STATISTIKA KOHT KEELEMUDELIS

HEIKI-JAAN KAALEP

Käesolev artikkel arutleb, kuidas keelematerjal ja selle põhjal tehtav statistika võiks olla seotud keele kui süsteemi (ja mitte pelgalt keelekasutuse) kirjeldusega.

„Keele struktuuri sõltumatus keele kasutusest on eeldus, mille laialt levinud omaksvõttu keeleteaduses võib pidada strukturalismi-traditsiooni üheks pärandiks. See eeldus on sõnastatud mitmesuguste teoreetiliste eritlustena, nagu Saussure'i *langue* ja *parole* või Chomsky *keelepädevus* ja *-kasutus*." (Hopper, Bybee 2001: 1) Seevastu Hans-Jörg Schmidi sõnul on korpuslingvistikas tavaks oletada, et keeleüksuste sagedusjaotused on – väga üldiselt öeldes – mingil moel olulised. Täpsemalt: arvatakse, et sagedamad struktuurid mängivad olulisemat rolli kui harvemad ka keeles kui süsteemis, mitte ainult keele kasutuses. Kognitiivse suunitlusega korpuslingvistid lähevad veelgi kaugemale, püüdes leida psühholoogiliselt usutavaid keelekirjeldusi, mis põhineksid keelekorpusest saadavatel kvantitatiivsetel andmetel. Täpsemalt: nad püüavad siduda keelenähtuste esinemissagedusi nende olulisuse või juurdumusega kognitiivses süsteemis. Selle ettekujutuse üheks kaasnevaks oletuseks on, et leksikaalse ja/või grammatilise variandi esinemissagedus vastab selle juurdumusele kognitiivses protsessis või protsessiga seotud esitusviisis. (Schmid 2010: 101–102)

Artikkel puudutab kahte probleemi: esiteks, kas ja kuidas sobituvad keelekasutusest, nt tekstikorpusest saadud arvandmed keele kui süsteemi (mitte kasutuse) oletatavasse mudelisse. Ja teiseks, milliseid statistilisi seoseid oleks üldse vaja tekstide põhjal otsida, st millele tähelepanu pöörata, mida täpselt tuleks kokku lugeda ja arvutada (kõigest lugemist ja arvutamist võimaldavast, mis keelematerjalis olemas on). Küsimus seostub üldise ettekujutusega selle kohta, mis on lingvisti uurimisvaldkond – kas keelekasutus (tekstid) ise või hoopis inimese viis tekste produtseerida. Esimesel juhul on otsitavaks mudeliks, millele andmed peaksid vastama, midagi teksti struktuuriga seotut, teisel juhul aga midagi teksti genereerimise mehhanismiga seotut. Tundub, et rahulolematus teksti struktuurseid aspekte kirjeldavate tõenäosuslike mudelitega tekibki juhul, kui need mudelid ei ole (kas või spekulatiivsete oletuste kaudu) genereerimise mehhanismiga seostatavad.

Artiklis vaadeldakse kolme varasemat uurimistöö näidet, kus kasutussagedust on püütud rakendada keele kui süsteemi kirjeldamiseks ja seletamiseks. Esimeseks näiteks (Kaalep 2009, 2010, 2012) on eesti keele käändsõnade vormimoodustussüsteem; kasutati realistlike keeleandmete (tekstikorpuse) põhjal lihtsalt arvutatavat statistikat, mis on seejuures teooria seisukohalt kergesti tõlgendatav. Teises näites (Klavan 2012; Siiman 2016) olid samuti kasutusel realistlikud keeleandmed, kuid nende põhjal tehti keerulisi arvutusi, mis teooria seisukohalt on raskesti tõlgendatavad; seejuures oli eesmärgiks kirjeldada faktoreid, mis mõjutavad keelekasutajat eesti keele

sünonüümsete grammatiliste vahendite valikul. Seda näidet kasutatakse käesolevas artiklis kui hoiatavat juhtumit. Kolmandaks näiteks (Saffran jt 1996; Wonnacott jt 2008) on mitte realistlike, vaid väljamõeldud keelte peal tehtavad rangelt kontrollitud õppimiskatsed. Seejuures pole tehtavad arvutused keerulised, kuid algandmete – mis on see, millele statistilisi arvutusi rakendada – valimine on väga teooriakeskne. Selle juhtumiuuringu esitamisel on rahvavalgustuslik, isegi propagandistlik eesmärk. Autori arvates väärivad nii küsimuseasetus kui ka saadud tulemused suurt tähelepanu, kuna osutavad järgnevustõenäosustele ja ignoreeritavatele keeleüksustele, millele tuleks reaalsete keelekasutusandmete korral suuremat tähelepanu pöörata.

Kvantitatiivsete meetodite rakendamine on kunst, mida õpitakse edukate ja läbikukkunud katsete kaudu. Kunst on ka läbi näha, millised keeleteoreetilised küsimused on tõlgitavad kvantitatiivse uuringu jaoks sobivasse esitusviisi. Artikkel kirjeldab mõttekäike, mis konkreetsetel juhtudel keelekasutusstatistikat keeleteoreetiliste küsimustega seovad, lootuses, et see meetodite kasutamise kunsti edendab.

1. Eesti keele morfoloogia ja kasutussagedus

1.1. Paralleelvormid

Eesti keele morfoloogia keerulisus seisneb selles, et kui on teada väljendamist vajav grammatiliste kategooriate komplekt (nt et sihitise väljendamiseks tuleks konkreetsel juhul kasutada ainsuse osastavat käänet), siis tuleb otsustada, kuidas peaks asjakohase sõna vorm selle väljendamiseks muutuma.

Protsess, mis vastutab õige sõnavormi genereerimise eest, saab sisendiks sõna algvormi, eesti keele käändsõnade puhul ainsuse nimetava vormi. Väljundiks peaks olema üks vorm ühe grammatiliste kategooriate komplekti kohta. Eesti keeles on mitu erinevat muuttüüpi, st mitu erinevat viisi sama grammatiliste kategooriate komplekti väljendamiseks, nt ainsuse osastava jaoks kas tüve tugev aste + vokaal (*äppi*) või tüvemuutuseta + *t* (*soolot*) jpt. Need erinevad muuttüübid on seotud sõna algvormi kujuga, kusjuures olulised on lõpuhäälikud, rõhu olemasolu eelviimasel silbil, viimase silbi pikkus ja teatavatel juhtudel ka eelviimase silbi pikkus (Kaalep 2012). Võiks eeldada, et algvormi kuju määrab muuttüübi üheselt ära, st ei ole olukorda, kus sama kujuga sõnad käänduvad erinevalt, ega sellist olukorda, kus üks sõna käändub mitmel moel. Tegelikult tuleb eesti keeles ette nii seda kui ka teist. Nt kahesilbilise CVC*e* struktuuriga sõna käänamisviis võib olla: *kõne* : *kõne* : *kõnet*, *ruse* : *ruskme* : *ruset*, *rase* : *raseda* : *rasedat*, *vise* : *viske* : *viset*, *pide* : *pideme* : *pidet*; mitmel pika vokaaliga lõppeval sõnal on mitu võimalikku mitmuse osastava vormi, nt *puu* : *puid* / *puusid*, *idee* : *ideid* / *ideesid*. Küsimus on, milline vormimoodustuse mehhanism selliseid vorme genereeriks.

Sama sõna erinevaid muutmisviise võiks põhjustada asjaolu, et vormimoodustusmehhanismi mõjutab veel miski muu peale sõna algvormi kuju, näiteks lause kontekst. Sellist seletust on pakutud sõnade puhul, millel on võimalik nii lühike kui ka pikk sisseütlev: verbidega nagu *suhtuma*, *kiinduma*, *armuma* eelistatavat kasutada pikka vormi (EKK: 247). Tegeliku kee-

lekasutuse andmed lükkavad sellise väite siiski ümber (Hasselblatt 2000; Kio 2006; Kaalep 2009). Paljude muude rohkem või vähem süstemaatiliselt sünonüümsete muutevormide puhul ei ole lause kontekstil kindlasti mingit osa vormivaliku mõjutajana, olgu näitena toodud kas või mitmuse osastava käände variandid *id* / *sid idee*-tüüpi sõnades, *e* / *id muuseum*-tüübis, vokaal/*sid sepp-*, *koon-*, *saar-*, *taud*-tüüpides või tugeva/nõrga astme valik *rabelema*-tüüpi verbide umbisikulises tegumoes (*rabeletud* / *rabeldud*), isikulise tegumoe käskivas kõneviisis (*rabelegu* / *rabelgu*) jpt.

Vormimoodustusmehhanismi kõrvalekaldumist ideaalmallist „üks algvormi kuju – üks muutevorm" võib seletada algvormi mitmeti tõlgendatavus, aga ka keele ajalugu ja kasutussagedus. Algvormi mitmeti tõlgendatavus ilmneb näiteks sõnatüüpides *ümbrik* ja *muuseum*. *Ümbrik*-tüüpi sõnu iseloomustab ühelt poolt kahesilbilisus, mille kohaselt käänamismalli eeskujuks sobiks *virsik*; teiselt poolt sarnanevad nad C*ik*-liitega sõnadega, nagu *elanik* (ainsuse osastav vastavalt *ümbrikut* / *ümbrikku*); *muuseum* on hääldatav kui kolmesilbiline [muu-se-um] või kahesilbiline [muus-jum]: esimesel juhul oleks käänamismalli eeskujuks *seminar* (ainsuse osastav *muuseumi*), teisel juhul *redel* (ainsuse osastav *muuseumit*).

Keele ajalugu ja kasutussagedus tulevad mängu seetõttu, et kui keele areng toob kaasa sõnade muutmisviisi teisenemise, siis ei toimu see kõigi sõnadega ühekorraga ja ühesugusel kiirusel. Sagedamad sõnad säilitavad oma varasema, nüüdseks uutele reeglitele mitte vastava muutmisviisi veel mõnda aega. Seetõttu võime näha välise kuju poolest sarnaseid, kuid erinevalt käänatavaid-pööratavaid sõnu, kusjuures enamikul neist on reaalses keelekasutuses ainult üks paralleelsetest võimalustest, mis tähendab seda, et nad on juba kas oma tüübi vahetanud või pole sellele teele veel asunud. Vaid väike arv sõnu on selliseid, mis hetkel on tüüpi vahetamas, ja seejuures on väga lühike periood, mil see sünonüümsete vormide kasutussagedus on võrdne; valdavalt näeme keelekasutuses olukorda, kus üks sünonüümsetest vormidest on palju sagedam.

1.2. Kasutussagedus muuttüübi lisatunnusena

Erinevalt käsitlusest, mille kohaselt paralleelvormid on eesti keeles üleüldine ja sage nähtus (ÕS 2013), näitavad korpusepõhised kasutusandmed, et on üsna vähe sõnu, millel on tõepoolest kasutusel sünonüümsed vormid (Kaalep 2009, 2012). ÕS 2013 normingu kohaselt esindab *ee*-lõpulisi sõnu tüüpsõna *idee*, mitmuse osastavas paralleelvormina *ideid* / *ideesid*. Tõepoolest, mõlemat vormi võib kohata ka tegelikkuses. Kuid ÕS-i kohaselt kuuluvad samasse muuttüüpi ka *maantee* ja *varietee*, ehkki *maanteesid* on kasutusel üliharva ja vormi *varieteid* on õnnestunud leida ainult Inglise Kolledži 1934. aasta kodukorrast: „Õpilane ei tohi külastada joogikohti, varieteid ja kohvikuid [---]". Seega *idee*, *maantee* ja *varietee* esitamine samas muuttüübis on eksitav. Õigem oleks jagada muuttüüp kaheks: sagedased sõnad nagu *idee* ja *maantee* ühte, haruldased nagu *varietee* teise. Selliste sõnade eeskujuks võiks võtta ÕS 2013 viisi liigitada homonüümi *tee*: tähendused 'liikumisrada' ja 'jook' erinevad oma käänamisviisilt, mis juhtumisi järgib sõnade sagedusi: sagedam

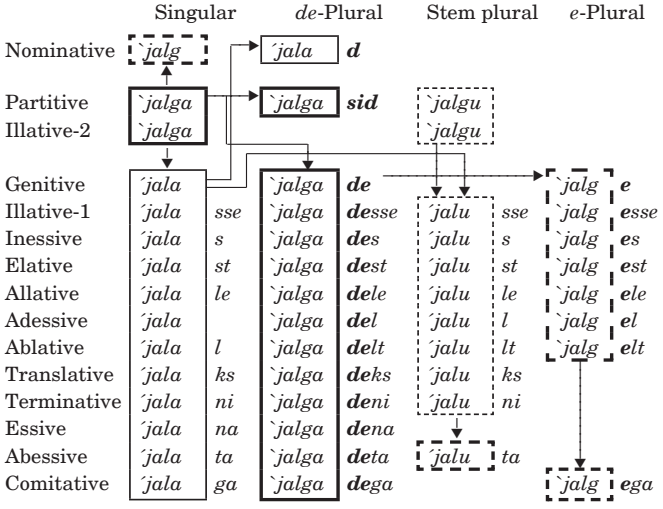
sõna – liikumisrada – on tüübis *idee*, harvem aga tüübis *koi*. Et põhjendatult otsustada, millised *ee*-lõpulised sõnad kummassegi tüüpi kuuluvad, tulekski kasutusele võtta lisaparameeter, mis võimaldab sõna käänamisviisi ennustada: sõna kasutussagedus.

Analoogiline on olukord selliste ühesilbiliste kaashäälikulõpuliste sõnadega nagu *siil*, *piim*, *eit*, *võik*. Produktiivne ainsuse tüvevokaal on *-i* (*siil* : *siili*) ja selle tüübi sõnade mitmuse osastav on *e*-lõpuline (*siile*). Ebaproduktiivsed ainsuse tüvevokaalid on *-a* (*piim* : *piima*), *-e* (*eit* : *eide*) ja *-u* (*võik* : *võigu*). Juhul kui selline sõna on sage (ja seega ka mitmuse osastav on sage), on selle mitmuse osastav tavaliselt vokaalilõpuline, nt *kuiv* : *kuivi*, *aas* : *aasu*, *kurg* : *kurgi*, *känd* : *kände*, aga haruldase (või mitmuse vorme vältiva) sõna ainsaks võimalikuks lõpuks on *-sid*, nt *piim* : *piimasid*, *eit* : *eitesid*, *võik* : *võikusid*. Ebaproduktiivse ainsuse vokaaliga sõnad jagunevad seega omaette väikestesse tüüpidesse, kus sagedaste sõnade muuttüübid on mitmuse osastavas vokaalilõpulised, harvade omad aga *sid*-lõpulised.

Sageduse kui lisaparameetri arvestamine muuttüübi määratlemisel realiseerub praktikas sel moel, et sagedased sõnad (mida on vähe) esitatakse loendina, haruldased aga häälikulisele kujule rakenduva reegli kaudu. Eraldi meelespidamist vajava ja seejuures väikese loendi postuleerimine on kooskõlas ka nõudega, et muutevormide genereerimise mehhanism oleks psühholoogiliselt usutav.

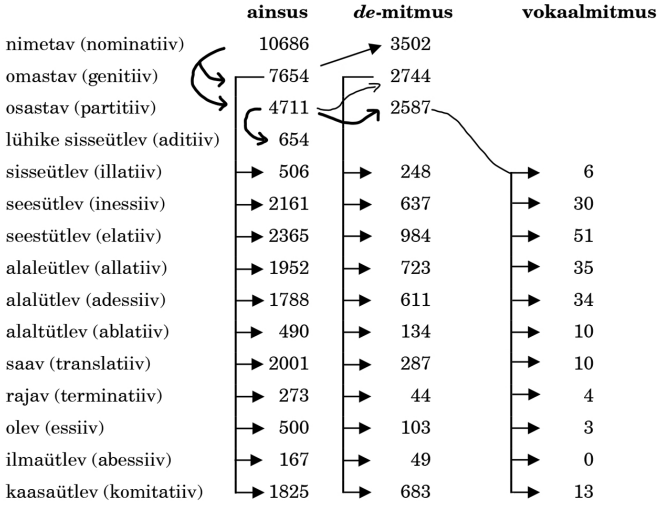
1.3. Kasutussagedus paradigma struktuuri määrajana

Muuteparadigma ei ole võimalike vormide lihtne loend, vaid hierarhiline struktuur: on olemas põhivormid, mille alusel saab moodustada teisi, analoogiavorme. Joonis 1 esitab ühe võimaliku skeemi, kuidas põhi- ja analoogiavormid on omavahel seotud.



Joonis 1. Nõrgeneva astmevaheldusega käändsõna paradigma struktuur Tiit-Rein Viitso (2003: 38) järgi.

Joonis 2 esitab alternatiivse vormidevahelise sõltuvuse skeemi, mis põhineb Kaalep 2010 ja 2012 andmetel. Erinevat tüüpi nooled tähistavad seoste üldisust ja absoluutsust: kandilised nooled seovad omastava käände ja sellest alati tulenevad muutevormid, peened kaarjooned seovad osastava käände ja sellest sageli tulenevad muutevormid, jämedad kaarjooned aga vorme, mida nimetatakse põhivormideks ja mille tuletatavus muudest vormidest võib mitteproduktiivsete muuttüüpide korral olla võimatu. Kuidas põhjendada, kumb skeem on õige? Abiks võiks olla asjaolu, et sõltuvuse suund ja sagedusinfo peaksid olema omavahel kooskõlas: ainult selline vorm, mis on sagedasem, st mida on juba nähtud, saab olla harvema ehk veel mitte nähtud vormi moodustamise aluseks. Joonisel 2 on käändsõnavormide sõnastikusagedused 500 000-sõnalises tekstikorpuses, st erinevate sõnade arv, mis konkreetses vormis esinevad. Joonise 2 tuletussuhted on sagedusinfoga kooskõlas, joonise 1 omad mitte.



Joonis 2. Vormide seosed ja sõnastikusagedus 500 000-sõnalises tekstikorpuses Kaalep 2010 ja 2012 põhjal.

On morfoloogiast sõltumatu fakt, et sõna kõiki muutevorme ei kasutata võrdselt tihti. Vormide kasutussageduse erinevus tingibki vajaduse varem kasutamata vorme tuletada ja teha seda läbinähtaval moel, sagedamate vormide põhjal. Eriti ilmne on see vähekasutatavate sõnade puhul. Nii et paradigmasisene vormihierarhia ise on keele funktsioneerimise paratamatu tulemus (Bybee 1995: 237).

Muutevormide kasutussagedusega arvestamine on üks oluline (ehkki mitte ainus) argument, mis võimaldab otsustada, kas pakutav teoreetiline paradigmahierarhia võib olla tegelikkusega kooskõlas või mitte. Käesoleval juhul moodustavad kvantitatiivsed andmed teooria loomuliku osa.

Vahekokkuvõtteks võib öelda, et ülalesitatud vormimoodustusmudelisse, mille aluseks on ettekujutus sisendi põhjal väljundisse sõnavorme genereerivast mehhanismist, mis töötab nii reeglite alusel kui ka mälus erandite loen-

deid omades, sobitub keelekasutuse statistika täiesti loomulikul viisil. Mudel on põhimõtteliselt läbipaistev, st selle toimemehhanism pole varjatud (ehkki kõik detailid pole selged). Mudel on ka testitav. On võimalik sõnastada konkreetseid reegleid ja erandite loendeid ning kontrollida, kas mudel ennustab sõnavorme samal moel, nagu seda teeb inimene.

Kuigi mudel on mõeldud genereerimiseks, on ta kergesti ümberpööratav ja kasutatav sõnavormide analüüsimiseks: sõnavormi võib (kas või ekslikult või lihtsustavalt) teisendada, näiteks eemaldades sõna lõpu, tüvevokaali või muutes sõna tugevaastmelisest nõrgaastmeliseks. Sel viisil leitakse üks võimalik algvorm, mille olemasolu oma mälus kontrollida. Kui selline algvorm on mälus olemas, siis oligi analüüs edukas. Taoline mehhanism on ka psühholoogiliselt usutav, kajastades võõrkeele õppimisel sageli tajutavat probleemi, et sõnavormist aru saada on kergem kui seda ise moodustada. See tähendab, et kui sõnavorm on juba olemas, siis on keeleõppija võimeline oletama, milline oli selle võimalik moodustamisviis, kuid ise ei suuda ta seda mehhanismi veel õiges kontekstis rakendada.

Mudeli kirjeldamisel jäi küll seletamatuks, kuidas keeleõppijad omandavad teadmise häälikulis-tuletusliku malli ja käänamisviisi seosest, st kuidas nad õpivad ära, et sõnu tuleb just sellisteks klassideks jagada just neid tunnuseid (lõpuhäälik, rõhulise silbi asukoht jne) kasutades, mitte näiteks algushääliku järgi. Vastuse sellele küsimusele võiks anda statistiline õpe, mida kirjeldatakse 3. osas.

2. Faktorite leidmine keelekorpusest

Teaduse ajaloos on korduvalt juhtunud, et uues valdkonnas võetakse kasutusele meetodid, mis on välja töötatud teistsuguste probleemide jaoks. Teiste sõnadega, rakendatakse analoogiat, eeldades, et uuritav valdkond ja küsimused on mingis mõttes sarnased nendega, millel laenatav metoodika töötas.

Elektrooniliste tekstikorpuste olemasolu on tekitanud mõtte, et selle keelematerjali peal võiks kasutada statistilisi meetodeid – neid kasutatakse mitmes eri valdkonnas, st nad on näidanud ennast kergesti laenatavatena. Keeles on palju nii reeglipäraseid kui ka erandlikke, nii muutumatuid kui ka varieeruvaid elemente, nii et tõenäosuslik vaatenurk näib keele uurimiseks hästi sobivat. Alatihti näemegi, et korpuslingvist loeb materjalist mingeid sagedusi kokku, seejärel paneb saadud arvud võrranditesse ja arvutab tulemused. Metoodika laenamine võib olla väga efektiivne viis saada uusi huvitavaid teadustulemusi, aga nagu analoogiatega ikka, tuleb kontrollida, kas laenatava metoodika rakendamise eeldused kehtivad ka laenavas valdkonnas – see ei ole lihtne ülesanne.

Lauseid ja väljendeid ei looda kaootiliselt, vaid juba öeldul on mõju sellele, mida ja kuidas veel öeldakse. See annab aluse vaadelda korpust kui mingil moel seotud elementide hulka ja eeldada, et faktorid, mis väljendi vormi määratlevad, korpuses ennast ka ilmutavad. Seega näib mõistlik analüüsida korpusematerjali statistiliste meetoditega, et uusi teadmisi saada. Kuid korpusi statistiliselt uurivate kognitiivsete lingvistide hulgas tõstatub alatasa küsimus, kas korpuse statistilised analüüsid annavad vastuseid meid huvi-

tavatele küsimustele või on tulemuseks lihtsalt arvud (Klavan, Divjak 2016). Divjak jt (2016a) täheldavad kvantitatiivsete korpusepõhiste käsitluste hulga suurenemist ja seejuures ka kahtluste ning kriitika kasvu nende meetodite vastu.

Alljärgnevas vaatame laenamise näitena regressioonimudeli meetodit, mida korpuslingvistikas palju kasutatakse. Et arutleda meetodi ülekantavuse üle, oleks hea teada, millist abstraktsiooni meetodi rakendamisel ette kujutatakse. Regressioonimudeli kui meetodi puhul eeldatakse, et (tekstis) on mingid tunnused ja osa neist (sõltumatud tunnused) mõjutavad meid huvitava (st sõltuva) tunnuse väärtust. Uuritav valdkond on regressioonimudeli jaoks *n*-mõõtmeline ruum; tunnuste väärtused on punktid selles ruumis, kusjuures sõltumatute tunnuste väärtused asuvad kindlasti ruumi mõnel teljel. Kui meil oleks kasutada kõik võimalikud tekstid (ka need, mida veel loodud pole) ja oleksime sõltumatute tunnustena osanud arvesse võtta neid ja ainult neid, mis tõepoolest sõltuva tunnuse väärtusi mõjutavad, siis moodustaksid viimase väärtused mingi(d) joone(d) selles ruumis. See oleks mingi matemaatilise funktsiooni graafik, mille valem seob sõltumatute tunnuste väärtused sõltuva tunnuse väärtusega. Tegelikult on meil muidugi ainult osa tekste, seetõttu ei ole punkte väga tihedalt; me ei arvesta kõiki ja eranditult õigeid sõltumatuid tunnuseid (sest ei oska), mistõttu meie ruumis pole õige arv mõõtmeid ja ruumi teljedki on arvatavasti valesti valitud; ka punktid ei asu lihtsa valemiga väljendatava funktsiooni graafikul, vaid moodustavad midagi pilvesarnast. Kuid ka sellise pilvesarnase punktihulga korral on võimalik tõmmata joon(ed), mis võimalikult hästi järgiks pilve kuju, st tuletada sellist graafikut kirjeldava funktsiooni valem. Funktsiooni valemi tuletamine on sisuliselt optimeerimisülesande lahendamine, st parima lahenduse otsimine olemasolevate kitsenduste raames: eesmärk on, et punktide summaarne kaugus graafikust oleks minimaalne, st maksimeeritakse tõenäosust, et funktsiooni valemi järgi arvutatu on lähedane andmetest ilmneva tegelikkusega. Lisaks loodetakse, et sõltumatute muutujate väärtuste panemine funktsiooni valemisse ennustab sõltuva muutuja väärtust hästi ka varem mittenähtud teksti puhul, st valem on üldine, mitte kitsalt treeningmaterjalile kohandatud.

Ideed, et huvitava tunnuse väärtust kujundab mingi mõõdetav faktor, ei pea muidugi tingimata realiseerima regressioonitehnika kaudu. Selleks võib kasutada ka tagasihoidlikumaid võtteid, nt arvutades paarikaupa sõltuva tunnuse korrelatsiooni iga sõltumatu tunnusega või võrreldes sõltuva tunnuse sagedusjaotust ükshaaval iga sõltumatu tunnuse omaga, kasutades hii-ruut testi, nagu on teinud näiteks Ann Siiman (2016).

Kõigi selliste statistiliste seoste (mida parimal juhul võiks tõlgendada sõltuvustena) leidmise katsete eelduseks on ettekujutus, et keeleväljendite produtseerimise käigus mingid faktorid mõjutavad väljundit; ning kõigi faktorite koosmõju ongi see, mis tulemuse ära määrab. Regressioonanalüüsiga on võimalik leida automaatselt, kui suure osa sõltuva muutuja väärtuse varieerumisest seletab ühe või teise sõltumatu muutuja variatsioon. Kui jätta mõni sõltumatu muutuja regressioonvõrrandi koostamisel kõrvale, siis teoreetiliselt peaks sõltuva muutuja mitteseletatav variatsioon suurenema samas ulatuses kui see, mida antud sõltumatu muutuja varem seletas, ja ülejäänud sõltumatute muutujate osakaal varieerumise seletamisel ei tohiks suureneda. Kui see

aga nii ei ole, siis viitab see sellele, et mudeli kasutamise eeldused ei kehti ja sõltumatud muutujad ei olegi tegelikult üksteisest sõltumatud. Keelestatistikas ei ole selline olukord üllatav, sest paljud faktorid on sõltuvuses kas üksteisest või hoopis mõnest kolmandast, varjatud faktorist. Endas kahtlev uurija saab regressioonimudelit kasutada mitte otse faktorite määratlemiseks ja nende osakaalu hindamiseks, vaid alles eeltööna tõeliste mõjurite leidmisel.

Eelnevast peaks olema näha, et raskused regressioonimudeli tõlgendamisel tulenevad selle aluseks olevast abstraktsioonist – *n*-mõõtmelises ruumis määratud funktsioon –, mille puhul pole selge, kuidas täita selliseid formaalseid nõudeid nagu ruumi telgede sõltumatus ning kas selline tekstisiseste tunnuste seostamine üldse aitab mõista seda, kuidas inimene keelt kasutab. Kriitikute (Milin jt 2016) sõnul ei ole ülalesitatud mudeldamisviis üldse sobiv inimkäitumise ja õppimise modelleerimiseks. Divjak jt (2016b: 78) järgi „[a]lgoritmid, millele standardsed statistilised klassifitseerijad nagu regressioonitehnikad toetuvad, ei ole loodud inimõpet imiteerima. [---] Nad esitavad kognitiivselt ebarealistlikke mudeleid, mis pakuvad kasutuspõhisele lingvistikale vähe huvi."

Eraldi küsimus on, milline on põhimõtteline eelhoiak probleemi suhtes, et keeles on grammatilise tähenduse väljendamiseks kasutusel mitu samaväärset viisi, mis ei erine tajutavalt tähenduse ega stiili poolest. Keele kui suhtlusvahendi funktsiooni silmas pidades ei ole selliste paralleelväljendusviiside olemasolu ootuspärane. Efektiivse suhtluse nõue suunab väljendeid genereerima alati ühel viisil ja ainult juhul kui seda miski takistab, on tulemus tavapärasest erinev. Ootuspärane oleks sel juhul, et sünonüümseid väljendusviise ei kasutata võrdselt, vaid üks neist on valdav, tavaline.

Kui aga samaväärsete paralleelväljendusviiside olemasolu tundub ootuspärane, siis ainus keele genereerimise mudel, mis on sellise ootusega kooskõlas, eeldab, et väljendeid genereeritakse juhuslikult, piltlikult münti või täringut visates, ning uurija ülesanne on leida faktorid, mis kallutavad mündi või täringu kas ühele või teisele poole. Sel juhul oleks ootuspärane, et sünonüümseid väljendusviise kasutatakse võrdselt, ükski pole valdav ega tavalisem kui teised. Kui vaatluse all on morfoloogilised paralleelvormid (näiteks sellised, mida käsitleti osades 1.1 ja 1.2), siis oleks potentsiaalselt paralleelvorme omavate sõnade sagedusjaotus ootuspäraselt normaaljaotuse sarnane: kõige rohkem oleks sõnu, millel samatähenduslikke vorme kasutatakse ühepalju, ning ainult üht vormi eelistavaid sõnu oleks vähe. Tegelik paralleelvormide kasutus on mõlema ootusega vastuolus.

Eelhoiak paralleelsuse põhjusesse võib kanduda üle ka uurimisküsimuse asetusse ja materjali valikusse. Kui arvatakse, et paralleelsus on ootuspärane, siis valitakse materjal nii, et ta kajastaks paralleelvariantide jaotust ühtegi varianti eelistamata, st nii, nagu algaks variantide valik mündiviskega. Näiteks Jane Klavan (2012: 103), uurides kaht samatähenduslikku konstruktsiooni (kaassõna *peal* ja alalütlev kääne), peab oluliseks, et mõlemat oleks arvutustele allutatavas materjalis võrdselt. Ta võtab konstruktsioone sisaldavad väljendid eri korpustest ega hooli omaenese tähelepanekutest, et esiteks, samatähenduslikke konstruktsioone ei ole korpuses võrdselt, ning teiseks, korpuses kasutatud kaassõna *peal* saaks asendada alalütleva käändega palju sagedamini, kui saaks teha vastupidist asendust, st et konstruktsioonide kasutus ei

ole sümmeetriline (Klavan 2012: 104). Analoogiliselt, uurides lühikest ja pikka sisseütlevat käänet, ütleb Siiman (2015: 211) kooskõlas mündiviskamise-eeldusega: „Seega peaksid siinse materjali teoreetilise ideaali järgi eesti keeles olema illatiiv ja aditiiv võrdselt kasutusel ehk mõlema esinemistõenäosus võiks olla 50%." Vaadanud korpusest läbi 2190 sisseütleva vormi sõnadelt, millel ÕS-i normingu kohaselt võiks olla lühike või pikk sisseütlev, ja leidnud kõigest 45 sõna, mis tõepoolest mõlemas vormis esinevad, ignoreerib ta sõnade ilmset tendentsi olla ainult kas ühes või teises vormis ning võtab edaspidiste arvutuste aluseks 840 sõna, millest ainult 41 esinevad ka tegelikult korpuses mõlemas vormis. Antud juhul näib, et teoreetiline eelhoiak paralleelvormide tekkemehhanismi suhtes on uurijaid arvandmete osas pimestanud.

Et tekstipõhiseid statistilisi mudeleid sisukamalt kasutada, oleks võib-olla huvitav lisada võrranditesse tekstiväliseid sõltumatuid parameetreid, nt kirjutaja päritolu või teksti loomisaja kohta. Sellega nihkuks vaatenurk keelekasutuse enesekesksuselt rohkem tekstiloomise kontekstile.

3. Statistiline õppimine: sagedus ja lingvistilise struktuuri kujunemine

Keeles leidub tõenäosuslikke mustreid, mida inimesed suudavad ilma juhendamiseta omandada. Kõnes ehk keelelises sisendis võiks eristada väga palju elemente, mille seoste kohta saaks tõenäosuslikke arvutusi teha. See tähendab, et olukord on täpselt vastupidine Noam Chomsky postuleeritud andmete vähesuse (ingl *poverty of stimulus*) probleemile (Saffran, Kirkham 2018).

Kui eeldada, et tähtis on ära õppida (ja inimesed panevad neid tähele) elementidevahelised seosed, mitte üksikelemendid, siis oleme silmitsi kombinatoorse plahvatusega. Kui elementide arv suureneb lineaarselt, siis nende kombinatsioonide arv eksponentsiaalselt. Sõnadevaheliste suhete mehaaniline meeldejätmine vajaks ebarealistlikult palju mälu. Pealegi on keelekasutusest (tekstikorpustest) näha, et kaugeltki kõik sõnad ei kombineeru teistega, st väga paljusid võimalikke sõnajadasid tegelikkuses ei esine, nii et väga paljude sõnade seoseid ei olegi justkui võimalik ära õppida.

Eelnev arutlus sõnadega eeldab lihtsustavalt, et õppija teab ette, millised on need elemendid, mis omavahel peaksid suhestuma. Tegelikult peab kõigepealt kindlaks tegema, millised tunnused eristavad olulisi üksusi, mille vahelisi suhteid tuleks tähele panna, ja millistel tunnustel pole tähtsust, st mis pole muud kui müra.

Kuidas oleks võimalik sellistes tingimustes, nii tõsiste raskuste puhul statistilist õppimist (SÕ) läbi viia? Ei tundu usutav, et lühimälu oleks piisavalt efektiivne kiiresti tulvavat sisendit säilitama ja töötlema või et pikaajaline mälu oleks piisavalt võimas, et säilitada või arvutada keerulisi sagedusjaotusi, ei täiskasvanutel ega ammugi lastel. Aga kuigi SÕ on põhimõtteliselt võimatu, on küllalt näiteid, et inimesed saavad sellega hakkama. Korpusuuringud ja eksperimendid tehiskeeltega on näidanud, et inimesed suudavad arvestada vähemalt kõrvuti olevate (ja teatud tingimuste korral ka mitte kõrvuti olevate) elementide tinglikke tõenäosusi ja sagedusjaotusi (Saffran 2009). See ei tähenda muidugi, et ka absoluutsagedused või -tõenäosused ei oma tähtsust.

SÕ uurimisparadigma seisneb tillukeste tehiskeelte loomises ja katsetes, mille eesmärgiks on uurida, milliseid statistilisi seaduspärasusi on inimesed võimelised ära tundma ja õppima. Uurimise algtõukeks on mingi loomulikus keeles avalduv nähtus, mille kohta ei ole selge, „kuidas selline asi võimalik on". Algul konstrueeritakse väike tehiskeel ja luuakse selles keeles piiratud hulk väljendeid, mida katsealustel lastakse laboritingimustes õppida. Õppematerjalis esinevate keele elementide sagedusjaotused on hoolikalt kontrolli all: üksikuid elemente ja nende sagedusjaotusi varieerides saab tekitada erinevaid keeli. Pärast õppimist lastakse katseisikutel moodustada oma lauseid või hinnata varem mitte kuuldud väljendite vastavust kuuldud keele reeglitele. Sel moel tuleb välja, milline on see keel, mille inimesed õppematerjali näinuna enda jaoks on tuletanud. Eksperimendid näitavad, milliseid struktuure on inimesed võimelised õppima ja millisest sisendist nad võiksid seda teha, ehkki nad ei näita, kas päris keeli ka tegelikult niimoodi omandatakse; viimast tuleks eraldi uurida.

Allpool refereeritakse kahte katset SÕ vallast, et anda lugejale konkreetsem tunnetus SÕ potentsiaalsest olulisusest korpuslingvistika jaoks. Uurimisküsimused ja -tulemused on arvatavasti üle kantavad ka eesti keele uurimisse, ehkki pole päris ilmne, kuidas seda teha. Sellest on tingitud ka juhtumikirjelduste detailsus: inspireerivaid aspekte on mitu ja milline neist oluliseks osutub, ei ole praegu selge, seega tuleb tähele panna paljusid asju.

3.1. Klassifitseerimine: järgnevustõenäosused

Reaalses kõnes ei ole pausid tingimata seal, kus asuvad sõnade vahed. Kuidas siis õpivad inimesed sõnu eraldama, st häälikuid õigesti grupeerima? Tavaliselt on keeles nii, et häälikule teise hääliku järgnemise tõenäosus on suurim sõna sees, aga sõnavahet ületav häälikute järgnemise tõenäosus on suhteliselt väike. Sõnavahe puhul pole tähtis tõenäosuse täpne suurus, vaid see, et ta on ümbritsevatest väiksem. Saffran jt (1996) näitasid, millise kergusega on inimesed võimelised järgnevustõenäosusi arvutama ja kasutama. Nad andsid 8-kuustele lastele, teises katses aga täiskasvanutele, kõigest mõne minuti vältel kuulata CV silbijadasid, milles puudusid pausid, kuid mis olid moodustatud 3-silbilistest „sõnadest" nii, et silpide järgnevustõenäosused erinesid: „sõna"-siseselt 1,0 ja „sõnade" vahel 0,33 (see saavutati sõnade järjekorra varieerimise abil). Seejärel mängisid nad ette nii varem kuuldud kui ka varem kuulmata „sõnu". Viimased koosnesid tuttavatest silpidest, kuid silbijada sisaldas järgnevust, mis varem esines ainult kahe „sõna" vahel. Näiteks kui varem oli kõlanud kahe sõna jada *da*(1,0)*ro*(1,0)*pi*(0,33)*go*(1,0)*la*(1,0)*tu*, siis uueks „sõnaks" oli *pi*(0,33)*go*(1,0)*la*. Katsealused oskasid eristada kuuldud „sõnu" uutest „mittesõnadest", st tegid kindlaks sisendi struktuuri, ehkki sellest ei andnud märku mitte miski peale järgnevustõenäosuste, st ei olnud mingit tüüpilist algus- või lõpusilpi.

3.2. Klassifitseerimine: sisendi piisavus

Wonnacott jt (2008) märgivad, et keeleoskus hõlmab oskust tasakaalustatult kasutada nii sõnapõhiseid kui ka üldisemaid grammatilisi mustreid. Mustrite keerukas vahekord tuleb eriti selgelt välja verbide ja nende argumendistruktuuride puhul. Üheks palju kasutatud näiteks on siin inglise ditransitiivne konstruktsioon. Paljude verbidega võib seda kasutada eessõnakonstruktsiooni asemel, nt *John gave a toy to Mary* asemel võib öelda *John gave Mary a toy* ('John andis Maryle mänguasja'). Samasugune kahetine väljendus osutub võimalikuks ka väljamõeldud verbidega, nagu on näidanud katsed (Gropen jt 1989). Kuid sellist asendust ei saa teha kõigi verbidega, nt *carried* ('kandis') seda ei luba. See tähendab, et verbid jagunevad alamliikidesse: mõned on piiratud selle poolest, millistes konstruktsioonides nad võivad esineda.

Keeleõppega seotud probleem on siin järgmine: kui keelekasutajad üldistavad ditransitiivi lubatavust uutele verbidele, siis mille järgi nad teavad, et mõne verbi puhul oleks see üleüldistus ja seega ekslik? See, et nad pole viimaseid ditransitiivses konstruktsioonis näinud kasutatavat, ei pruugi ju tähendada seda, et neid ei tohikski seal kasutada, vaid võib-olla tähendab lihtsalt seda, et seni pole neid ette tulnud. Sellist õppimisprobleemi nimetatakse Bakeri paradoksiks (Baker 1979). Ka lastel on keele õppimise ajal üleüldistamise periood, kuid viimaks nad Bakeri paradoksi mingil viisil lahendavad.

Üks võimalus on oletada, et liigitamise aluseks on verbi tähendus. Lila Gleitman ja Barbara Landau (2012) näitavad siiski, et verbi argumendistruktuuri omandamiseks ei ole tingimata vaja, et lapsed verbi tähendust teaksid. Seega tuleb oletada, et argumendistruktuur omandatakse tähendusest sõltumatult, tuginedes verbi ja struktuuri kooseksisteerimisele sisendis.

Katseks loodi tehiskeel (Wonnacott jt 2008), mis koosneb kahest lausemallist: 1) verb alus sihitis (VAS), 2) verb sihitis alus partikkel (VSA_ka), näiteks 1) *glim tombat nagid* ja 2) *glim nagid tombat ka* ('kaelkirjak lõi elevanti'). Keeles on viis nimisõna ja kaksteist ühesilbilist verbi. Nimisõnad võivad esineda vabalt mõlemas konstruktsioonis ja nii aluse kui ka sihitisena. Sõnavormid ei muutu, seega kuni partikli *ka* nägemiseni ei ole teada, kas *tombat nagid* tähendab 'kaelkirjak elevanti' või 'elevant kaelkirjakut', v.a juhul, kui verb *glim* on võimalik ainult ühes kahest võimalikust konstruktsioonist. Verbid võivad olla kasutusel kas ainult VAS-konstruktsioonis, ainult VSA_ka-konstruktsioonis või nii ühes kui ka teises. Eri katsetes olid muutujateks verbide jaotus eri verbiklassidesse (st verbiklasside suhteline suurus), verbi üldine esinemissagedus ja konstruktsioonis esinemise sagedus.

Esimeses katses esines neli verbi ainult VAS-mallis, neli verbi ainult VSA_ka-mallis, neli verbi pooltel juhtudel ühes, pooltel juhtudel teises mallis. Seejuures kaks verbi VAS-mallis ja kaks VSA_ka-mallis esinesid sagedamini kui mõlemas mallis esinevad verbid, kaks verbi nii VAS-mallis kui ka VSA_ka-mallis aga harvemini kui mõlemas mallis esinevad verbid.

Teises katses esines kaks verbi ainult VAS-mallis, kaks verbi ainult VSA_ka-mallis, kaheksa verbi pooltel juhtudel ühes, pooltel juhtudel teises mallis; kahe-malli-verbe esines kokku kaks korda sagedamini kui ühe-malli-verbe ja kõigis klassides esinesid eri verbid eri sagedusega.

Tulemuseks oli esiteks see, et peaaegu kõik katsealuste moodustatud lausedolid VAS- või VSA_ka-mallis (mitte VSA ega VAS_ka), mis tähendab, et võimalikud lausemallid omandatigi õppimise käigus. Esimeses katses õpiti õigeks pidama just neid lauseid, mille verbi oli õppimise faasis nähtud täpselt samas mallis, kuigi sisendis oli verbe, mis esinesid mõlemas mallis ja oleksid seega võinud kallutada üleüldistavale arvamusele, et verbe võib kasutada ka varem mitte nähtud mallis. Sagedaste ühe-malli-verbide klassifitseerimises olid katseisikud kindlamad kui harvade ühe-malli-verbide puhul.

Seevastu teises katses peeti sobivateks ka selliseid lauseid, mille verb õppimise faasis samas mallis ei esinenud. Kuigi ühe-malli-verbid esinesid õppematerjalis täpselt sama moodi eranditult oma spetsiifilises mallis kui esimeses katses, usaldati seekord seda infot vähem. Lisaks ilmnes, et nii sagedamaid ühe-malli-verbe kui ka harvemaid liigitati kahe-malli-verbideks võrdsel määral, erinevalt katsest 1, kus sagedamad liigitati kindlamalt ühe-malli-verbideks. Kahe-malli-verbide suurem hulk muutis suhtumist sellesse, kuivõrd ammendavaks saab pidada ühe-malli-verbi kohta nähtud infot, st (umb)usaldati ühe-malli-verbe selles osas, kas nad õppeperioodi jooksul ikka ilmutasid oma täieliku sagedusjaotuse.

Kõigil katsealustel kujunes keeletaju, mis peegeldas õppematerjali sagedusjaotusi. Nad õppisid ära nii verbispetsiifilised tõenäosused, et verb on mingis mallis, kui ka mallide tõenäosused, millega need keeles esinevad. Õpitu sõltus verbi enese sagedusest – verbispetsiifilist statistikat usaldati harvaesinevate verbide puhul vähem – ja verbimallide jaotusest keeles – verbispetsiifilist infot ignoreeriti rohkem sel juhul, kui keeles oli suur mitut malli lubavate verbide klass. Seega verbide käsitlemine sõltus lisaks verbi enese käitumisele ka teiste verbide käitumisest; õppija pidi tasakaalustama oma teadmisi ühe verbi kohta teadmistega verbide kohta üldiselt. Tasakaalustamine sõltus sisendi sagedusjaotusest, mis tekitas õppijas suurema või väiksema kindluse, et nähtud andmed on teatud järelduste tegemiseks piisavad.

Need katsed (Saffran jt 1996; Wonnacott jt 2008) võiksid anda suuniseid, millist liiki statistikat korpusuuringutes maksaks arvesse võtta. Näiteks eesti sõnamuutmise osas: kuidas eesti lapsed õpivad ära, mil moel jagada sõnu muutkondadesse, et muutmismallil on üldse seos sõna häälikulise ja tuletisliku kujuga ning et tähele tuleb panna just lõpuhäälikuid ja rõhulise silbi asukohta.

4. Arutelu

Strukturalistide eeskujul maksab eristada ühelt poolt keelt kui süsteemi ja teiselt poolt selle süsteemi toimimise kaudu tekkivat keelematerjali, tekstikorpust. Samas keel kui süsteem ise on õpitav sel moel, et õppija sisendiks on seesama keelematerjal, st miski, milles keele kui süsteemi toimimise jäljed on selgelt nähtavad. Kuid jälgede olemasolu ei tähenda seda, et nad oleksid lingvistiliselt kergesti äratuntavad või tõlgendatavad. Korpuslingvist, kes püüab näha olevate keeleüksuste ja -struktuuride põhjal mõista, milline on neid tekitav mehhanism, tugineb loomulikule eeldusele, et korduvus on tähenduslik, olles tihedalt seotud ennustatavusega. Oma keelematerjali tõlgendamiseks

saab lingvist laenata võtteid statistikast ja tõenäosusteooriast, st teadusharudest, mis tegelevad just nimelt eri liiki korduvuste ja ennustatavusega. Seejuures tuleb aga meeles pidada, et need distsipliinid ei anna juhiseid algandmete valimise kohta (v.a algandmetele seatud formaalsed nõuded, nt et sagedusjaotus peaks mõnel juhul olema normaaljaotuse sarnane või et mitme muutujaga regressioonvõrrandi sõltumatud muutujad ei tohiks omavahel korreleeruda).

Morfoloogiauuringutest on ilmnenud, et sõna muutmisviisi reeglipära/erandlikkust võimaldab seletada sõnavormide teksti- ja sõnastikusageduste suhe, kusjuures arvesse tuleb võtta nii konkreetset sõnavormi kui ka terve keelekorpuse kõiki sõnu. Statistilise õppimise raames läbi viidud katsed väikeste tehiskeeltega on näidanud, et arvesse tulevad veel järgnevustõenäosus ja leksikaalsete üksuste sagedusjaotus terves korpuses. Näib, et eri reeglite äraõppimiseks peab õppija märkama mitte ainult erinevaid keeleüksusi, vaid ka erineval moel määratletud tõenäosuslikke sündmusi, millele ta oma arvutusi rakendab.

Kunst on need arvutuste aluseks olevad sündmused ära tunda: kas peaks tähele panema lingvistilise elemendi esinemissagedust või tervet jaotust korpuses või leksikonis, või hoopis järgnevustõenäosusi või kontekstide osalist kattuvust? Niisiis pole raske mitte keerulise(ma) arvutusmudeli, vaid õigete üksuste leidmine, mida kokku lugeda ja mille peal statistikat teha.

Artikli valmimist on toetanud Euroopa Liidu Euroopa Regionaalarengu Fond (Eesti-uuringute Tippkeskus) ja HTM-i uurimistoetus IUT20-56 „Eesti keele arvutimudelid".

Kirjandus

Baker, Carl L. 1979. Syntactic theory and the projection problem. – Linguistic Inquiry, kd 10, nr 4, lk 533–581.

Bybee, Joan L. 1995. Diachronic and typological properties of morphology and their implications for representation. – Morphological Aspects of Language Processing. Toim Louis B. Feldman. Hillsdale, NJ: Lawrence Erlbaum Associates, lk 225–246.

Divjak, Dagmar, Levshina, Natalia, Klavan, Jane 2016a. Cognitive linguistics: Looking back, looking forward. – Cognitive Linguistics, kd 27, nr 4, lk 447–463.

Divjak, Dagmar, Arppe, Antti, Baayen, Harald 2016b. Does language-as-used fit a self-paced reading paradigm? (The answer may well depend on how you model the data.) – Slavic Languages in Psycholinguistics: Chances and Challenges for Empirical and Experimental Research. Toim T. Anstatt, A. Gattnar, C. Clasmeier. Tübingen: Narr Francke Attempto Verlag, lk 52–82.

EKK = Mati Erelt, Tiiu Erelt, Kristiina Ross 2007. Eesti keele käsiraamat. Kolmas, täiendatud tr. Tallinn: Eesti Keele Sihtasutus.

Gleitman, Lila R., Landau, Barbara 2012. Every child an isolate: Nature's experiments in language learning. – Rich Languages from Poor Inputs. Toim Massimo Piattelli-Palmarini, Robert C. Berwick. Oxford: Oxford University Press, lk 91–104.

Gropen, Jess, Pinker, Steven, Hollander, Michelle, Goldberg, Richard, Wilson, Ronald 1989. The learnability and acquisition of the dative alternation in English. – Language, kd 65, nr 2, lk 203–257.
Hasselblatt, Cornelius 2000. Eesti keele ainsuse sisseütlev on lühike. – Keel ja Kirjandus, nr 11, lk 796–803.
Hopper, Paul J., Bybee, Joan L. 2001. Introduction to frequency and the emergence of linguistic structure. – Frequency and the Emergence of Linguistic Structure. Toim J. L. Bybee, P. J. Hopper. Amsterdam–Philadelphia: John Benjamins, lk 1–24.
Kaalep, Heiki-Jaan 2009. Kuidas kirjeldada lühikest sisseütlevat kasutusandmetega kooskõlas? – Keel ja Kirjandus, nr 6, lk 411–425.
Kaalep, Heiki-Jaan 2010. Mitmuse osastav eesti keele käändesüsteemis. – Keel ja Kirjandus, nr 2, lk 94–111.
Kaalep, Heiki-Jaan 2012. Eesti käänamissüsteemi seaduspärasused. – Keel ja Kirjandus, nr 6, lk 418–449.
Kio, Kati 2006. Sisseütleva käände kasutus eesti kirjakeeles. Magistritöö. Tartu. http://dspace.ut.ee/handle/10062/865
Klavan, Jane 2012. Evidence in Linguistics: Corpus-linguistic and Experimental Methods for Studying Grammatical Synonymy. (Dissertationes linguisticae Universitatis Tartuensis 15.) Tartu: Tartu Ülikooli Kirjastus.
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.
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.
Saffran, Jenny R. 2009. What is statistical learning, and what statistical learning is not. – Neoconstructivism: The New Science of Cognitive Development. Toim Scott Johnson. New York: Oxford University Press, lk 180–195.
Saffran, Jenny R., Aslin, Richard N., Newport, Elissa L. 1996. Statistical learning by 8-month-old infants. – Science, kd 274, nr 5294, lk 1926–1928.
Saffran, Jenny R., Kirkham, Natasha Z. 2018. Infant statistical learning. – Annual Review of Psychology, kd 69, lk 181–203.
Schmid, Hans-Jörg 2010. Does frequency in text instantiate entrenchment in the cognitive system? – Quantitative Methods in Cognitive Semantics: Corpus-driven Approaches. Toim Dylan Glynn, Kerstin Fischer. Berlin–New York: De Gruyter Mouton, lk 101–136.
Siiman, Ann 2016. Ainsuse sisseütleva vormi valiku seos morfosüntaktiliste ja semantiliste tunnustega – materjali ning meetodi sobivus korpusanalüüsiks. – Emakeele Seltsi aastaraamat, kd 61 (2015). Tallinn: Teaduste Akadeemia Kirjastus, lk 207–232.
Viitso, Tiit-Rein 2003. Structure of the Estonian language. – Estonian Language. (Linguistica Uralica. Supplementary series 1.) Toim Mati Erelt. Tallinn: Estonian Academy Publishers, lk 9–129.
Wonnacott, Elizabeth, Newport, Elissa L., Tanenhaus, Michael K. 2008. Acquiring and processing verb argument structure: Distributional learning in a miniature language. – Cognitive Psychology, kd 56, nr 3, lk 165–209.

ÕS 2013 = Eesti õigekeelsussõnaraamat ÕS 2013. Toim Maire Raadik. Koost Tiiu Erelt, Tiina Leemets, Sirje Mäearu, M. Raadik. Tallinn: Eesti Keele Sihtasutus, 2013.

Heiki-Jaan Kaalep (*snd 1962*)*, PhD, Tartu Ülikool, vanemteadur,*
Heiki-Jaan.Kaalep@ut.ee

Place of statistics in a language model

Keywords: morphology, corpus linguistics, linguistic variation, text statistics

The article speculates on how quantitative data may fit into a theoretical model of language. It argues that the language model should include an idea about the generation procedure at play, albeit a speculative one. A concrete example shows how quantitative data form an integral part of a model of Estonian morphology, another concrete example shows how corpus-based statistical models may result in dubious statistical calculations, and two descriptions of old experiments in statistical learning show a potential path worth following in corpus linguistics in the future: one should pay more attention to some not-so-obvious features that play a role in human language learning, namely, transitional probabilities and linguistic units that should be left out from computations.

Heiki-Jaan Kaalep (*b. 1962*)*, PhD, University of Tartu, Senior Researcher,*
Heiki-Jaan.Kaalep@ut.ee