast vastava lepingu sõlmimisel (vt täpsemalt http://sys130.psych.ut.ee/~linds/).

Varem morfoloogiliselt ühestatud tekstidest on mahukaim George Orwelli romaani „1984” terviktekst, seda tööd kirjeldab Keeles ja Kirjanduses ilmunud artikkel „Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti kirjakeele testkorpuse morfoloogilise märgendamise kogemusest“ (Kaalep jt 2000).

2. Mis on morfoloogiline ühestamine?

Paljudel sõnavormidel on eraldiseisvaina, ilma kontekstita, mitu võimalikku morfoloogilist analüüsi. Näiteks sõnavorm *poodi* võib olla verbi *pooma* lihtmineviku impersonaali vorm ja noomeni *pood* ainsuse osastav või lühike sisseütlev. Kuid lauses *Ema*

läks poodi piima tooma on võimalik ainult üks neist kolmest tõlgendusvõimalusest. Teksti automaatsel analüüsil lisatakse tavaliselt igale sõnavormile kõik tema võimalikud morfoloogilised tõlgendused ja siis valitakse nende hulgast antud konteksti sobiv. Meie kasutasime kõigi võimalike analüüside saamiseks OÜ Filosofti morfoloogiaanalüsaatorit (selle kohta täpsemalt vt Kaalep, Vaino 2000). Õige analüüsi väljavalimiseks ehk morfoloogiliseks ühestamiseks kasutatakse morfoloogilisi ühestajaid, eesti keele jaoks on olemas nii statistikal põhinev (vt nt Kaalep, Vaino 1998) kui ka reeglipõhine (vt nt Puolakainen 2001) morfoloogiline ühestaja. Kuid meie projektis tegi ühestaja tööd inimene, veelgi enam – ühte teksti ühestas paralleelselt kaks inimest, hiljem võrreldi tulemust ja erinevused lahendati ühiste arutelude käigus. Sellega lootsime tagada tulemuse sajaprotsendilise täpsuse ja korrektsuse.

2.1. Kasutatud morfoloogiliste kategooriate süsteemist

Morfoloogiline analüsaator eristab traditsioonilist üheksat sõnaliiki: substantiiv, adjektiiv, pronoomen ja numeraal kui käändsõnad, verb kui pöördsõna ja adverb, pre- ja postpositsioon, konjuktsioon ning interjektsioon kui muutumatud sõnad. Eesti keele teaduslikus grammatikas (EKG I: 18) on eristatud 12 sõnaliiki; adverbi alt on eraldi sõnaliikidena välja toodud proadverbid, afiksaaladverbid ja modaaladverbid. Neid morfoloogiline analüsaator praegu ei tunnista, kuid nii statistilise ühestaja kui ka teksti edasise analüüsi huvides tuleks suur ja heterogeenne adverbide klass väiksemateks ja homogeensemateks rühmadeks tükeldada küll.

Verbid on jagatud põhi- (_V_ main), abi- (*olema* liitaegades, *ei* verbi eitavates vormides, *ära* käskiva kõneviisi eitavates vormides (_V_ aux)) ja modaalverbideks (_V_ mod); nimisõnad jagunevad üld- (_S_ com) ja pärisnimedeks (_S_ prop), arvsõnad põhi- (_N_ card) ja järgarvsõnadeks (_N_ ord), kaassõnad ees- (_K_ pre) ja tagasõnadeks (_K_ post). Asesõnu (_P_) liigitati varasemal morfoloogilisel ühestamisel EKG I põhjal kaheksasse alaliiki (EKG I: 26–31). Juba „1984” ühestamise kogemuse põhjal võis öelda: „See [asesõnade detailsed semantilis-funktsionaalsed rühmad] muutis otsustamise lausekonteksti arvestades sageli raskeks, sest mida täpsem on semantiline jaotus, seda enam kaalu on sisulistel otsustustel, mis võib suurendada määratluste subjektiivsust.“ (Kaalep jt 2000: 629). Nii otsustasime asesõnade alaliikidesse jaotamisest loobuda. Kasutatud morfoloogiliste märgendite ja kategooriate täiskomplektiga saab tutvuda arvutuslingvistika uurimisrühma kodulehekülje vahendusel.

3. Milleks meile väga korralikult morfoloogiliselt märgendatud korpus?

Täielikult vigadeta märgendatud korpust läheb vaja näiteks:

- süntaktilise analüüsi sisendiks;
- semantilise analüüsi sisendiks;
- automaatsete morfoloogiliste ühestajate arendamiseks;

- sagedusloendite ja sagedussõnastike tegemiseks;
- lingvistiliseks uurimistööks.

Nende rakenduste huvid võivad olla mõnevõrra erinevad. Nii on statistilise ühestaja treenimiseks vaja sellist korpust, kus igale sõnavormile on jäetud vaid üks analüüs. Ent lingvistika seisukohalt pakuks rohkem selline lähenemine, kus kahe sõnaliigi hajusal piirialal asuvale sõnavormile on jäetud mõlemad võimalikud analüüsid. Selle korpuse märgendamisel oleme siiski seadnud esikohale just statistilise ühestaja huvid, sest kui õnnestub selle korpuse abil tema tööd piisavalt heaks muuta, siis saab veelgi suuremat hulka tekste automaatselt piisavalt kvaliteetselt ühestada.

4. Valik kerkinud lingvistilistest probleemidest

Nagu osas 2 juba näidati, tuleb morfoloogilise ühestamise käigus määrata konkreetse sõnavormi sõnaliigiline kuuluvus; edasi käändsõnal arv ja kääne ning pöördsõnal kõigepealt finiitsus/infiniitsus; infiniitsetel vormidel tegumood ja kääne, finiitsetel vormidel arv, isik, tegumood, aeg, kõneviis ja kõnelaad (jaatav/eitav). Suur osa neist kategooriatest on meie õnneks vormi põhjal üheselt määratavad. Ka suur osa mitteühesustest on inimesele lihtsalt lahendatavad (nt 2. osas toodud mitteühene näide *poodi* on konteksti põhjal hõlpsasti määratav). Edaspidi käsitleme aga mõningaid selliseid mitteühesuse tüüpe, mis käsitsiühestajagi (s.o inimese) mõtlema panid ja millest on ilmselt oodata probleeme ka automaatsel ühestamisel.

Põhiline probleem, mis järjekindlalt morfoloogilise ühestamise käigus teadvustus, on teoreetilise keeleteaduse arusaamine sõnaliikidest kui hajusate piiridega hulkadest vastandatuna ühestamisel kehtivale nõudele jätta igale sõnavormile ainult üks analüüs. Järelikult tuleb ühestatud tekstide kasutamisel arvestada asjaoluga, et hajusad piirialad on jagatud nn kirvemeetodil.

Sõnade liigitamisel lähtusime üldtunnustatud põhimõtetest, mis Fred Karlssoni sõnastuses kõlavad järgmiselt: „Sõnaliikide määramise kõige selgemad kriteeriumid põhinevad vaatluse all oleva lekseemi struktuuriomadustel, milleks on sõna morfoloogilised omadused, distributsioon, võimalikud süntaktilised funktsioonid ja rektsioon. Neid võib täiendada sõna semantiliste omadustega.“ (Karlsson 2002: 218). Kuna varem oli morfoloogilise käsitsiühestamise eesmärgiks luua test- ja treeningkorpus automaatsete morfoloogilise ja süntaktilise analüsaatori jaoks, siis sõna semantiliste omadustega pole ühestamisel ega ka morfoloogiliste märgendite süsteemi väljatöötamisel eriti arvestatud. See on nüüd, mil morfoloogiliselt ühestatud tekste kasutatakse ka semantilise ühestamise sisendiks, tekitanud mitmeid probleeme, mida selles kirjutises edaspidi ka käsitletakse.

Probleemidest morfoloogilisel ühestamisel (sõnaliikide määramisel) on juttu ka „1984” morfoloogilist ühestamist käsitlevas artiklis (Kaalep jt 2000) ja peame tõdema, et probleemid on põhijoones ikka samaks jäänud:

- käändsõnade ja verbide vormidest on arenemas kaassõnu ja adverbe;

- partitsiibid paiknevad verbi ja adjektiivi piirimail ja vahel ei piisa ka lausekontekstist otsustamaks, millisesse sõnaliiki mingi vorm kuulub.

Lisaks sõnade sõnaliigilise kuuluvuse määramisele tekitas vahel harva probleeme ka nimisõna käände kindlakstegemine. Kas näiteks lauses (1) on *tähelpanu* nimetavas või osastavas käändes? Ja kas sama sõna on omastavas või osastavas käändes lauses (2)?

(1) Meie tähelepanu juhiti asjaolule, et …

(2) … juhib keegi tähelepanu sellele, et …

Või on oluline ainult see, et lugeja saaks aru, et nimisõna *tähelepanu* on siin formaalselt sihitise rollis ja püsiühendi *juhib tähelepanu millelegi* osa ja kääne polegi oluline?

Kui aga rääkida sõnaliikidest, siis on tegeliku teksti morfoloogilisel ühestamisel üheks raskemaks probleemiks kaassõnade ja määrsõnade pidev tekkimine nimisõnade ja verbide teatud vormidest. Lingvisti seisukohast on väga huvitav jälgida grammatikaliseerumisprotsessi avaldumist tekstides, aga morfoloogilisele ühestajale on see nuhtluseks.

Verbist kujuneva kaassõna näiteks võib tuua verbi *algama* vormi *alates*, mis on kirjalikes tekstides küllaltki sage (selle ja muudegi *des*- ja *mata*-vormide kaassõnastumise kohta vt nt Uuspõld 2001). Meie kasutatava morfoloogilise analüsaatori aluseks olnud Ülle Viksi „Väikeses vormisõnastikus" (Viks 1992) pole vorm *alates* eraldi märksõna staatust saanud, EKSS-is (EKSS I: 78–79) on ta esitatud alamärksõnana verbi *algama* juures, kuid tema sõnaliigilist kuuluvust pole seal märgitud. Morfoloogiaanalüsaatori järjekordse täiendamise käigus lisati see vorm kaassõnade hulka, nii et lausetes (3), (4) on ta meie märgendatud korpuses kaassõnaks analüüsitud.

(3) alates uuest nädalast võib ta oma jõuvarud töösse suunata

(4) Uuest arginädalast alates võivad Kaljukitse suhted veelgi tiheneda

Aga mida teha verbi *lõpetama* des-vormiga *lõpetades* lauses (5)?

(5) alates Tallinna Masinatehase direktorist ning lõpetades kohaliku tööstuse ministriga …

Kaassõnalaadsena kasutatakse vormi *lõpetades* ainult selliselt, paaris vormiga *alates* ja sellise ühendi kasutus on piiratud mingi loendi või skaala kahe otsa märkimisega, nt (6).

(6) Varustatus on küllaltki rikkalik, alates roolivõimendist ja lõpetades istmete soojendusega.

Ajalise või ruumilise vahemiku piiritlemiseks see konstruktsioon ei sobi, selle asemel kasutatakse ühendit *alates … kuni* (mida kasutatakse ka samas funktsioonis nagu ühendit *alates … lõpetades*), nt

(7) …toimingud, mis on tehtud alates vara õigusvastasest võõrandamisest kuni seaduse kehtestamiseni;

(8) … alates teisest korrusest kuni pööninguni saab ruumid kinnistusamet.

(9) …anda võimalused eri mudelite rakendamiseks, alates sajaprotsendilisest eesti õppekeelest kõikides ainetes kuni kakskeelse mudelini.

On see küllaldane alus vormile *lõpetades* alati võimaliku kaassõna analüüsi lisamiseks? Arvestada tuleb, et liiased märgendid teevad automaatse morfoloogilise analüüsi

keerukamaks. Praegu on meie ühestatud tekstides *lõpetades* analüüsitud alati verbi vormina, kuigi paariskonstruktsiooni *alates … lõpetades* märgendatakse selle tulemusena ebajärjekindlalt.

Huvitavate probleemidega puutusime kokku ka postpositsioonide *pool*, *poole* ja *poolt* puhul. Valdavalt kasutatakse neid genitiivi nõudvate postpositsioonidena, nt (10)–(12).

(10) …kaitseala valitsemine on antud kümmekond kilomeetrit põhja pool paikneva Matsalu looduskaitseala pädevusse.

(11) …kes pöörduvad arsti poole ning räägivad tervisehäirest …

(12) Murdi ametist vabastamise poolt hääletas 26 volikogu liiget

Lisaks esinevad nad mitmetes püsiühendites, milles saab kirjeldada vorme *pool*, *poole* ja *poolt* kui vastavalt adessiivi, allatiivi või ablatiivi nõudvaid postpositsioone (nagu on tehtud EKSS-is *poole* ja *poolt* puhul (EKSS IV: 450 ja 457–458). Alternatiivne võimalus on kirjeldada neid ühendeid sellistena, kus nimisõnavorm on haploloogia tulemusena lühenenud, näiteks ühend *mõnel poolel* on lühenenud väljendiks *mõnel pool*. (Vrd EKSS näide *Madalpistetikand jääb riide mõlemal poolel ühesugune* (EKSS IV: 448)). Meie oleme ühestades neid kolme sõnavormi sellistes konstruktsioonides märgendanud kaassõnadeks, nt (13)–(23).

(13) ühel pool on Himaalaja hiidahelik koos 30 kilomeetri kaugusel valendava Everesti tipuga.

(14) nagu igal pool mujalgi

(15) Mitmel pool on arvutamise probleem juba lahendatud

(16) ühel ja teisel pool on abivajajatele ka putru ja teed jagatud

(17) Sind jätkub igale poole

(18) vaatasin sammasportikuse all ühele ja teisele poole.

(19) ühelt poolt häirib rohke elavhõbedasisaldus ainevahetust

(20) Teiselt poolt tähendab see ühiseid õppekavasid

(21) mida remondimeestel polnud õnnestunud igalt poolt välja tõrjuda

(22) … ent nende loomus uhkas kahelt poolt talle peale

(23) … vastab sellele omalt poolt Türgi

Kolmanda ja kõige problemaatilisema grupi moodustavad sellised näited, kus eelmise rühma konstruktsioon (s.o vorm *pool*, *poole* ja *poolt* koos talle eelneva sõnaga adessiivis, allatiivis või ablatiivis) käitub tervikuna prepositsioonina, nõudes temaga liituvalt sõnalt partitiivi, nt (24)–(28).

(24) kes siin minu vastas teisel pool lauda istus

(25) kitsastel kõnniteedel kahel pool tänavat liikus hämmastavalt palju rahvast

(26) Nad oleksid nagu olnud kinnitatud teine teisele poole hobust

(27) Hommiku poole ööd helises telefon

(28) … kes teda teiselt poolt kassaautomaati hädalisena jälgis

Näidetes domineerivad ühendid *teisel pool, teisele poole ja teiselt poolt*, kuid on ka teisi esikomponente, ühend on vähemalt osaliselt produktiivne. Lingvistikateoreetilisest vaatepunktist võime öelda küll, et tegu on adpositsioonifraasi perifeeriasse kuuluva nähtusega ja erinevus liitsõnalistest adpositsioonidest *allpool*, *pealpool*, *siinpool* jne seisneb peamiselt ortograafias, aga jooksva teksti ühestamisel tuleb lisada analüüs igale sõnavormile. Lisaks peab morfoloogiliselt ühestatud tekst sobima süntaktilise analüüsi sisendiks. Kuidas aga defineerida nende sõnavormide omavahelisi süntaktilisi suhteid? Kas kaassõnal on täiend või seob üks kaassõna endaga kahte noomenit, nõudes neilt erinevaid käändeid? Praktilise lahendusena kaetakse produktiivsemad juhud (*teisel pool midagi* jne) ilmselt mitmesõnaliste leksikaalsete üksuste leksikoniga, aga haruldasemate juhtude (nt *hommiku poole ööd*) puhul on viga garanteeritud.

5. Sõnaliigi probleemist sõnatähenduste ühestamisel

Eelnevast võis näha, et sõnaliigitus on tõsine morfosüntaktiline probleem eesti keele arvutianalüüsis. Morfoloogiast (ja süntaksist) alguse saanud kompromisslahendused tekitavad omakorda probleeme järgmises analüüsietapis – semantikas.

Sõnatähendust nagu sõnaliikigi pole olemas *ad hoc*, vaid see on määratletav kontekstis.

Formaalsete (morfoloogiliste ja süntaktiliste) määratluste jäikuse tõttu on kõrvale heidetud liigitusvõimalused, mis selguvad alles semantilise analüüsi etapis.

Mõnede sõnavormide substantiivikasutuse ja adverbikasutuse piiri on samuti raske määrata, sellised on näiteks somaatilised sõnad *pea*, *jalg*, *selg*, *käsi*, mille vorme kolmes kohakäändes on liigitatud erinevates allikates nii substantiiviks kui adverbiks.

Näiteks vormid *kätte : käes : käest* esinevad nii „Eesti kirjakeele seletussõnaraamatus”, „Väikeses vormisõnastikus” kui „ÕS 1999-s” omaette märksõnadena, kahes esimeses on märksõna sõnaliigiline kuuluvus adverbina samuti ära toodud.

Kuna sõnaliik peab arvutianalüüsis olema täpselt määratletav ja sõnaliigilise kuuluvuse piiri on eelnimetatud juhtudel formaalselt raske määrata, on morfoloogilisel ühestamisel otsustatud adverbivõimalus pigem välistada.

Milles avaldub probleem sõnatähenduste ühestamise (STÜ) seisukohalt? Sõnastikus, mida STÜ jaoks kasutame, on praegu kasutamiseks märksõnad (sõnatähendused), mille sõnaliik on substantiiv (S) või verb (V), ehk täistähenduslikud, avatud klassi sõnaliigid. Adverb (D) on samuti avatud klassi sõnaliik, kuid selle sõnaliigi oluline eristav tunnus – muutumatus – komplitseerib sõnaliigi probleemi ülalmainitud morfoloogilises analüüsis.

Vaatleme kõnealust probleemi ühe konkreetse sõna näitel. Nimisõna *käsi* tähistab tüüpiliselt inimese ülajäset, pole isegi oluline kui pikalt (kas ainult käelaba või koos õlavarrega). Sõnavormide kolmik *kätte : käes : käest* on sisekohakäändeis, kuigi kognitiivselt võiks kasutada väliskohakäändeid, aga see meid sõnatähenduse seisukohalt väga ei häiri. Küsimus on pigem, kas nende kolme vormi pärast substantiivina tuleks:

1) STÜ sõnastikku lisada märksõnale KÄSI (S) mingi uus tähendus, mis esineb ainult vaegparadigmas kolmes sisekohakäändes või

2) on võimalik mingite formaalsete tunnuste või kontekstireeglite kirjeldamise teel neid vajalikul juhul määrsõnadeks liigitada.

Paljudel allpool toodud juhtudel võib ka väita, et tegemist on mitmesõnaliste üksustega, kus tähendus ei ole niivõrd ühendi komponentide tähenduste summa vaid ühendil tervikuna on oma tähendus. Praktilise lahendusena tuleb kõne alla mitmesõnaliste üksuste tekstis tuvastamise programmi loomine, mis kasutaks juba olemasolevat verbikesksete püsiühendite andmebaasi (selle andmebaasi kohta vt nt Kaalep, Muischnek 2003).

Sõnavormi grammatikaliseerumise peamine tunnus on leksikaalse tähenduse „lahjenemine“. Konkreetse (ehk leksikaalse) tähendusega sõna vormi kasutatakse ka abstraktsema (ehk grammatilisema) tähenduse väljendamiseks (Heine, Kuteva 2002: 1–14). Meid huvitavate sõnavormide konkreetne tähendus viib need kokku nimisõnaga *käsi*, abstraktsema tähendusega kasutusjuhud tuleks aga liigitada omaette määrsõnadeks.

Tähenduse konkreetsuse või abstraktsuse hindamine on praegu jõukohane inimesele, empiirilise materjali detailsem uurimine annab aga ehk ka võimalusi sõnastada vajalikke reegleid arvutianalüüsi jaoks.

Vaatleme sõnavormide *kätte : käes : käest* esinemist koos lause predikaadiks oleva verbiga. Kõik näitelaused on võetud morfoloogiliselt ühestatud korpusest, pole seega teoreetilised võimalikud konstruktsioonid, vaid tegelik keelekasutus. Iga predikaadi ja uuritava sõnavormi ühendi kohta on artiklis näiteks toodud skaala KONKREETNE ↔ ABSTRAKTNE äärmuslikumad juhud.

Näide 1. KÄTTE

\+ **tooma**

kui ta kinda **kätte** toob

= konkreetne, *toob kuhu?*

mis toob **kätte** kõik maailma uudised

= abstraktne, väljend *kätte tooma* tähenduses ‘edastama’

\+ **saama**

Meie kaudu saavad nad oma raha kõige lihtsamalt **kätte**,

Pudelis peitunud kiri saadi **kätte** alles Scankristalli tsehhis,

Varas saadi **kätte**,

= konkreetselt käte vahele või käte abil, tähenduses ‘kinni püüdma, tabama’

Poole tunni pärast on Endrik direktori telefoni teel **kätte** saanud.

Tugeva nokapeitliga muströhn saab puudest **kätte** ka need maiuspalad,

= abstraktsemalt, tegelikult ilma käteta, tähenduses ‘kinni püüdma, tabama’

et kuumas vees hoitud sidruneid on kergem koorida ja neist saab ka mahla paremini **kätte**.

ise midagi head lugeda ja „vaimsus” niimoodi siiski **kätte** saada.

oleksin endas midagi olulist **kätte** saanud või õieti sellest vabanenud.

= abstraktselt, tähenduses ‘omandama’

+ **andma, jagama**

annab mu teenija teile selle **kätte**.

kellele raudrehasid **kätte** jagati.

= konkreetne

Seepärast ei anna kirjeldatud meetod otseselt doosi väärtust **kätte**.

See andis **kätte** suuna edasiseks,

just see naiste halvustamine annabki tegelikult igale algajale joodikule **kätte** alatise õigustuse ise samuti teha: naisi kiruda,

= abstraktsem

+ **võtma**

Ta pöördus tagasi malelaua poole ja võttis valge ratsu uuesti **kätte**.

Ta valas kõigile veini ja võttis oma klaasi **kätte**.

pani kuue selga ning võttis pange **kätte**.

võta vardad **kätte** ja koo uus.

= väga konkreetne, saab küsida *võtma kuhu?*

mil Antti ühtegi erialast teost **kätte** ei võtnud.

= sama konkreetne, situatsioon ja järeldused nõuavad tähendusnüansi tundmist, mis kuulub hoopis teise tähendusvälja

Kui nad **kätte** võtaksid,

Pärtel võttis **kätte** ja pesi panni puhtaks.

= abstraktne, ei saa küsida: *võtma kuhu?*

Esineb muidugi ka hulk verbe, millega koos esinedes *kätte* konkreetset tähendust ei saagi.

+ **maksma**

Kuidas täpselt kavatseb Türgi oma NATO liitlasele **kätte** maksta,

Keeldumise eest oli talle julmalt **kätte** makstud.

nüüd tahtis ta **kätte** maksta kõigi oma piinade eest.

= väljend kätte maksma

+ **võitma, võitlema**

et teadusajakirjade „turul” valitseb äärmiselt tugev konkurents ja uuel väljaandel on raske oma kohta teaduspäikese all **kätte** võita.

mille sa kuradilt nii raske eneseületamise hinnaga **kätte** oled võidelnud.

mil meie vanaisad ja isad iseseisvuse **kätte** võitlesid ja seda kindlustada püüdsid,

= ei saa küsida: *võitma/võitlema kuhu?*

+ **jõudma**

Ja siis jõudis see päev **kätte**,

Tasahilju jõuab **kätte** aeg,

et kevad on tõepoolest **kätte** jõudnud.

aga varem või hiljem jõuab alati **kätte** hetk,

= väljend *kätte jõudma*, seotud ainult ajaga

+ **tulema**

kuid mõni aeg hiljem tuli minek tõesti **kätte**.

et elus tuleb ikka kõik **kätte** väikese hilinemisega nagu see piletki

= väljend *kätte tulema*, seotud aja kulgemisega

+ **liikuma, käima**

mille autor oli Goldstein ja mis liikus siin-seal salaja käest **kätte**.

See käib osade kaupa pidevalt käest **kätte**,

Jaava ja Tseilon võivad küll lugematuid kordi käest **kätte** käia,

= väljend käest kätte käima/liikuma

Näide 2. KÄES

+ **hoidma**

Ta hoiab **käes** lühikese varrega oda.

Tema ainus soov oli seda fotot veel kord **käes** hoida või vähemalt seda nähagi.

= konkreetne

+ **olema**

Rongkäigus on üks loosungi ots alati minu **käes**.

Tal oli **käes** jämedast pruunist riidest tööriistakott,

= konkreetne

30. ringil oligi häda **käes** ning viies Kanada GP võit libises sakslasel peost.

et tal juba peremehelt ööluba **käes** oli.

Varsti on **käes** riigieksamid

Kaks võitu on **käes**,

kui auto ja selle võõras kasutaja olid politseil juba **käes**.

on see muutus juba **käes**.

= järjest abstraktsem

Aga südaöö on **käes**,

kui aeg **käes** on.

Kui see lõpuks **käes** oli,

Ent kui see hetk on **käes**,

= abstraktne, seotud ajaga, aga kelle või mille käega?

Näide 3. KÄEST

+ **võtma, kahmama, haarama, saama** (+ kinni)

Ta võttis Edgaril **käest** kinni,

Keegi võttis tal **käest** ja püüdis aidata tal edasi ronida.

Nagu demonstratsiooniks kahmas üks neist tüdruku **käest** kinni.

Ta jäi kohmetult ukse kõrvale seisma ja haaras Tehvanil **käest** kinni.

aga Erlend sai tal **käest** kinni ja tiris tagasi.

= konkreetne

+ **panema, pillama, libisema, lendama …**

Panin raamatu **käest**.

ja pani sulepea **käest**.

et ta noa **käest** pillas ja õlgade vappudes naerma puhkes.

et ta võib tema kaotada ja et see noor valge keha võib tal **käest** libiseda!

kandik oli **käest** lennanud,

= konkreetne

+ **hoidma, lahti laskma**

kui O'Brien ta **käest** lahti laskis; aga kuigi ta ei saanud seda enam tagasi,

ja ta ei hoia enam endise kindlusega Iivul **käest**.

= konkreetne

\+ **tirima, kiskuma, krabama, krahmama …**

krahmasin ma ta **käest** pudeli.

olid kahmanud ühe ja sama kastruli ja tirisid seda teineteise **käest** ära.

püüdis kastrulit ja lusikat ema **käest** ära kiskuda,

Siis krabas ta äkki õe **käest** šokolaaditüki ja oli ainsa hüppega uksest väljas.

= konkreetne, kuid eeldab, et predikaadi objekt on eelnevalt kellegi või millegi käes või valduses

\+ **andma, laskma, pudenema**

et Pärtel ei anna krunti sellepärast **käest**,

siis lasevad võimalused **käest**

Alguse ja lõpu hetk pudeneb **käest**.

= abstraktne, väljend *käest andma/laskma/pudenema*, tähenduses ‘loovutama või loobuma’

Tuleb tunnistada, et uuritavate sõnavormide konkreetsest abstraktseks ülemineku selget punkti tuvastada ei õnnestunud. Lihtsam oli eristada tähenduse muutumist kindlate verbide mõjuväljas. Kui tunnistada kõik *kätte : käes : käest* + V konstruktsioonid morfoloogilises (ja ka süntaktilises) analüüsis püsiühenditeks, lahendab see enamikul juhtudel nende sõnavormide semantiliselt vale sõnaliigi (substantiiv) probleemi, kuid koormab sõnastikku rohkete uute püsiühenditega, mille tähenduslik või fraseoloogiline ühtsus on vaieldav. Jääb alles ka vajadus teatud juhtudel määratleda vaadeldud sõnavorme iseseisva tähendusega substantiivi *käsi* vormidena (eriti kehtib see *kätte* puhul).

6. Kokkuvõtteks

Käesolevas artiklis kirjeldati eestikeelse teksti morfoloogilist märgendamist, selle töö tulemust kui uut keeleressurssi ja märgendamise käigus kerkinud probleeme.

Nendest probleemidest vaadeldi lähemalt verbide ja noomenite muutevormide järk-järgulist üleminekut adpositsioonideks ja adverbideks ning sellest tulenevaid raskusi nii morfoloogilise ühestamise enda kui ka morfoloogiliselt ühestatud teksti erinevate rakendusvaldkondade – süntaktilise ja eriti semantilise analüüsi jaoks.

Sõnaliikide määramise semantiliste kriteeriumite kohta võib öelda, et sõnavormide kasutuse semantilisest analüüsist on abi fraseoloogiliste üksuste leidmisel. Võimalust või võimatust küsida mingi lause üksuse kohta iseseisvat küsimust saavad lauseanalüüsis kasutada inimesed, kuid tõene (ja ühene?) vastus pole leitav puhtformaalsete tunnuste kaudu ei morfoloogilises ega süntaktilises analüüsis.

Kirjandus

EKG I = Erelt, Mati & Kasik, Reet & Metslang, Helle & Rajandi, Henno & Ross, Kristiina & Saari, Henn & Tael, Kaja & Vare, Silvi 1995. Eesti keele grammatika I. Eesti Teaduste Akadeemia Eesti Keele Instituut. Tallinn.

EKSS I = Eesti kirjakeele selestussõnaraamat 1988. 1. vihik. Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut. Tallinn: Valgus.

EKSS IV = Eesti Kirjakeele Seletussõnaraamat 1996. IV köide. Eesti Teaduste Akadeemia Eesti Keele Instituut.

Heine, Bernd & Kuteva, Tanja 2002. World Lexicon of Grammaticalisation. Cambridge University Press.

Kaalep, Heiki-Jaan & Muischnek, Kadri & Müürisep, Kaili & Rääbis, Andriela & Habicht, Külli 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti keele testkorpuse morfosüntaktilise märgendamise kogemusest. – Keel ja Kirjandus 9, 623–633.

Kaalep, Heiki-Jaan & Muischnek, Kadri 2003. Püsiühendite leidmine suurtest tekstikorpustest. – Toimiv keel I. Töid rakenduslingvistika alalt. Eesti Keele Instituudi toimetised 12. Tallinn: Eesti Keele Sihtasutus, 101–136.

Kaalep, Heiki-Jaan & Vaino, Tarmo 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Arvutuslingvistikalt inimesele. Toim Tiit Hennoste. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu, 87–99.

Kaalep, Heiki-Jaan & Vaino, Tarmo 1998. Kas vale meetodiga õiged tulemused? Statistkale tuginev eesti keele morfoloogiline ühestamine. – Keel ja Kirjandus 1, 30–38.

Karlsson, Fred 2002. Üldkeeleteadus. Tõlkinud ja kohandanud Renate Pajusalu, Jüri Valge ja Ilona Tragel. Tallinn: Eesti Keele Sihtasutus.

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

Uuspõld, Ellen 2001. *des*- ja *mata*-vormide kaassõnastumine ja eesti komareeglid. – Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks 12. märtsil 2001. Koost ja toim Reet Kasik. Tartu Ülikooli eesti keele õppetooli toimetised 17. Tartu, 306–321.

Viks, Ülle 1992. Väike vormisõnastik. Tallinn.

ÕS 1999 = Eesti keele sõnaraamat. ÕS 1999. Toimetanud T. Erelt, koostanud T. Leemets, S. Mäearu, M. Raadik ja T. Erelt. Tallinn: Eesti Keele Sihtasutus.

The problems of word class disambiguation in the automatic analysis of Estonian

Kadri Muischnek, Kadri Vider

University of Tartu

This article presents a new language resource – a morphologically annotated text corpus and discusses some linguistic problems that rose during the process of manual morphological disambiguation.

The research group of computational linguistics of the University of Tartu has developed a morphologically disambiguated corpus of Estonian. This work was supported by the national program „Eesti keel ja Rahvuskultuur (Estonian Language and Culture)“.

During that project 500 000 running words were manually morphologically disambiguated. The disambiguated texts belong to the following text classes:

- newspaper texts (100 000 words)
- fiction (100 000 words)
- legal texts (100 000 words)
- texts from the scientific magazine „Horisont“
- and 100 000 words of transcribed spoken language texts.

The disambiguated (written language) texts are available at the home page of the Research group for computational linguistics www.cl.ut.ee.

In this article we describe the tagset and process of the manual disambiguation, but the focus is on the linguistic problems that the annotators encountered during the process of manual disambiguation.

Although in some rare cases a human annotator had difficulties i.e. in determing the case form of a noun the main difficulties were encountered in the domain of wordclass disambiguation. The borderlines between wordclasses are known to be fuzzy, in Estonian even the closed wordclasses are not really closed as new pre- and postpositions (as well as adverbs) are constantly developing from inflectional forms of nouns and verbs – a phenomenon that is of extreme interest for a linguist but most depressing for a human morphological annotator.

The morphologically disambiguated texts are used as an input for word sense disambiguation (in addition to many other applications). The word sense disambiguation deals only with content words, so the exact border between the inflectional forms of nouns and verbs on the one hand and the adpositions on the other is important for this further task.

KADRI VIDER (sünd 1969) on lõpetanud Tartu Ülikooli eesti filoloogia erialal. Kaitses magistrikraadi Tartu Ülikoolis 1999. a üldkeeleteaduse alal. Alates 1995. a töötab Tartu Ülikoolis üldkeeleteaduse õppetooli juures arvutuslingvistika uurimisrühmas. On tegelenud leksikaalse semantikaga, semantiliste suhetega tesaurustes (eesti wordnet-tüüpi tesaurus), sõnatähenduste ühestamisega tekstides.

kadri.vider@ut.ee

KADRI MUISCHNEK (sünd 1965) on lõpetanud Tartu Ülikooli eesti filoloogia erialal. Kaitses magistrikraadi Tartu Ülikoolis 1998. a üldkeeleteaduse alal. Töötab Tartu Ülikoolis üldkeeleteaduse õppetooli juures arvutuslingvistika uurimisrühmas. On tegelenud korpuslingvistika, eesti keele automaatse morfoloogilise ja süntaktilise analüüsi ning püsiühendite ja nende tekstis tuvastamisega.

kadri.muischnek@ut.ee