

Keelekorpus kui leksikograafi abiline kõnekeelsuse tuvastamisel

LYDIA RISBERG, MARIA TUULIK, MARGIT LANGEMETS,
KRISTINA KOPPEL, ENE VAINIK, ESTA PRANGEL, ELERI AEDMAA

Leksikograafi ehk sõnaraamatukoostaja tööriistu arendatakse pidevalt, misjärel tasub tööprotsesse kogu aeg üle vaadata ja kohendada (vt ülevaadet Klosa-Kückelhaus, Tiberius 2024). Keele uurimiseks kasutatakse tänapäeval mahukaid keelekorpusaid (vt ka Baayen 2024; Rundell 2024; Davies 2025), millele Eesti leksikograafias on tuginetud juba 1990. aastatest. Spetsiaalsete leksikograafiliste tööloikude jaoks on Eesti Keele Instituudis (EKI) loodud ka korpusandmete analüüsi tööriistu, näiteks kalkulaatoreid sõnaliigi muutumise tuvastamiseks (vt nt Vainik jt 2021; Paulsen jt 2023). Siin artiklis keskendume sellele, kuidas saab kasutada korpusanalüüsi sõnaraamatumärgendite, konkreetsemalt kõnekeelsuse märgendi üle otsustamisel.

Märgendite esitamise järele sõnaraamatus on alati olnud ühiskondlikku nõudlust, nii ka praegu. Näiteks otsivad keeleõppijad sõnastikust infot sõnade kasutusregistrite kohta (Pool jt 2025), ent on väljendatud ka soovi, et 2025. aastal ilmuvas õigekeelsussõnaraamatus (ÕS) oleks kirjakeele kiht välja sõelutud (Päärt 2023), või ka, et ÕS esitaks üksnes neutraalset sõnavara (Nemvalts 2023). Nagu teistes sõnastikes, nii on seni ÕS-iski olnud erisugust keelematerjali – seda on väljendanud ka ÕS 1999 toimetaja Tiiu Erelt (2002: 74): „Sõnaraamatu plussiks julgen pidada rohket väljendite tõlkimist ühest allkeelest, registrist, stiilist teise, hrl neutraalkirjakeelde.”

Keskendume kõnekeelele kui neutraalsest keelest irduval stiilitasandile. Valisime kõnekeele praktilisel kaalutlusel, sest erinevalt näiteks vulgaarsest või lastekeelest, mida on sõnaraamatutes samuti märgendatud, on kõnekeelsust olnud leksikograafil keerulisem tabada, sest see on subjektiivsem ja vähem äärmuslik register. Peale selle on ÕS-is mõnele sõnale lisatud kõnekeelsusele osutav märgend (ARGI) keelekorralduslikel kaalutlustel (vt ka Raag 2008: 11–12; Risberg 2024a: 53–57). Süsteemselt tasub üle vaadata ka EKI ühendsõnastikus (ÜS; keeleportaalis Sõnaveeb) olevad kõnekeelsuse märgendiga keelendid, sest normatiivse ÕS 2018 integreerimine on võinud mõjutada ÜS-i keeleinfot, olgugi et järgmise, ÕS 2025 koostamise käigus on ÕS 2018 infot üle analüüsitud ja ajakohastatud (nt Paet 2023; Risberg 2024a; vt Langemets 2025).

Uurisime, kuidas saab leksikograaf korpusinfost tuge kõnekeelsuse märgendi lisamise või eemaldamise üle otsustamisel. Millal on alust sõna või väljendi tähendusele sõnaraamatus kõnekeelsusele osutav märgend lisada?

1. Registrid ja sõnaraamatumärgendid

Terminite *register*, *žanr*, *allkeel*, samuti *kirjakeel* ja *kõnekeel* kasutuses ning definitsioonides pole teoreetilistes käsitlustes olnud ühist kokkulepet. Terminikasutus on põhinenud pigem autorite eelistustel, samas ongi need mõisted hāgusate piiridega ja osaliselt kattuvad. (Kuzman, Ljubešić 2023; Lindström jt 2023; Vaik 2024) Jārgnevalt avame seda mõistestikku, taotlemata ebaūhtluse lahendamist (alapeatūkk 1.1), ning kirjeldame sõnaraamatumārgendite tausta (alapeatūkk 1.2).

1.1. Mõistestik

Tiit Hennoste ja Karl Pajusalu (2013: 13) on *allkeele* defineerinud kui keele variandi, „mida kasutatakse järjekindlalt teatud kasutussituatsioonides või inimrühmades ning millel on omad keelelised kasutusnormid või -normingud”. Nad on kirjeldanud kolme võimalust allkeeli liigitada: 1) normitud keele, 2) kasutussituatsioonide ja 3) kasutajate (murreteūlene ūhiskeel vs. murded) jārgi. Teise variandi, st kasutussituatsioonide omaduste kaudu on nad mõtestanud just *registreid*: keelelised valikud olenevad sellest, kas tegu on suulise või kirjaliku, avaliku-ametliku või kõnekeelse suhtlusega (Hennoste, Pajusalu 2013: 13). Siin artiklis jārgime sama mõtestust, seda toetab ka Anke Lüdelingi jt (2022: 3) tõdemus, et registrid on keelteūlesed ja registri all mõeldakse sageli selliseid kategooriaid nagu ametlik keel ja kõnekeel.

Mõisteid *register* ja *žanr* on eristanud mitmed autorid: registrit iseloomustavad tekstist sõltumatud keelelised tunnused, samas kui *žanr* iseloomustab just teksti (vt Lee 2001; ka Vaik 2024: 24–26). Douglas Biberi ja Susan Conradi (2009: 10) jārgi peab registrianalūūs tuginema tekstivalimile, mis esindab registrit võimalikult teraviklikult, see ei tohi põhineda ainult ūhel tekstil või ūhe autori stiilil.

Artiklis osutab *žanr* konkreetsele tekstiliigile, nāiteks blogitekstile. Blogipostitus vōib olla nii informaalses registris, sisaldades muu hulgas palju vārvikaid kõnekeelseid sõnu, kui ka neutraalses registris. Mōlemal juhul vōidakse blogipostituses jārgida kirjakeele norme. Register aga paigutab keelelise vāljenduse informaalse-neutraalse-formaalse skaalale, sõltumata *žanrist*. Ūhe *žanri* sees vōivad tekstid registriliselt varieeruda.

Artikli fookuses oleva kõnekeelega seostub termin *kirjakeel*, mida on eri autorid samuti aja jooksul kasutanud (kohati vāga) erineva sisu tāhistamiseks. Olenevalt käsitlusest on mōeldud nāiteks 1) neutraalset keelt vastandina sageli vārvikale kõnekeelele, 2) ranget normikirjakeelt ehk seadusega ette nāhtud kirjakeele normile vastavat ametlikku keelekasutust vastandina muule vabamale keelekasutusele, kus jārgitakse kasutuses kujunenud norme, vōi 3) kirjutatud keelt vastandina kōneldud keelele (Lindström jt 2023: 8–9). Artiklis keskendume kõnekeelele kui neutraalkeelest vārvikamale registrile.

Terminit *standardkeel* on samuti kasutatud eri sisu tāhistamiseks, vastandades seda kas 1) rangetest normingutest vabamale kõnekeelele vōi 2) murretele. Artiklis kasutame sõna *ūhiskeel*, kui mōtleme murreteūlest ūhiskeelt, mida sajand tagasi loodi ja mida tānapāevaks ūle Eesti rāāgitakse. Mōistet *ūldkeel*, mida kasutatakse

vanusest, ametist, haridusest, elupaigast olenemata, vastandame erialakeeltele, kus kasutatakse erialaspetsiifilisi termineid.

1.2. Märgendid sõnaraamatutes

Sõnaraamatumärgendite lisamine on sõltunud sellest, kas sõnastik on deskriptiivne või preskriptiivne. Anke Lüdelingi jt (2022: 4, 9) järgi peavadki keeleallikad (nagu sõnaraamat) osutama, et teatud keelendid või grammatilised vormid sobivad kindlatesse olukordadesse. Samas on nad nentunud, et preskriptiivsetes sõnastikes ja grammatikates on selliseid väiteid esitades toetunud vaid väiksele ja piiratud tekstivalimile, mistõttu on esitatud vähe või juhuslikke tõendeid. (Lüdeling jt 2022) Teisalt ei pruugi ka deskriptiivsed allikad alati olla objektiivsed (Wit, Gillette 1999: 2; Pullum 2023).

Leksikograafias pole välja kujunenud üldist ühtset süsteemi, kuidas sõnakasutust märgendite abil täpsustada: märgendite valik võib eri keeltes ja eri sõnakogudes varieeruda väga suurel määral. Enamasti on sõnastikes kasutusel registri- (nt *kõnekeelne*) ja erialamärgendid (nt *bioloogia*, *õigusteadus* või selgitus „oskuskeeles täpsem”), aga märgend võib piiritleda ka kasutussagedust (nt *harv*), aega (nt *uus*, *vanamoodne*), stiili (nt *stiilitundlik*, *luulekeelne*), suhtumist (nt *halvustav*), geograafilist piirkonda (nt *Briti inglise*, *murdes*) või anda muud infot. Neutraalset üldkeelset sõna pole tavaks märgendiga varustada, nii on enamik sõnu sõnaraamatus märgendita.

Sõnaraamatumärgend osutab sellele, et sõna kaldutakse kasutama mujal kui neutraalses kontekstis, näiteks kõnekeelsuse märgendit saab käsitleda informaal-sele registrile osutavana. Sue Atkinsi ja Michael Rundelli (2008: 185) käsitluses on osad keelendid neutraalkeelest formaalsemad ja teised informaaalsemad, mistõttu oleks sõnaraamatus nende hinnangul otstarbekas eristada keelendi formaalsuse astet: neutraalsest ametlikumat registrit märgiks näiteks *ametlik* (ingl *formal*, vahel isegi *official* 'ametlikult kehtiv'), neutraalsest informaaalsemat registrit aga *mitte-ametlik* (*informal* või selle variatsioonid, nagu *familiar*, *casual*, *relaxed*). Samas on sõnaraamatuesituse üldisem probleem see, et iga kirjeldatav sõna on loomulikust kontekstist väljas ja nii ei saa märgend määrata universaalset normi selle sõna igale kasutusjuhule, ehkki seda on kipunud nõnda tõlgendada (Langemets jt 2024: 753). Samuti on kasutajad märgendeid võtnud kui (normatiivset) märguannet või hoiatust sõna kasutuspiirangu kohta (vt nt Trap-Jensen 2002). Briti entsüklopeedia Britannica kodulehel on lausa hoiatatud, et inimesed, kes (sõnastikumärgendite najal) jagevad keelekasutust „õigeks” ja „valeks”, ei mõista, kuidas keel tegelikult toimib (Read s. a.; vt ka Baayen 2024: 634). Seepärast on seal soovitatud olla märgendite lisamisel pigem säästlik: paremini illustreerivad keelekasutust hoopis (autentsed) näitelauseid.

Eesti seletussõnaraamatuid (EKSS 1988–2007, 2. trükk 2009) on alati koostatud keelenäidete põhjal: kuni 1990-ndateni kirjakeele arhiivi najal (ca 4 mln sedelit), alates 1990-ndate algupoolest on kasutatud elektroonilisi tekstikogusid (EKSS 2009: 5–6). 2010. aastatest on kasutusel mahukas eesti keele ühendkorpus (Koppel, Kalas 2022) ja spetsiaalsed korpustööriistad, nagu Sketch Engine. EKI ühendsõnastiku (ÜS) selgrooks saanud „Eesti keele sõnaraamat” (EKS 2019) toetus korpusanalüüsile

(Langemets jt 2018). Märgendite valikul on EKI ühendsõnastikus järgitud EKSS-i tava, sh on kasutatud märgendit KÕNEKEELNE (mitte ARGİ, nagu ÕS-is). Kõnekeelseks märgendatud keelendeid (täpsemalt, nende tähendusi) on seal 2025. aasta alguse seisuga umbes 5000 (kokku on ÕS-is 180 000 märksõna).

Leksikograafid on keeleandmeid ja nende toel registrist kuuluvust analüüsinud ja hinnanud küll näidete varal, ent varem pole olnud ühist suunist, kuidas seda teha, ning iga otsuse taga on olnud konkreetne inimene oma keeletajuga. Seletussõnaraamatu otsuseid on osaliselt mõjutanud ka varasemates ÕS-ides kujundatud keelekorralduslikud ettekirjutused, näiteks on 2009. aasta EKSS-is sõna *õieti* tähenduse 'õigesti' juures kõnekeelsusele osutav märgend, sest alates 1999. aasta ÕS-ist jäeti see tähendus välja (vt lähemalt Risberg, Langemets 2021: 912–914).

ÕS 2018-s on neutraalkeelest irduvad sõnad tähistatud märgendiga ARGİ ('argikeelne'), aga ka MURDES, VULG ('vulgaarne'), LASTEK ('laste- või hoidjakeelne') jt. ÕS 2018-s on argi- ehk kõnekeelseid sõnu ja väljendeid ligi 2300 (kokku on ÕS-is 50 000 märksõna). Kuigi ÕS-ide avaldamise järel on autorid märgenditest siin-seal rääkinud (nt Erelt 2002; Raadik 2014; ka Kerge 2004; vt analüüsi Langemets, Risberg 2023), siis seda, miks konkreetsed sõnad on argikeelseks määratletud, peab uurima üksiksõnade kaupa eri allikatest (nt Henn Saari kogumikest, vt Risberg 2024a: 51). Kõiki valikuid pole tõenäoliselt ka kirjeldatud ega põhjendatud.

Tõlgendamisraskusi on tekitanud see, et ÕS 1999-s ja ÕS 2006-s kasutati ARGİ-märgendit kahes tähenduses: kõnekeele eristamiseks kirjakeelest (nt sõna *kõiksugune* märgendita, *kõiksugu* märgendiga ARGİ) ning oskuskeele eristamiseks üldkeelest (nt märgendiga AJ ('ajalugu') *bürgermeister* 'linnapea, linnavalitsuse eesistuja' vs. samatähenduslik *pürjermeister* märgenditega VMO ('vanamoeline') ja ARGİ) (vt ka Erelt 2002; kriitika Vare 2001).¹ Krista Kerge (2004: 17) on ARGİ-märgendit tõlgendanud kui osutust „vaba registri stiililt neutraalsele sõnale”, mis ei sobi ametlikku teksti. Samas on ARGİ-märgendiga tähistatud ka värvikaid sõnu, mis pole stiililt neutraalsed (nt *ahnepäits*, *lābustama*, *munajoodik*). Ametnike formaalse, pahatihti kantseliitliku keelega seostub ÕS-i märgend PABERL ('paberlik, kuivametlik') (81 juhtu, neist enamus näitelauses), mis suunab vastavaid väljendeid ametlikus keeles vältima.

Kokkuvõttes on eesti leksikograafias senised märgendiga sõnad põhjalikult ja süsteemselt läbi uurimata. Samas tõlgendatakse märgendeid objektiivse informatsioonina, mitte ei peeta subjektiivseks (Pajusalu 2009: 6). Keeleallikates, eriti vaneimates, on paljusi siiski olnud tegemist subjektiivsete otsustega, nagu on Rudolf Karelson osutanud ka eesti kirjakeele seletussõnaraamatu koostamise kohta: „Ent kõigest hoolimata võivad kõik meie poolt sõnade **kõnekeelsusele**, humoristlikule, halvustust väljendavale jm. kasutusele viitavad märgendid osutada täiesti subjektiiv-

¹ Alates ÕS 2013-st vähendati ARGİ-märgendi koormust ja hakati kasutama üldkeelsõna juurest termini juurde juhatavat osutust *oskuskeeles täpsem* (Raadik 2014: 72–73), ent mitte kõikide selliste sõnade juures. Peale selle on erialamärgendeid ÕS-is aja jooksul vähendatud, muutmata sõnartiklis muud infot, ja nii on sõna kirjeldus jäänud lähtuma erialakeelest (Paet, Risberg 2021: 969). ÕS 2018-s on erialamärgendeid endiselt rohkesti, 13 000 sõna juures, samas muid märgendeid on märgatavalt vähem: kokku umbes 6600 sõna või väljendi juures.

seteks, sest nad **rajanevad peamiselt sõnaraamatu autorite isiklikul keeletundel**” (Karelsen 1990: 34, meie rõhutused). Sellest tuleneb ka vajadus süsteemsemate ja valideeritavamate meetodite ja ühise suunise järele, millele saaks leksikograaf märgendi lisamise töös otsust langetades tugineda.

2. Korpustööriista kasutamise võimalused

Sõnaraamatutöö jaoks on viimastel aastatel katsetama hakatud ka suuri keelemudeleid (SKM-id; vt ülevaadet de Schryver 2023; ka EKI-s tehakse katsetusi, vt nt Jürviste jt 2025; Tuulik jt (ilmumas)), mida pidevalt arendatakse ja mis annavad vastuseid üha paremas eesti keeles. SKM-id võivad öelda töönaosusliku hinnangu oma treeningandmetele tuginedes, ent nad ei anna usaldusväärseid viiteid ega võimalust vaadata reaalselt esinenud kasutusjuhtumeid sellises kontekstis, nagu neid saab vaadelda korpuses. Küll aga saab kasutada SKM-e korpusedmetele analüüsimiseks. (Vt nt Davies 2025)

Korpuseid on kasutatud registri või stiili uurimiseks tekstitasandil (vt nt Cvrček 2024; Eder 2025), meie lähtusime aga keelenditest, sest leksikograafi ülesanne on kirjeldada just keelendit selle kõigis tähendustes. Eesti keele uurimiseks on leksikograafil vaja kiiret ning tõhusat vahendit, mis aitaks hinnata, millises registris sõna(tähendust) tüüpiliselt kasutatakse.

2.1. Tekstiliigid keelekorpuses

EKI ühendsõnastiku koostamisel on algusest peale toetunud korpusanalüüsile (Langemets jt 2018): 2010. aastatest on kasutatud tarkvarafirma Lexical Computing Ltd. arendatavat korpustööriista Sketch Engine (Kilgariff jt 2014). Korpustööriista kaudu ligipääsetavas eesti keele ühendkorpuste sarjas (Koppel, Kallas 2022) on 2025. aasta seisuga ilmunud viis versiooni. Hiljutisimas, eesti keele ühendkorpuses 2023, on 13 allkorpust, millest suurimad on veebikorpused. Erinevalt traditsioonilistest (nt tasakaalus ja erialastest) korpustest, millesse valitakse tekstid kindlaksmääratud tekstiliike (žanreid) silmas pidades, pole veebist kogutud korpuse puhul teada, mida see täpsemalt sisaldab. Kuna veebitekstid pole üldlevinud taksonoomia järgi klassifitseeritud, tuleb neid korpusesse lisades klassifitseerida kas käsitsi või (pool)automaatsete meetoditega. (Vaik jt 2020) Tarkvarafirma kasutab veebitekstide klassifitseerimiseks poolautomaatseid meetodeid ning ka eesti keele ühendkorpuses on kasutusele võetud nende pakutud klassifikatsioon (muude klassifitseerimisvõimaluste kohta vt Vaik jt 2020; Vaik 2024).

Alates 2021. aasta ühendkorpuse versioonist on kasutusel kahetasandiline klassifikatsioon: žanrid (ingl *genre*, laiem klass) ja teemad (*topic*, kitsam klass, nt sport, poliitika, tervis) (Koppel, Kallas 2022). Teema määrab teksti sisu, mida on võimalik ära tunda peamiselt teksti leksikaalsete omaduste, st selles kasutatavate sõnade kaudu. Žanrit on võimalik määratleda teksti süntaktiliste ja leksikaalsete omaduste järgi ehk lihtsamalt öeldes kirjutamisstiili põhjal (näiteks kirjutatakse akadeemilisi

tekste hoopis teises stiilis kui blogipostitusi).² (Suchomel, Kraus 2022) 2023. aasta ühendkorpuse tekstide klassifitseerimisel võeti kasutusele kolmetasandiline liigitus, jagades teatud teemad omakorda allteemadeks (nt üldise *sport* kõrval ka *jalgpall*, *korvpall* jm). Uurisime, kuidas tuleb tekstide poolautomaatne liigitus kasuks sõnaraamatute kõnekeelsuse märgendi üle otsustamisel ja mida tuleks seejuures arvesse võtta.

2.2. Märgendi määramise katse

2024. aasta kevadel tehtud katses lähtusime konkordantsipõhisest korpusanalüüsist (st vaatasime sõnu kontekstis), lisades juurde žanrianalüüsi. Võtsime valimisse sõnaraamatus juba kirjeldatud, mitte uued sõnad, et meil oleks võrdlusalusena võtta varem tehtud leksikograafiline otsus. Nii pärinevad katsesse valitud sõnad kahest üldkeelsesõnaraamatust: „Eesti õigekeelsussõnaraamatust ÕS 2018”, mis oli katse ajal eesti kirjakeele normi aluseks, ning üle vaadatud ja uuemat infot sisaldavast EKI ühendsõnastikust 2024 (ÜS). Uurimust tegime sõnatasandil, mitte (kõnekeelsete) tähenduste kaupa, sest korpusest ei saa sõnatähendusi automaatselt kätte, neid tuleb uurijal ise tõlgendada.

Kokku oli materjaliks 240 juhuvalimiga saadud sõna. Täpsemalt pärines 50 KÕNEKEELNE-märgendiga sõna ÜS-ist (nt *maarott*) ja 50 ARG-märgendiga sõna ÕS-ist (nt *eraots*). Kontrollrühmaks võtsime 50 märgendita sõna ÜS-ist (nt *ehitismälestis*) ja 50 märgendita sõna ÕS-ist³ (nt *instrueering*). Võrdluseks kaasasime katsesse ka murdesõnu: 20 murdemärgendiga sõna ÜS-ist (nt *serbak*) ja 20 ÕS-ist (nt *kuukama*).⁴ Valimis leidis sõnu eri sõnaliikidest (nt *tüübiline*, *toss*), samuti sõnaühendeid (nt *vastu lonti andma*), kuna sõnaraamatu märksõnadeks ongi erisugused keelendid.

Korpuse infot analüüsis EKI teadusprojekti EKKD-III1 registrite tööühma üheksa liiget, kelle vahel jagati analüüsitavad sõnad võrdselt ära.⁵ Ülesandeks oli

² Veebitekstide klassifitseerimist tehakse URL-i põhiselt. See tähendab, et kõik tekstid, mis on alla laaditud ühelt URL-ilt, klassifitseeritakse samasse tekstiliiki. Näiteks kõik dokumendid, mis on alla laaditud URL-ilt postimees.ee, on märgendatud kui *perioodika*, samas saavad URL-ilt sport.postimees.ee alla laaditud tekstid juurde teemamärgendi *sport*. Käsitsi klassifitseeritud tekste kasutab Lexical Computing omakorda treeningmaterjalina, mille abil õppinud klassifikaator liigitab ülejäänud tekstid žanritesse ja/või teemadesse, juhul kui klassifikaator on teksti omaduste põhjal kindel, et tegu on just seda tüüpi tekstiga (Suchomel, Kraus 2021, 2022). Kui klassifikaator ei ole teksti omaduste põhjal kindel, et tegemist on teatud žanri või teemaga, ei klassifitseeri see tekste ainult selleks, et mingit silti külge panna. Seepärast on korpuses žanrite ja teemade all ka *none* 'määramata'.

³ Kuna ÕS 2018 andmebaasis pole registri- ja erialamärgendid sama klassifikaatoriga, sattus sellesse valimisse erialamärgendiga termineid. Jätsime need siiski valimisse, sest terminid ei asetu informaalsetesse registrisse.

⁴ Murdesõnad kaasasime katsesse, kuna murdekeele ja kõnekeele piiril on leksikograafias ajaga nihkumisi toimunud, näiteks sõna *pudru* oli EKSS 2009-s märgendatud murdekeelseks, ent EKI ühendsõnastikus kannab see kõnekeelsuse märgendit.

⁵ Selle ülesande eesmärk oli imiteerida leksikograafi reaalsel tööd, kus sõnaga tegeleb üldiselt üks inimene. Sõnaraamatutööle on iseloomulik see, et valikuid tuleb teha kiiresti, jäämata üht sõna analüüsima või kolleegidega arutama mitmeks kuuks. Kuna leksikograafide keeletaju ja tõlgendused võivad erineda, tegime hiljem kontrollkatse, et välja selgitada hinnangute varieerumise

tutvuda Sketch Engine'i tööriista abil eesti keele ühendkorpuse (ÜK 2023) keeleandmetega. Korpuse põhjal märgiti analüüsitabelisse sõna sagedusinfo (absoluutne ja *per mln*) ning viis sagedamat žanrit ja viis sagedamat teemat, mis selle sõna puhul esildusid. Iga žanri ja teema juures määrati sõna tüüpilisus neis (tekstiliigi suhtelise sageduse järgi).⁶ Niisiis uurisime, kui sage on sõna ise ning millistes žanrites ja teemades see kõige rohkem esineb ning kas ta on seal tüüpiline või mitte. Kui sõna esineb tüüpiliselt määramata žanris (ingl *none*), saab leksikograaf tõlgendada seda määrgina, et sõna kasutatakse mitmesugustes tekstides, kaldumata mõnda konkreetse žanri. Soovi korral sai leksikograaf tutvuda ka konkordantsiinfoga ja seejärel pidi andma hinnangu, kas korpuse põhjal võiks sõna olla ÜS-is märgendita või märgendiga – ja kui, siis millisega. (Analüüsijale ei olnud ette antud infot, kas tema uuritava sõnal on ÜS-is või ÕS-is märgend küljes.) Andmete põhjal võis ükskõik mis märgendit pakkuda. Tahtsime näha, kas ja kui palju pannakse kõnekeelele osutav märgend. Võimalus oli teha märkmeid ka vabasse kommentaarilahtrisse.

Analüüsi põhjal tegid kaks töörühma liiget tulemustest kokkuvõtte. Esiteks tuli iga sõna puhul välja tuua (*jah/ei*-vormis), kas korpus oli informatiivne ehk kas korpusinfo põhjal oli leksikograafil võimalik märgendi kohta otsus langetada. Kui vastuseks märgiti „ei”, tuli leksikograafil täita selgituse lahter. Seda lahtrit analüüsisime kvalitatiivselt. Üldanalüüsis kõrvutasime leksikograafi korpuse põhjal tehtud otsust ka originaalallika märgendiga. Eraldi vaatlesime juhtumeid, kus originaalallikas oli keelend märgendatud kõnekeelseks, kuid korpusinfo põhjal otsustas leksikograaf selle märgendita jätta.

3. Tulemused

3.1. Korpuse informatiivsus

Informatiivsuse all peame silmas seda, kas žanriandmetega täiendatud korpusinfo põhjal oli leksikograafil võimalik otsustada, milline märgend sõnale lisada. Üldanalüüsisist selgus, et korpust hinnati märgendi määramisel üldiselt informatiivseks (82,1%). Näiteks saab sõnale *fanatt* korpusandmete põhjal lisada märgendi KÕNE-

suurusjärk. Väiksema valimi peal (24 sõna 240-st, 10%) viisime läbi lihahindamise, mis näitas, et kui sama sõna hindab kolm leksikograafi, siis 21 sõna korral olid hindajad täpselt sama meelt. Ülejäänud kolmest juhust kattus kahel juhul arvamus osaliselt ja ainult ühe sõna puhul jäid hindajate arvamusd ja vastukäivaks.

⁶ Tekstiliigi suhteline sagedus (ingl *relative text type frequency*) on mõõdik, mis näitab sõna suhtelist esindatust konkreetsetes tekstiliigis võrreldes kogu korpusega ehk seda, kui iseloomulik on sõna mingile tekstiliigile. Tekstiliigi suhtelise sageduse arvutamiseks jagatakse päringu suhteline sagedus (*query relative frequency*) konkreetse tekstiliigi suhtelise suurusega. Tekstiliigi suhteline sagedus alla 100% viitab sellele, et sõna esineb selles konkreetsetes tekstiliigis harvem kui kogu korpus, mistõttu see ei ole selle tekstiliigi puhul tüüpiline ega spetsiifiline. Sagedus 100% näitab, et sõna on teatud tekstiliigis sama sage kui kogu korpus. Sagedus üle 100% osutab, et sõna esineb selles konkreetsetes tekstiliigis sagedamini kui kogu korpus, mistõttu on see sõna seal tekstiliigis eriti tüüpiline. Vt https://www.sketchengine.eu/my_keywords/relative-text-type-frequency/.

KEELNE, sest see esineb enamasti foorumites ja blogides ning on mõlemas tüüpiline, erinevalt näiteks perioodikast, kus see esineb palju vähem ja pole tüüpiline. Korpust ei peetud aga informatiivseks näiteks väga homonüümsete sõnade puhul (nt *siin*, *corp*), kuna korpusanalüüs nõudnuks ajamahukat uurimist, ega korpus valisti märgendatud sõnade puhul (nt verb *jäärama* esines enamasti nimisõna *jäär* vormidena). Samuti ei peetud korpusandmeid informatiivseks pärisnimega kattuvate sõnade puhul, mille kohta leidis vaid nimeinfot: *oskar* 'siga' asemel nimi Oskar ja murdesõna *rait* 'hiiglasuur asi' asemel nimi Rait. Selline tulemus ei olnud üllatav, kuna *oskari* tähendus 'siga' on keelekasutuses väga harv, mistõttu on ootuspärane, et see pole korpusse jõudnud või ei esildu seal piisavalt. Korpuspäringusüsteem võimaldab otsida sõna küll väikese esitähega, ent see ei välistanud tulemustest lausealguse suurtähte. Samuti on väga tõenäoline, et sellised sõnad nagu *oskar* ja *rait* on korpusse märgendamise faasis valisti lemmatiseeritud, kuna lahendamata jäänud mitmesuste puhul valib ühestaja analüüsitud lemmade loendist ainult esimese, st kõige sagedama (Koppel 2020: 54). Niisiis peab leksikograaf teadma korpusanalüüsi puudujääke, et ta oskaks nendega arvestada.

Ilmnes seos sõna korpusageduse ja korpusse informatiivseks pidamise vahel. Korpust peeti märgendite määramisel informatiivseks pigem kõrgema korpusagedusega sõnade puhul, mitte siis, kui sõna esines liiga vähe (alla kümne korra), et selle põhjal saaks järeldusi teha (nt *uhjutama*, kolm korda). Informatiivsuse ja sageduse vahelise seose kindlaksmääramiseks arvutasime nendevahelise Spearmani korrelatsioonikordaja (ρ), mille põhjal saab kinnitada, et korrelatsioon on korpusse informatiivsuse puhul sagedusega statistiliselt olulises ($p < 0,001$) positiivses (kuid mitte tugevas) seoses ($\rho \approx 0,44$). Üldistatult saab öelda, et mida sagedasem on sõna, seda suurema tõenäosusega peetakse korpust märgendi määramisel informatiivseks.

Katse sõnade hulgas olnud murdesõnad olid ühiskeelesõnadega võrreldes korpusse vähem esindatud. Isegi kui sõnakuju ise oli sage, ei pruukinud see olla õiges tähenduses, näiteks sõna *hoon* 'halb, vilets' esineb korpuses 537 korda, aga leksikograaf ei suutnud tuvastada ühtegi murdetähenduses vastet, vaid tegu oli kas perekonnanimega, homonüümiga tähenduses 'lõikeriist metallpindade täppisviimistluseks' või lemmatiseerimis- või trükiveaga. Niisiis ei olnud 2024. aasta mai seisuga ühendkorpusest murdesõnade asjus erilist abi.

3.2. Kõnekeelsus sõnaraamatute vs. korpusandmete põhjal

Võrdlesime üldanalüüsis sõnade originaallikates (ÕS või ÜS) olnud märgendeid nende märgenditega, mida leksikograafid pakkusid välja žanrit arvestava korpusanalüüsi põhjal. Selgus, et originaallikka märgendid langesid korpusse alusel pakutud märgendiga koguvalimis kokku 59%, seega 41%-l juhtudest jõudsid leksikograafid korpust analüüsides teistsugusele tulemusele (vt tabelit 1). Viimaste hulgas oli ka 6 sõna, mille puhul leksikograafi hinnangul osutas korpusse info kõnekeelsusele, ehkki originaallikkas märgendit ei olnud. Näiteks sõna *tõrts* kohta, mis on nii ÕS-is kui ÜS-is märgendita, kommenteeris leksikograaf korpusandmete uurimise

põhjal, et „on nii kõnekeelseid kui ka just toidukontekste, kus sama kõnekeelsena ei mõju, ehkki tõrts ikka”, ja otsustas, et varustaks selle sõna kõnekeelsuse märgendiga, ehkki võib-olla mitte kokanduse tähenduses.

Tabel 1. Märgendite kattuvus üldvalimis

Originaalallika ja korpusandmete järgi <i>sama</i> märgend või ilma märgendita	142 sõna 240 sõnast	59%
Originaalallika ja korpusandmete järgi <i>erinev</i> märgend või ilma märgendita	98 sõna 240 sõnast	41%
Originaalallika ja korpusandmete järgi kõnekeelsuse märgend	39 sõna 100 sõnast	39%
Originaalallikas kõnekeelsuse märgend, korpusandmete järgi mitte	61 sõna 100 sõnast	61%

Nende 100 katsesõna puhul, millel originaalallikas oli kõnekeelsusele osutav märgend (50 sõna ÜS-ist, 50 sõna ÖS-ist), pakuti kõnekeelsuse märgendit 39%-l juhtudest. Seega korpusinfo põhjal ei pakutud kõnekeelsuse märgendit 61%-l juhtudest (vt tabelit 1). Täpsemalt kattus leksikograafi kõnekeelsuse märgendi kasuks tehtud hinnang ÖS-i ARG1-märgendiga 18 sõna puhul 50-st (36%) ning ÜS-i KÕNEKEELNE-märgendiga 21 sõna puhul 50-st (42%). Kuna leksikograaf sai otsustada ka muude märgendite kasuks, ei tähendanud see alati, et mõnd märgendit ei oleks pandud. Näiteks võis sõna paista pigem harv, vanamoodne või halvustav kui kõnekeelne ning saada vastava märgendi.

Uurisime, miks oli valimis esinenud sõnadel sõnastikes kõnekeelsuse märgend, et mõista, miks originaalallika ja korpusandmete järgi oli märgendi kattuvus ainult 39%. Kattuvate sõnade hulgas oli mitmeid selliseid sõnu, mille puhul eesti keelt emakeelena kõnelejana ei tekkinud küsimust, miks see sõnaraamatus kõnekeelseks on märgendatud: näiteks sõna on lühendatud (*fanatt*, vrd *fanaatik*) või kõnekeelsus tuleneb tuletusliitest (*näpukas* 'trükiviga', vt *kas*-liite kohta Kasik 2015: 264–265).

Mittekattuvate sõnade hulgas oli selliseid sõnu, mille puhul paistis, et on osutatud, et tegu ei ole terminiga (ÖS-is *pasunapoiss*, *mänguriist*, *kompleks*; ÜS-is *võrdõigus*, vt ÖS-i ja oskuskeele suhte kohta Vare 2001; Risberg 2024a: 52). Osa sõnu või mõne nende tähendustest saanuks kõnekeelsuse asemel või kõrval märgendada ka vulgaarseks või halvustavaks (nt *lits*, *pehmo*), teised ülekantud tähenduseks (*kassiahastus* 'pohmelus; masendusmeeleolu'). Mõnel juhul paistis tegu olevat keelekorraldusotsustega, näiteks varieeruvuse vähendamine (*takkaotsa* ARG1, vrd *tagantjärele* märgendita) või muu põhjus, mida tuleks lähemalt uurida (nt *kõiksugu*, vt Risberg 2024b). Originaalallika ja meie töörühma pakutud märgendite erinevus võib olla tingitud ka sellest, et ÖS-i ja ÜS-i koostajatel pole olnud märgendite määramisel keeleandmete arvestamiseks ühist suunist, samuti võib esineda märgendeid, mis on määratud tänapäeva keeleandmetega tutvumata.

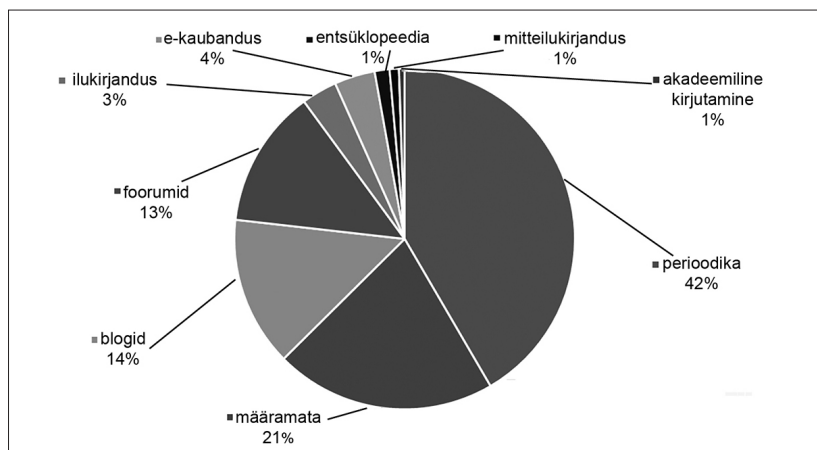
Teadvustame, et ka korpusandmete tõlgendamine jätab ruumi subjektiivseteks otsusteks ja nii mängis selleski katses materjali tõlgendamisel rolli konkreetse uurija keeletunnetus. Kommentaaridest ilmnes, et samuti võisid keeleinfo tõlgendamist segada või siis sellele kaasa aidata taustteadmised keelekorralduslike otsuste kohta. Korpusanalüüsist on siiski subjektiivsuse vähendamisel abi, kuna see pakub infot laiema kõnelejaskonna keelekasutusmustrite kohta ja nii ei pea uurija tuginema

üksnes enda sisetundele. Puuduseks on seesuguse kontrolli ajamahukus, samuti võib žanrite ja teemade ühtlase jaotumise korral otsuse langetamine olla keeruline.

Kirjeldame järgmiseks, milline oli žanriline jaotumine korpuses puhkudel, kui leksikograaf pidas sõna kõnekeelseks. Analüüsi põhjal pakume välja orientiiri, mida saab edaspidi ühe alusena teiste seas kõnekeelsuse hindamisel arvestada.

3.3. Kõnekeelsuse mõõtmine korpusanalüüsi põhjal

Leksikograafidel kerkis katset tehes küsimus, kui palju peaks sõna esinema kõnekeelsuse poole kaldu olevates tekstiliikides (eraldi mainiti blogisid ja foorumeid), et oleks põhjendatud kõnekeelsuse märgendi lisamine. Ehkki rangeid piire seada ei ole võimalik, määrasime edasisi arvutusi tehes just blogid ja foorumid kõnekeelsusele kalduvateks žanriteks, kuna üldiselt on nende tekstide keelekasutus rangete reeglite järgimisest vabam ja enamasti neid tekste teine inimene, harilikult keeleprofessiional, ei toimetata.



Joonis 1. Žanrite jaotumine eesti keele ühendkorpuses (2023).⁷

Tegelikult ei saa korpuse žanreid väga lihtsasti liigitada informaalseks, neutraalseks või formaalseks, sealjuures ei ole tekstide toimetatuse aste ära määratud. Nii nagu on hägune kõnekeele ja kirjakeele piir, ei eksisteeri (korpuseski) üksnes ja ainult kõnekeelseid žanreid – ka blogide hulgas võib leiduda selliseid, kus autor ennast ise toimetab ja kirjakeele norme järgib. Perioodika all võib samuti leiduda nii toimetamata veebiartikleid kui ka mitu toimetamisringi läbinud paberajalehe artikleid, nii neutraalseid uudiseid kui ka värvikamat seltskonnakroonikat – seega ei saa seda žanrit formaalsuse skaalale paigutada. Rohkem formaalseks peetavaid žanreid, nagu teadusartiklid, seadusetehtid jm, on korpuse algmaterjalis ka vähem. Joonis 1 annab ülevaate žanrite jaotumisest eesti keele ühendkorpuses 2023 (vt ka Koppel, Kallas 2022).

⁷ Kirjandus on ühendkorpuses jaotatud kaheks: ilukirjandus ja mitteilukirjandus. Viimase alla kuuluvad näiteks käsiraamatud, reisikirjad, esseed, õpikud.

Möötsime blogides ja foorumites esinemise osakaalu (ehk nendes žanrites esinemise protsenti sõna üldsagedusest) korpuses kõigil uuritud sõnadel: nii neil, mille leksikograaf oli otsustanud korpusanalüüsi põhjal kõnekeelseks määrata, kui ka katsesõnadel, millele seda märgendit ei lisatud. Selleks arvutasime kõnekeelsuse osakaalu (ko), mis näitab, kui suure osa konkreetse sõna kõikidest esinemistest korpuses moodustavad tema esinemised blogides ja foorumites. Žanrite analüüsist jätsime kõrvale 18 sõna (nt *pudrupööris*), millel oli korpuses alla 10 vaste, kuna nii vähesel materjalil põhjal ei ole võimalik žanrilise jaotumise kohta usaldusväärseid järeldusi teha ning need sõnad mõjutanuks liialt kokkuvõtvat analüüsi. Lisaks jäi kõrvale 13 sõna, mille suur homonüümsus ja polüseemsus takistas otsitava tähenduse usaldusväärset analüüsi ja kiiret leidmist, nt *siin* (kardinaasiini tähenduses). Kokku jäi kõnekeelsuse osakaaluga seotud analüüsiosa aluseks 209 sõna.

3.3.1. Katsesõnade kõnekeelsuse osakaal

Leksikograafid otsustasid korpusandmete põhjal kõnekeelse märgendi lisada 42 sõnale 209 katsesõnast. Nende sõnade keskmine blogide ja foorumite koondosakaal ehk kõnekeelsuse osakaal oli 52% (väikseim ko 13%, kõrgeim 99%). 167 sõna puhul otsustasid leksikograafid korpusel põhjal kõnekeelsuse märgendi lisamata jätta. Selliste sõnade keskmine kõnekeelsuse osakaal oli 23% (väikseim 0%, kõrgeim 79%) – üle poole madalam kõnekeelseks märgendatud sõnade rühma keskmisest.

Korpusel põhjal kõnekeelseks määratud sõnade rühm varieerus nii kõnekeelsuse osakaalu kui ka üldise korpussageduse poolest. Rühma kuulus kõrge sagedusega sõnu, millel oli ühtlasi kõrge kõnekeelsuse osakaal, näiteks *spets* (sagedus 23 551, ko 79%). Kuid ka kõrge sagedusega sõna võis olla spetsiifilise kasutusfääriga, näiteks *akva* 'akvaarium' esines korpuses 27 952 korral, millest 99% blogides ja foorumites. Rühma kuulus ka madalama korpussagedusega sõnu, millel oli kõrge kõnekeelsuse osakaal, näiteks *tulenui* (sagedus 115, ko 74%). Vaatasime eraldi madala ko-ga sõnu, millele siiski oli otsustatud kõnekeelsuse märgend lisada. Näiteks sõnal *silmakae* oli ko vaid 13%, ent tõenäoliselt mõjutas leksikograafi töik, et *silmakae* on rahvapärane nimetus *läätsekae* kohta ja see pole terminina aktsepteeritud. Otsus väljendab lähemist, et sõna, mis pole ametlik termin, on järelikult kõnekeelne, ehkki otstarbekam oleks ehk sellised juhud märgendada näiteks rahvapäraseks, et eristada termineid üldkeelsesõnadest. Kõnekeelsuse määramise juurde lisatud kommentaaridest järelduski, et otsust võis mõjutada terminoloogiline korrektsus, aga ka küsimus, kas igasugune rahvapärane sõna (nt *ridvamees*, *süstamees*) vajab kõnekeelsuse märgendit. Samuti mõjutas märgendi määramist sõna tundlikkus (nt *kapihomo* juurde võiks paremini sobida ehk muu märgend, nt HALVUSTAV). Niisiis tekitasid leksikograafidele kõhkluskohti nii kõnekeelsuse-kirjakeelsuse suhe, slängi-kõnekeelsuse suhe, kõnekeelsuse-murdekeelsuse suhe, kõnekeele-erialakeele suhe kui ka sõna tundlikkuse aspekt.

Madalama ko-ga, ent siiski kõnekeelseks märgendatud võisid olla ka polüseemsed sõnad, millel esines kõnekeelse tähenduse kõrval korpuses teisigi tähendusi ja mida leidis seetõttu rohkem ka muudes tekstiliikides. Näiteks sõnal *sara* (ko 39%)

esineb nii neutraalne tähendus 'kerge ehitusega kõrvalhoone nt tööriistade või küttepude hoidmiseks' kui ka kõnekeelseks märgitud tähendus 'viletsa sooja mittepidaava elamu kohta või vanarauaks sõidetud auto kohta'. Polüseemsel sõnal võis olla ka kõrge ko (nt *sebima*, ko 70%) ja analüüsiv leksikograaf otsustas liikumistähendusele märgendit mitte lisada (nt „putukas sebib jalgadega”), kuid märkis, et võrgutamise tähendusele ta kõnekeelsuse märgendi lisaks.

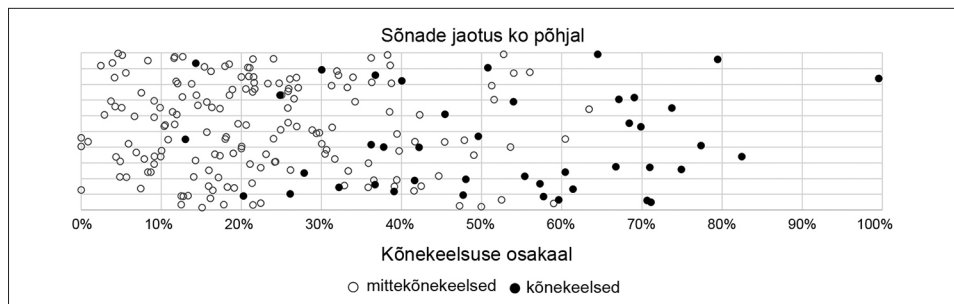
167 katsesõna, mille puhul leiti, et kõnekeelsuse märgend ei ole vajalik, olid samuti varieeruva kõnekeelsuse osakaaluga. Madal ko oli näiteks sõnadel *järjend* (3%) ja *tugimaantee* (5%). Kõnekeelsuse üle otsustamisel mõjutas leksikograafe ka sõna tüüpilisus žanriti, näiteks sõna *käitumine* juures toodi välja, et sõnal on „kõigis žanrites ebatüüpiline kasutus, see näitab ka ühtlast jaotumist, nii et märgendit ei paneks”.

Sõna kalduvust esineda teatud žanrites (blogides ja foorumites) väljendab ka meie pakutav korpusorientiir, mida tutvustame järgmiseks.

3.3.2. Kõnekeelsuse orientiir eesti keele ühendkorpuses 2023

Pakume välja korpusorientiiri, mis võiks abistada leksikograafe kõnekeelsuse märgendi lisamisel. Selle väljatöötamisel arvestasime, et see peaks olema leksikograafide reaalset kasutatav vahend, mis ei oleks üleliia töömahukas. Orientiiri arvutamisel arvestasime sõnade kõnekeelsuse osakaalu (blogides ja foorumites esinemise osakaalu sõna kõikidest kasutustest) ja seda, kas leksikograaf märgendas sõna katses kõnekeelsuse märgendiga või mitte. Eesmärk oli leida kõnekeelsuse osakaalu piir, mille puhul võimalikult suur osa kõnekeelseks märgendatud sõnu jääks piirist kõrgemale ja võimalikult suur osa sõnadest, mida kõnekeelseks ei märgendatud, jääks piirist allapoole. Et võrdselt käsitleda kõnekeelsete ja mitte kõnekeelsete sõnade rühma, kasutasime arvutuses rühmade osakaale. Absoluutsageduste puhul mõjutaksid kõnekeelsuse orientiiri määramist rohkem mitte kõnekeelsed sõnad, kuna neid oli materjalis suurem hulk (lisaks 100 registreeritud sõnale oli valimis ka 40 murdekeelset sõna).

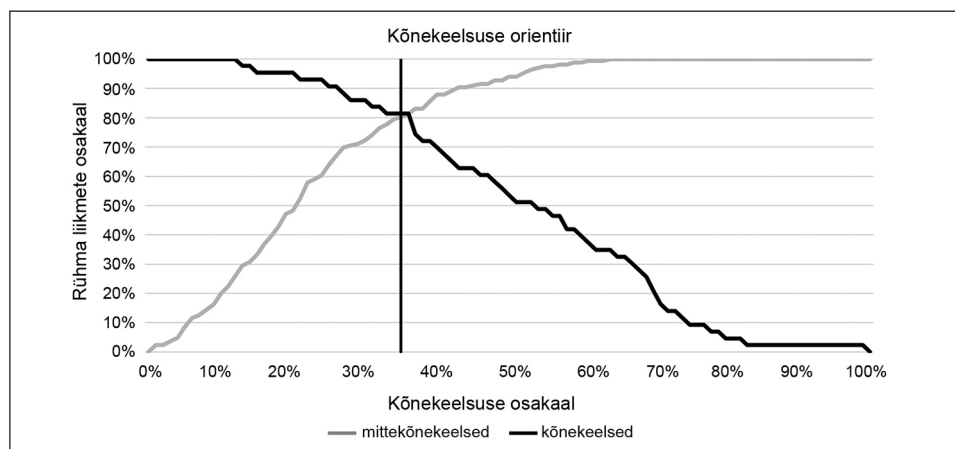
Joonis 2 kujutab 209 katsesõna jaotust kõnekeelsuse osakaalu põhjal (x-teljel). (Y-telje aluseks on Excelis genereeritud juhuslikud väärtused, et oleks võimalik vaadelda katsesõnade jaotumist.)



Joonis 2. Katsesõnade jaotus kõnekeelsuse osakaalu põhjal.

Mustade täppidega on joonisel tähistatud kõnekeelseks märgendatud sõnad ja valgete täppidega⁸ sõnad, millele kõnekeelsuse märgendit ei lisatud. Jooniselt nähtub, et kõnekeelsuse osakaalu tõustes suureneb ka kõnekeelseks märgendatud katsesõnade arv. Kõnekeelsuse osakaalu vahemikku 0–10% ei paigutu ühtegi kõnekeelseks märgitud katsesõna. Kõnekeelsuse osakaalu vahemikku 70–100% ei paigutu ühtegi katsesõna, mida ei märgitud kõnekeelseks.

Joonis 3 illustreerib kõnekeelsuse orientiiri määramiskäiku. Must püstjoon kujutab kõnekeelsuse osakaalu piiri, millest suurim võimalik osa kõnekeelsete sõnade rühmast (tähistatud musta joonega) jääb piirist kõrgemale ($\geq$) ja suurim võimalik osa mitte-kõnekeelsete sõnade rühmast (tähistatud halli joonega) jääb piirist allapoole ($<$).⁹



Joonis 3. Kõnekeelsuse osakaalu piir (must püstjoon, 36%).

Kõnekeelsuse osakaalu piir 36% jätab kõige suurema võimaliku osa kõnekeelseks analüüsitud sõnu (81%) piirist ülespoole, nii et kõige suurem võimalik osa mitte-kõnekeelseks analüüsitud sõnu (81%) jääb piirist allapoole. Võrreldes kõnekeelseks märgitud sõnade rühma keskmise ko-ga (52%) võib orientiir tunduda madal, ent kui arvestada, et korpustekstide koguhulgast on blogide ja foorumite osakaal kokku 27% (vt joonist 1), on orientiir siiski neljandiku võrra kõrgem osakaalust, mis oleks juhul, kui sõna esineks korpuses kõikides žanrites võrdselt. Kindlasti mõjutab saadud orientiiri suurust võimalik polüseemia ehk neutraalsete sõnatähenduste esinemine kõnekeelsete tähenduste kõrval, mis toob kaasa katsesõnade sagedama kasutamise ka muudes tekstiliikides. Pakutud orientiir ei olegi range piir, vaid abistav suunis, mis võib leksikograafi korpusanalüüsil toetada. Lõplik otsus jääb ikkagi spetsialistile, kellel on võimalik tutvuda korpusmaterjaliga vajadusel süvitsi ja võtta arvesse ka muid aspekte (nagu erialakeel, tundlik teema jm).

Pakutud orientiiri puhul (36%) peab arvestama, et arvutuse aluseks on olnud eesti keele ühendkorpus 2023 ja uue versiooni valmides ei pruugi tekstiliikide vahe-

⁸ Artikli veebiversioonis on mustade täppide asemel punased ja valgete asemel sinised täpid.

⁹ Artikli veebiversioonis on musta joone asemel punane ja halli asemel sinine joon.

korrad jääda täpselt samaks. Seetõttu tuleb orientiiri järgmise korpusversiooni jaoks ajakohastada, ent seniks on olemas ühtne suunis, mida varem pole olnud, ja sarnase arvutuskäigu saab järgmise korpusversiooni avaldudes uue orientiiri leidmisel aluseks võtta.

4. Arutelu ja kokkuvõte

Seni pole süstemaatiliselt üle uuritud, kas ja kuidas eesti sõnaraamatutes esitatud märgendid vastavad sõna tegelikule kasutusele. Käesolevas artiklis keskendusime kõnekeelele. Leksikograaf peab oma töös langetama konkreetse otsuse, kas (kõnekeelsuse) märgend lisada või mitte. Tal pole olnud vahendit, mis ütleks üheselt ära, kas sõna – õigupoolest selle mõni tähendus – on kõnekeelne, ning keeles ei olegi alati võimalik konkreetseid piire tõmmata. Kõnekeelsuse märgendi määramine on leksikograafile ühtlasi mõnevõrra keerukam kui kas või näiteks vulgaarsuse märgendi lisamine, kuna kõnekeelsus jääb tihtilugu pigem informaaalse-neutraalse-formaalse skaala keskpaika, võrreldes äärmuslikumate väljendustega, ja seega on kõnekeelsust raskem tabada.

Meie deskriptiivse lähenemisega uurimus oli osaliselt tehtud preskriptiivse sõnaraamatu tarbeks, sest lisaks EKI ühendsõnastiku koostamisprotsessi hõlbustamisele soovisime oma uurimusega aidata kaasa järgmise ÕS-i (2025) koostamisele. ÕS 2025-s on normitud ortograafia ja morfoloogia, ka välde, kui see mõjutab sõnamuutmist. Muu keeleinfo on deskriptiivne. Niisiis võiksid ka ÕS-i lisatavad märgendid olla deskriptiivsed, tugineda kasutusandmetele.

Keeleteadlased on pidanud korpusanalüüsil põhinemist sõnaraamatute koostamisel oluliseks (nt Lippus, Lindström 2024) ja märgendeid lisades ongi eri aegadel toetunud kasutusnäidetele (enne tekstikorpuseid sedelkartoteekidele), ent alati ei ole olnud saadaval samavõrd mahukalt erinevaid keeleandmeid, mida süstemaatiliselt uurida. Artiklis vaatasimegi, kuidas on võimalik kasutada praegust korpus tööriista Sketch Engine, et otsustada võimalikult tõhusalt sõna (tegelikult tema tähenduse) kõnekeelsetes kontekstides esinemise üle ja langetada otsus, kas sõnaraamatusse märgend lisada või mitte.

EKI ühendsõnastikku koostades on korpus tööriist olnud kasulik abivahend, sest annab infot laiema kõnelejaskonna keelekasutusmustritest. Siiski on seni olnud puudu ühine suunis, kuidas korpus märgendi määramisel kasutada. Artiklis kirjeldatud katse tulemusena selgus, et Sketch Engine'i uuem funktsioon, mis võimaldab vaadata žanreid ja teemasid, on leksikograafi jaoks märgendi lisamisel informatiivne 82,1%-l juhtudest. Töötasime tehtud analüüsi põhjal välja ka korpusorientiiri, millele kõnekeelsuse märgendi lisamisel toetuda. Leitud korpusorientiir (36%), mis tugineb sõna esinemisele blogides ja foorumites kui kõnekeelsuse poole kalduvates tekstiliikides, on siiski üksnes abistav suunis otsuse toetamiseks, mitte range juhise, ja seda ei saa rakendada muid aspekte kaalumata. Kõnekeelsuse määramisel tekitas leksikograafidele kõhkluskohti näiteks kõnekeele suhe rahvapärasusega ehk vastandumine erialakeelele, aga samuti sõna tundlikkus – kas tundlik sõna on tingimata kõnekeelne või peaks seda märgendama muul viisil?

Selle töö võimalik tulevikusuund on töötada välja registrikalkulaator, mis aitaks teha automaatselt kiire kokkuvõtte korpuses leiduvast infost, mis on just registri-märgendite üle otsustamisel asjakohane. Pakutud kõnekeelsuse korpusorientiir võiks olla kalkulaatori üks osa, aga lisaks saaks kaasata veel teisi žanreid ja teemasid (nt akadeemilised, entsüklopeedilised ja seadusetekstid), milles esinemine kahan-daks sõna kõnekeelsuse tõenäosust. Samuti võiksid kalkulaatorisse järgemööda lisanduda ka korpuskokkuvõtete tegemise võimalused teiste märgendite jaoks (nt LUULEKEELNE, LASTEKEELNE).

Kasutajauuringutest on teada, et sõnaraamatu puhul peetakse üheks olulisimaks tunnuseks keeleinfo usaldusväärsust (Müller-Spitzer, Koplenig 2014: 168; Lange-mets jt 2024). Meie praegu välja töötatud orientiir võib aidata leksikograafi otsuse langetamisel ning muuta otsustamist kiiremaks ja süsteemsemaks. Lõpliku valiku teeb aga siiski leksikograaf, keeleekspert. See on ühtlasi nii eelis (ta saab andmeid vajadusel juurde vaadata) kui ka puudus (mõjutama hakkavad eelteadmised, keele-tunnetus jne). Teisalt, sõnaraamat ei saagi kajastada kogu tõe keele kohta, vaid sel-les saab esitada üksnes sõnakasutuse üldistusi, vihjata, et sõna kiputakse mõnes tema tähenduses kasutama neutraalsest irduvas ümbruses.

Artikli valmimist on toetanud teadusprojekt EKKD-III1 „Suurte keelemudelite rakenda-mine leksikograafias: uued võimalused ja väljakutsed”. Täname katsesse panustamise ja analüüsi tegemise eest Sirli Zuppingut ja Hanna Pooki.

VEEBIVARAD

EKS 2019 = Eesti keele sõnaraamat. Koost Katrin Kuusik, Külli Kuusk, Margit Langemets, Mai Tiits, Udo Uibo, Tiia Valdre, Piret Voll. Toim M. Langemets, M. Tiits, U. Uibo, T. Valdre, P. Voll. Tallinn: Eesti Keele Instituut. <http://www.eki.ee/dict/eks/>

Sketch Engine. <https://www.sketchengine.eu>

ÜK 2023 = Eesti keele ühendkorpus 2023. Koost Jelena Kallas, Helen Kaljumäe, Kristina Koppel. <https://doi.org/10.1515/3-00-0000-0000-08C04M>

ÜS = EKI ühendsõnastik 2024. Eesti Keele Instituut. Sõnaveeb. <https://sonaveeb.ee>

KIRJANDUS

Atkins, B. T. (Sue); Rundell; Michael 2008. The Oxford Guide to Practical Lexicography. Oxford–New York: Oxford University Press.

<https://doi.org/10.1093/oso/9780199277704.001.0001>

Baayen, R. Harald 2024. The wompom. – Corpus Linguistics and Linguistic Theory, kd 20, nr 3, lk 615–648. <https://doi.org/10.1515/cllt-2024-0053>

Biber, Douglas; Conrad, Susan 2009. Register, Genre, and Style. (Cambridge Textbooks in Linguistics.) Cambridge: Cambridge University Press.
<https://doi.org/10.1017/CBO9780511814358>

- Cvrček, Václav 2024.** Enhancing national corpus infrastructure with multidimensional model of register variation. – Ettekanne Eesti Rakenduslingvistika Ühingu kevadkonverentsil, 18. IV. <https://trost.korpus.cz/~cvrcek/prezentace/mda/tallinn/#/>
- Davies, Mark 2025.** Corpora and AI/LLMs: Overview. English-Corpora.org. <https://www.english-corpora.org/ai-llms/corpora-vs-llms.html>
- de Schryver, Gilles-Maurice 2023.** Generative AI and lexicography: The current state of the art using ChatGPT. – International Journal of Lexicography, kd 36, nr 4, lk 355–387. <https://doi.org/10.1093/ijl/ecad021>
- Eder, Maciej 2025.** Text analysis is easy, unless it is not: Reliability issues in measuring textual similarities. – DHNB2025 konverentsiettekanne, 5. III. <https://computationalstylistics.github.io/presentations/dhnb2025/>
- EKSS 1988–2007** = Eesti kirjakeele seletussõnaraamat. 26 vihikut. Peatoim Rudolf Karelson, Valve Kullus, Erich Raiet, Mai Tiits, Tiia Valdre, Leidi Veskis. Tallinn: Eesti Keele Instituut, Eesti Keele Sihtasutus.
- EKSS 2009** = Eesti keele seletav sõnaraamat. Kd I–VI. „Eesti kirjakeele seletussõnaraamatu” 2., täiendatud ja parandatud tr. Toim Margit Langemets, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks, Piret Voll. Tallinn: Eesti Keele Sihtasutus.
- Erelt, Tiit 2002.** Mida ÕSist leida on. – Oma Keel, nr 2, lk 62–75.
- Hennoste, Tiit; Pajusalu, Karl 2013.** Eesti keele allkeeled. Õpik gümnaasiumile. Tallinn: Eesti Keele Sihtasutus.
- Jürviste, Madis; Paet, Tiina; Soosaar, Sven-Erik 2025.** Eesti vanade sõnakujude tuvas-tamisest suurte keelemudelitega. – Eesti Rakenduslingvistika Ühingu aastaraamat, kd 21, lk 63–84. <https://doi.org/10.5128/ERYa21.04>
- Karelson, Rudolf 1990.** „Eesti kirjakeele seletussõnaraamat” tegija pilgu läbi. – Keel ja Kirjandus, nr 1, lk 24–34.
- Kasik, Reet 2015.** Sõnamoodustus. (Eesti keele varamu 1.) Toim Katrin Kern. Tartu: Tartu Ülikooli Kirjastus.
- Kerge, Krista 2004.** Keelenormi tänapäevane olemus (2). Normi liigid ja sõnavaranorm. – Õiguskeel, nr 1, lk 11–20.
- Kilgarriif, Adam; Baisa, Vít; Bušta, Jan; Jakubíček, Miloš; Kovář, Vojtěch; Michel-feit, Jan; Rychlý, Pavel; Suchomel, Vít 2014.** The Sketch Engine: Ten years on. – Lexicography, kd 1, nr 1, lk 7–36. <https://doi.org/10.1007/s40607-014-0009-9>
- Klosa-Kückelhaus, Annette; Tiberius, Carole 2024.** The lexicographic process revisited. – International Journal of Lexicography, kd 38, nr 1, lk 1–12. <https://doi.org/10.1093/ijl/ecae016>
- Koppel, Kristina 2020.** Näitelausete korpuspõhine automaattuvastus eesti keele õppe-sõnastikele. (Dissertationes linguisticae Universitatis Tartuensis 38.) Tartu: Tartu Ülikooli Kirjastus.
- Koppel, Kristina; Kallas, Jelena 2022.** Eesti keele ühendkorpuste sari 2013–2021: mahu-kaim eestikeelsete digitekstide kogu. – Eesti Rakenduslingvistika Ühingu aastaraamat, kd 18, lk 207–228. <https://doi.org/10.5128/ERYa18.12>
- Kuzman, Taja; Ljubešić, Nikola 2023.** Automatic genre identification: A survey. – Language Resources and Evaluation, kd 59, lk 537–570. <https://doi.org/10.1007/s10579-023-09695-8>
- Langemets, Margit 2025.** Vilets keel, vilets meel. ÕSiga või ÕSita. – Plenaarettekanne Eesti Rakenduslingvistika Ühingu kevadkonverentsil, 24. IV. https://www.youtube.com/watch?v=ozi34zbQ_4g

- Langemets, Margit; Risberg, Lydia 2023.** Mis on ÕSi sõna? – Sirp 13. X, lk 20–21.
<https://www.sirp.ee/mis-on-osi-sona/>
- Langemets, Margit; Risberg, Lydia; Algvere, Kristel 2024.** To dream or not to dream about 'correct' meanings? Insights into the user experience survey. – Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress. Toim Kristina Š. Despot, Ana Ostroški Anić, Ivana Brač. Cavtat: Institut za hrvatski jezik, lk 741–760.
- Langemets, Margit; Tiits, Mai; Uiibo, Udo; Valdre, Tiia; Voll, Piret 2018.** Eesti keel uues kuues. Eesti keele sõnaraamat 2018. – Keel ja Kirjandus, nr 12, lk 942–958.
<https://doi.org/10.54013/kk733a2>
- Lee, David Y. W. 2001.** Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. – Language Learning & Technology, kd 5, nr 3, lk 37–72.
- Lindström, Liina; Risberg, Lydia; Plado, Helen 2023.** Language ideologies and beliefs about language in Estonia and Estonian language planning. – Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics, kd 14, nr 1, lk 7–48.
<https://doi.org/10.12697/jeful.2023.14.1.01>
- Lippus, Pärtel; Lindström, Liina 2024.** ÕS-ist ning teaduse demoniseerimisest. – ERR, 23. IV. <https://www.err.ee/1609320843/partel-lippus-ja-liina-lindstrom-os-ist-ning-teaduse-demoniseerimisest>
- Lüdeling, Anke; Alexiadou, Artemis; Adli, Aria; Donhauser, Karin; Dreyer, Malte; Egg, Markus; Feulner, Anna Helene; Gagarina, Natalia; Hock, Wolfgang; Jannedy, Stefanie; Kammerzell, Frank; Knoeferle, Pia; Krause, Thomas; Krifka, Manfred; Kutscher, Silvia; Lütke, Beate; McFadden, Thomas; Meyer, Roland; Mooshammer, Christine; Müller, Stefan; Maquate, Katja; Norde, Muriel; Sauerland, Uli; Solt, Stephanie; Szucsich, Luka; Verhoeven, Elisabeth; Waltereit, Richard; Wolfgruber, Anne; Zeige, Lars Erik 2022.** Register: Language users' knowledge of situational-functional variation. Frame text of the First Phase Proposal for the CRC 1412. – Register Aspects of Language in Situation (REALIS), kd 1, nr 1, lk 1–58.
<https://doi.org/10.18452/24901>
- Müller-Spitzer, Carolin; Koplenig, Alexander 2014.** Online dictionaries: Expectations and demands. – Using Online Dictionaries. (Lexicographica. Series Maior 145.) Toim C. Müller-Spitzer. Berlin: Walter de Gruyter, lk 143–188.
<https://doi.org/10.1515/9783110341287.143>
- Nemvalts, Peep 2023.** Sõnatsunami. – Sirp 22. XII, lk 45–46.
<https://www.sirp.ee/sonatsunami/>
- Paet, Tiina 2023.** Võõra keele kinnistumine eesti keeles: keelekorralduslik ja leksikograafiline vaade. (Dissertationes philologiae estonicae Universitatis Tartuensis 51.) Tartu: Tartu Ülikooli Kirjastus.
- Paet, Tiina; Risberg, Lydia 2021.** Võõrsõnade tähendussoovitused ja nende esitus üldkeele sõnaraamatus. – Keel ja Kirjandus, nr 11, lk 965–984. <https://doi.org/10.54013/kk767a2>
- Pajusalu, Renate 2009.** Sõna ja tähendus. Toim Tiiu Ereli. Tallinn: Eesti Keele Sihtasutus.
- Paulsen, Geda; Lohk, Ahti; Tuulik, Maria; Vainik, Ene 2023.** From experiments to an application: The first prototype of an adjective detector for Estonian. – Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference. Toim Marek Medveď, Michal Měchura, Carole Tiberius, Iztok Kozem, Jelena Kallas, Miloš Jakubiček, Simon Krek. Brno: Lexical Computing CZ, s.r.o., lk 476–500.

- Pool, Railii; Teral, Marika; Kallas, Jelena 2025.** Keeleminutid. Eesti keele õppijad sõnaraamatute maailmas. – ERR Kultuur 24. III. <https://kultuur.err.ee/1609642106/keeleminutid-eesti-keele-oppijad-sonaraamatute-maailmas>
- Pullum, Geoffrey K. 2023.** Why grammars have to be normative – and prescriptivists have to be scientific. – *The Routledge Handbook of Linguistic Prescriptivism*. (Routledge Handbooks in Linguistics.) Toim Joan C. Beal, Morana Lukač, Robin Straijer. London: Routledge, lk 3–16. <https://doi.org/10.4324/9781003095125-2>
- Päärt, Villu 2023.** Keelenõunik Helika Mäekivi: lihtne keel ei ole sama mis lihtsakoeline keel. – *Universitatis Tartuensis*, nr 4, lk 22–25.
- Raadik, Maire 2014.** Mida uut on uues õigekeelsussõnaraamatus? – *Oma Keel*, nr 1, lk 67–75.
- Raag, Raimo 2008.** Talurahva keelest riigikeeleks. Tartu: AS Atlex.
- Read, Allen Walker s. a.** Features and problems. – *Encyclopedia Britannica: Dictionary*. <https://www.britannica.com/topic/dictionary/Features-and-problems>
- Risberg, Lydia 2024a.** Sõnatähendused ja sõnaraamat. Kasutuspõhine sisend eesti keelekorraldusele. (Dissertationes philologiae estonicae Universitatis Tartuensis 25.) Tartu: Tartu Ülikooli Kirjastus.
- Risberg, Lydia 2024b.** Keelesäuts. Kõiksugu(sed) säutsud tulevad pähe. – *Vikerraadio*, ERR, 20. XII. <https://vikerraadio.err.ee/1609557077/keelesauts-koiksugu-sed-sautsud-tulevad-pahe>
- Risberg, Lydia; Langemets, Margit 2021.** Paronüümide probleem eesti keeles. – *Keel ja Kirjandus*, nr 10, lk 903–926. <https://doi.org/10.54013/kk766a4>
- Rundell, Michael 2024.** Automating the creation of dictionaries: Are we nearly there? – *Humanising Language Teaching*, kd 26, nr 1. <https://www.hltmag.co.uk/feb24/automating-the-creation-of-dictionaries>
- Suchomel, Vit; Kraus, Jan 2021.** Website properties in relation to the quality of text extracted for web corpora. – *Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN 2021)*, lk 167–175.
- Suchomel, Vit; Kraus, Jan 2022.** Semi-manual annotation of topics and genres in web corpora, the cheap and fast way. – *Proceedings of the Sixteenth Workshop on Recent Advances in Slavonic Natural Languages Processing (RASLAN 2022)*. Toim Aleš Horák, Pavel Rychlý, Adam Rambousek. Brno: Tribun EU, lk 141–148.
- Trap-Jensen, Lars 2002.** Descriptive and normative aspects of lexicographic decision-making: The borderline cases. – *Proceedings of the 10th EURALEX International Congress*. Toim Anna Braasch, Claus Povlsen. København: Center for Sprogteknologi, lk 503–509.
- Tuulik, Maria; Vainik, Ene; Prangel, Esta; Langemets, Margit; Aedmaa, Eleri; Koppel, Kristina; Risberg, Lydia (ilmumas).** Tähenduste seletamine leksikograafias: kui-võrd on abi suurtest keelemudelitest? – Eesti ja soome-ugri keeleteaduse ajakiri. *Journal of Estonian and Finno-Ugric Linguistics*.
- Vaik, Kristiina 2024.** Beyond Genres: A Dimensional Text Model for Text Classification. (Dissertationes linguisticae Universitatis Tartuensis 47.) Tartu: Tartu Ülikooli Kirjastus.
- Vaik, Kristiina; Sirts, Kairit; Muischnek, Kadri 2020.** Dimensionaalne tekstimudel. Teoreetiline ülevaade. – *Keel ja Kirjandus*, nr 10, lk 875–898. <https://doi.org/10.54013/kk755a4>
- Vainik, Ene; Lohk, Ahti; Paulsen, Geda 2021.** The distribution index calculator for Estonian. – *Electronic Lexicography in the 21st Century (eLex 2021): Post-editing Lexicography. Proceedings of the eLex 2021 Conference*. Toim Iztok Kosem, Michal Cukr,

Miloš Jakubíček, Jelena Kallas, Simon Krek, Carole Tiberius. Brno: Lexical Computing CZ, s.r.o, lk 121–138.

Vare, Silvi 2001. Üldkeele ja oskuskeele nihestunud suhe. – Keel ja Kirjandus, nr 7, lk 455–472.

Wit, Ernst-Jan C.; Gillette, Marie 1999. What is Linguistic Redundancy? Technical Report. Chicago: The University of Chicago.

ÕS 1999 = Eesti keele sõnaraamat ÕS 1999. Toim Tiiu Erelt. Koost T. Erelt, Tiina Leemets, Sirje Mäearu, Maire Raadik. Tallinn: Eesti Keele Sihtasutus.

ÕS 2006 = Eesti õigekeelsussõnaraamat ÕS 2006. Toim Tiiu Erelt. Koost T. Erelt, Tiina Leemets, Sirje Mäearu, Maire Raadik. Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus.

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

ÕS 2018 = Eesti õigekeelsussõnaraamat ÕS 2018. Toim Maire Raadik. Koost Tiiu Erelt, Tiina Leemets, Sirje Mäearu, M. Raadik. Eesti Keele Instituut. Tallinn: EKSA.

Lydia Risberg (snd 1988), PhD, Eesti Keele Instituudi teadur-keelekorraldaja (Munga 18, 50088 Tartu), Tartu Ülikooli eesti keele teadur, lydia.risberg@eki.ee

Maria Tuulik (snd 1985), PhD, Eesti Keele Instituudi vanemteadur (Roosikrantsi 6, 10119 Tallinn), maria.tuulik@eki.ee

Margit Langemets (snd 1961), PhD, Eesti Keele Instituudi juhtivleksikograaf (Roosikrantsi 6, 10119 Tallinn), margit.langemets@eki.ee

Kristina Koppel (snd 1985), PhD, Eesti Keele Instituudi vanemarvutileksikograaf (Roosikrantsi 6, 10119 Tallinn), kristina.koppel@eki.ee

Ene Vainik (snd 1964), PhD, Eesti Keele Instituudi juhtivteadur (Roosikrantsi 6, 10119 Tallinn), ene.vainik@eki.ee

Esta Prangel (snd 1985), MA, Eesti Keele Instituudi tootejuht (Roosikrantsi 6, 10119 Tallinn), esta.prangel@eki.ee

Eleri Aedmaa (snd 1989), PhD, Eesti Keele Instituudi keeletehnoloog (Roosikrantsi 6, 10119 Tallinn), eleri.aedmaa@eki.ee

Using corpus data to support lexicographers in identifying informal language

Keywords: corpus query system, lexicography, registers, labels, informal language

This study examines how new corpus analysis tools can assist lexicographers in determining whether to assign a word an informal register label in a dictionary. Labelling words in dictionaries is necessary for language users seeking register information.

Moreover, there have been calls for the upcoming *Dictionary of Standard Estonian* (DSE, 2025) to clearly distinguish standard language from other linguistic varieties.

Informal language was chosen for analysis because it is more difficult to define than other marked registers. In DSE 2018, some words were labelled as informal based on language planning decisions rather than empirical analysis. As register labels should be data-driven and based on corpus evidence, a systematic review of these words is necessary for the revised edition.

Our study investigates how corpus genre data can support lexicographers in deciding whether to add or remove the informal label. We found that corpus data provided useful insights in 82.1% of cases. Based on our experiment, we developed a guideline to assist in labelling word meanings as informal. Namely, if a word occurs in blogs and forums in 36% or more of its total corpus occurrences, it may be considered as tending towards informal usage. This guideline is not a rigid rule but a supportive tool, as additional factors should be considered based on the lexicographer's linguistic expertise.

Users value reliable linguistic information in dictionaries. Our proposed guideline helps lexicographers make more systematic decisions while maintaining expert judgment as the ultimate determinant.

Lydia Risberg (b. 1988), PhD, Institute of the Estonian Language, Researcher and Language Planner (Munga 18, 50088 Tartu), University of Tartu, Research Fellow in Estonian Language, lydia.risberg@eki.ee

Maria Tuulik (b. 1985), PhD, Institute of the Estonian Language, Senior Researcher (Roosikrantsi 6, 10119 Tallinn), maria.tuulik@eki.ee

Margit Langemets (b. 1961), PhD, Institute of the Estonian Language, Leading Lexicographer (Roosikrantsi 6, 10119 Tallinn), margit.langemets@eki.ee

Kristina Koppel (b. 1985), PhD, Institute of the Estonian Language, Senior Computational Lexicographer (Roosikrantsi 6, 10119 Tallinn), kristina.koppel@eki.ee

Ene Vainik (b. 1964), PhD, Institute of the Estonian Language, Leading Research Fellow (Roosikrantsi 6, 10119 Tallinn), ene.vainik@eki.ee

Esta Prangel (b. 1985), MA, Institute of the Estonian Language, Product Manager (Roosikrantsi 6, 10119 Tallinn), esta.prangel@eki.ee

Eleri Aedmaa (b. 1989), PhD, Institute of the Estonian Language, NLP Engineer (Roosikrantsi 6, 10119 Tallinn), eleri.aedmaa@eki.ee