SÕNALIIGITUSE KÜSIMUSI EESTI MURRETE KORPUSE PÕHJAL

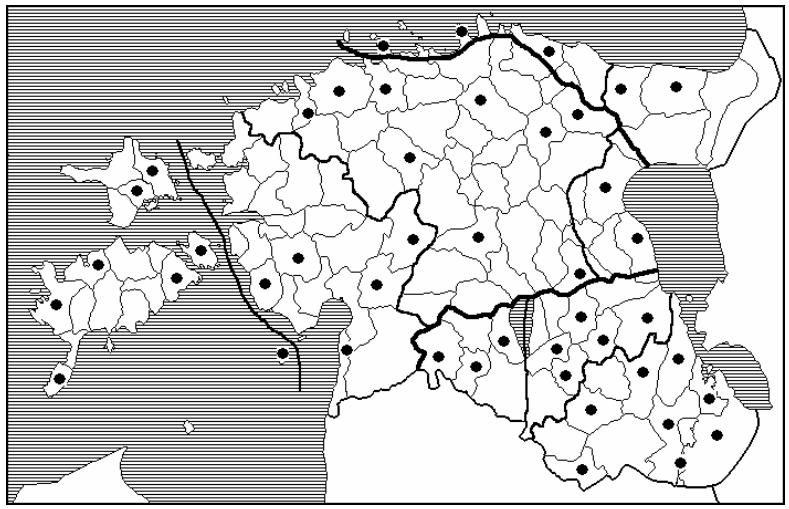
Liina Lindström, Liisi Bakhoff, Mari-Liis Kalvik, Anneliis Klaus, Rutt Läänemets, Mari Mets, Ellen Niit, Karl Pajusalu, Pire Teras, Kristel Uiboaed, Ann Veismann, Eva Velsker
Tartu Ülikool, Eesti Keele Instituut

1. Sissejuhatus

Eesti murrete korpus on kõiki eesti murdeid sisaldav elektrooniline tekstikogum, mida Tartu Ülikooli ja Eesti Keele Instituudi koostöös on koostatud alates 1998. aastast. See koosneb foneetilises transkriptsioonis murdetekstidest, lihtsustatud transkriptsioonis tekstidest morfoloogiliseks ja süntaktiliseks analüüsiks (murdekorpuse transkriptsiooni kohta vt nt Lindström 2004, 2005) ning morfoloogiliselt märgendatud tekstidest. Kõigi tekstide kohta on olemas ka lindistus, mille põhjal transkriptsiooni on kontrollitud. Suurem jagu lindistusi on praeguseks olemas ka digitaalsel kujul. Murdekorpus koosneb seega autentsetest murdetekstidest. Murdekorpuse põhjalikumat kirjeldust ja varasemat seisu vt Lindström jt 2001, Lindström, Pajusalu 2003, Lindström 2004, Lindström 2005.

Murdekorpusesse lisandub tekste igal aastal *ca* 100 000 sõna, eesmärk on jõuda vanemate tekstidega (lindistatud 1960.–1970. aastatel) vähemalt 1 miljoni tekstisõnani. 2006. aasta jaanuaris oli murdetekste foneetilises transkriptsioonis sisestatud üle 700 000 sõna, neist morfoloogiliselt märgendatud oli ligi 270 000 sõna.

Joonis 1. **Murdekorpuses esindatud murrakud veebruaris 2006**



Põhirõhk on viimastel aastatel olnud murdetekstide morfoloogilisel märgendamisel. Märgendamiseks on kasutatud abiprogrammi Mark. Märgendatud tekstid on xml-formaadis ja need on loetud MySQL-andmebaasi, mida on võimalik kasutada Interneti kaudu (www.murre.ut.ee --> otsing), andmebaasist otsimiseks vt juhiseid Lindström 2005. Praegusel hetkel on kasutusel kaks otsingumootorit, millest esimene (search.php) võimaldab detailsemat analüüsi koos kontekstiga, milles otsitav sõna/vorm paikneb, ning teine (search2.php) lisab ka statistikat. Andmebaasi on praeguseks laetud umbes 270 000 märgendatud tekstisõna. Iga tekstisõna puhul on märgendatud järgmised väljad:

- **SNE**: sõne originaalkujul, nii nagu see tekstis esineb;
- **MSN**: märksõna kirjakeelestatud kujul (kasutatud kirjakeele ortograafiat, kaotatud on vokaalharmoonia). Kui kirjakeeles on sama tüve ja tähendusega sõna olemas, on märksõnana esitatud kirjakeelne sõna;
- **FRA**: fraas, märgendatud fraasistunud ühendite puhul, nt ühend- ja väljendverbide puhul;
- **TAH**: tähendus. Tähendus esitatakse ainult siis, kui see erineb kirjakeelest või kui kirjakeeles vastav sõna puudub;

• **SLK**: sõnaklass. Sõnaklassid ja nende lühendid on esitatud tabelis 1;
• **MRF**: morfoloogiline info muutevormide kohta;
• **KHK**: kihelkond, millest sõne pärit.

2. Murdekorpuses kasutatud sõnaklassid

Morfoloogilisel märgendamisel on esimeseks tööks sõnade klassifitseerimine sõnaliikidesse.

Sõnaliike määratletakse tavaliselt sõnade morfoloogilise ja süntaktilise käitumise alusel, aga ka semantika alusel. Morfoloogilise käitumise poolest vaadatakse, milliseid lõppe ja tunnuseid võib sõnale lisada (nt verbile saab lisada aja, kõneviisi, arvu ja isiku tunnused, adjektiivile aga näiteks käände, arvu, komparatsiooni tunnuse). Süntaktilise käitumise poolest vaadatakse, kas sõna võib lauses olla iseseisva lauseliikmena või mitte ning missuguses süntaktilises funktsioonis ta lauses prototüüpselt on. Näiteks substantiiv on lauses tüüpiliselt argumendi peasõna (nt *Mees läks metsa*), adjektiivi kasutatakse aga tüüpiliselt substantiivi modifitseerijana (nt *Paks mees läks metsa*), adverbi aga adjektiivi või adverbi modifitseerijana (nt *väga ilus*, *väga valusasti*) (vt nt Schachter 1985, Hengeveld 1992, Anward 2001).

"Eesti keele grammatikas" (EKG I) eristatakse sõnaliike tähenduse alusel (täistähenduslikud ja mittetäistähenduslikud sõnaliigid), süntaktilise iseseisvuse alusel (iseseisvad sõnad, mis võivad esineda üksi lauseliikmena, ja mitteiseseisvad sõnaliigid, mis ei esine lauseliikmena) ning morfoloogilise muutumise alusel (muutuvad ja muutumatud sõnaliigid) (EKG I: 14–41). Selle jaotuse alusel eristatakse eesti keeles 12 sõnaliiki.

Keele kirjelduses eristatud sõnaliikide süsteem ei pruugi aga olla piisavalt eristav ja piisavalt selge praktiliste eesmärkide saavutamiseks. Et eristada n-ö keelesüsteemi sõnaliike (mida esindab EKG) konkreetsel eesmärgil loodud sõnaliikidest, nimetame viimaseid sõnaklassideks. Sõnaklasse saab vajadusel lisada ja vähendada, sõltuvalt uurija eesmärgist ja vajadustest.

Murdekorpuse puhul on sõnaklasside määratlemisel olnud põhieesmärgiks teha selline klassifikatsioon, mis ühelt poolt oleks murdekeele uurijale võimalikult arusaadav ning piisavalt detailne, teiselt poolt aga morfoloogilise märgenduse tegijatele piisavalt

selgete piiridega. Seega oleme üritanud eristada sõnaklasse, mis küllalt hõlpsalt eristuvad (nt küsivad-siduvad sõnad eristuvad selgelt muudest *pro*-sõnadest), teiselt poolt aga leppinud teatava hägususega seal, kus piire on raske tõmmata – näiteks adverbi eri liikide vahel on vahetegemine praktikas kaunis keeruline.

Morfoloogilise märgendamise aluste väljatöötamisel oleme osalt lähtunud EKIs tehtud Hargla murraku morfoloogilise andmebaasi eeskujust (vt http://www.eki.ee/dict/hargla/). Selle põhjuseks oli esialgne idee, et need andmebaasid peaksid olema ühitatavad ning seetõttu peaksime kasutama samu märgendeid. Esialgse märgendite süsteemi võtsimegi sealt üle, ent sellesse tuli teha üksjagu muudatusi. Kui Hargla murraku andmebaas sisaldab peamiselt üksiksõnu, siis murdekorpuses on tegu pikkade suuliste tekstidega. Paratamatult tuleb arvestada suulise kõne eripäraga: leida lahendus, kas ja kuidas märgendada poolikuid sõnu, üneeme (*ee, ää*), diskursuspartikleid (*noh, mh, jajah* jne). Murdekorpuse jaoks oleme EKIs kasutatud liigitust detailsemaks muutnud ning lisanud suulise kõne eripärast tingitud sõnaklasse (diskursuspartikkel, suhtlussõna; aluseks on Hennoste käsitlus suulise kõne sõnaklassidest, vt nt Hennoste 2002). Hargla andmebaasist oleme üle võtnud aga märgendid mõningate grammatiliste üksiksõnade jaoks. Nii käsitleme omaette sõnaklassina verbi juurde kuuluva eitussõna (*ei, es, mitte, ühti* jne; märgend Mn), samuti võrdlussõna liitsuperlatiivis (*kõige ilusam*; märgend Ms). Hargla andmebaasi eeskujul saavad omaette sõnaklassi märgendi ka adjektiiv komparatiivis (*ilusam*; Ak) ja adjektiiv superlatiivis (*ilusaim*; As). Tagantjärele tarkusena võib öelda, et see langeb meie üldisest süsteemist siiski välja ning tekitab segadust märksõnastamisel (nt kas märksõna panna *noor* või *noorem*). Hargla andmebaasi eeskujul eristame omaette sõnaliigina ka abiverbi (Va) liitajavormides, piirdudes siiski vaid *olema*-verbiga (nt *oli läinud, on tehtud*). See on osutunud vajalikuks andmebaasist liitajavormide (täis- ja enneminevik) ülesleidmisel – liitajavormides kasutatavad partitsiibid mingit spetsiaalset märgendit ei saa (vaid märgendid *nud*, *tud*).

Hargla murraku andmebaasi sõnaklasside süsteem esindab eesti traditsioonis sõnaliikide varasemat, EKG-eelset käsitlust: siin ei eristata erinevaid adverbitüüpe (täistähenduslik adverb, modaaladverb, proadverb, afiksaaladverb jne). Oleme lähtunud sõnaklasside määratlemisel loomulikult ka EKG ja EKKi sõnaliikide

süsteemist. Nii eristame EKG ja EKKi eeskujul tavalistest leksikaalsetest adverbidest proadverbe ning afiksaaladverbe. Võrreldes EKG ja EKKga on murdekorpuse sõnaliikide süsteemis siiski mõningaid muudatusi.

Iseseisvate täistähenduslike sõnaliikide puhul oleme eristanud näiteks pärisnime muudest nimisõnadest, kuigi nad käituvad nimisõnadega nii morfoloogiste tunnuste ja lõppude lisandumise kui ka süntaktilise funktsiooni mõttes ühtemoodi. Tekstis eristub pärisnimi aga selgelt ortograafiliselt (ka murdekorpuse lihtsustatud transkriptsioonis kasutatakse suurt algustähte), samuti võib uurijatel olla spetsiifiline huvi pärisnimede vastu. Seega tundub pärisnimede eristamine muudest nimisõnadest mõistlik ja vajalik. Sama on tehtud ka Hargla murraku andmebaasis.

Mittetäistähenduslike sõnaliikide hulgas oleme teinud muudatusi veelgi enam.

Pronoomenite sõnaliigi oleme teinud tükkideks selle alusel, mis sõnaliiki kuuluvat sõna see asendab: prosubstantiiv (nt *see, tema, mina*), proadjektiiv (nt *selline, sihuke, niisugune*), pronumeraal (nt *mitu*; kahjuks märgendusprogrammi ebatäiuslikkuse tõttu on praegu märgendatud kategooriasse Muu). Selline jaotus on olemas ka EKGs, ent pronoomenite sõnaliigi sees. Samasse seeriasse kuulub ka proadverb, mis on ka EKGs omaette sõnaliigina esitatud (nt *siin, seal, nii, siis*).

Omaette sõnaklassina on murdekorpuses käsitletud küsisõna (märgend Intr), mis hõlmab nii interrogatiiv-relatiivpronoomeni (*kes, mis*) kui ka muud interrogatiivid (*kas, võ*(*i*)*, kus, millal, missugune* jne), seda nii küsisõnalises kui küsivas-siduvas funktsioonis. Küsivad-siduvad sõnad on muudest *pro*-sõnadest hõlpsalt eristatav sõnaklass, millel on ka üsna selged kasutamistingimused (paikneb küsilause või kõrvallause algul). Nende eristamine sõnaklassina võimaldab korpusest hõlpsamini otsida küsi- ja kaudküsilaused, samuti relatiivlauseid.

Katse suulise kõne partikleid integreerida eesti keele sõnaliikide süsteemi on teinud Tiit Hennoste (2002). Hennoste soovituste põhjal oleme loobunud hüüdsõna (interjektsiooni) kategooriast ning eristanud selle asemel diskursuspartikleid (Par; siia hulka kuuluvad ka modaaladverbid), onomatopoeetilisi sõnu (Ono, nt *auh-auh, soll-soll*) ja suhtlussõnu (Suht, nt *tere, aitäh*).

Hüüdsõna on muudest sõnaliikidest tavaliselt eristatud selle põhjal, et ta võib üksi moodustada lause, väljendada terviklikku situatsiooni. Seda on peetud n-ö *root*-kategooriaks (vt Anward 2001). Näiteks *Oh!* võib moodustada omaette lause või terviksituatsiooni vastusena küsimusele *Kas sul on artikkel juba lõpetatud?* Enam-vähem sama annab vastusena edasi ka näiteks eituspartikkel *ei*. Üksi võivad vähemalt suulises vestluses lause moodustada aga ka teised sõnaliigid, näiteks väga omane on see adjektiividele: *Jube! Kohutav! Suurepärane*! Praktikas on sageli raske selle omaduse alusel eristada interjektsioone muudest muutumatutest sõnadest, sest spontaanses kõnes on hulk muutumatuid sõnu, mis võivad käituda nii interjektsioonina kui ka modaaladverbina. Üheks selliseks on näiteks hesitatiivne *noh*, mis sageli ei seostu muu kontekstiga ning mida kasutatakse pausitäitjana, ent vahel siiski annab edasi mingit modaalset tähendusvarjundit ning funktsioneerib pigem modaaladverbina: *`leibä muiduk'k'i ja siss noh peenet `leibä ka küdsät't'i* (HAR); *siis veel veikke poisikke noh esimest kord näind koa* (JUU).

Seega on kategooriat Par (diskursuspartikkel) kasutatud korpuses väga ulatuslikult: siia hulk kuulub hulk hüüdsõnu ning modaaladverbid, mis sageli on samad sõnavormid. Kuna süntaktiliselt ei ole tegemist lauseliikmetega, vaid üldlaienditega, siis tundub selline lahendus olevat sobiv.

Tabel 1. **Morfoloogilisel märgendamisel eristatud sõnaklassid ja nende lühendid**

Sõnaklass	Lühend	Näited, kommentaarid
Substantiiv	S	*kas's*
Pärisnimi	H	*Jüri, Pärnumaa*
Verb	V	*ostma*, v.a *olema* abiverbina
• Abiverb	Va	ainult *olema*-verbi vormid liitaegades: ***oli*** *läinud*
Adverb	Adv	*täna, ilusasti*
• Afiksaaladverb	Adva	ühendverbi koostisesse kuuluvad adverbid, nt *välja* (*mõtlema)*
Numeraal		
• Põhiarvsõnad	Nump	*kaks*
• Järgarvsõnad	Numj	*teine*

Sõnaklass	**Lühend**	**Näited, kommentaarid**
Adjektiiv	A	*vana*
• Adjektiiv komparatiivis	Ak	*vanem*
• Adjektiiv superlatiivis	As	*vanim*
Pro-sõnad		
• Prosubstantiiv	ProS	*see, too, tema, mina*
• Proadjektiiv	ProA	*niisuke, sihuke*
• Proadverb	ProAdv	*siin, seal*
Kaassõnad		
• Postpositsioon	Post	*maja* ***taga***
• Prepositsioon	Pre	***pärast*** *sööki*
Diskursuspartiklid	Par	*no, noh, oh, nagu, ikka*
Suhtlussõnad	Suht	*aitäh, palun, tere*
Onomatopoeetilised sõnad	Ono	*mürts, soll-soll*
Küsisõna	Intr	*kas*, relatiiv-interrogatiivpronoomenid *kes, kelle, keda*, *mis* jne, küsivad-siduvad adverbid *miks, millal, kus* jne
Konjunktsioon	Konj	nt *ja*, sh piiripartiklid
Eitussõna liitvormides	Mn	***ei*** *ole* ***mitte, ühti***
Võrdlussõna liitsuperlatiivis	Ms	***kõige*** *ilusam*
Muud	Muu	ei oska määrata või ei sobi ühtegi eespool toodud sõnaklassi
Määramata	X	ei ole võimalik määrata (poolikud sõnad)

Tabelis 2 on esitatud andmed selle kohta, kui palju on märgendatud teksti hulgas esindatud erinevatesse sõnaklassidesse kuuluvaid sõnesid. Teistest selgelt enam on verbe ja substantiive. Väga suur on ka prosubstantiivide, proadverbide ja sidesõnade osakaal – selle põhjuseks on ilmselt teksti suulisus. Teksti suulisusest on tingitud ka diskursuspartiklite suur osakaal.

Palju on murdetekstides kasutatud eitussõnu. Raske on öelda, kas selle põhjuseks on eestlaste suur eituselembus või mitmest eitussõnast koosnevad eituskonstruktsioonid, mis on levinud eriti eesti läänepoolsemates murretes (nt *killel* ***ess*** *ole maad* ***`mette*** *siit `Atla külast `antti*, KHK), igal juhul on üllatav, et vaid mõnest liikmest koosnev sõnaklass tekstides nii sagedasti esineb.

Tabeli lõpuosast leiame võrdlussõna, mida on kasutatud 130 korral, omadussõna ülivõrdes pole aga murdetekstidest üldse leida. See näitab kujukalt, et murdekeelele on iseloomulik just ana-

lüütilise ülivõrde kasutamine (nt *pohja tuul oli ikka keige ägedamb*, JÕE).

Kategooria X sisaldab peamiselt poolelijäänud sõnu ja tundmatuid sõnu, mida pole võimalik ühegi sõnaklassiga seostada. Kategooria Muu alla on koondatud juhtumid, mis ei mahu olemasolevasse sõnaklasside süsteemi. Põhiosa sellesse kategooriasse kuuluvatest sõnadest on sõna *mitu* kasutusjuhtumid, sest tehnilistel põhjustel pole praegu märgendusprogrammis sõnaklassi asearvsõna (pronumeraal), kuhu *mitu* peaks kuuluma.

Tabel 2. **Sõnede hulk sõnaklasside kaupa murdekorpuse andmebaasis veebruaris 2006**

Sõnaliik	**Märgendatud sõnu**	**Osakaal (%)**
Verb	46754	20,0
Substantiiv	41390	17,7
Prosubstantiiv	26347	11,2
Konjunktsioon	25276	10,8
Proadverb	20745	8,9
Adverb	17151	7,3
Diskursuspartikkel	15928	6,8
Adjektiiv	6707	2,9
Eitussõna	5115	2,2
Määramata	4718	2,0
Afiksaaladverb	4673	2,0
Postpositsioon	3867	1,7
Küsisõna	3853	1,6
Pärisnimi	3240	1,4
Põhiarvsõna	2809	1,2
Abiverb	2319	1,0
Proadjektiiv	1423	0,6
Adjektiiv komparatiivis	641	0,3
Prepositsioon	433	0,2
Järgarvsõna	338	0,1
Muu	211	0,1
Võrdlussõna	130	0,1
Onomatopoeetiline sõna	126	0,1
Suhtlussõna	54	0,0
Kokku	234248	100,0

3. Sõnaklassid, grammatikaliseerumine ja leksikaliseerumine

Kuna keel on pidevas muutumises, ei ole sugugi lihtne leida sobivat sõnaklassi kõigile sõnedele, ükskõik kui täiuslik sõnaklasside süsteem ka ei tunduks.

Morfoloogilise märgendamise käigus on kõige keerukamaks küsimuseks osutunud mitte niivõrd murdekeele arhailisus, kuivõrd keele varieerumine ja muutumine: grammatikaliseerumine ja leksikaliseerumine. Seda eeskätt seoses sõnaklasside määratlemisega. Kõige sagedasem küsimus, millele teksti märgendaja peab vastama, on: kas murdes X on sõnavorm Y pigem substantiiv või adverb?

3.1. Substantiiv või adverb ajamäärusena

Kõige sagedasem probleem märgendamisel on seotud ajamääristega, täpsemalt mõningate substantiivide kasutamisega ajamäärusena. Substantiivid nagu *öö, päev, kevad, sügis, talv, suvi, hommik, õhtu* jne on kõige sagedamini kasutuses just ajamäärusena, seejuures on sageli tegemist ka selle substantiivi erandliku vormiga.

Näiteks *öö* kõrval esineb väga sageli ajamäärusena vorm *ööse~öösse, öösi~öössi*, nt *`päävä ol'i seppa+baeas* ***öös'se*** *`tehti puu+`särkka* (JUU), *tunnõ `üttegi `raskust ei* ***`üüse*** *ei `päivä* (KAM). Kui ajamääruse moodustab sõnaühend, kasutatakse enamasti *öö*-sõna, *sie olli sääl* ***ütte üü*** *üleven* (HLS), ***tõsõ üü*** *ol'li rihi* ***tõsõ üü*** *oll'i nii kotti+`däitmine* (RÕN). *ööse~öösi* eristub muust paradigmast, ka kasutuses on näha erinevusi, niisiis on tegu selgelt adverbistunud sõnavormiga. Mõnikord aga võib ka *ööse~öösi* saada täiendeid nagu substantiiv: *saa õi `maada* ***tuu `üüse*** *ei* (Lääne-Setu), *ee rüi olnd pruut+paari pääl* (.) ***esimese `öösse*** *magada* (KÄI). Praegu on *ööse~öösi* märgendatud siiski alati adverbiks. Sõnade *öö* ja *ööse* kohta on ilmumas artikkel Eva Velskrilt (Velsker, ilmumas).

Samamoodi käituvad ka paljud muud ajaväljendid. Sõna *päev* puhul on läänepoolsetes murretes adverbistunud tugevas astmes vorm *`päeva* ‘päeval’: *kohe* ***`päävä*** *ol'i seppa+baeas* (JUU), *tunnõ `üttegi `raskust ei `üüse ei* ***`päivä*** (KAM), ***`pääva*** *`kuivast sääl ära=ja* (PHL). Tõenäoliselt on tegemist vana essiivi vormiga (< *päivänä*), mis on adverbistunud. Selle sõna puhul tundub, et

ajaväljendid, milles on lisaks sõnale *päev* ka täiend, käituvad morfoloogiliselt substantiivina regulaarselt ning on tüüpiliselt genitiivis: *ja siiss* ***teise pääva*** (.) *meid aetti `jälle `sönna* (KHK).

Ajaväljendite interpreteerimise mingisse sõnaklassi kuuluvaks teeb keerukaks see, et pole alati selge, mis käändes mingit ajaväljendit kasutatakse, sellest tulenevalt kas tegu on adverbistunud kasutusega või substantiiviga. Näiteks mis käändes on sõna *õhtu~õhta* ajamäärusena: *kudust* ***`öhta*** *tule tule `valgel* (PHL).

Tavaliselt on ajaväljendid kas genitiivis, nt *jahh* ***tõise päevä kolmanda päevä*** *`tul'li `perrä* (RÕN), nominatiivis, nt *mina pole* ***see aeg*** *küll näin* (PHL) või adessiivis, nt ***ühel korral*** *neid sai soolat `rohkem* (PHL). On ka juhtumeid, kus on võimatu vormi tuvastada: *sääl siis maamehed käisid `ikka* ***hommigu*** *`vastas kaa* (PHL). Varieerumine toimub ühe murde sees väljendite kaupa, aga mingil määral ka murrete vahel.

Laiemalt on tegemist ajaväljendite leksikaliseerumisega ajamääruslikus kasutuses: grammatiline seos tüve ja käändelõpu vahel on ähmastunud ja muutunud ebamääraseks (käänet pole võimalik määrata), vorm on sageli muust paradigmast eemaldunud, sõnet interpreteeritakse tervikuna, mitte analüütiliselt (vt nt Lehmann 2002). Põhjuseks on võib pidada selle sõnavormi sagedast (peamist) kasutamist just selles funktsioonis.

3.2. Substantiiv või kaassõna

Kaassõna on suletud sõnaklass, millel on selge grammatiline funktsioon ning mis kuulub käändsõna juurde. Siiski ei ole kaassõnade hulk mingi väga kindlate piiridega sõnade kogum, sest grammatikaliseerumisprotsessi tulemusena võib sinna pidevalt uusi sõnu lisanduda. Kaassõnade peamine tekkeallikas on substantiiv, aga ka mõningad verbivormid (verbivormide kaassõnastumise kohta eesti keeles vt Uuspõld 2001). Üheks oluliseks ruumi väljendavate kaassõnade allikaks on kehaosade nimetused (Heine, Claudi, Hünnemeyer 1991: 217, eesti ja soome keele kohta vt Ojutkangas 2000), aga ka muud. Kaassõnade loend võib eri murretes olla erinev ning seetõttu on märgendajal aeg-ajalt väga keeruline otsustada, kas tegemist on kaassõna või nimisõnaga.

Kaassõna võib pidada n-ö valmisolevaks kaassõnaks, kui ta seostub vabalt väga erinevate substantiividega ning kaassõnale ei

saa lisada täiendeid. Piir ei ole alati selge, kui substantiivi ja kaassõna vahel toimib osa-terviku seos nagu näites *sääl tuu tõse tii* ***veeren*** *miss jõe puul't lätt* (HAR): teel on tõesti olemas veer, seega siin ei pruugi sõna *veeren* pidada kaassõnaks. Lisaks saab ühendis *tii veeren* tõenäoliselt laiendada sõna *veeren* omadussõnalise täiendiga (*tii tõsõn veeren*) – ka see näitab, et *veeren* kuulub siin veel nimisõna paradigmasse (murdekorpuses küll ühtki sellist näidet polnud). Näidetes *olli küll tönni kaivo* ***veeren*** *rian* (TRV), *paa* ***veeren*** 'paja ääres' (VÕN) on ilmselt aga teatav semantiline pleekimine toimunud ning kasutus on sarnasem kaassõnalisele kasutusele.

Samamoodi võib näites *mere* ***ääres*** *old üks `väikkene maea* (PHL) sõna *ääres* pidada substantiiviks, sest merel on äär: semantilist pleekimist ning tähenduse abstraktsemaks muutumist ei ole selles näites näha. Samas ei saa sõnale *ääres* lisada tõenäoliselt omadussõnalist täiendit (vähemalt murdetekstides pole). Ka tähenduse poole pealt vaadatuna ei väljenda neil juhtudel *ääres*-vorm niivõrd mitte kohta, kui juures- või lähedalolekut. See-eest näites *kaa sii sii `Elda+maa* ***ääres*** *sääl maa ole käind* (KHK) on ühendi *Eldamaa ääres* ainukeseks tähenduseks, et keegi on käinud Eldamaa lähedal – sõna *ääres* siin hakanud substantiivi *äär* tähendusest kaugenema. Uus tähendus annab tunnistust kaassõnastumisest.

Kui *veeres* ja *ääres* on kirjakeeleski tuntud kaassõnastumise juhud, siis *perrä* on vaid lõunaeesti murretes kasutusel olev kaassõna. Selgelt väljakujunenud kaassõnana funktsioneerib *perrä* järgmistes näidetes: *lät's'ivä siss leevä* ***`perrä*** (SE), *lät's' imä* ***`perrä*** (PSE), *`lät's'ivä kas's'ikkõsõga `mõtsa puijõ* ***`perrä*** (PSE). Sama sõnavorm on kasutuses ka (afiksaal)adverbina: *jahh tõise päevä kolmanda päevä `tul'li* ***`perrä*** (RÕN). Näites *lei pin'i jala* ***`perrä*** 'lõi peni jala pihta' (SE) võiks sõna *`perrä* tõlgendada aga pigem kui illatiivivormi substantiivist *perä*, kogu fraasi tähendus oleks umbkaudu 'lõi peni jala tagaossa'.

4. Kokkuvõte

Esitasime artiklis ülevaate eesti murrete korpuse märgendamisest ning märgendamisel kasutatud sõnaklassidest. Praeguse korpuse

põhjal on juba võimalik teha uurimusi eesti murrete grammatika (eriti morfoloogia) kohta.

Vaatlesime artiklis ka mõningaid sõnaliigiprobleeme, mis on seotud substantiivide sagedaste vormidega. Üldiselt on eesti keele sõnaliikidesse jagamine olnud suhteliselt vormikeskne. See eeldab paradigma olemasolu, meil peab olema eelteadmisi selle kohta, kas ja kuidas sõna muutub. Tegelikes tekstides on pilt selline, et neid täisparadigmasid meil tekstist leida ei ole võimalik, meil on vaid üksikud sõnavormid, mida kasutatakse mingis kindlas süntaktilises funktsioonis (nt substantiivid ajamäärusena jne). Kui jätta kõrvale morfoloogia ning läheneda sõnaliikidele süntaksi ja semantika poole pealt, torkavad silma hoopis teatud sõnade teatud tähenduses ja funktsioonis kasutatavad kindlad vormid. Keele muutumise käigus võivad need vormid leksikaliseeruda või grammatikaliseeruda.

Leksikaliseerumine ja grammatikaliseerumine ongi praegu need teemad, mida murdekorpuse tegijad oma magistri- ja doktoritöödes enim käsitlevad. Eva Velsker uurib doktoritöös ajaväljendite leksikaliseerumist eesti keeles (vt ka Velsker, ilmumas). Kaassõnade tekkimist eesti keeles uurib oma magistritöös Liisi Bakhoff, kaassõnade semantikaga kognitiivse keeleteaduse aspektist tegeleb doktorant Ann Veismann. Murdekorpuse põhjal on valmimas veel magistritöö Mari-Epp Tirkkonenil, kes uurib personaal- ja demonstratiivpronoomenite kasutust ja varieerumist kirde- ja rannamurdes. Kõne vahendamist väljendavaid konstruktsioone eesti murretes vaatleb magistrant Anneliis Klaus. Alustanud oleme ka uurimust ainsuse esimesele isikule viitamise kohta verbilõpuga ja/või personaalpronoomeniga võrdlevalt neis murretes, kus verbi lõpust on *-n* kadunud (*ma lähe*) ja neis, kus see on säilinud (*ma lähen*). Murdekorpuse põhjal on Võru ja Setu *nud-* ja *tud*-partitsiipide varieerumist uurinud Mari Mets (2005). Murdekorpust kasutatakse ka ETFi grandi "Eesti murrete grammatika" (ETF 5968) täitmisel.

KIRJANDUS

Anward, Jan 2001: Parts of speech. – M. Haspelmath, E. König, W. Oesterreicher, W. Raible (eds.), Language typology and language universals: an international handbook. Vol. 1. Berlin/New York: W. de Gruyter. 726–735.

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

EKK = Mati Erelt, Tiiu Erelt, Kristiina Ross 2000. Eesti keele käsiraamat. Teine, täiendatud trükk. Tallinn: Eesti Keele Sihtasutus.

Heine, Bernd, Ulrike Claudi, Friederike Hünnemeyer 1991: Grammaticalization: A Conceptual Framework. Chicago: The University of Chicago Press.

Hengeveld, Kees 1992: Parts of Speech. – M. Fortescue, P. Harder, L. Kristoffersen (eds.), Layered Structure and Reference in a Functional Perspective. Amsterdam: Benjamins. 29–56.

Hennoste, Tiit 2002: Suulise kõne uurimine ja sõnaliigi probleemid. – R. Pajusalu, I. Tragel, T. Hennoste, H. Õim (toim.), Teoreetiline keeleteadus Eestis. (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4.) Tartu: Tartu Ülikooli Kirjastus. 56–73.

http://www.eki.ee/dict/hargla/

Lehmann, Christian 2002: New reflections on grammaticalization and lexicalization. – I. Wischer, G. Diewald (eds.), New Reflections on Grammaticalization. Amsterdam/Philadelphia: Benjamins. 1–18.

Lindström, Liina, Karl Pajusalu 2003: Corpus of Estonian Dialects and the Estonian Vowel System. – Linguistica Uralica 4. 241–257.

Lindström, Liina 2004: Lõunaeesti keelematerjalid eesti murrete korpuses. – Tartu Ülikooli Lõuna-Eesti keele- ja kultuuriuuringute keskuse aastaraamat III. Tartu. 85–94.

Lindström, Liina 2005: Ülevaade Eesti murrete korpusest. Käsikiri. http://www.murre.ut.ee/EMK.PDF

Lindström jt = Lindström, Liina, Varje Lonn, Mari Mets, Karl Pajusalu, Pire Teras, Ann Veismann, Eva Velsker, Jüri Viikberg 2001: Eesti murrete korpus ja kolme murde sagedasema sõnavara võrdlus. – R. Kasik (toim.), Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks 12. märtsil 2001. (Tartu Ülikooli eesti keele õppetooli toimetised 17.) Tartu: Tartu Ülikooli Kirjastus. 186–211.

Mets, Mari 2005: Võru ja Setu kõnekeele mineviku kesksõnade tunnused: kas tegelik keelekasutus vastab võru kirjakeele normile? – Tartu Ülikooli Lõuna-Eesti keele- ja kultuuriuuringute keskuse aastaraamat IV. Tartu. 65–77.

Ojutkangas, Krista 2000: Ruumiinosannimien kieliopillistumisesta suomessa ja virossa. – Virittäjä, nr 1. 2–22.

Schachter, Paul 1985: Parts-of-speech systems. – T. Shopen (ed.), Language Typology and Syntactic Description, vol. 1: Clause Structure. Cambridge: Cambridge University Press. 3–61.

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

Velsker, Eva, ilmumas: *Öö* ja *ööse*: murded, ühiskeel, kirjakeel. – Emakeele Seltsi aastaraamat nr 51.

Kihelkondade lühendid

HAR	Hargla
HLS	Halliste
JUU	Juuru
JÕE	Jõelähtme
KAM	Kambja
KHK	Kihelkonna
KÄI	Käina
PHL	Pühalepa
PSE	Põhja-Setu murrakurühm
RÕN	Rõngu
TRV	Tarvastu
SE	Setu
VÕN	Võnnu