Eesti keele väliskohakäänete kasutus poolspontaanses kõnes automaatse transkriptsiooni põhjal

JANE KLAVAN, TANEL ALUMÄE, ARVI TAVAST

1. Sissejuhatus: korpuslingvistika

Kasutuspõhise keeleteaduse seisukohast on keelesüsteemi kirjeldamisel kandvaks jõuks keelekasutus (Diessel 2017). Kasutuspõhine keeleteadus on välja arenenud 1980.–1990. aastatest pärit funktsionaalse ja kognitiivse keeleteaduse uurimistöödest, mille rõhuasetus oli pragmaatilistel ja kontseptuaalsetel teguritel. Praeguseks on see-eest kasutuspõhises keeleteaduses keskne esinemissagedus ja selle mõju keelesüsteemile (nt Arnon, Snider 2010; Bybee, Hopper 2001; Bybee 2006, 2007, 2010; Divjak 2019; Hay 2001). Sageduse mõju ilmneb keeles igal tasandil (Bod jt 2003: 10). Tsiteerides Martin Haspelmathi (2006: 45): „[---] sagedus/harvaesinevus tekstis (st peamiselt igapäevakõnes) on peamine nähtus, mis seletab paljusid teisi keelelisi nähtusi [---]." Siinne artikkel vaatlebki sagedust igapäevakõnes raadiosaadete näitel. Eesmärk on automaatse kõnetuvastuse rakendamine eesti keele morfosüntaksi uurimiseks, kasutades näidismaterjalina väliskohakäändeid ja nendega väljendatud funktsioone.

Eesti keele väliskohakäändeid on varem uuritud eri perspektiividest, kuid käsitsi märgendatud väliskohakäänete funktsioonide analüüsi mahuka valimi põhjal põhjal pole autoritele teadaolevalt tehtud. Seni kõige laiaulatuslikuma uurimuse väliskohakäänetest on kirjutanud Ene Vainik (1995), kelle põhiprobleemiks oli eesti keele väliskohakäänete polüseemia. Varasemates töödes on tähelepanu pälvinud adessiivi ja allatiivi abstraktsed kasutusjuhud, näiteks omaja ja kogeja vormistamiseks. Liina Lindström ja Virve-Anneli Vihman (2017) ning Lindström jt (2014) on vaadelnud predikaatidega *tarvis/vaja olema* moodustatud konstruktsioonide varieerumist ja leidnud, et kogejat tähistatakse sagedamini adessiiviga kui allatiiviga. Lindströmi ja Vihmani (2017: 816) uurimus lisab kaalu väitega, et „adessiiv on eesti keeles rohkem grammatiseerunud kui argumenti markeeriv kääne võrreldes rohkem semantilise allatiivi käändega". Lindström ja Ilona Tragel (2007, 2010) on vaadelnud adessiivi rolli impersonaali ja seisundipassiivi konstruktsioonides. Daativile sarnaseid adessiivi funktsioone on kirjeldanud Kazuto Matsumura (1994). Mõnevõrra eraldi seisev, kuid samuti huvitav uurimisteema on väliskohakäänded ja nendega paralleelselt kasutatavad kaassõnad. Eesti keele grammatika kohaselt on kaassõnade tähendus konkreetsem ja täpsem kui kohakäänete tähendus (EKG I: 33–34; Erelt jt 2007: 191). Saami ja soome keele kohta on Raija Bartens (1978) ja Krista Ojutkangas (2008) leidnud, et analüütiline kaassõnakonstruktsioon võrreldunа sünteetilise käändekonstruktsiooniga rõhutab rohkem asukohta ja seda kasutatakse koos väiksemate,

käega liigutatavate esemetega. Sarnasele tulemusele on jõudnud ka uurimused, kus on võrreldud adessiivi ja kaassõna *peal* kasutust kirjakeeles (Klavan 2012, 2017) ja eesti keele murretes (Klavan jt 2015).

Tulles väliskohakäänetega seotud sagedusandmete juurde, on oluline mainida üldist eesti keele käänete sagedusloendite kohta käivat infot. Kõige asjakohasemad on ehk kaks allikat: Jüri Valge 1970. aastast pärinev uurimus eesti keele käänete sagedusest ning Tartu Ülikooli Tasakaalus korpuse põhjal koostatud käändsõna grammatiliste kategooriate sagedusloendid (Sõnaliikide sagedusloend...). Valge püüdis oma uurimuses teha pärast statistilist analüüsi järeldusi eesti keele käänete süsteemi kvantitatiivsete iseärasuste kohta, analüüsides kolme funktsionaalset stiili: ilukirjanduslikku proosat (autorikõne), kõnekeelt draamateoste tegelaste kõne alusel ja ajalehekeelt. Valge (1970: 146–147) uurimuse valimi moodustab kokku 15 000 käändsõna, 5000 käändsõna iga funktsionaalse stiili kohta (ilukirjandusliku proosa ja draamateoste osas on mõlemast analüüsitud viie autori teost, 1000 sõna autori kohta). Analüüsi tulemused on esitatud iga funktsionaalse stiili kohta eraldi, arvutatud on iga käände suhtelise sageduse nii kolmes erinevas stiilis üldiselt kui ka iga autori kohta. Valge statistilise analüüsi tulemused näitavad, et adessiiv asub kõigis kolmes funktsionaalses stiilis käänete esinemissageduste järgi neljandal kohal; stabiilne on ka ablatiivi positsioon, mis kuulub koos essiivi, translatiivi ja abessiiviga nelja kõige vähem kasutatava käände hulka. Allatiivi positsioon sõltub funktsionaalsest stiilist: ilukirjanduse ja ajalehekeele valimites asetub see 7. kohale, kõnekeeles aga 5. kohale. (Valge 1970: 151–156) Tasakaalus korpuse põhjal tehtud käändekategooria sagedusloend annab samalaadse järjestuse: adessiiv asub Tasakaalus korpuses tervikuna 4. kohal, allatiiv 7. kohal ja ablatiiv viimase nelja käände hulgas (Sõnaliikide sagedusloend...: tabel 9).

Peale eesti keele käänete väga stabiilse esinemissageduse erinevates stiilides (nt adessiiv ja ablatiiv) näitab Valge (1970) uurimus ka teatud käänete varieeruvust (nt nimetav, omastav ja osastav), mille põhjal saab eristada nii funktsionaalseid stiile kui ka autoreid. Olulise järeldusena märgib ta, et draamateostes, mida kasutati kõnekeele esindajana, avaldab kõnele mõju autori stiil, mistõttu „oleks kõnekeele uurimisel otstarbekas kasutada lindistatud (juhuslikku) kõnet" (Valge 1970: 161). Siinses artiklis ei ole küll kasutatud juhuslikku kõnet, kuid eeldatavasti on raadiosaadetes tegemist poolspontaanse suulise keelega, mis erineb ilukirjanduse keelest, sh draamateoste kõnest. On selge, et keelenähtuste statistilisel uurimisel on väga olulisel kohal materjali representatiivsus ja tulemuste usaldatavus.

Artikli järgnevates osades keskendume kõigepealt eesti keele väliskohakäänete olemusele ja nende erinevatele funktsioonidele. Seejärel kirjeldame 2. osas poolspontaanse kõne automaatset transkriptsiooni ja morfoloogilist märgendamist. 3. osas analüüsime sagedusandmete põhjal väliskohakäänete kasutust poolspontaanses kõnes. 4. osas arutame väliskohakäänete rolli üle kohasuhte väljendamisel.

1.1. Eesti keele (välis)kohakäänded

Eesti keele väliskohakäänded moodustavad kolmest liikmest koosneva kohakäänete sarja: allatiiv, adessiiv ja ablatiiv. Peale väliskohakäänete kuuluvad eesti keele kohakäänete süsteemi veel sisekohakäänded, mis moodustavad samuti omaette sarja: illatiiv, inessiiv ja elatiiv. „Eesti keele grammatika" järgi on kohakäänete esmane funktsioon vormistada kohta väljendavat adverbiaali või atribuuti. Peale kohatähenduse tarvitatakse kohakäändeid ka aega väljendavates substantiivides, nt *sel nädalal*, *tol hetkel*, *eelmisel aastal*. Veel võivad kohakäänetes vormistatud sõnad väljendada „spetsiifilisi, kontekstuaalselt avalduvaid tähendusi". (EKG I: 54–55) Väliskohakäänded kuuluvad kohustusliku laiendi vormina mitmesse rektsioonistruktuuri, mistõttu vastavad nad paljudel juhtudel indoeuroopa keelte daativi funktsioonidele (Vainik 1995: 162). Mõne uurija jaoks näib väliskohakäänete tähendus olevat nii abstraktne, et neid on raske kohakääneteks pidada (nt Matsumura 1994).

Siinse kirjutise eesmärk ei ole analüüsida eesti keele väliskohakäänete tähenduste omavahelist seotust ja nende ajaloolist teket. Eesti keele väliskohakäänete polüseemia kohta võib pikemalt lugeda Ene Vainiku monograafiast, mis käsitleb väliskohakäänete semantikat just kognitiivse grammatika vaatenurgast. Oluline on rõhutada, et ka siinne artikkel võtab käände kui grammatilise kategooria käsitluses eeskujuks kognitiivse semantika ja lähtub arusaamast, et väliskohakäänetel on eksplitsiitselt kirjeldatavad tähendused, mis moodustavad sisemiselt liigendatud tähenduskimbu (Vainik 1995: 164). Tähenduskimp on radiaalse iseloomuga ning selles esinevad eri tähendused on vähem või rohkem prototüüpsed. See on ka põhjus, miks empiirilise materjaliga töötades ei ole alati selge, millist konkreetset tähendust väliskohakäändega parasjagu väljendatakse, kuna tähenduste vahele ei saa alati selget piiri tõmmata. Lokalismi teooria järgi (ingl *localist theory*; Andersson 1971, 2006: 95–96; Lyons 1977: 718–724) võime eeldada, et väliskohakäänete kohatähendused on esmasemad kui abstraktse(ma)d tähendused. Nii näiteks põhineb eesti keele omaja konstruktsioon, kus omaja on väljendatud adessiiviga, lokatiivsel skeemil: *Y asetseb X-l > X-l on Y* (Heine 1997: 47); adessiivi kasutus omaja väljendamiseks on kujunenud välja adessiivi kohasuhte väljendamise kasutusest.

1.2. Eesti keele väliskohakäänete funktsioonid

Artikli eesmärk on uurida eesti keele väliskohakäänete funktsioonide sagedusjaotust poolspontaanses kõnes. Selleks on vajalik esmalt anda ülevaade, millised on väliskohakäänete erinevad funktsioonid. Funktsioonide loetlemisel on aluseks võetud „Eesti keele grammatika" (EKG I); empiirilises osas on loetelusid lihtsustatud, sest artikli eesmärk on võrrelda võrrelda kohasuhte väljendamist teiste kohakäänete funktsioonidega.

Alaleütleva käände ehk allatiivi funktsioonid (EKG I: 58):

a) latiivse[1] tähendusega koht: *Panin raamatu lauale*;
b) latiivse tähendusega aeg: *Asja otsustamine lükati järgmisele päevale*;
c) latiivse tähendusega olukord või seisund: *Ilm kisub sajule*;
d) piir (sünonüümne terminatiiviga): *Temperatuur langes 30 kraadile*;
e) adressaat (ühenduses teatud kindlate verbidega): *Mees kinkis naisele lilli*;
f) kogeja (ühenduses teatud kindlate verbidega): *Mulle meeldib see raamat*;
g) sünonüümne kaassõnaühenditega 'substantiivi genitiivivorm + *peale/vastu/jaoks*': *Ära ole mulle kuri*;
h) ilma selge tähenduseta (teatud kindlate verbide rektsioonilise laiendina): *Manitsused ei lugenud poisile midagi*.

Alalütleva käände ehk adessiivi funktsioonid (EKG I: 58–59):

a) lokatiivse tähendusega koht: *Raamat lebas laual*;
b) lokatiivse tähendusega aeg: *Tule järgmisel kolmapäeval meile*;
c) lokatiivse tähendusega olukord või seisund: *Pildilt vaatas naerul nägu*;
d) viis või määrus: *Tüdruk vaatas meid suuril silmil*;
e) vahend: *Ma oskan klaveril ainult paari lihtsat viisijuppi mängida*;
f) infiniittarindite ja nominalisatsioonide tegevussubjekt: *Ema käskis lastel tuppa tulla*;
g) omaja (verbi *olema* laiendina): *Poisil on kaks koera*;
h) finiitses vormis predikaadi tegevussubjekt (teatud kindlate verbide rektsioonilise laiendina): *See asi ununes mul täielikult*;
i) ilma selge tähenduseta (teatud kindlate verbide rektsioonilise laiendina): *Millel su arvamus tugineb?*

Alaltütleva käände ehk ablatiivi funktsioonid (EKG I: 59):

a) separatiivse tähendusega koht: *Raamat kukkus laualt põrandale*;
b) separatiivse tähendusega aeg: *Koosolek lükati reedelt esmaspäevale*;
c) teate või asja lähteallikas (ühenduses teatud kindlate verbidega): *Emalt varastati trammis vihmavari*;
d) sünonüümne kaassõnaühenditega 'substantiivi genitiivivorm + *poolest*': *Ta on ametilt torulukksepp*.

[1] Artiklis kasutatakse „Eesti keele grammatika" (EKG I) termineid *latiivne*, *lokatiivne* ja *separatiivne*, mis viitavad vastavalt sihtkohale, asukohale ja lähtekohale.

2. Poolspontaanse kõne automaatne transkriptsioon ja morfoloogiline märgendamine

Töö praeguses etapis oli eesmärk kasutada kirjeldavat statistikat (peamiselt sagedusandmeid) hindamaks, kas automaatse kõnetuvastuse ja morfoloogilise märgendamise kasutamine sellisel kujul, nagu seda on tehtud siinses artiklis, lubab tulevikus ka ulatuslikumaid kvantitatiivseid ja kvalitatiivseid uurimusi. Kokku on automaatselt transkribeeritud materjali 2681 raadiosaate jagu (15 318 158 transkribeeritud tekstisõna). Materjali kokkupanemisel võtsime episoode viiest raadiosaatest, täpsemad andmed on esitatud tabelis 1. Käsitsi märgendatud väliskohakäänete analüüs põhineb pilootuurimusel, mille valimisse kuulub 15 raadiosaate episoodi (kokku 101 575 transkribeeritud tekstisõna).

Tabel 1. Poolspontaanse kõne (raadiosaadete) materjal.

Saade	Episoodide/failide arv	Transkribeeritud sõnade arv
„Huvitaja”	531	1 583 007
„Kultuurikaja”	332	1 647 034
„Räägivad”	314	2 104 949
„Rahvateenrid”	364	2 806 265
„Reporteritund”	1140	7 176 903
Kokku	2681 faili	15 318 158 sõna

Raadiosaadete valiku oluline kriteerium oli see, et kõnelejaid oleks rohkem kui üks ja tegemist oleks võimalikult spontaanse kõnega. Andmeanalüüsi tehti järgmistes etappides: 1) andmete transkribeerimine automaatset kõnetuvastust kasutades; 2) andmete automaatne morfoloogiline märgendamine; 3) vajalike andmete (väliskohakäänete kasutus) väljavõtmine kogu andmestikust; 4) pilootuurimuse valimi moodustamine; 5) väliskohakäänete (funktsioonide) kasutussagedustel põhinev analüüs. Allpool kirjeldame igat andmeanalüüsi etappi veidi lähemalt.

2.1. Raadiosaadete automaatne transkriptsioon

Raadiosaadete transkribeerimiseks kasutatav kõnetuvastussüsteem koosneb mitmest komponendist (Alumäe jt 2018). Transkribeerimise esimene samm on töödeldava kõnesalvestuse segmenteerimine kuni 20 sekundi pikkusteks kõnelõikudeks ning saadud kõnelõikude klasterdamine kõneleja järgi (ingl *speaker diarization*). Kõnetuvastuse esimeses faasis genereeritakse igale kõnelõigule kõige tõenäolisemaid tuvastustulemusi kodeeriv sõnade graaf, kasutades närvivõrgupõhist akustilist mudelit ning suhteliselt väikest statistilist keelemudelit. Saadud sõnade graaf töödeldakse kõigepealt suurema statistilise keelemudeli ning seejärel rekurrentsel närvivõrgul põhineva keelemudeli abil. Seejärel leitakse graafist parima skooriga tee, mis vastabki lausungi tuvastustulemusele. Tundmatud (st keelemudelis mitteolevad) sõnad tuvastatakse häälikutel põhineva n-gramm-mudeli abil. Sellised sõnad on näiteks keelemudelisse veel mitte jõudnud nimed, paljude harvemini esinevate nimi-

sõnade „eksootilisemad" käändevormid jms. Kõnetuvastuse sõnavigade määr raadio vestlussaadetes on umbes 9%.

Tuvastatud tekst järeltöödeldakse automaatse kirjavahemärgistamise tehnoloogia abil, mis oskab tuvastatud tekstis leida kõige tõenäolisemad lausepiirid, leida lausepiirile parim kirjavahemärk (punkt või küsimärk) ning lisada sobivatesse kohtadesse komad (Tilk, Alumäe 2016). Lingvistiliste reeglite asemel kasutatakse siingi sügavaid kahesuunalisi rekurrentseid närvivõrke, mis on treenitud suurte tekstikogumite peal. Peale selle kasutab mudel sõnadevaheliste pauside pikkuse infot, toetudes intuitsioonile, et pikemate pauside kohal on lause lõpu ja koma tõenäosus suurem. Kirjavahemärkide lisamise täpsus ja saagis on u 80%, st u 20% kirjavahemärke jääb leidmata ja u 20% süsteemi pakutud kirjavahemärke on valed.

Töös kasutatud kõnesalvestuste transkribeerimise süsteem on avalikult ja tasuta saadaval veebirakendusena (Veebipõhine kõnetuvastus). Kuna salvestuste hulk oli suhteliselt suur, kasutasime transkribeerimisel selleks eksperimendiks kohandatud lahendust, mis võimaldas transkribeerida kõik saatesalvestused kõrge võimekusega arvutiklastri abil kõigest mõne päevaga. Näites 1 on esitatud väljavõte automaatselt transkribeeritud „Huvitaja" saate episoodist „Tehnoloogia aasta 2017".

(1)

1 Vikerraadio tere, kell on neli ja pool minutit kümme läbi algab aasta viimane tehnoloogiateemaline Huvitaja ja tänases saates me võtamegi tehnoloogia aasta kokku räägime virtuaalsete assistentide võidukäigust, krüptorahast, aga ka tõe ja valepiiride hägustumisest tehnoloogia vallas. Aga need on vaid mõned märksõnad, mis jäävad iseloomustama. Lõpevad tehnoloogia aastat, kokkuvõtteid teeme koos tehnoloogia asjatundja Sander Saarega ja saate teises pooles tuleb nagu reedeti ikka aastat kokkuvõtvate uudistega stuudiosse Jakob Rosin, mina olen saatejuht Juhan Kilumets, head kuulamist.

2 Aasta viimase tehnoloogiateemalise Huvitaja saate esimene külaline on tehnoloogiaentusiasti asjatund ja Sander Saar, tere Sander. Tere. Tere. Sa oled Huvitaja saates ka enne külas käinud, aga nüüd on sul auväärne ülesanne siis minu abiga võtta see lõppev tehnoloogia aasta kokku tuua välja mõned märksõnad, mõned suunad, midagi olulisemat. Hakkame kohe ühest otsast minema, ma küsin sissejuhatuseks nii üldiselt kui üldse saab küsida, milline see aasta on tehnoloogia mõttes olnud tavalise inimese jaoks, neid tehnoloogilisi arenguid tuleb nii palju ja nii kiiresti, et see kõik pea totaalselt ringi käima, et kas siin kahe tuhande seitsmeteistkümnes aasta ei erinenud millegi poolest, et kõik jätkub.

3 Ma arvan täpselt samamoodi nagu tavatarbija tunnevad ka tehnoloogiaentusiastid ja, ja kirglikud jälgijaid, et uusi tehnoloogiaid-lahendusi ja murranguid on praegu toimuvas nii palju ja nii kiiresti enamus need arengud ei ole mittelineaarsed, nii-öelda, mida me saame etteennustatavaid eksponentsiaalselt, nad kasvavad ruttu ja tekivad ruttu ja mõnedki neist kaovad ruttu. Et see aasta on kindlasti olnud taaskord väga-väga uudisterohke ja tormiline tehnoloogia valdkonnas.

4 Ma palusin sul mõelda teemadele, mis sind ennast huvitavad ja mis jäävad seda aastat iseloomustama, on iseloomustanud. Hakkame peale virtuaalsete assisten-

tide läbimurre, oled ka sina meie saates virtuaalsetes tassistentidest juttu teinud, mis siis nüüd sel lõppeval aastal juhtus, selles vallas?

5 Kaks tuhat seitseteist. Aasta ilmselt võib juba tagasiulatuvalt pidada tõeliseks murranguks virtuaalassistentide seas, eriti vaadates Amazoni poolt looduda leksad ja Google'i poolt loodud Google assistenti. Eelkõige sellepärast, et need seadmed, mis võimaldavad siis tavatarbijal kodus kasutada virtuaalset assistenti, on omahinnas tulnud alla kümnekordselt. Ehk kui nad esimest korda turule tulid näiteks Amazoni Eco seade, mis võimaldas Reksat kasutada, maksis kakssada dollarit siis täna need seadmed on pühade ajal tulnud müüki kahekümne üheksa dollariga ehk siis need on muutunud nii kättesaadavaks ja, ja see on paljude inimeste kinkekoti jõudnud ja Amazon on ka teada andnud, et Seli esimene aasta küll täpselt numbreid täpsustamata, kus nende müük ulatus kümnete miljonite seadmeteni, mida siis koguarvus võib pidada, et kasutajate arv ilmselt jõuab siis sadadesse miljonitesse, mis muudab sellest tehnoloogiast ja eksis häälegi juhitavast virtuaalsest asist, nendest kui platvormist tõeliselt tõsiseltvõetava lahenduse.

6 Sinul neid virtuaalseid assistente on olnud erinevaid, sa oled proovinud ja oled seda usku, mees. Miks sa oled seda usku, mida siis see virtuaalne assistent nüüd aastal kaks tuhat seitseteist pakkuda? Tegelikult suudab?

2.2. Andmete morfoloogiline märgendamine

Automaatselt transkribeeritud teksti analüüsisime morfoloogiliselt Vabamorfi abil EstNLTK versiooni 1.6 kaudu (vt ka Orasmaa jt 2016). Kasutasime parameetrite vaikeväärtusi, sh tekstitasemel ühestamist, liitsõnaosade märgendamist ning tundmatute vormide ja pärisnimede oletamist (guess = True, propername = True, disambiguate = True, compound = True, phonetic = False). Analüüsiks pärisime tulemuse hulgast nimisõna või asesõna[2] ainsuse või mitmuse allatiiviks, adessiiviks või ablatiiviks märgendatud sõnu koos sõnaliikide ja neid sisaldanud lausega. See päring ehk andmete lugemine EstNLTK andmestruktuurist pärast teksti morfoloogilist analüüsi võttis aega ootamatult kaua: 2681 failis sisalduva 15 318 158 tekstisõna peale kokku üle 130 tunni. Põhjust ei õnnestunudki välja selgitada, ka EstNLTK autorite abil mitte. Rohkem tundusid aega võtvat pikemates lausetes asuvad sõnad, mis ongi üks võimalik aegluse põhjuse hüpotees. Nagu näitest 1 näha, on transkriptsioonis kohtadel, kuhu harjumuspärases kirjalikus tekstis tahaks panna lauselõpumärgi, sageli kas koma või pole üldse märki, mis annab tulemuseks ebatavaliselt pikki lauseid.

Tekstitasemel ühestamine töötas pistelise kontrolli põhjal ootuspäraselt paremini kui analüüs ilma ühestamiseta. Sobivateks vormideks lugesime kõik tekstisõnad, millel oli analüüside hulgas märgend [sg all], [pl all], [sg ad], [pl ad], [sg abl], [pl abl] ja sõnaliikide hulgas _s_ või _p_. Mitmesusi ehk mitme märgendiga vorme jäi sisse väga vähe. Analüsaator ei osanud neis määrata arvu, nt „kellele" [_p_ pl all, _p_ sg all] või sõnaliiki, nt „ühelt" [_n_ sg abl, _p_ sg abl], „teisele" [_o_ sg all, _p_ sg all],

[2] Otsustasime esialgu analüüsida nimi- ja asesõnu, kuid edaspidises töös tasuks vaadata ka omadussõnu ja arvsõnu.

või oli analüsaatori leksikonis mitu vormiga sobivat lemmat, nt „kevadel" [_s_ sg ad, _s_ sg ad] („kevad" ja „kevade"), „kohal" [_s_ sg ad, _s_ sg ad] („koht" ja „koha"). Kuna kääne oli mitmestel analüüsidel sama, siis edasist uurimist need mitmesused ei mõjutanud.

3. Väliskohakäänete kasutus poolspontaanses kõnes

Alustame väliskohakäänete sagedusandmete kirjeldamist kogu materjalis (15 318 158 tekstisõna) esinenud kasutuste põhjal. Tabelis 2 on toodud väliskohakäänete absoluutne ja suhteline sagedus poolspontaanses kõnes. Absoluutne sagedus näitab, mitu korda kääne või käändefunktsioon korpusvalimis esines. Suhteline sagedus võimaldab võrrelda keelenähtuste kasutust erineva sõnade arvuga korpustes. Suhteline ehk normaliseeritud sagedus on arvutatud järgneva valemi põhjal (Brezina 2018: 43):

$$suhteline\ sagedus = \frac{absoluutne\ sagedus}{sõnade\ arv\ korpuses} \times normaliseerimisbaas$$

Tabelist 2 on näha, et andmed kinnitavad kenasti varasemate uurimuste tulemusi väliskohakäänete järjestuse kohta (nt Valge 1970), kus teistest märksa sagedamini esineb adessiiv, millele järgneb allatiiv ja märkimisväärselt vähem esineb ablatiivi kasutust. Et natukene paremini aimu saada väliskohakäänete sagedusandmetest poolspontaanses kõnes ja kuidas see sarnaneb või erineb ajakirjanduse, ilukirjanduse ja teaduse tekstides, panime Tasakaalus korpuse põhjal tehtud sagedustabelite andmetest (vt Sõnaliikide sagedusloend) kokku ülevaatliku tabeli (tabel 3), kus on esitatud väliskohakäänete absoluutne ja suhteline sagedus Tasakaalus korpuses kokku ja igas kolmes tekstiliigis eraldi.

Tabel 2. Väliskohakäänete sagedusjaotus poolspontaanses kõnes (suhteline sagedus on arvutatud 1 000 000 sõna kohta, korpuse kogumaht 15 318 158 sõna).

Väliskohakääne	Absoluutne sagedus	Suhteline sagedus
Allatiiv	172 540	11 263,8
Adessiiv	294 727	19 240,4
Ablatiiv	22 318	1456,9

Tabelist 3 on näha, et väliskohakäänete sagedus erineb tekstiliigiti (sellele juhtis tähelepanu ka Valge (1970) uurimus): adessiivi esineb ajakirjanduses ja teadustekstides märkimisväärselt rohkem kui ilukirjanduses, esineb jällegi ilukirjanduses mõnevõrra rohkem kui ajakirjanduses või teaduses. Kõigi kolme väliskohakäände suhteline kasutussagedus poolspontaanses kõnes on mõnevõrra erinev nende käänete suhtelisest sagedusest Tasakaalus korpuses: kõiki kolme käänet esineb poolspontaanses kõnes vähem kui Tasakaalus korpuses üldiselt. Allatiivi ja ablatiivi suhteline sagedus poolspontaanses kõnes on pigem sarnane allatiivi ja ablatiivi suhtelise sagedusega teadustekstides. Ent adessiivi suhteline sagedus poolspontaanses kõnes on sarnane pigem adessiivi suhtelise sagedusega ilukirjanduses. Siinse arutelu kon-

tekstis tuleb siiski meeles pidada, et vaadatud on ainult nimisõnade ja pronoomenite kasutust väliskohakäänetega. Erineda võib ka viis, kuidas on saadud käesoleva uurimuse väliskohakäänete sagedused ja kuidas seda on tehtud Tasakaalus korpuse andmete põhjal. Mõnevõrra teistsuguse pildi saab ilmselt siis, kui lisada ka teised sõnaliigid, millel esineb väliskohakäände tunnus.

Tabel 3. Väliskohakäänete sagedusjaotus Tasakaalus korpuses[3] (suhteline sagedus on arvutatud 1 000 000 sõna kohta, korpuse kogumaht 15 000 000, osakorpuste maht 5 000 000 sõna).

Väliskoha-kääne	Tasakaalus korpus kokku		Ajakirjandus		Ilukirjandus		Teadus	
	Abs. sagedus	Suht. sagedus	Abs. sagedus	Suht. sagedus	Abs. sagedus	Suht. sagedus	Abs. sagedus	Suht. sagedus
Allatiiv	210 541	14 036,1	70 427	14 085,4	75 200	15 040,0	64 914	12 982,8
Adessiiv	322 328	21 488,5	121 192	24 238,4	91 434	18 286,8	109 702	21 940,4
Ablatiiv	28 176	1878,4	10 567	2113,4	9765	1953,0	7844	1568,8

3.1. Pilootuuringu valimi moodustamine

Väliskohakäänete funktsioonide analüüsimiseks on vaja materjal käsitsi märgendada. Selleks korraldasime pilootuuringu, mille tarvis valisime 2681 automaatselt transkribeeritud raadiosaatefaili hulgast viis pikemat episoodi saadetest „Huvitaja", „Kultuurikaja" ja „Räägivad" (kokku 101 575 tekstisõna) ja analüüsisime neis väliskohakäänete kasutust käsitsi. Valimi täpsemad andmed on esitatud tabelis 4.

Kokku analüüsisime pilootuuringu raames käsitsi 3566 väliskohakäände kasutust. Oluline on veel kord rõhutada, et analüüsisime ainult nimisõnade ja pronoomenite kasutust, välja jäid pärisnimed ja muud sõnaliigid. Tabelist 5 on näha, et väliskohakäänete enamuse moodustasid adessiivi kasutused (2147 kasutust), millele järgnesid allatiiv (1214 kasutust) ja ablatiiv (205 kasutust).

Pilootuuringu valim on käsitsi puhastatud. Kuna sagedusandmed kajastavad nii nimisõnalisi kui ka asesõnalisi kasutusi, on kogu materjali (ja eeldatavasti ka Tasakaalus korpuse) sagedusandmetes korduvkasutusi. Võtame ühe näidislause, kus on kolm adessiivset fraasi alla joonitud: *meil*, *Läti riigi sellisel kõige hapramal hetkel*, *sellisel viisil* (näide 2).

(2) Meil on siin kõrval naaberriik Läti, kus nüüd kui ma ei eksi, siis umbes poolteist aastat tagasi Läti riigi sellisel kõige hapramal hetkel tegelikult parlamenti otsustas haldusterritoriaalse reformi ära sellisel viisil („Räägivad riigist, omavalitsusest ja kodanikust")

[3] Tabelis 3 on autorite kohandatud algsed andmed, mis pärinevad Tasakaalus korpuse põhjal koostatud sagedustabelist sõnaliigi, arvu ja käänete kaupa (Sõnaliikide sagedusloend...: tabel 3). Oleme esialgsed absoluutsed sagedused summeerinud järgnevate vormide põhjal: [_p_ pl abl], [_p_ pl ad], [_p_ pl all], [_p_ sg abl], [_p_ sg ad], [_p_ sg all], [_s_ pl abl], [_s_ pl ad], [_s_ pl all], [_s_ sg abl], [_s_ sg ad], [_s_ sg all].

Tabel 4. Pilootuuringu valim.

Saade	Episoodi pealkiri	Transkribeeritud sõnade arv
„Huvitaja”	„Tehnoloogia-aasta 2017”	7417
	„Öko- ja köögikosmeetika”	6919
	„Eesti oma supertoit ja santpooliad”	6851
	„EV100. Meditsiin 100 aastat tagasi”	6124
	„Idufirmade finantspool”	5806
Kokku		33 117
„Kultuurikaja”	10.03.2012	7681
	24.09.2011	5877
	21.04.2007	5296
	23.03.2013	5083
	15.12.2007	6362
Kokku		30 299
„Räägivad”	„Kes me oleme?”	8314
	„Mis saab sotsiaaldemokraatiast?”	6823
	21.10.2005	7449
	„Räägivad erakonnad”	7565
	„Räägivad riigist, omavalitsusest ja kodanikust”	8008
Kokku		38 159
Kokku		101 575

Tabel 5. Väliskohakäänete sagedusjaotus raadiosaadete valimis (valimi suurus 101 575 sõna).

Väliskohakääne	Absoluutne sagedus	Suhteline sagedus (1 000 000 sõna kohta)
Allatiiv	1214	11 951,8
Adessiiv	2147	21 137,1
Ablatiiv	205	2018,2

Pilootuuringu adessiivi sagedusloendis esineb sellest lausest kolm kasutust: *meil* (tegevussubjekt), *hetkel* (aeg) ja *viisil* (adverbiaalne). Kuna pilootuuringus oli oluline kohakäände semantiline funktsioon, ei oleks olnud mõtet analüüsida *sellisel hetkel* kaheks erinevaks aja kasutuseks või *sellisel viisil* kaheks adverbiaalse kasutuse esinemiseks. Kogu materjali põhjal tehtud sagedusandmetes on aga näitest 2 võetud viis adessiivi kasutust: *meil* (pronoomen), *sellisel* (pronoomen), *hetkel* (substantiiv), *sellisel* (pronoomen), *viisil* (substantiiv). Pärast andmete puhastamist jäi analüüsitava materjali kogumahuks 2848 väliskohakäände kasutust, millest adessiivi kasutusi oli 1687, allatiivi 990 ja ablatiivi 172. Puhastatud andmete pealt on näha, et kuigi kohakäänete suhteline sagedus on madalam, kehtib endiselt kohakäänete sagedusjärjestus: adessiiv on märkimisväärselt sagedasem kui allatiiv ja ablatiivi esineb kolmest käändest märkimisväärselt kõige harvem.

Tabelis 6 on toodud väliskohakäänete esinemissagedus kolmes raadiosaates. Kuigi üldiselt on kohakäänete suhteline sagedus raadiosaadete lõikes sarnane, esineb siin mõningaid erinevusi. Märgata võib näiteks ablatiivi vähest kasutust „Huvitaja” saates, samal ajal kui adessiivi kasutus selles saates on sagedasem kui kahes teises

saates. Erinevusi esines kolme raadiosaate lõikes ka kohakäänete funktsioonides, mille juurde tuleme tagasi artikli 4., aruteluosas.

Tabel 6. Väliskohakäänete sagedusjaotus pilootuuringu valimis raadiosaadete lõikes.

Väliskohakääne	Saade	Absoluutne sagedus	Suhteline sagedus (1 000 000 sõna kohta)
Allatiiv	„Huvitaja”	279	8493,4
	„Kultuurikaja”	334	11 090,5
	„Räägivad”	377	9991,8
	Kokku	990	9746,5
Adessiiv	„Huvitaja”	603	18 356,7
	„Kultuurikaja”	500	16 602,5
	„Räägivad”	584	15 477,9
	Kokku	1687	16 608,4
Ablatiiv	„Huvitaja”	43	1309,0
	„Kultuurikaja”	54	1793,1
	„Räägivad”	75	1987,8
	Kokku	172	1693,3

3.2. Väliskohakäänete funktsioonide sagedusjaotus poolspontaanses kõnes

Kuna pilootuuringu peamine eesmärk oli keskenduda just kohafunktsiooni osakaalule võrreldes väliskohakäänete teiste funktsioonidega, otsustasime märgendamisskeemi lihtsustada võrreldes „Eesti keele grammatika” käsitlusega (EKG I). Eristasime kolm funktsiooni: aeg, koht ja muud kasutused.

Tabelis 7 on toodud allatiivi kasutuse funktsioonide sagedusjaotus. Sihtkoha väljendamine ei ole allatiivi peamisi funktsioone, moodustades 15% kasutusjuhtudest. Väga väikse osakaaluga on latiivse tähendusega aja väljendamine (ainult kolm kasutust 990 juhu kohta). Muud kasutused, sh kogeja ja adressaadi väljendamine, moodustavad kokku 85% kasutusjuhtudest.

Tabel 7. Allatiivi funktsioonide sagedusjaotus poolspontaanses kõnes.

Käändefunktsioon	Absoluutne sagedus	Osakaal
Aeg	3	0%
Sihtkoht	147	15%
Muud kasutused	840	85%
Kokku	990	100%

Tabelis 8 on toodud adessiivi kasutuse funktsioonide sagedusjaotus. Sarnaselt allatiiviga on kõige sagedasemad funktsioonid kohakäände abstraktsed kasutused: aja väljendamine (29% kasutusjuhtudest) ja muud kasutused (sh omaja väljendamine, kokku 53% kasutusjuhtudest). Võrrelduna allatiivi funktsioonidega on adessiivi kasutuste hulgas veidi suurema osakaaluga koha väljendamine: 18% adessiivi kasutusjuhtudest.

Tabel 8. Adessiivi funktsioonide sagedusjaotus poolspontaanses kõnes.

Käändefunktsioon	Absoluutne sagedus	Osakaal
Aeg	481	29%
Koht	296	18%
Muud kasutused	910	53%
Kokku	1687	100%

Tabelis 9 on toodud ablatiivi kasutuse funktsioonide sagedusjaotus. Selgelt domineerib lähtekoha väljendamine (58% kasutusjuhtudest), millele järgnevad muud kasutused, sh teate või asja lähteallika väljendamine (42% kasutusjuhtudest).

Tabel 9. Ablatiivi funktsioonide sagedusjaotus poolspontaanses kõnes.

Käändefunktsioon	Absoluutne sagedus	Osakaal
Aeg	0	0%
Lähtekoht	100	58%
Muud kasutused	72	42%
Kokku	172	100%

4. Arutelu ja kokkuvõte

Artiklis lähtume kasutuspõhisest keeleteadusest, mis rõhutab keelekasutuse, sh just kvantitatiivsete sagedusandmete uurimist, et teha (kvalitatiivseid) järeldusi keelesüsteemi kohta. Uurimuse fookuses on eesti keele väliskohakäänete sagedusandmed igapäevases, poolspontaanses kõnes. Töös on kasutatud vabavaralist kõnesalvestuste transkribeerimise süsteemi (Veebipõhine kõnetuvastus). Uuringus kasutasime 2681 raadiosaate salvestusi, kokku 15 318 158 transkribeeritud tekstisõna. Materjali kogumisel võtsime episoode viiest raadiosaatest: „Huvitaja", „Kultuurikaja", „Räägivad", „Rahvateenrid", „Reporteritund". Raadiosaadete valiku kriteerium oli, et kõnelejaid oleks rohkem kui üks ja et tegu oleks võimalikult spontaanse kõnega. Kuna raadiosaadete hulk oli suhteliselt suur, kasutasime seetõttu transkribeerimisel selleks eksperimendiks kohandatud lahendust, mis võimaldas transkribeerida kõik saatesalvestused kõrge võimekusega arvutiklastri abil mõne päevaga. Raadio vestlussaadete kõnetuvastuse sõnavigade määr on umbes 9%. Automaatselt transkribeeritud teksti analüüsisime morfoloogiliselt Vabamorfi abil EstNLTK 1.6 kaudu. Väliskohakäänete sagedusandmete analüüsiks valisime kõik tekstisõnad, millel oli analüüside hulgas vorm [sg all], [pl all], [sg ad], [pl ad], [sg abl], [pl abl] ja mis kuulusid substantiivide (S) või pronoomenite (P) sõnaliiki. Kuna väliskohakäänete eri funktsioonid tuli märgendada käsitsi, moodustasime pilootuuringu jaoks valimi, millesse kuulub 15 episoodi (kokku 101 575 tekstisõna).

Artiklis tutvustatud pilootuuringu eesmärk oli heita pilk väliskohakäänete sagedusjaotusele poolspontaanses suulises kõnes. Erilist huvi pakkus väliskohakäänete kasutus kohasuhte väljendamisel. Kõige suurema kohasuhte väljendamise osakaaluga oli ablatiiv (58% kõigist ablatiivi kasutusjuhtudest), millele järgnesid adessiiv (18% kõigist adessiivi kasutusjuhtudest) ja allatiiv (15% kõigist allatiivi kasutusjuhtudest).

On selge, et allatiivil ja adessiivil on eesti keeles väga suur roll muude funktsioonide väljendamisel, nt kogeja ja adressaat ning tegevussubjekt (sh omaja). Suhtelise sageduse järgi võiks eeldada, et koha väljendamine on adessiivi funktsioonide seas suurema osakaaluga kui allatiivil (umbkaudu 2914 *vs.* 1447 kasutust miljoni sõna kohta). Sellise oletuse paikapidavuse kontrollimiseks oleks vaja teha teistlaadi empiiriline uuring. Väliskohakäänete rolli koha väljendamisel on keeruline hinnata, sest ei ole teada, kas väliskohakäänete suhteliselt madal kasutussagedus koha väljendamisel on tingitud sellest, et kohasuhet väljendatakse muude keeleliste vahenditega, näiteks kaassõnadega, või ei ole kohasuhte väljendamine üldiselt raadiosaadete materjalis sage. Võib eeldada, et ilukirjanduses esineb kohasuhte väljendamist rohkem ja sellest tulenevalt on sagedasem ka väliskohakäänete kasutus koha väljendamiseks. Huvitav oleks väliskohakäänete funktsioonide sagedusjaotuse võrdlemine eri registrite vahel. Adessiivi kohta on näiteks Klavan (2012: 104–108) morfoloogiliselt ühestatud korpuse ilu- ja ajakirjanduse allkorpuste põhjal esitanud järgmised sagedusandmed: 4168 adessiivi kasutusjuhust on kõige sagedasem aja väljendamine (1502 kasutust, 36%), sellele järgneb pea sama sagedasti tegevussubjekti väljendamine (1468 kasutust, 35%). Vähem sagedad on koha väljendamine (656 kasutust, 16%) ja adessiivi adverbilised kasutused (542 kasutust, 13%). Arvestades, et morfoloogiliselt ühestatud korpuse ilu- ja ajakirjanduse allkorpuste kogumaht on 215 000 sõna, saame adessiivi koha funktsiooni suhteliseks sageduseks 3051 kasutust miljoni sõna kohta. Raadiosaadete korpuses on see umbkaudu sama: 2914 kasutust.

Pilootuuringu põhjal võib järeldada, et üldiselt on kohakäänete suhteline sagedus kolme raadiosaate (tabel 6) lõikes sarnane. Ühe erinevusena selgus, et ablatiivi oli võrdlemisi vähe kasutatud saates „Huvitaja". Ka konkreetselt kohasuhte väljendamises võib märgata sama tendentsi. „Huvitaja" viies saates leidus kokku 925 väliskohakäände kasutust, millest 16% moodustas kohasuhte väljendamine (147 kasutust, suhteline sagedus umbes 45 kasutust 10 000 sõna kohta). Saadete „Kultuurikaja" ja „Räägivad" puhul moodustas kohasuhte väljendamine proportsionaalselt veidi suurema osa väliskohakäänete funktsioonidest. „Kultuurikaja" viies saates leidus kokku 888 väliskohakäände kasutust, millest 21% moodustas kohasuhte väljendamine (188 kasutust, suhteline sagedus umbes 62 kasutust 10 000 sõna kohta). „Räägivad" valimis oli jaotus järgmine: kokku 1036 väliskohakäände kasutust, 20% sellest moodustas kohasuhte väljendamine (208 kasutust, suhteline sagedus umbes 55 kasutust 10 000 sõna kohta).

Sellise väikese valimi pealt on raske midagi põhjapanevat järeldada, aga on selge, et kui vaadata kohakäänete kasutust ja keskenduda just kohasuhte väljendamisele, võib esineda raadiosaateid ja episoode, kus räägitakse „kohast" rohkem kui teistes saadetes. See tulemus haakub ka Valge (1970) uurimusega, kus oli käänete kasutuse sageduste kohta märgitud, et kõikumine on eri autorite vahel kohati väga suur.

Sellele, et eri žanrides on grammatiliste kategooriate kasutus mõnevõrra erinev, viitavad nii Tasakaalus korpuse põhjal tehtud üldised väliskohakäänete sagedusloendid (nt tabel 3 ülalpool, vt ka Sõnaliikide sagedusloend...), Valge (1970) uurimus kui ka siinne poolspontaanse kõne materjali uurimus. Kui vaadata väliskohakäänete üldisi sagedusi, siis nii siinse uurimuse kõnekeele tulemused kui ka varasemad kirja-

keele uurimused kinnitavad, et kõige sagedasem väliskohakääne on adessiiv (kõigi eesti keele käänete arvestuses sageduselt neljandal kohal), sellele järgneb allatiiv ning seejärel ablatiiv (viimane kuulub nelja kõige harvema eesti keele käände hulka). Ent suhteline sagedus on kolmel kohakäändel eri tekstitüüpide lõikes erinev. Märgatavalt sagedasem on näiteks adessiivi kasutus ajakirjanduses (umbes 24 000 kasutust miljoni sõna kohta) võrreldes poolspontaanse kõne (umbes 19 000 kasutust miljoni sõna kohta) ja ilukirjandusega (umbes 18 000 kasutust miljoni sõna kohta). Samal ajal võib sagedusloendite põhjal väita, et ilukirjanduses on allatiivi kasutus veidi sagedasem (umbes 15 000 kasutust miljoni sõna kohta) kui ajakirjanduses (umbes 14 000 kasutust miljoni sõna kohta) või poolspontaanses kõnes (umbes 11 000 kasutust miljoni sõna kohta). Ablatiivi esineb poolspontaanses kõnes märgatavalt vähem (umbes 1400 kasutust miljoni sõna kohta) võrreldes nii ilukirjanduse (umbes 1900 kasutust miljoni sõna kohta) kui ka ajakirjandusega (umbes 2100 kasutust miljoni sõna kohta). Üldistavalt võib väita, et väliskohakäändeid esineb raadiosaadete poolspontaanses kõnes märgatavalt vähem kui ajakirjandus- või ilukirjandustekstides.

Viimase arutelupunktina tõstatub küsimus väliskohakäänete üldisest sagedusest – tekstisõna sagedus (ingl *token frequency*) – ja nende käänete sõnatüübi sagedusest (ingl *type frequency*). Tekstisõna sagedus näitab, mitu korda esineb nähtus tekstis, meie uurimuse kontekstis – mitu korda esineb väliskohakääne koos nimi- või asesõnaga poolspontaanse kõne materjalis. Sõnatüübi sagedus näitab, kui palju erinevaid sõnu selle nähtusega tekstis esineb, meie uurimuse kontekstis – mitu erinevat nimi- või asesõna esineb väliskohakäänetega poolspontaanse kõne materjalis. Meie uurimuse tulemused kinnitavad varasemaid tulemusi: adessiivi tekstisõna sagedus on kordades suurem kui allatiivi tekstisõna sagedus. Meie uurimus näitab ka seda, et allatiivi sõnatüübi sagedus on võrreldes adessiivi sõnatüübi sagedusega suurem. Kokku esines poolspontaanses kõnes (kokku ~15 000 000 tekstisõna) umbes 295 000 adessiivi kasutust, kusjuures erinevate nimi- ja asesõnade arv on ligikaudu 11 000 (tekstisõnade ja sõnatüübi suhe on 0,04). Allatiivi esines kokku umbes 173 000 korda, erinevate nimi- ja asesõnade arv oli 14 000, mis on palju suurem kui adessiivil (tekstisõnade ja sõnatüübi suhe on 0,08). Mida suurem on tekstisõnade ja sõnatüübi suhe, seda suurem on leksikaalne varieeruvus nimi- ja asesõnade kasutuses kohakäänetega. Morfoloogiaalastest uurimustest on teada (nt Bybee 1995), et sõnatüübi sagedus on üks peamisi faktoreid, mis määrab morfoloogilise produktiivsuse. Kas ja kuidas seostub allatiivi ja adessiivi tekstisõnade ja sõnatüübi erinev suhe morfoloogilise produktiivsusega ja kas sellel näitajal võiks olla seos väitega, et allatiiv on semantilisem kääne kui adessiiv (vrd Lindström, Vihman 2017: 816), jääb edaspidiste uurimuste mõttekohaks. Tegemist on huvitava tendentsiga väliskohakäänete sagedusandmetes, mis vajab põhjalikku kvalitatiivset uurimust, et kvantitatiivsetest andmetest midagi teoreetiliselt usaldusväärset eesti keele väliskohakäänete süsteemi kohta järeldada.

Usume, et töö praegune eesmärk kasutada kirjeldavat statistikat (peamiselt sagedusandmed) hindamaks, kas automaatse kõnetuvastuse ja morfoloogilise märgendamise kasutamine võimaldab põhjalikumaid uurimusi, sai täidetud: uurimuse tulemusena esitasime mitmesuguseid väliskohakäänete sagedusandmeid, mis anna-

vad palju materjali edaspidiseks kvalitatiivseks uurimistööks. Uurimuse põhjal võime kinnitada, et kõnesalvestuste automaatne transkribeerimise süsteem ja transkriptsioonide automaatne morfoloogiline märgendamine töötavad piisavalt hästi, et selle põhjal uurida eesti keele morfosüntaksi eripära poolspontaanses kõnes.

Artikli valmimist on toetanud Eesti Teadusagentuur (PUT1358 „Mudelite loomine ja lõhkumine: Klassifitseerimismudelite valideerimine keeleteaduses") ja Euroopa Liit Euroopa Regionaalarengu Fondi kaudu (Eesti-uuringute Tippkeskus).

VEEBIVARAD

EstNLTK. Vabavara eestikeelsete tekstide töötluseks. https://github.com/estnltk/estnltk

Sõnaliikide sagedusloend ning käändsõna grammatiliste kategooriate sagedusloendid Tasakaalus korpuse põhjal. https://www.cl.ut.ee/ressursid/gram-kat

Vabamorf. Eesti keele morfanalüsaator. https://github.com/Filosoft/vabamorf

Veebipõhine kõnetuvastus. http://bark.phon.ioc.ee/webtrans

KIRJANDUS

Alumäe, Tanel; Tilk, Ottokar; Asadullah 2018. Advanced rich transcription system for Estonian speech. – Human Language Technologies: the Baltic Perspective. Proceedings of the Eighth International Conference, Baltic HLT 2018. (Frontiers in Artificial Intelligence and Applications 307.) Toim Kadri Muischnek, Kaili Müürisep. Amsterdam: IOS Press, lk 1–8.

Andersson, John M. 1971. The Grammar of Case: Towards a Localistic Theory. (Cambridge Studies in Linguistics 4.) Cambridge: Cambridge University Press.

Andersson, John M. 2006. Modern Grammars of Case. Oxford: Oxford University Press.

Arnon, Inbal; Snider, Neal 2010. More than words: Frequency effects for multi-word phrases. – Journal of Memory and Language, kd 62, nr 1, lk 67–82.

Bartens, Raija 1978. Synteettiset ja analyyttiset rakenteet Lapin paikanilmauksissa. (Suomalais-ugrilaisen seuran toimituksia 166.) Helsinki: Suomalais-ugrilainen seura.

Bod, Rens; Hay, Jennifer; Jannedy, Stefanie 2003. Introduction. – Probabilistic Linguistics. Toim R. Bod, J. Hay, S. Jannedy. Cambridge–Massachusetts–London: The MIT Press, lk 1–10.

Brezina, Vaclav 2018. Statistics in Corpus Linguistics: A Practical Guide. Cambridge: Cambridge University Press.

Bybee, Joan L. 1995. Regular morphology and the lexicon. – Language and Cognitive Processes, kd 10, nr 5, lk 425–455.

Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. – Language, kd 82, nr 4, lk 711–733.

Bybee, Joan L. 2007. Frequency of Use and the Organization of Language. Oxford: Oxford University Press.

Bybee, Joan 2010. Language, Usage, and Cognition. Cambridge: Cambridge University Press.

Bybee, Joan; Hopper, Paul J. (toim) 2001. Frequency and the Emergence of Linguistic Structure. (Typological Studies in Language 45.) Amsterdam: John Benjamins.

Diessel, Holger 2017. Usage-Based Linguistics. Oxford Research Encyclopedia of Linguistics. https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-363 (20. III 2020).

Divjak, Dagmar 2019. Frequency in Language: Context, Memory and Attention. Cambridge: Cambridge University Press.

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

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

Haspelmath, Martin 2006. Against markedness (and what to replace it with). – Journal of Linguistics, kd 42, nr 1, lk 25–70.

Hay, Jennifer 2001. Lexical frequency in morphology: Is everything relative? – Linguistics, kd 39, nr 6, lk 1041–1070.

Heine, Bernd 1997. Possession: Cognitive Sources, Forces, and Grammaticalization. (Cambridge Studies in Linguistics 83.) Cambridge: Cambridge University Press.

Klavan, Jane 2012. Evidence in Linguistics: Corpus-Linguistic and Experimental Methods for Studying Grammatical Synonymy. (Dissertationes linguisticae Universitatis Tartuensis 15.) Tartu: Tartu University Press.

Klavan, Jane 2017. Pitting corpus-based classification models against each other: A case study for predicting constructional choice in written Estonian. – Corpus Linguistics and Linguistic Theory. https://doi.org/10.1515/cllt-2016-0010

Klavan, Jane; Pilvik, Maarja-Liisa; Uiboaed, Kristel 2015. The use of multivariate statistical classification models for predicting constructional choice in spoken, non-standard varieties of Estonian. – SKY Journal of Linguistics, nr 28, lk 187–224.

Lindström, Liina; Tragel, Ilona 2007. Eesti keele impersonaali ja seisundipassiivi vahekorrast adessiivargumendi kasutamise põhjal. – Keel ja Kirjandus, nr 7, lk 532–553.

Lindström, Liina; Tragel, Ilona 2010. The possessive perfect construction in Estonian. – Folia Linguistica, kd 44, nr 2, lk 371–399.

Lindström, Liina; Vihman, Virve-Anneli 2017. Who needs it? Variation in experiencer marking in Estonian 'need'-constructions. – Journal of Linguistics, kd 53, nr 4, lk 789–822.

Lindström, Liina; Uiboaed, Kristel; Vihman, Virve-Anneli 2014. Varieerumine *tarvis-*/*vaja*-konstruktsioonides keelekontaktide valguses. – Keel ja Kirjandus, nr 8–9, lk 609–630.

Lyons, John 1977. Semantics. Kd 2. Cambridge: Cambridge University Press.

Matsumura, Kazuto 1994. Is the Estonian adessive really a local case? – Journal of Asian and African Studies, nr 46/47, lk 223–235.

Ojutkangas, Krista 2008. Mihin suomessa tarvitaan *sisä*-grammeja. – Virittäjä, nr 3, lk 382–400.

Orasmaa, Siim; Petmanson, Timo; Tkachenko, Alexander; Laur, Sven; Kaalep, Heiki-Jaan 2016. EstNLTK – NLP toolkit for Estonian. – Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož: European Language Resources Association, lk 2460–2466.

Tilk, Ottokar; Alumäe, Tanel 2016. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. – Proceedings of the INTERSPEECH 2016: Understanding Speech Processing in Humans and Machines. San Francisco: International Speech Communication Association, lk 3047–3051.

Vainik, Ene 1995. Eesti keele väliskohakäänete semantika kognitiivse grammatika vaatenurgast. Tallinn: Eesti Keele Instituut.

Valge, Jüri 1970. Eesti keele käänete sagedused kolmes funktsionaalses stiilis. – Keel ja struktuur 4. Tartu, lk 145–162.

Jane Klavan (snd 1983), PhD, Tartu Ülikooli anglistika osakonna inglise keele ja lingvistika lektor (Lossi 3, 51003 Tartu), jane.klavan@ut.ee

Tanel Alumäe (snd 1976), PhD, Tallinna Tehnikaülikooli tarkvarateaduse instituudi vanemteadur (Akadeemia tee 21B, 12618 Tallinn), tanel.alumae@taltech.ee

Arvi Tavast (snd 1969), PhD, Eesti Keele Instituudi direktor (Roosikrantsi 6, 10119 Tallinn), arvi@tavast.ee

Analysis of Estonian external locative cases in semi-spontaneous speech using an automatic transcription system

Keywords: corpus linguistics, semi-spontaneous speech, external cases, Estonian language

We proceed from the tenets of usage-based linguistics which stresses the importance of studying language use, especially quantitative frequency measures, in order to make (qualitative) inferences about linguistic knowledge. We focus on the frequency counts of Estonian external cases (allative, adessive, ablative) and their different functions in semi-spontaneous everyday speech. The automatic transcription system applied for the present study can be accessed free of charge via the web application http://bark.phon.ioc.ee/webtrans. In our study we used the recordings of 2,681 radio broadcasts containing 15,318,158 transcribed words in total. The average word error rate for the automatic transcription system is around 9%. The transcriptions were morphologically analysed via EstNLTK 1.6 using Vabamorf. For the analysis of exterior locative cases, we extracted data about nouns (S) and pronouns (P) whose word form included one of the following tags: [sg all], [pl all], [sg ad], [pl ad], [sg abl], [pl abl]. The second part of the study focused on the use of the different functions of the locative cases. Since the annotation of functions needs to be done manually, a smaller sample was analysed as a pilot study (15 broadcasts, 101,575 transcribed words in total).

The results of the present study confirm earlier results about the overall frequency of use of external locative cases. The most frequent case is the adessive (it ranks fourth in the overall ranking of Estonian cases), followed by allative and then ablative (the latter belongs to the four least frequently used Estonian cases). Very broadly speaking, our study indicates that in semi-spontaneous speech external cases are used less frequently than in newspaper texts and fiction. As for the different functions expressed by external locative cases, we were interested in finding out the overall proportion of uses where the cases express a spatial relation. As expected, ablative has the highest proportion of spatial relations (58% out of 172 uses), followed by adessive (18% out of 1,687 uses) and allative (15% out of 990 uses). It is clear that for the adessive and allative, expressing a spatial relation is not the most frequent function. These two cases carry other important functional loads in the Estonian language, e.g. expressing the experiencer, addressee or possessor. Very broadly, it seems that expressing a spatial relation is proportionally more frequent for the adessive case than for the allative case. Overall, our study presents a number of different frequency counts pertaining to the use of Estonian external locative cases, which can serve as input for further qualitative studies. Based on the results of the present study we confirm that using an automatic transcription system for recorded speech and automatic morphological analysis of the transcriptions are accurate enough to serve as basis for studying Estonian morphosyntax in semi-spontaneous speech.

Jane Klavan (b. 1983), PhD, University of Tartu, Faculty of Arts and Humanities, College of Foreign Languages and Cultures, Lecturer in English Language (Lossi 3, 51003 Tartu), jane.klavan@ut.ee

Tanel Alumäe (b. 1976), PhD, Tallinn University of Technology, School of Information Technologies, Department of Software Science, Senior Researcher (Akadeemia tee 21B, 12618 Tallinn), tanel.alumae@taltech.ee

Arvi Tavast (b. 1969), PhD, Institute of the Estonian Language, Director (Roosikrantsi 6, 10119 Tallinn), arvi@tavast.ee