KOGNITIIVNE KEELETEADUS ARVUDE RÄGASTIKUS

JANE KLAVAN

Keeleteaduses on põnevad ajad: elame digihumanitaaria ajastul ning uurijatena on meil võimalik kasutada üha suuremat hulka erinevaid meetodeid, mille abil keeleandmeid koguda, ja üha rohkem (üha keerulisemaid) andmeanalüüsi vahendeid, mille abil andmetes korda luua ja neist tähenduslikke tulemusi kätte saada. Teadlasena olen veendunud, et suund keeleteaduse „kvantifitseerimise" poole on midagi, mis aitab tagada, et ka tähendust ja grammatikat uurides peame kinni heast teadustöö tavast. Iga teooria ja väide võiks olla empiiriliselt falsifitseeritav (Popper 1965). Keeleteadlasena huvitab mind küsimus erinevate keeleuurimis- ja andmeanalüüsimeetodite rakendamisvõimalustest ja piirangutest. Sukeldudes pidevalt oma töös arvude rägastikku, on hea end vahepeal sealt välja kiskuda ja mõtiskleda selle üle, kas valitud uurimismetoodika ja andmeanalüüsi vahendid annavad ikkagi valiidse ja usaldusväärse vastuse teoreetilisele uurimisküsimusele, mille olen püstitanud. Viitan siinkohal Pennsylvania ülikooli keeleteaduse professori William Labovi sõnavõtule konverentsil „New Ways of Analysing Variation 46", kus ta väga ilmekalt tõdes, et „arvud on meie vahendid, mitte meie kupjad" (D'Arcy 2017). Suhtumist arvudesse kui kubjastesse võib täheldada teisteski ingliskeelsetes fraasides, mida leidub kasutuspõhises keeleteaduslikus erialakirjanduses: *doing numbers just for numbers' sake* (Langacker 2016), *number-crunching* (Nesset 2016), *empirical imperialism* (Geeraerts 2016; Schmid 2016). Ma ei taha lugejat õhutada empiirilisele imperialismile ega propageerida arusaama, nagu peaks iga keeleteadlane tegelema süvitsi statistikaga. Küll aga on eesmärk illustreerida võimalust, kuidas keelt uurida nii, et selle aluseks olevad teooriad ja väited oleksid empiiriliselt falsifitseeritavad.

Keeleteadus on suur ja lai ning igal uurijal on vabadus valida teoreetiline raamistik, mis peegeldab just tema parimat arusaama keelest ja selle olemusest. Lähtuvalt autori keeleteaduslikest huvidest ja tõekspidamistest, on artikkel kirjutatud ühe spetsiifilise keeleteaduse haru – kasutuspõhise keeleteaduse (Barlow, Kemmer 2002), veelgi kitsamalt kognitiivse keeleteaduse (Dąbrowska, Divjak 2015; Dancygier 2017) – kontekstis. Kognitiivseid keeleteadlasi huvitab küsimus, kuidas keeleandmete põhjal jõuda keelekasutuse taga peituvate kognitiivsete protsessideni. Jõudmine selle teadmiseni keelekasutust jälgides ei ole lihtne ja seda välditakse, või hullem – järeldusi tehakse liialt kergekäeliselt. Püüan visandada mõningaid erinevaid metodoloogilisi lähenemisi ja nendega kaasnevaid teoreetilisi küsimusi ühe eesti keele morfosüntaktilise alternatsiooni näitel.

Järgnevas osas annan ülevaate kognitiivse keeleteaduse hetkeseisust suurte andmehulkade ja empiirilise keeleteaduse kontekstis. Taustaks on 2016. aastal ilmunud ajakirja Cognitive Linguistics erinumber, kus vaadeldi kognitiivse keeleteaduse erinevaid teoreetilisi suundi. Suuna algusaastatest peale on kognitiivne keeleteadus olnud suhteliselt populaarne, aga kvantitatiivse kasvuga on kaasnenud kvalitatiivne eristumine. Teisisõnu on suund nüüdseks väga heterogeenne ja selle sees on esile kerkinud erinevad – ja nii mõnigi kord vastandlikud – vaated kognitiivse keeleteaduse teoreetilise ja metodoloogilise olemuse kohta. Keskendun metodoloogilisele teljele ja püüan teha kokkuvõtte, milline on olnud ja on metodoloogia panus kognitiivse keeleteaduse kui teooria arengusse. Teises osas tutvustan ülevaatlikult ühe keelelise katse tulemusi ning võrdlen neid korpuspõhise uurimuse tulemustega adessiivi ja *peal*-konstruktsiooni rööpse kasutuse kohta tänapäeva kirjakeeles.

1. Kognitiivse keeleteaduse metodoloogiline telg

Ajalooliselt võib prototüüpseks kognitiivseks keeleteaduseks pidada suunda, mis on peamiselt mentalismi poole kaldu, põhineb introspektsioonil, on spetsialiseerunud analüüsile, mis lähtub sünkroonsest vaatepunktist, mille keskmes on Lääne-Euroopa keelte (eriti inglise keele) andmed ja mis kuigipalju ei vaatle suhtluse sotsiaalset ja multimodaalset aspekti (Divjak jt 2016a: 3). Viimastel aastatel on sellest prototüüpsest kognitiivsest keeleteadusest välja kasvanud rida laiendusi ning Dagmar Divjaki jt (2016a) nägemuses on laiendused toimumas kolmel teljel: kognitiivsel, sotsiaalsel ja metodoloogilisel teljel. Järgnevas arutlen, milliseid võimalusi ja väljakutseid empiiriline lähenemine keelele kognitiivse keeleteaduse kui teooria seisukohast esitab.

Kuigi kognitiivne keeleteadus, lähtudes kasutuspõhisuse printsiibist, on alati olnud empiiriline ja kognitiivsed keeleteadlased on alati kasutanud erinevaid andmete kogumise viise, pole siiski kahtlust, et introspektsioonil on kognitiivses keeleteaduses väga eriline koht. Introspektsioon on kognitiivse keeleteaduse harusse sügavalt juurdunud ning selle privilegeeritud staatus tuleneb nii ajaloolistest kui ka teoreetilistest põhjustest. Kognitiivse keeleteaduse teerajajate teadustöö oli, arusaadavalt, pigem teoreetiline kui empiiriline: enne kui hakata teooria paikapidavust kontrollima, on teooria vaja luua. Kognitiivse keeleteaduse sünniaeg langes 1970-ndate algusesse vastureaktsioonina äärmuslikule biheivioristlikule empirismile, mis 1950. ja 1960. aastatel keeleteaduses, eriti formaalse süntaksi vallas, valitses. 1990-ndatel toimus aga (kognitiivses) keeleteaduses uus paradigma muutus – kvantitatiivne pööre, mis väljendus keeleandmete analüüsis statistiliste meetoditega. Laura Janda (2013: 2) sõnul märgib ajakirja Cognitive Linguistics jaoks seda muutust aasta 2008. 2000. aastateks on eksponentsiaalselt kasvanud uurimuste arv, kus kasutatakse korpuspõhiste ja katselisel teel kogutud andmete uurimiseks statistilisi meetodeid (Klavan, Divjak 2016); ilmunud on rida monograafiaid ja kogumikke andmete kogumise ja andmeanalüüsi meetodite kohta (Gonzalez-Marquez jt 2007; Glynn, Fischer 2010; Rice, Newman 2010; Janda 2013; Glynn, Robinson 2014) ning hulk tähelepanuväärseid õpikuid, mis on suunatud just keeleteadlastele (nt Baayen 2008; Johnson 2008; Gries 2009,

2013; Levshina 2015). Dirk Geeraertsi (2006: 31) sõnul teevad nii kognitiivse keeleteaduse alustõed (selle kognitiivne telg) kui ka kasutuspõhine lähenemine just sellest keeleteooriast väga hea kandidaadi keeleteaduse üldise metodoloogilise progressi lipulaeva positsioonile.

Nii nagu avalõigus on mainitud hirmu, et arvudest on saanud kupjad, võib küsida, kas oleme jõudnud olukorda, kus pendel on liikunud teise äärmusesse: kas oleme kaotanud „inimnäolise"[1] kognitiivse keeleteaduse? Tuntakse muret selle pärast, et (kognitiivne) keeleteadus on muutumas liialt empiiriliseks ja suur osa kvantitatiivsetest uurimustest, mida kognitiivse keeleteaduse katuse all publitseeritakse, ei pööra piisavalt tähelepanu keelele endale ega keeleteooriale (nt Langacker 2016). Kognitiivse keeleteaduse arengu seisukohalt on oluline roll nii neil uurijatel, kes postuleerivad hüpoteese ja loovad teooriaid, kui ka neil, kes kontrollivad hüpoteeside ja teooriate paikapidavust. Ohtlik on olukord, kus üht tüüpi uurimusi ja uurijaid peetakse teistest parematekst – lõppkokkuvõttes ei ole selline suhtumine produktiivne ega edasiviiv. Loodetavasti näeme kognitiivses paradigmas edaspidi rohkem mitme autoriga kaastöid – ebarealistlik on oodata, et üks uurija oleks võimeline täitma kõiki laiapõhjalise uurimuse jaoks vajaminevaid rolle.

Läbi aegade on kognitiivse keeleteaduse ridades märgata ideoloogilist vastasseisu teooria sees – empiristid *vs.* introspektsionistid. Jordan Zlatev (2016: 567) on tabavalt sõnastanud kognitiivse keeleteadlase eksistentsialistliku küsimuse: olla empiiriline või introspektiivne? Pidades silmas kognitiivse keeleteaduse kui teooria edasist arengut, on mõlemad lähenemised iseenesestmõistetavalt olulised ja propageerida tuleks mõlemat. Tsiteerin siinkohal Ronald Langackerit (2016: 472): „Kvalitatiivsed kirjeldused on aluseks sellistele kvantitatiivsetele meetoditele nagu katsed, neuropilditehnikad, statistiline mudeldamine – kvalitatiivsed kirjeldused soovitavad, mida otsida, ja lubavad andmete tõlgendamist." Muidugi need, kes teevad introspektsioonil põhinevat (kvalitatiivset) uurimistööd, väidavad, et sellelaadseid uurimusi ei ole kognitiivses keeleteaduses piisavalt (Langacker 2016), ja need, kes teevad empiirilisi uurimusi, väidavad, et kognitiivne keeleteadus toetub endiselt liialt palju introspektsiooniga kogutud andmetele (Dąbrowska 2016).

Empiiriliste meetodite kasutamine, eriti just labori tingimustes tehtud katsete tegemine tundub paljudele kognitiivsetele keeleteadlastele intuitiivselt vastumeelne. Nad näevad keele uurimist pigem teiste inimolendite ja nende kultuuriliste tõekspidamiste, mitte füüsilise objekti uurimisena. Väidetakse isegi, et kognitiivne keeleteadus on mitteobjektiivne teooria, mis on vastuolus paljude meetodite kasutamisega, kus püütakse maksimeerida keeleliste kirjelduste objektiivset alust (Geeraerts, Cuyckens 2010: 5). Kognitiivse keeleteaduse üldise arengu seisukohast tuleks siin näha ohu märki: kas vastasseis leeveneb või viib üha suurenev heterogeensus ühtse teooria lagunemiseni.

[1] Idee pärineb Ene Vainikult (2017).

2. Kognitiivne keeleteadus praktikas: korpuspõhiste ja katseliste meetodite kombineerimine konstruktsiooniliste alternatsioonide uurimisel

Endiselt on päevakorral üleskutse empirismile kognitiivse keeleteaduse sees, mis öeldi välja juba 20 aastat tagasi (nt Sandra, Rice 1995; Cuyckens jt 1997). Olen püüdnud oma uurimistöös kognitiivsete keeleteadlaste üleskutset empirismile arvesse võtta ja artikli teises osas arutlen, kas ja kuidas on võimalik kombineerida erinevaid meetodeid kognitiivse keeleteaduse teoreetilistelt alustelt lähtuvas uurimuses. Siiani olen metodoloogilistele ja teoreetilistele küsimustele vastuste leidmiseks uurinud konstruktsioonilisi alternatsioone. Erilist huvi pakub mulle adessiivi ja kaassõna *peal* paralleelne kasutus eesti keeles. Järgnevalt annan lühikese ülevaate senistest tulemustest adessiivi ja *peal*-konstruktsiooni rööpse kasutuse kohta tänapäeva kirjakeeles. Seejärel tutvustan ülevaatlikult ühe keelelise katse tulemusi ning võrdlen neid korpuspõhise uurimuse tulemustega, arendades edasi oma varasemat uurimistööd (Klavan 2012, 2014; Klavan, Veismann 2017). Põhirõhk on katseliste andmete võrdlusel korpusandmetel põhineva mudeliga.

2.1. Adessiivi ja *peal* kasutus tänapäeva kirjakeeles

Näited 1 ja 2 illustreerivad kahte võimalikku viisi, kuidas eesti keeles väljendada olukorda, kus üks entiteet (kognitiivse keeleteaduse terminoloogiat kasutades trajektoor, ingl *trajector*, edaspidi TR) asetseb teise entiteedi (kognitiivse keeleteaduse terminoloogias orientiir, ingl *landmark*, edaspidi LM) peal. Üheks võimaluseks sellist ruumilist paiknemist väljendada on adessiivi kasutamine (näide 1: *laual*). Käändelist vormi nimetatakse kirjanduses ka sünteetiliseks vormiks, siinses töös nimetan seda adessiivi konstruktsiooniks. Teiseks võimaluseks on kasutada omastavas käändes nimisõna või asesõna koos kaassõnaga *peal* (näide 2: *laua peal*). Kaassõnaga vormile viidatakse kirjanduses kui analüütilisele vormile, siinses töös nimetan seda *peal*-konstruktsiooniks.

(1) *Raamat on laual.* = adessiivi konstruktsioon
(2) *Raamat on laua peal.* = *peal*-konstruktsioon

Mõlema konstruktsiooniga väljendatakse peale kohasuhete sagedasti ka muid funktsioone. Nii näiteks saab eesti keeles adessiiviga väljendada toimumisaega, seisundit, omajat või muud tegevussubjekti pöördelise verbivormi juures, samuti vahendit ja viisi (Erelt jt 2007: 250). Varasemad uurimused ongi näidanud, et tegelikult on adessiiviga väljendatud funktsioonidest sageduselt esikohal hoopis toimumisaja ja muude abstraktsete suhete väljendamine, mitte asukoha väljendamine (Klavan 2012: 108; Matsumura 1994). Eesti keele grammatika kohaselt on kaassõnade tähendus konkreetsem ja täpsem kui kohakäänete tähendus (Erelt jt 1995: 33–34; Erelt jt 2007: 191). See ühtib väidetega, mida võib kaassõnade ja käändelõppude vaheliste erinevuste kohta käivas kirjanduses laiemalt kohata (Comrie 1986; Hagège 2010;

Lestrade 2010). Saami ja soome keele kohta on vastavalt Raija Bartens (1978) ja Krista Ojutkangas (2008) leidnud, et analüütiline kaassõnakonstruktsioon võrrelduna sünteetilise käändekonstruktsiooniga rõhutab rohkem asukohta ja seda kasutatakse koos väiksemate, manipuleeritavate entiteetidega. Sarnasele tulemusele olen jõudnud ka oma varasemates uurimustes, kus võrdlen adessiivi ja kaassõna *peal* kasutust (Klavan 2012, 2014; Klavan jt 2015; Klavan, Veismann 2017).

2.2. Korpuspõhine mudel

Korpuspõhise uurimuse moodustab 900-lauseline valim, mille peal on rakendatud binaarset logistilise regressiooni segamudelit. Mudeli eesmärgiks on ennustada, kumb kahest konstruktsioonist – adessiiv või *peal*-konstruktsioon – on kirjakeele lausetes tõenäolisem. Valimi laused pärinevad morfoloogiliselt ühestatud korpuse (2015, kokku 215 000 sõna) ja tasakaaluskorpuse (2015, kokku 10 miljonit sõna) aja- ja ilukirjandustekstide (kokku 108 autorit) allkorpustest aastatest 1980–2000. Detailset andmekogumise protseduuri kirjeldust vt Klavan 2012: 100–108. Korpusuurimuse jaoks koguti 450 juhuslikult valitud lauset konstruktsiooni kohta. 900 lauset märgendati käsitsi vastavalt 12 semantilisele ja 10 morfosüntaktilisele tunnusele (vt tabelit 1), mis annavad lausetasandil informatsiooni erinevate lauseosaliste omaduste kohta (täpsemat infot märgendatud tunnustest koos näidetega vt Klavan 2012: 70–92).

Tabel 1.

Märgendamisskeem (tähestikulises järjekorras)

Tunnus	Tunnuse väärtused
ELUSUS (LM elusus)	elus, elutu
KOMPLEKSSUS (LM morfoloogiline kompleksus)	lihtsõna, liitsõna
KONSTRUKTSIOON (sõltuv muutuja)	adessiiv, peal
LAUSE	pealause, kõrvallause
LEMMA (LM sõna lemma)	397 lemmat
LMNR (LM arv)	ainsus, mitmus
LMSL (LM sõnaliik)	asesõna, nimisõna
LMTRSUURUS (TR & LM omavaheline suurus)	ebakonventsionaalne, konventsionaalne, ühesuurune
MOBIILSUS (LM fraasi mobiilsus)	mobiilne, staatiline
PIKKUS (LM fraasi pikkus silpides)	1–41 silpi (log. transf.)
POSITSIOON (TR & LM positsioon üksteise suhtes)	lm_tr, tr_lm
SJPOSITSIOON (LM fraasi positsioon)	alguses, keskel, lõpus
SUHTETÜÜP (LM & TR vaheline suhe)	abstraktne, ruumiline
SÜNFUN (LM süntaktiline funktsioon)	adverbiaal, täiend
TRANIM (TR elusus)	elus, elutu
TRKÄÄNE (TR käändevorm)	nom., part., muu, ei kehti

TRMOBIILSUS (TR mobiilsus)	mobiilne, staatiline
TRNR (TR arv)	ainsus, mitmus
TRSL (TR sõnaliik)	nimisõnafraas, muu
TRTÜÜP (TR tüüp)	abstraktne, objekt
TÜÜP (LM tüüp)	koht, asi
VERBIRÜHM	tegevus, olemasolu, liikumine, asetsemine, verb puudub

Analüüsi järgmine samm on välja selgitada, kas mõni märgendatud 22 tunnusest mängib statistiliselt olulist rolli selles, kumb konstruktsioon osutub teatud lausekontekstis valituks ja millised tunnused on olulisemad kui teised. Vajalike vastuste leidmiseks koostasin statistikaprogrammi R abiga binaarse logistilise regressiooni segamudeli (Harrell 2001; Pinheiro, Bates 2002). Segamudelis on nii fikseeritud faktorid (st kõik selle faktori võimalikud väärtused) kui ka juhuslikud faktorid (st faktoril on väga suur hulk võimalikke väärtusi, ent valimisse on kaasatud ainult mõned väärtused). Siinse uurimuse kontekstis on juhuslikuks faktoriks LM-sõna lemma, milleks võivad olla kõik eesti keele nimi- ja asesõnad, mis korpuses esinevad ja mida on võimalik koos adessiivi või kaassõnaga *peal* kasutada, kuid valimisse on juhuslikkuse printsiibil sattunud vaid 397 sõna lemma.

Mitme tunnusega regressioonimudeli leidmisel võib tunnuste valik mudelisse osutuda keeruliseks, näiteks käesolevas uurimuses on mudeli koostamisel võimalik valida 22 tunnuse hulgast. Vastavalt Baayeni jt (2013) juhtnööridele kasutasin parima, st lihtsaima ja samal ajal täpseima mudeli leidmiseks hüpoteeside kontrollimisel põhinevat strateegiat. Täpsemalt järgisin mudeli ehitamisel automaatset sammregressiooni ja valisin kahaneva valiku strateegia (ingl *backward*). See tähendab, et alustasin täismudelist, kuhu olid kaasatud kõik 22 tunnust, ja liikusin sammhaaval minimaalse lihtsaima mudeli poole. Igal sammul eemaldasin argumendi, mille väljajätmine suurendas *F*-statistiku väärtust. Tunnus jäi mudeli valemisse sisse ainult juhul, kui see parandas statistiliselt oluliselt mudeli headust. Minimaalsesse parimasse mudelisse jäi lõpuks neli tunnust (üks semantiline ja kolm morfosüntaktilist) ja LM-sõna lemma kui juhuslik faktor; binaarse regressiooni segamudelit kirjeldab järgnev valem:

KONSTRUKTSIOON ~ PIKKUS + KOMPLEKSSUS + MOBIILSUS + TRSL + (1 | LEMMA)

Mudeli klassifitseerimistäpsuse hindamiseks arvutasin õigesti klassifitseeritud sündmuste osakaalu. Mudeli täpsust hindasin kahe muutuja järgi: üldine täpsuse protsent ja *C*-statistik (Hosmer jt 2013: 173–182). Mudeli üldise klassifitseerimistäpsuse hindamiseks tuleb risttabeli kujul vaadata kahe tunnuse tegelikku väärtust andmestikus (adessiivi ja *peal*-konstruktsiooni esinemine) koos mudeli poolt prognoositud väärtuse tõenäosusega. Mudeliga leitud tõenäosushinnangute lõikepunktiks (ingl *cut-off point*) on 0,5. Seega ei ole mudel klassifitseerimisviga teinud juhul, kui mudeli poolt prognoositud *peal*-konstruktsiooni esinemise tõenäosus on $\geq$ 0,5 ja *peal*-konstruktsioon on ka tegelik uuritava tunnuse väärtus. *C*-statistik jääb alati 0,5 ja 1 vahele ning

selle muutuja väärtus peegeldab seda, kui hästi suudab mudel kahte võimalikku konstruktsiooni eristada.[2] Kõnesoleva mudeli klassifitseerimistäpsus on 80% ja mudeli *C*-väärtus 0,88 näitab, et mudelil on väga hea eristusvõime.

Mudeli headuse hindamiseks võib veel vaadata paranemise määra, mille arvutasin, jagades mudeli täpsuse protsendi mudeli baastasemega. Kuna andmestikus oli mõlemat uuritavat konstruktsiooni korpusvalimis ühepalju (kumbagi 450, kokku 900), siis on siinse arvutuse baastase 50%. Mudeli headuse näitajate põhjal võib öelda, et mudel klassifitseerib üpris täpselt, kas andmestikus esineb *peal*-konstruktsioon või adessiiv. Siinse uurimuse raames ei ole mudelit ristvalideeritud, mis kindlasti on statistiliste mudelite ehitamises väga oluline aspekt. Antud juhul on mudelit treenitud ja testitud ühe ja sama andmestiku peal. Edasise uurimistöö käigus suurendan korpusvalimit ja testin sama mudelit uute andmete peal. Kirjandusest ei ole õnnestunud leida juhiseid logistiliste segamudelite ristvalideerimise kohta. Seega pakub käesolev uurimus alternatiivse viisi, kuidas mudeli headust hinnata: võrdlus keelelise katsega, mille tulemusena saab leida keelekasutajate klassifitseerimistäpsuse ülem- ja alampiiri.

Mudeli headuse hindamisel tuleb silmas pidada veel tõsiasja, et käesolevas andmestikus on kaks paralleelselt kasutatavat konstruktsiooni. Seetõttu on ootuspärane, et mudel prognoosib suhteliselt sarnased tõenäosushinnangud, kuna põhimõtteliselt sobivad mõlemad konstruktsioonid igasse konteksti, mis valimis on esindatud (vrd valimi moodustamise kriteeriume Klavan 2012: 105–108). Siinkohal võiks tõstatada küsimuse, kui täpne on keskmise keelekasutaja ennustus, kumb kahest konstruktsioonist lauses esineb ja kas mudeli klassifitseerimistäpsus paigutub keelekasutajate klassifitseerimiskäitumise ülem- ja alampiiri vahemikku. Siin tulevad mängu keelelised katsed. Järgnevalt kirjeldan sunnitud valiku katset, mille järgi saan leida ühed võimalikud keelekasutajate klassifitseerimiskäitumise ülem- ja alampiirid uuritava keelenähtuse kontekstis.

2.3. Katseline semantika: sunnitud valiku katse

Järgnevalt võrdlen korpuspõhise mudeli ennustustäpsust emakeelsete keelekasutajate valikutega. Selleks et saada korpusandmetega võrreldavad katsepõhised andmed, otsustasin kasutada sunnitud valiku katset.[3] Katsetulemuste võrdlemine korpuspõhise mudeliga viib kolme võimaliku stsenaariumini (vt ka Divjak jt 2016b). Juhul kui 1) keelekasutajate valikud ühtivad korpuspõhise mudeli ennustustega, saame kinnitust, et valitud mudel kirjeldab andmetes esinevat variatsiooni piisavalt hästi; 2) keelekasutajate valikud ei ühti korpuspõhise mudeli ennustustega ja keelekasutajad ei saavuta samaväärset ennustustäpsust, on üheks võimalikuks järelduseks, et mudeli valem on keerulisem, kui tegelik keelekasutus seda lubaks; 3) keelekasutajate ennustustäpsus on

[2] Hosmer jt (2013: 177) pakuvad välja järgnevad juhised *C*-statistiku tõlgendamiseks: $C = 0{,}5$ – kaks sündmust ei eristu üldse; $0{,}5 < C < 0{,}7$ – vähene eristatus; $0{,}7 < C < 0{,}8$ – vastuvõetav eristatus; $0{,}8 < C < 0{,}9$ – väga hea eristatus; $C \geq 0{,}9$ – suurepärane eristatus.

[3] Eeskujuna olen kohandanud Joan Bresnani (2007) inglise keele daativi alternatsiooni katset.

aga korpuspõhise mudeli ennustustäpsusest parem, on põhjust kahtlustada, et mudeli valemist on puudu olulised tunnused, mida uurijal ei ole õnnestunud tuvastada. Isegi sellisel juhul, kui masin ja inimene eelistavad sama konstruktsiooni samas kontekstis, on täiesti võimalik, et nad küll jõuavad sama valikuni, kuid erinevat teed pidi.

2.3.1. Meetod: sunnitud valiku katse

Katse ülesehitusest, katseüksustest ja üldistest tulemustest on pikemalt ja detailsemalt kirjutanud Klavan ja Veismann (2017); järgnevalt annan katsest vaid lühiülevaate. Katse koosnes 30 korpuslausest, kus konstruktsiooni asemel oli lünk, millele järgnesid mõlemad alternatiivsed variandid, adessiivi konstruktsioon ja *peal*-konstruktsioon. Katsesse valiti laused juhuslikkuse printsiibil igast viiest tõenäosusvahemikust, mis olid määratletud binaarse logistilise regressioonimudeliga (Klavan 2012: 176–181). Laused esindavad seega kogu võimalikku tõenäosusskaalat. Katses olid nii sellised laused, kus korpuspõhise mudeli järgi võivad esineda mõlemad konstruktsioonid võrdse tõenäosusega, kui ka sellised, kus korpuspõhise mudeli järgi esineb üks konstruktsioon palju suurema tõenäosusega kui teine. Katseüksused olid pooljuhuslikus järjestuses, järgides põhimõtet, et järjestikku ei satuks samast tõenäosusvahemikust pärinevad laused. Katsest koostati neli erineva järjestusega versiooni. Samuti jälgiti, et varieeruks esimese valikuna esitatud konstruktsioon: näiteks esitati A-katseüksuse versioonides 1 ja 3 esimesena *peal*-konstruktsioon ning versioonides 2 ja 4 adessiivi konstruktsioon.

Katses osales 96 eesti keelt emakeelena kõnelejat, kes värvati sotsiaalmeedia ja meililistide kaudu. 47 meessoost ja 49 naissoost katseisiku vanus varieerus 18 ja 54 eluaasta vahel (keskmine vanus oli 29, standarhälve 9,5). Katsealused jagati juhuslikkuse printsiibil nelja katseversiooni vahel: esimeses versioonis osales 22, teises 29, kolmandas 24 ja neljandas 21 inimest. Katseversioon ei osutunud statistiliselt oluliseks faktoriks, mis aitaks selgitada andmetes leiduvat variatiivsust. Katses osalejatel paluti valida, milline kahest konstruktsioonist sobiks lünka kõige paremini. Mõlemad variandid esitati üksteise kõrval horisontaalselt. Katsealused nägid korraga vaid ühte lauset ja neil ei olnud võimalik tagasi liikuda eelmise lause juurde ega oma algset vastust muuta. Katse koostati ja viidi läbi veebipõhises küsimustikuplatvormis PsychData. Keskmiselt sooritati katse 10 minutiga.

2.3.2. Tulemused: sunnitud valiku katse *vs.* korpuspõhine mudel

Esitan siin ühe võimaliku viisi, kuidas sunnitud valiku katse tulemusi analüüsida. Kindlasti oleks edasises töös vajalik katsetulemuste mudeldamine, kuid seda ei ole siin ette võetud. Põhjuseks on see, et artikli eesmärk on välja tuua keelekasutajate ennustustäpsus võrrelduna korpuspõhise mudeliga. Sellest eesmärgist lähtuvalt olen andmete analüüsimisel arvutanud lihtsalt katsealuste klassifitseerimistäpsuse. Siinses analüüsis võtsin kõigepealt korpusvalimist välja need 30 lauset, mida kasutasin sunnitud valiku katses. Seejärel treenisin korpuspõhist mudelit järele jäänud 870 lause peal ja kasutasin kujunenud mudelit, et välja arvutada mõlema konstruktsiooni esinemise tõe-

näosus 30 katselause jaoks. Kuna katses oli 30 lauset ja valida sai kahe konstruktsiooni vahel, on valik juhuslik siis, kui valitakse õigesti 15 lauset 30-st. Selleks et korpuspõhiseid ennustusi võrrelda keelekasutajate valikutega, lähtusin eeldusest, et kõrgema tõenäosuse saanud konstruktsiooni võiksid valida ka keelekasutajad. Siinse analüüsi kontekstis pean õigeks vastuseks konstruktsiooni, mis esines algses korpuslauses. Omaette huvitav on õige konstruktsiooni küsimus. Nii näiteks võiks õigeks konstruktsiooniks pidada hoopis seda, mida keelekasutajad sunnitud valiku katses kõige rohkem valisid. Siiski on siin jäädud praktilistel kaalutlustel selle juurde, et õige konstruktsioon, millega nii korpuspõhiseid tõenäosushinnanguid kui ka katsealuste valikuid võrrelda, on see, mis esines algses korpuslauses.

Tabelis 2 on toodud korpuspõhise segamudeli täpsus ja keelekasutajate keskmine skoor, kõrgeim skoor ja madalaim skoor 30 katselause puhul – see tähendab, mitmes lauses 30-st osutus kõige enam valituks algses korpuslauses olnud konstruktsioon. Viimases veerus on esitatud paranemise määr võrreldes baastasemega, milleks katselausetes on 53%. Nii keskmine keelekasutaja kui ka korpuspõhine segamudel on võimelised algset konstruktsiooni ennustama baastasemest palju suurema täpsusega. Korpuspõhise segamudeli ennustustäpsus on muljet avaldavalt kõrge: 93%. Sama skoor on ka laeks, mille olid võimelised saavutama kõige parema ennustustäpsusega keelekasutajad sunnitud valiku katses. Emakeele kõnelejate keskmine skoor jääb korpuspõhise mudeli skoorile alla: keskmiselt olid keelekasutajad võimelised algset konstruktsiooni ennustama 23 lauses 30-st (täpsus = 77%, mediaan = 23, standardhälve 2,5).

Tabel 2.

Korpuspõhise segamudeli ja keelekasutajate ennustustäpsus

Korpuspõhine mudel *vs.* kõnelejad		**Üleüldine täpsus**	**Paranemise määr**
Korpuspõhine segamudel	tulemus	28/30 = 93%	1,8 võrra
Emakeele kõnelejad	keskmine skoor	23/30 = 77%	1,5 võrra
	kõrgeim skoor	28/30 = 93%	1,8 võrra
	madalaim skoor	14/30 = 47%	0,9 võrra

Sarnaselt Divjaki jt (2016b) uurimusele võib ka siinses uurimuses täheldada küllaltki suurt individuaalset varieeruvust: ennustustäpsus varieerub 14-st õigesti valitud lausest (madalaim skoor) 28 õigesti valitud lauseni (kõrgeim skoor). Mõned keelekasutajad ei saavuta juhusliku valiku määra, milleks antud katses oli 15/30, samal ajal kui teised on võimelised ennustama sama hästi kui korpusandmetel treenitud segamudel. Divjak jt (2016b) pakuvad üheks võimalikuks põhjuseks, miks selline variatiivsus rühma keskmise skoori ja individuaalsete skooride vahel esineb, selle, et erinevad keelekõnelejad toetuvad valikute tegemisel eri tunnustele ja et kollektiivselt on neil ligipääs suuremale hulgale tunnustele. Sama põhjus võiks olla ka üks võimalikke tegureid, mis selgitaks siinses andmestikus esinevat rühma keskmise ja individuaalsete skooride varieeruvust. Eelnevate uurimuste valguses tundub selline selgitus üpris usutav, kuna varem on näidatud, et erinevatel keele-

kasutajatel on erinev grammatika (hea ülevaate sellesuunalistest uurimustest pakub näiteks Dąbrowska 2015).

Seega olen saanud tulemuse, et 30 katselause puhul ennustab korpuspõhine mudel võrreldes keskmise keelekasutajaga palju täpsemini, kumb konstruktsioon mingis kindlas kontekstis võiks esineda. Oluline on märkida, et antud uurimuses on keelekasutaja klassifitseerimiskäitumise ülem- ja alampiiride määramisel piirdutud sunnitud valiku katsega. Pole aga selge, millist katseliiki pidada inimkäitumise ideaaliks, ja edaspidi loodetakse uurimises tööd laiendada erinevat liiki katsetega (vt nt Klavan, Veismann 2017, kes võrdlevad sunnitud valiku katse tulemusi vastuvõetavuse hinnangutega sama katse stiimulile).

2.4. Korpuspõhised statistilised mudelid *vs.* keelekasutajad

Eelnevast analüüsist selgus, et üksikasjalikult märgendatud korpusandmete põhjal treenitud statistiline mudel klassifitseerib märksa paremini kui keskmine keelekasutaja. Sama tulemuseni on jõudnud ka teised uurijad, kes on võrrelnud korpuspõhist regressioonimudelit keelekasutajate klassifitseerimiskäitumisega (nt Bresnan 2007; Arppe, Abdulrahim 2013; Divjak jt 2016b). Nii siinse uurimuse kui ka varasemate uurimuste puhul täheldame, et inimkäitumist iseloomustab variatiivsus (Baayen 2011: 313; Divjak jt 2016b), samal ajal kui masinkäitumine (vähemalt regressioon) püüdleb matemaatilise täpsuse poole. Niisiis saab järeldada, et see, kuidas statistiline mudel jõuab jaotusmustrite kvantitatiivse struktuurini, erineb olemuselt suuresti sellest, kuidas keelekasutajad jõuavad teadmiseni sellest struktuurist (Baayen 2011: 317).

Logistilise regressioonimudeliga, mida kasutasin korpusandmete mudeldamiseks, on võimalik väga täpselt ennustada, sest sellega leitakse tundmatuid parameetreid ja hinnatakse neid suurima tõepära meetodil (ingl *maximum likelihood method*) – eesmärgiks on maksimaalse tõepära saavutamine. Pole aga teada, kas ka keelekasutajate klassifitseerimiskäitumist iseloomustab optimaalsus (Milin jt 2016: 508). Kognitiivse keeleteaduse seisukohast tuleks seega eelistada statistilisi mudeldamistehnikaid, mis peegeldavad inimkäitumist ja põhinevad bioloogiliselt ning psühholoogiliselt usutavatel õppimisalgoritmidel. Klassifitseerimismudelite kontekstis peaksime uurijatena eelistama mudeldamistehnikaid, mis näiteks ei kitsenda tunnuste valikut selliselt, et ei lubata mitme kollineaarse (st omavahel tugevalt seotud) tunnuse kooskasutamist ühes mudeli valemis, nii nagu seda regressioonitehnikad sätestavad. Üheks võimalikuks alternatiiviks on NDL-mudel (ingl *Naive Discriminative Learning*; Baayen 2011; Milin jt 2016; Rescorla, Wagner 1972). Mudeldamistehnikad, mis lubavad mitme kollineaarse tunnuse samaaegset mudeldamist, peegeldavad tegelikku keelekasutust paremini, kuna just liiasus on see, mis teeb inimeste jaoks keele õppimise võrdlemisi lihtsaks (Baayen 2011: 309) ja seletab, miks on võimalik koos esineda individuaalsetel erinevustel ja ühtsusel (Divjak jt 2016b).

Teine võimalik tõlgendus tulemusele, kus korpuspõhine mudel ennustab täpsemini kui keskmine keelekasutaja, on see, et korpuspõhine mudel on liialt kompleksne. Korpuspõhise mudeli valemis on parameetreid, millele keskmisel

keelekasutajal puudub ligipääs. Samas tuleb märkida, et mõni keelekasutaja suudab saavutada korpuspõhise mudeliga sama täpsusskoori, mis näitab, et teoreetiliselt on ka inimesel võimalik saavutada sama kõrge skoor kui masinal. Siinkohal tuleb aga kindlasti ära märkida, et meil pole piisavalt infot selle kohta, mil viisil keelekasutajad jõuavad oma valikuteni ning kas inimene ja masin jõuavad sama tulemuseni sama teed pidi. Keskseks küsimuseks on, kas keskmine keelekasutaja jõuab korpuspõhise regressioonimudeliga samale tulemusele, kasutades neidsamu tunnuseid, mis on olnud korpuspõhise regressioonivalemi sisendiks. Näiteks on Divjak jt (2016b: 11) oma sunnitud valiku katse tulemustele tuginedes välja toonud, et keelekasutajad võisid pakkuda vastuseid huupi, sest valikutes oli suur varieeruvus. Seega võis mõnel katses osalenul „õige" vastuse äraarvamisel lihtsalt vedada.

Alternatiivselt võib arutleda, kas on usutav, et inimestel on erinev keeleline kogemus ja nad opereerivad erinevate keeleliste tunnustega. Eesti keele kõnelejad, kes saavutasid korpuspõhise mudeliga samaväärselt kõrge skoori, võivadki kasutada samu tunnuseid, mis on korpusmudeli valemis. Samas võib olla tõenäoline, et samaväärse tulemuse annaks kahe konstruktsiooni valiku ennustamisel hoopis teiste tunnuste kombinatsioon kui see, mida on kasutatud siin artiklis esitatud korpuspõhise regressioonimudeli valemis. Lisaks võiks see tunnuste kombinatsioon peegeldada sama hästi ka keelekasutajate klassifitseerimiskäitumist sunnitud valiku katses. Divjak jt (2016b) näitavad, kuidas mitu erinevat mudelit, kuhu on juhuslikkuse printsiibi põhjal valitud erinevad tunnused, saavutavad inimkäitumisega sarnaseid tulemusi. Nende uurimus illustreerib, et oluline pole mitte see, milliseid tunnuseid keeleõppijad kasutavad, vaid see, et nad kasutaksid piisaval hulgal tunnuseid. Sarnasele järeldusele jõuab ka Harald Baayen (2011), kes näitab, et erinevate mudelite ja mudeldamistehnikate täpsus ei vähene oluliselt ka juhul, kui selle valemis olevaid tunnuseid permuteerida.[4] Kuna keelelised (ja ka keelevälised) tunnused on omavahel tihedalt seotud, pole ükski keeleline tunnus eraldiseisvana piisavalt oluline, et mudeli täpsust olulisel määral kõigutada (Baayen 2011: 306).

3. Kokkuvõtteks

Siinses artiklis tutvustasin ülevaatlikult ühe keelelise katse tulemusi ning võrdlesin neid korpuspõhise uurimuse tulemustega adessiivi ja *peal*-konstruktsiooni rööpse kasutuse kohta tänapäeva kirjakeeles. Empiiriline uurimus eesti keele adessiivi ja kaassõna *peal* kasutusest näitas, et käsitsi põhjalikult märgendatud korpusvalimil põhinev logistilise regressiooni segamudel on suurepärane klassifitseerija: kogu korpusandmestiku peal treenitud mudeli täpsus on 80%. Sunnitud valiku katse tulemus näitas, et keskmiselt on keelekasutajad võimelised ennustama algset konstruktsiooni 77%-lise täpsusega. Võrreldes aga korpuspõhise mudeli täpsust keskmise keelekasutaja katselausete täpsusega, peab tõdema, et korpuspõhine regressioonimudel on ehk liialt optimistlik ja tegelik keelekasutus on palju suurema variatiivsusega ja vähem

[4] Permuteerima tähendab hulga elementide järjestuse muutmist (kombinatoorikas).

täpne kui idealiseeritud, maksimaalsele täpsusele orienteeritud regressioonimudel.

Loodetavasti on käesoleva artikliga õnnestunud näidata, et võimalusel tuleks keelenähtuse põhjalikul uurimisel kaaluda statistiliste korpuspõhiste mudelite kasutamist koos keeleliste katsetega. Siinse uurimuse kontekstis oleks ainult korpuspõhise mudeliga piirdumine jätnud fookusest välja olulise info: korpuspõhised mudelid võivad olla rohkem või vähem täpsed ja nende võrdlemine keelekasutajate klassifitseerimistäpsusega keelelises katses aitab hinnata statistiliste mudelite suhtelist headust. Keel on üks osa inimesele omasest käitumisest; keeleteadlaste koostatud mudelid on vaikimisi äärmiselt ebamäärased ja suure ennustusvigade arvuga. Kognitiivses keeleteaduses, ja teoreetilises keeleteaduses laiemalt, võiks võtta eesmärgiks püüelda kognitiivse usutavuse poole, isegi juhul kui seetõttu tuleb teha järeleandmisi matemaatilises täpsuses. Statistikast on teada, et „kõik mudelid on valed" (Box 1976: 792), mõned mudelid on lihtsalt natukene paremad ja kasulikumad kui teised; paraku pole seda kõige õigemat mudelit kunagi võimalik täieliku tõsikindlusega leida (Crawley 2007: 339).

Kognitiivse keeleteaduse arengu perspektiivist on vaja multifaktoriaalse korpusanalüüsi kõrval teha katseid ja analüüsis rakendada peenekoelisi kognitiivselt usutava(ma)id mudeldamistehnikaid (nt NDL; Milin jt 2016). Laura Janda (2013: 6) on tabavalt märkinud, et kasutades keeleteaduses kvantitatiivseid meetodeid, tagame keeleteaduse kui teadussuuna arengu, tehes seejuures läbi sama arenguetapi, mille on läbinud juba psühholoogia ja sotsioloogia. Samas aga ei saa unustada introspektsiooni ja sellel põhinevaid teoreetilisi postulaate. Vastusena teemanumbris tõstatatud küsimusele „Mis saab teooriast?" nõustun emeriitprofessor Haldur Õimu (2017) sõnadega, et „teoorial pole häda midagi". Just praegu on keeleteaduses äärmiselt põnevad ajad, mil on võimalus kasutada järjest laiemat valikut meetodeid, mille abil teoreetiliste mõistete ja hüpoteeside paikapidavust kontrollida, andes seeläbi palju juurde keeleteaduse kui „päris" teaduse kuvandile.

Artikli valmimist on toetanud Sihtasutus Eesti Teadusagentuur (*PUT1358 „Mudelite loomine ja lõhkumine: Klassifitseerimismudelite valideerimine keeleteaduses"*).

Võrguviited

PsychData. https://www.psychdata.com (15. VIII 2018).

Kirjandus

Arppe, Antti, Abdulrahim, Dana 2013. Converging linguistic evidence on two flavors of production: The synonymy of Arabic COME verbs. – Ettekanne. Second Workshop on Arabic Corpus Linguistics, University of Lancaster, 22–26 July, 2013.

Baayen, Harald R. 2008. Analyzing Linguistic Data: A Practical Introduction To Statistics Using R. Cambridge: Cambridge University Press.

Baayen, Harald R. 2011. Corpus linguistics and naive discriminative learning. – Revista Brasileira de Linguística Aplicada, kd 11, nr 2, lk 295–328.
Baayen, Harald R., Endresen, Anna, Janda, Laura A., Makarova, Anastasia, Nesset, Tore 2013. Making choices in Russian: pros and cons of statistical methods for rival forms. – Russian Linguistics, kd 37, nr 3, lk 253–291.
Barlow, Michael, Kemmer, Suzanne (toim) 2002. Usage-Based Models of Language. Stanford: CSLI Publications / Center for the Study of Language and Information.
Bartens, Raija 1978. Synteettiset ja analyyttiset rakenteet Lapin paikanilmauksissa. Helsinki: Suomalais-Ugrilainen Seura.
Box, George E. 1976. Science and statistics. – Journal of the American Statistical Association, kd 71, nr 356, lk 791–799.
Bresnan, Joan 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. – Roots: Linguistics in Search of Its Evidential Base. Toim Sam Featherston, Wolfgang Sternefeld. Berlin–New York: Walter de Gruyter, lk 77–96.
Comrie, Bernard 1986. Markedness, grammar, people, and the world. – Markedness. Toim Fred R. Eckman, Edith A. Moravcsik, Jessica R. Wirth. New York: Plenum, lk 85–106.
Crawley, Michael J. 2007. Statistics: An Introduction Using R. Chichester: Wiley.
Cuyckens, Hubert, Sandra, Dominick, Rice, Sally 1997. Towards an empirical lexical semantics. – Human Contact Through Language and Linguistics. Toim Birgit Smeija, Meike Tasch. Bern: Peter Lang, lk 35–54.
Dąbrowska, Ewa 2015. Individual differences in grammatical knowledge. – Handbook of Cognitive Linguistics. Toim E. Dąbrowska, Dagmar Divjak. Berlin–Boston: De Gruyter Mouton, lk 650–668.
Dąbrowska, Ewa 2016. Cognitive Linguistics' seven deadly sins. – Cognitive Linguistics, kd 27, nr 4, lk 479–491.
Dąbrowska, Ewa, Divjak, Dagmar (toim) 2015. Handbook of Cognitive Linguistics. Berlin–Boston: De Gruyter Mouton.
Dancygier, Barbara (toim) 2017. The Cambridge Handbook of Cognitive Linguistics. Cambridge: Cambridge University Press.
D'Arcy, A. 2017, November 3. 'Numbers are our tools, not our masters.' – Labov #nwav46 #micdrop [Tweet]. https://twitter.com/LangMaverick/status/926400351091220480
Divjak, Dagmar, Levshina, Natalia, Klavan, Jane 2016a. Cognitive linguistics: Looking back, looking forward. – Cognitive Linguistics, kd 27, nr 4, lk 447–463.
Divjak, Dagmar, Dąbrowska, Ewa, Arppe, Antti 2016b. Machine meets man: Evaluating the psychological reality of corpus-based probabilistic models. – Cognitive Linguistics, kd 27, nr 1, lk 1–33.
Erelt, Mati, Erelt, Tiiu, Ross, Kristiina 2007. Eesti keele käsiraamat. Kolmas, täiendatud trükk. Tallinn: Eesti Keele Sihtasutus.
Erelt, Mati, Kasik, Reet, Metslang, Helle, Rajandi, Henno, Ross, Kristiina, Saari, Henn, Vare, Silvi 1995. Eesti keele grammatika I. Morfoloogia. Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut.
Geeraerts, Dirk 2006. Methodology in cognitive linguistics. – Cognitive Linguistics: Current Applications and Future Perspectives, kd 1. Toim Gitte Kristian-

sen, Michael Achard, René Dirven, Francisco J. Ruiz de Mendoza ibáñez. Berlin–Boston: Mouton de Gruyter, lk 21–50.
Geeraerts, Dirk 2016. The sociosemiotic commitment. – Cognitive Linguistics, kd 27, nr 4, lk 527–542.
Geeraerts, Dirk, Cuyckens, Hubert 2010. Introducing Cognitive Linguistics. The Oxford Handbook of Cognitive Linguistics. Oxford University Press.
Glynn, Dylan, Fischer, Kerstin (toim) 2010. Quantitative Methods in Cognitive Semantics Corpus-Driven Approaches. Berlin–New York: Walter de Gruyter.
Glynn, Dylan, Robinson, Justyna A. (toim) 2014. Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy. Amsterdam–Philadelphia: John Benjamins Publishing Company.
Gonzalez-Marquez, Monica, Mittelberg, Irene, Coulson, Seana, Spivey, Michael J. (toim) 2007. Methods in Cognitive Linguistics. Amsterdam: John Benjamins.
Gries, Stefan Thomas 2009. Quantitative Corpus Linguistics with R: A Practical Introduction. New York: Routledge.
Gries, Stefan Thomas 2013. Statistics for Linguistics with R: A Practical Introduction. Textbook. Berlin: De Gruyter Mouton.
Hagège, Claude 2010. Adpositions: Function-Marking in Human Languages. Oxford: Oxford University Press.
Harrell, Frank E. 2001. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer-Verlag.
Hosmer, David W., Lemeshow, Stanley, Sturdivant, Rodney X. 2013. Applied Logistic Regression. 3. tr. New York: John Wiley & Sons.
Janda, Laura A. 2013. Cognitive Linguistics: The Quantitative Turn. The Essential Reader. Berlin–Boston: De Gruyter Mouton.
Johnson, Keith 2008. Quantitative Methods in Linguistics. Malden, MA: Blackwell.
Klavan, Jane 2012. Evidence in Linguistics: Corpus-Linguistic and Experimental Methods for Studying Grammatical Synonymy. (Dissertationes linguisticae Universitatis Tartuensis 15.) Tartu: University of Tartu Press.
Klavan, Jane 2014. A multifactorial corpus analysis of grammatical synonymy. – Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy. Toim D. Glynn, J. A. Robinson. Amsterdam–Philadelphia: John Benjamins Publishing Company, lk 253–278.
Klavan, Jane, Divjak, Dagmar 2016. The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. – Folia Linguistica, kd 50, nr 2, lk 355–384.
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.
Klavan, Jane, Veismann, Ann 2017. Are corpus-based predictions mirrored in the preferential choices and ratings of native speakers? Predicting the alternation between the Estonian adessive case and the adposition *peal* 'on'. – ESUKA–JEFUL, kd 8, nr 2, lk 59–91.

Langacker, Ronald W. 2016. Working toward a synthesis. – Cognitive Linguistics, kd 27, nr 4, lk 465–477.
Lestrade, Sander 2010. Spatial Case. Berlin: Mouton de Gruyter.
Levshina, Natalia 2015. How to do Linguistics With R: Data Exploration and Statistical Analysis. Amsterdam: John Benjamins Publishing Company.
Matsumura, Kazuto 1994. Is the Estonian adessive really a local case. – Journal of Asian and African Studies, kd 46, nr 47, lk 223–235.
Milin, Petar, Divjak, Dagmar, Dimitrijević, Strahinja, Baayen, Harald R. 2016. Towards cognitively plausible data science in language research. – Cognitive Linguistics, kd 27, nr 4, lk 507–526.
Nesset, Tore 2016. Does historical linguistics need the Cognitive Commitment? Prosodic change in East Slavic. – Cognitive Linguistics, kd 27, nr 4, lk 573–585.
Ojutkangas, Krista 2008. Mihin suomessa tarvitaan sisä-grammeja. – Virittäjä, nr 3, lk 382–400.
Pinheiro, José C., Bates, Douglas M. 2002. Mixed-effects models in S and S-PLUS. New York: Springer.
Popper, Karl R. 1965. The Logic of Scientific Discovery. New York: Harper & Row.
Rescorla, Robert A., Wagner, Allan R. 1972. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. – Classical Conditioning II: Current Research and Theory. Toim A. H. Black, W. F. Prokasy. New York: Appleton-Century-Crofts, lk 64–99.
Rice, Sally, Newman, John (toim) 2010. Empirical and Experimental Methods in Cognitive/Functional Research. Stanford: CSLI Publications / Center for the Study of Language and Information.
Sandra, Dominiek, Rice, Sally 1995. Network analyses of prepositional meaning: Mirroring whose mind – the linguist's or the language user's? – Cognitive Linguistics, kd 6, nr 1, lk 89–130.
Schmid, Hans-Jörg 2016. Why Cognitive Linguistics must embrace the social and pragmatic dimensions of language and how it could do so more seriously. – Cognitive Linguistics, kd 27, nr 4, lk 543–557.
Zlatev, Jordan 2016. Turning back to experience in Cognitive Linguistics via phenomenology. – Cognitive Linguistics, kd 27, nr 4, lk 559–572.
Vainik, Ene 2017. Kas inimnäoline keeleteadus on võimalik. – Ettekanne. Teoreetiline keeleteadus Eestis V. Tartu, 23.–24. november.
Õim, Haldur 2017. Teoreetiline keeleteadus ja kvantitatiivsed meetodid. – Ettekanne. Teoreetiline keeleteadus Eestis V. Tartu, 23.–24. november.

Jane Klavan (*snd 1983*)*, PhD, Tartu Ülikool, inglise keele ja lingvistika lektor, jane.klavan@ut.ee*

Doing numbers and Cognitive Linguistics

Keywords: corpus linguistics, forced choice task, logistic regression, Estonian

The paper gives a short overview of the recent trends in Cognitive Linguistics. It focuses on the methodological aspects involved and exemplifies how the perfor-

mance of a corpus-based statistical model can be evaluated by comparing it against the behaviour of native speakers in a linguistic experiment. A mixed-effects logistic regression model is fitted to the corpus data of the Estonian adessive case and the adposition *peal* 'on' in present-day written Estonian. In order to evaluate the goodness of the corpus-based model, its performance is compared to the behaviour of native speakers in a forced choice task. In general, the results of the study reported in this paper show that an adequately constructed probabilistic model based on richly annotated corpus data can perform at a more or less equal level to human beings.

Jane Klavan (*b. 1983*)*, PhD, University of Tartu, Lecturer in English Language, jane.klavan@ut.ee*