doi:10.5128/ERYa18.12

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 18, 207–228

EESTI KEELE ÜHENDKORPUSTE SARI 2013–2021: MAHUKAIM EESTIKEELSETE DIGITEKSTIDE KOGU

Kristina Koppel, Jelena Kallas

Ülevaade. Eesti Keele Instituudi ja tarkvarafirma Lexical Computing Ltd. koostöös on valminud ühendkorpuste sari, milles on nüüdseks neli versiooni: eesti keele ühendkorpus 2013, 2017, 2019 ja 2021. Ühendkorpused on mahult suurimad eesti keele korpused ning nende rakendusvõimalused on laialdased, alates leksikograafia-alasest uurimistööst ning lõpetades masinõppe-otstarbeliste keelemudelite loomisega. Artiklis keskendume seni uusimale eesti keele ühendkorpusele 2021, mis koosneb suures osas veebist kogutud tekstidest. Kirjeldame veebitekstide kogumise, järeltöötluse ja puhastamise põhimõtteid ning ühendkorpuse allkorpusi, samuti anname ülevaate lähtetekstide klassifitseerimisest. Lisaks tutvustame korpuspäringusüsteemi Sketch Engine näitel korpusandemete uusi analüüsivõimalusi ning visandame korpusalase arendustöö edasisi perspektiive ja vajadusi.*

Võtmesõnad: eesti keele ühendkorpus, tekstikorpused, korpusleksikograafia, korpuspäringusüsteem, eesti keel

1. Sissejuhatus

Tekstikorpusi kasutatakse keele uurimisel, kirjeldamisel ja keelemuutuste jälgimisel, näiteks uute sõnade ja tähenduste tuvastamiseks, aga ka loomuliku keele töötluses (Jakubíček jt 2020). Eesti keele korpuste juhtiv looja on tänapäeval Eesti Keele Instituut (EKI), kus luuakse näiteks kaks- ja ükskeelseid valdkonnakorpusi, kõnekorpusi, eri keeleoskustasemetele suunatud õppekorpusi (loe lähemalt Koppel 2020) ning õpikute ja õppijakeele korpusi (loe lähemalt Kallas jt 2021). EKI korpuste visiitkaart on eesti keele ühendkorpuste (*Estonian National Corpus*) sari, mida on loodud koostöös tarkvarafirmaga Lexical Computing Ltd. alates aastast 2013. Nüüdseks on sarjas valminud neli versiooni: eesti keele ühendkorpus 2013 (ÜK 2013), 2017 (ÜK 2017), 2019 (ÜK 2019) ja 2021 (ÜK 2021). ÜK 2013 lähtetekstid olid lausestatud, morfoloogiliselt analüüsitud ja automaatselt ühestatud OÜ Filosofti

* Ühendkorpuste sari on valminud Haridus- ja Teadusministeeriumi Eesti Keele Instituudi baasfinantseerimise toel.

poolt analüsaatori Vabamorf (Kaalep, Vaino 2001) abil. Järgmiste versioonide märgendamiseks oleme kasutanud eesti keele töötlusprogrammi estNLTK teeke: ÜK 2017 märgendamisel estNLTK 1.4, ÜK 2019 märgendamisel estNLTK 1.6 ja ÜK 2021 märgendamisel estNLTK 1.6.9 teeki (Laur jt 2020). ÜK 2013, 2017 ja 2019 sõnestati, lausestati, osalausestati, lemmatiseeriti ning märgendati morfoloogiliselt, ÜK 2021 on aga esimene korpus, mis on märgendatud ka sõltuvussüntaktiliselt. See tähendab seda, et näidatakse lause puustruktuuri ehk seda, mis sõna millisele sõnale allub. Lisaks on tekstisõnadel süntaktilised märgendid, mis näitavad nende süntaktilist funktsiooni (nt subjekt, objekt, öeldisverb). Süntaktilisel märgendamisel on kasutatud EstNLTK meeskonna poolt treenitud Stanza mudeleid.[1]

Ühendkorpuste versioonid on kättesaadavad Lexical Computing Ltd. arendatud ja hallatavas korpuspäringusüsteemis Sketch Engine (Kilgarriff jt 2004, 2014) ning Eesti Keeleressursside Keskuse repositooriumis Entu. Kokku on 2022. aasta alguse seisuga Sketch Engine'is 20 avalikku eesti keele korpust (vt joonis 1).

Language	Name	↓ Words
Estonian	Estonian National Corpus 2021 (Estonian NC 2021)	2,410,296,919
Estonian	Estonian National Corpus 2019 (Estonian NC 2019)	1,500,284,681
Estonian	ELEXIS Estonian Web 2021	1,006,940,696
Estonian	Estonian Web 2021 (etTenTen21)	725,832,092
Estonian	Estonian Web 2017 (etTenTen17)	658,558,136
Estonian	Estonian Web 2019 (etTenTen19)	508,447,009
Estonian	Corpus of Estonian Web sentences 2021	473,455,876
Estonian	EUR-Lex Estonian 2/2016	291,077,511
Estonian	Corpus of Estonian Web sentences 2020	280,961,465
Estonian	Estonian Corpus for Learners 2020 (etSkELL)	280,572,215
Estonian	[DEV] Estonian Corpus for Learners 2018 (etSkELL)	248,203,200
Estonian	[DEV] Timestamped JSI web corpus 2014-2020 Estonian	212,608,965
Estonian	OpenSubtitles 2018 - Estonian	107,391,459
Estonian	[DEV] Estonian RSS Feed Corpus (Filosoft v2)	71,547,817
Estonian	OPUS2 Estonian	64,879,741
Estonian	DGT, Estonian	34,155,488
Estonian	EUR-Lex judgments Estonian 12/2016	15,029,608
Estonian	EUROPARL7, Estonian	11,171,727
Estonian	CHILDES Estonian Corpus	313,457
Estonian	Estonian coursebook corpus 2018	121,114

Joonis 1. Avalikud eesti keele korpused Sketch Engine'is

Osa korpustest on Sketch Engine'is arhiveeritud (vt joonis 2), kuid ka nendest on võimalik päringuid teha.

^ *Old versions of corpora found (**5**)*		
Estonian	Estonian Reference corpus 1990-2008 (EstonianRC)	203,267,951
Estonian	Estonian Web 2013 (etTenTen13)	260,559,829
Estonian	Estonian National Corpus 2013 (Estonian NC 2013)	463,827,780
Estonian	[DEV] Estonian RSS Feed Corpus	15,705,173
Estonian	Estonian National Corpus 2017 (Estonian NC 2017)	1,107,584,469

Joonis 2. Arhiveeritud eesti keele korpused Sketch Engine'is

[1] Täname EstNLTK meeskonda (Siim Orasmaa, Sven Laur, Kadri Muischnek, Kaili Müürisep) konsulteerimise eest ning Lexical Computing Ltd. tarkvaraarendajat Jan Michelfeiti ÜK 2021 märgendamise eest.

Järgnevas kirjeldame seni uusima ÜK 2021 kogumise, järeltöötluse ja puhastamise põhimõtteid allkorpuste kaupa, eraldi käsitleme lähtetekstide klassifitseerimist žanrideks ja teemadeks. Kuna veebist kogutud andmed moodustavad ühendkorpustes valdava osa, anname ülevaate veebikorpuste kogumise protsessist ning sellega kaasnevatest probleemidest. Eraldi kirjeldame ka uudisvoo (ingl *news feeds*) korpuste kogumise põhimõtteid. Artikli kolmandas osas anname Sketch Engine'i funktsioonide näitel ülevaate ühendkorpuse korpus(leksikograafia)alase analüüsi uutest võimalustest.

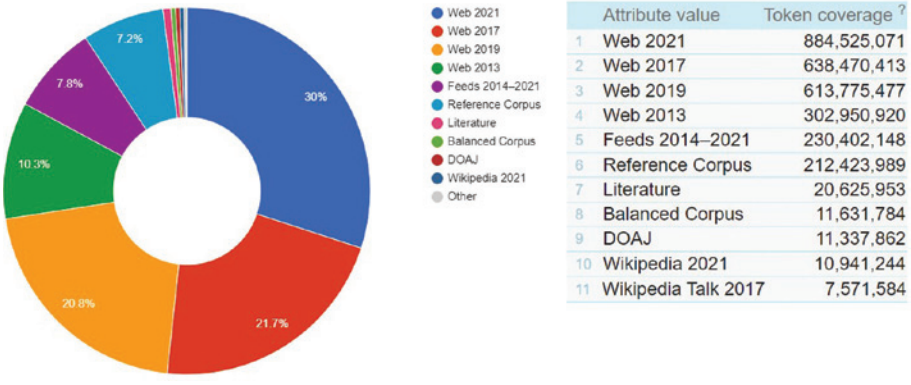
2. Eesti keele ühendkorpus ja selle allkorpused

Ühendkorpused on seni mahult suurimad eestikeelsete digitekstide kogud (vt tabel 1).

Tabel 1. Ühendkorpuste suurused

Korpus	Sõnesid	Sõnu	Lauseid	Lõike	Dokumente
ÜK 2013	563 mln	464 mln	38 mln	7,5 mln	700 tuh
ÜK 2017	1,3 mld	1,1 mld	88 mln	27 mln	3 mln
ÜK 2019	1,8 mld	1,5 mld	120 mln	35 mln	6 mln
ÜK 2021	2,9 mld	2,4 mld	197 mln	64 mln	12 mln

Ühendkorpuste kokkupanemisel oleme seadnud eesmärgiks tagada võimalikult laia keeleregistrite ja tekstitüüpide esindatuse. Selle tulemusena sisaldab ÜK 2021 ühteteist allkorpust (vt joonis 3): veebikorpused (*Web 2013*[2], *Web 2017*, *Web 2019*, *Web 2021*), uudisvood 2014–2021 (*Feeds 2014–2021*), Vikipeedia korpused (*Wikipedia 2021*, *Wikipedia Talk 2017*), avatud lähtekoodiga teadusartiklite korpus (DOAJ), kirjanduse korpus (*Literature*), koondkorpus (*Reference Corpus*) ja tasakaalus korpus (*Balanced Corpus*). Sketch Engine'is võib päringuid teha kas kogu korpuse ulatuses või selle allkorpuste kaupa. Joonis 3 illustreerib, et ÜK 2021 mahukaimad allkorpused on eri aastatel kogutud veebikorpused, millest suurim on 2021. aasta oma.



	Attribute value	Token coverage ?
1	Web 2021	884,525,071
2	Web 2017	638,470,413
3	Web 2019	613,775,477
4	Web 2013	302,950,920
5	Feeds 2014–2021	230,402,148
6	Reference Corpus	212,423,989
7	Literature	20,625,953
8	Balanced Corpus	11,631,784
9	DOAJ	11,337,862
10	Wikipedia 2021	10,941,244
11	Wikipedia Talk 2017	7,571,584

Joonis 3. ÜK 2021 allkorpuste suurused protsentides (diagrammil) ja sõnede arvu järgi (tabelis) [artikli veebiversioonis on diagramm värvilisena]

[2] Eesti keele veebikorpuse 2013 algupärane nimetus oli etTenTen13, mis viitas kuulumisele TenTen korpuste peresse (Jakubíček jt 2013). Alates 2017. aastast nimetatakse veebist kogutud tekstikorpusi veebikorpuseks (nt Web 2017).

2.1. Eesti keele veebikorpused 2013–2021

Traditsioonilisi korpusi, nagu nt British National Corpus (Leech 1992), koostatakse konkreetse eesmärgi jaoks ning teadlikult valitud allikatest, mille heas kvaliteedis ollakse kindlad. Selline tasakaalus ja kontrollitud sisuga korpuste loomine on väga aja- ja rahakulukas (Suchomel 2020) ja nii on tänapäeval põhiliseks tekstide allikaks saanud internet. Tekstide kogumine veebist on kiire, andmete mahud suured ning kulud madalad (Jakubíček jt 2020). Veebikorpusi saab tänapäeval luua sellistes mahtudes, mida traditsiooniliste korpuste loomise meetoditega on väga raske saavutada (Pomikálek jt 2012).

Osaliselt sel põhjusel moodustavad juba alates ÜK 2013-st veebikorpused suurima ühendkorpuse osa. 2013. aastal kogutud veebikorpuse suurus oli umbkaudu 313 mln sõnet, moodustades 56% ÜK 2013 mahust; 2017. aastal 763 mln sõnet, moodustades 80% ÜK 2017 mahust; 2019. aastal 615 mln sõnet, moodustades 87% ÜK 2019 mahust; ning 2021. aastal 951 mln sõnet, moodustades 91% ÜK 2021 mahust. Seega on ühendkorpus tervikuna olemuselt pigem veebikorpus koos selle tüüpiliste probleemidega.[3]

2.1.1. Veebikorpuste kogumise ja puhastamise üldpõhimõtted

Veebikorpuste kogumiseks kasutatakse veebi automaatseks lehitsemiseks loodud programmi ehk kroolijat (ingl *crawler*), mis liigub ühelt veebilehelt leitud URL-e mööda edasi, laadides neilt alla tekstilise materjali. Seda protsessi nimetatakse veebi kroolimiseks (*web crawl*).

Tarkvarafirma Lexical Computing Ltd. on spetsiaalselt veebikorpuste kogumiseks välja arendanud kroolija SpiderLing[4] (Suchomel, Pomikálek 2012, Suchomel 2020). Erinevalt teistest kroolijatest ei ole SpiderLingi eesmärk koguda kõiki andmeid kõikidelt veebilehtedelt, vaid otsida võimalikult palju grammatiliselt terviklikke lauseid sisaldavad dokumendid võimalikult lühikese ajaga. Selleks alustab SpiderLing kroolimist kindlatelt veebilehtedelt ehk võtmedomeenidelt (*seed domains*), mille puhul ollakse kindlad, et need sisaldavad kvaliteetset tekstilist materjali. SpiderLing laadib dokumendid alla ning kordab protsessi võtmedomeenidelt leitud URL-e mööda aina edasi liikudes. Nii on hiljem korpuse järeltöötluse faasis ka vähem müra.

Pärast dokumentide allalaadimist eemaldatakse programmiga jusText (Pomikálek 2011) mittetekstiline materjal (nt reklaamid, lingid, multimeedia), tabelid, veebilehtede päised ja jalused, HTML-i märgendus. Alles jäetakse vaid terviklikke lauseid sisaldavad lõigud. Samuti aitab JusText tuvastada keelt. Selleks toetub programm algoritmile, mis arvestab teatud keele grammatilisi vorme. Duplikaatide ehk identsete ja lähiduplikaatide ehk väga sarnaste tekstide eemaldamiseks kasutatakse programmi Onion (Pomikálek 2011). Duplikaatide tuvastamiseks võrreldakse dokumente lõigu tasandil, kuna lause on selleks liiga väike ning terve dokument liiga suur üksus (Jakubíček jt 2013). Duplikaatide eemaldamine hõlmab ka võrdlusi varasemate ühendkorpuse versioonidega – kui nt 2021. aastal

[3] Veebikorpuste tüüpilisi probleeme on kirjeldanud nt Maristella Gatto (2014), Jakubíček jt (2020), Koppel (2020).
[4] Kroolijat kasutatakse ka valdkonnakorpuste loomisel (vt lähemalt Kallas jt 2017).

kroolitud dokument oli olemas juba 2013. aasta korpuses, jäetakse alles varasem. Pärast korpuse puhastamist tekstid märgendatakse.

Keelte jaoks, mille kohta on veebis piisavalt materjali, saab eelnevalt kirjeldatud viisil koguda miljardeid sõnu päevas, nt koguti 12 mld ingliskeelset sõna sisaldav inglise keele korpus etTenTen12 ainult 12 päevaga (Jakubíček jt 2013). Selline korpuse kogumise protsess on küll väga kiire, aga toob kaasa ka palju müra (Manning jt 2008). Vít Suchomel (2020) on loetlenud veebikorpuste loomise peamised probleemid: veebi kroolimine piisaval hulgal, keele identifitseerimine ja sarnaste keelte eristamine[5], tekstikodeeringu tuvastamine, tekstide (HTML-ist) puhastamine, duplikaatide eemaldamine, ebakvaliteetse sisuga võitlemine, autorsuse ja loomevarguse tuvastamine, suurte tekstikogude varundamine ja indekseerimine. Sageli on probleeme väheste metaandmetega, näiteks ei saa usaldada teksti avaldamise kuupäeva, v.a teatud tüüpi tekstide (nt ajaleheuudis, pressiteade, blogipostitus) puhul (Jakubíček jt 2013).

Kõige põletavam probleem veebikorpuste loomisel on aga masintõlkelised tekstid ja veebispämm, mis mõjutavad oluliselt korpuse sisu ning need tuleb kroolimisele järgnevas korpuse töötluse faasis eemaldada. Tegelikult oskab ka SpiderLing URL-i omaduste järgi kindlaks teha ja kõrvale jätta potentsiaalsed ebakvaliteetse sisuga allikad, arvestades selleks URL-i kaugust võtmedomeenist ning veebiaadressi pikkust, kuid kroolitud dokumentidesse satub sellest hoolimata masina loodud sisu, eriti keelte puhul, millel on vähe rääkijaid ja mille väike turg inimtõlget lihtsalt ei võimalda (Jakubíček jt 2020).

Suur osa masintõlkelistest tekstidest näeb välja nagu inimloodud tekst ning neid on automaatselt keeruline tuvastada, mistõttu kasutatakse selleks poolautomaatseid meetodeid. See tähendab seda, et esmalt kontrollib keelt emakeelena rääkiv inimene käsitsi üle teatud arvu veebilehti, millelt on kogutud kõige rohkem tekste. Tavapraktika on, et suurte keelte (nt inglise, hispaania, saksa) puhul, mille korpuste suurused jäävad kümnetesse miljarditesse, kontrollitakse käsitsi u 2000–5000 URL-i; väiksemate keelte puhul, mille korpuste suurused jäävad miljarditesse, kontrollitakse u 300–500 URL-i. (Suchomel, Kraus 2021) Käsitsi kontrollitud URL-ide peal treenitakse välja klassifikaator, mis tuvastab sarnaste parameetritega veebilehed, kust alla laaditud dokumendid eemaldatakse (Suchomel 2020).

Ka spämmilehed toovad kaasa ebaloomulikku keelekasutust ja seega soovimatut sisu. Spämmilehtede alla liigituvad näiteks veebilehed, millele on sisse häkitud, sageli selleks, et kasutajaid meelitada illegaalsete teemadega (nt narkootikumid, pornograafia, hasartmängud, relvad). Tüüpiliselt on sellised spämmilehed üleval ajutiselt ning need eemaldatakse suhteliselt kiiresti, kuid kuna kroolija kroolib veebi teatud kuupäeva seisuga, satuvad ka spämmilehed alla laaditud dokumentide hulka. Üks viis, kuidas neid vältida, on alustada kroolimist usaldusväärsetelt URL-idelt ning veebikataloogidest.

Veebi kroolimist takistavad veel suletud, tasuline ja dünaamiline sisu. Näiteks on tänapäeval paljudel asutustel kodulehe asemel ainult sotsiaalmeediakonto, millele kroolijad ligi ei pääse. Samuti ei saa kroolijad kätte tasulisi uudiseid, mis tähendab, et veebist on võimalik kätte saada vaid murdosa keelelistest andmetest. On oht, et kui asutused kolivad oma kodulehed üleni sotsiaalmeediasse ning uudisteportaalid muudavad kõik postitused tasuliseks, siis ei ole veebi kroolimisel

[5] Näiteks on keeruline eristada Portugalis ja Brasiilias räägitavat portugali keelt, Ameerikas ja Ühendkuningriikides räägitavat inglise keelt (Jakubíček jt 2013), aga ka kirillitsas kirjutatud keeli (nt vene, valgevene ja ukraina) (Koppel jt 2019).

enam mõtet ning tuleb tagasi pöörduda traditsiooniliste korpuse loomise meetodite juurde. (Jakubíček jt 2020)

Järgnevalt kirjeldame probleeme, millega oleme kokku puutunud just eestikeelsete veebikorpuste kroolimisel.

2.1.2. Eesti keele veebikorpuste kogumine ja järeltöötlus

Veebikorpus 2021 krooliti 2021. aasta juunist septembrini (koos kuuajalise pausiga).[6] Kroolimist alustati usaldusväärsetelt URL-idelt, mille olime käsitsi tuvastanud juba 2019. aasta veebikorpuse jaoks (kokku 851 URL-i), ning *neti.ee* veebikataloogist pärit URL-idelt (kokku 20 718).

Kroolimise järeltöötluse faasis genereeris tarkvarafirma loendi 5000 URL-iga, millelt kõige rohkem dokumente alla laaditi. Käsitsi üle kontrollitud 4002 URL-ist oli masintõlkelisi 1036 ehk u 26%. Võrdluseks: 2019. aasta veebikorpuse 600 URL-iga loendist eemaldasime ebakvaliteetse sisu tõttu 110 ehk umbes 18% (Suchomel 2020). 2021. aastal oli esimese 600 URL-i seas masintõlkelisi tekste 310 ehk lausa 51,6%.[7] Vít Suchomeli ja Jan Krausi (2021) andmetel jõuab üks inimene tunnis kontrollida keskmiselt 50–70 URL-i, mis tähendab, et 4002 URL-i kontrollimiseks kulus 7–10 ühe inimese täispikka tööpäeva.

Selleks et tuvastada, kas tegemist on kvaliteetse või ebakvaliteetse materjaliga, piisas sageli lingi avamisest, kuna masintõlkelised veebilehed on sarnase välimusega (vt joonis 4). Lisavõimalus kvaliteedikontrolli teha oli analüüsida juhuslikke konkordantsiridu, mida igalt URL-ilt oli 70.

Tüüpiliselt pärinesid masintõlkelised tekstid URL-idelt, mis algavad eesti keelele viitava koodiga *et*, *ee* või *est* ning lõppevad rahvusvahelise domeeniga *com*, *net* või *org* (nt *et.e-pistolas.org*, *et.goodlifestudio.net*). Sarnaselt genereeritud domeenid on tüüpilised ka teistes keeltes ning nende tuvastamist on Miloš Jakubíčeki jt (2020) sõnul võimalik automatiseerida. Ent väga palju oli ka pealtnäha eestikeelseid URL-e, nagu *siitsealt.ee* ja *suhkrupatt.ee*, millest pelgalt välimuse tõttu võinuks eeldada kvaliteetset sisu, ent mis endas siiski masintõlget peitsid.

Järgnevalt analüüsisime veebikorpuse sisu täiendavalt võtmesõnade abil. Kasutasime selleks funktsiooni Keywords (Kilgarriff 2012), mis aitas tuvastada need märksõnad, mis oluliselt esildusid eesti keele veebikorpuses 2021 võrreldes ÜK 2019-ga. Enamasti olid nendeks prostutitutsiooni või tervise ja ravimitega seotud sõnad. Nii tegime kindlaks veel 352 masintõlkelisi tekste sisaldavat URL-i, millelt kogutud dokumendid korpusest eemaldati. Samas etapis eemaldasime ka vana kirjakeelt sisaldanud tekstid (allikad olid eesti vana kirjakeele korpus ja EKI piiblikonkordants) ning vigase tärktuvastusega PDF-failid (sõnadest olid kadunud täpitähed).

[6] Täname Lexical Computing Ltd. tarkvaraarendajat Vít Suchomeli eestikeelse veebi kroolimise ning ÜK 2021 kokkupanemise eest.

[7] Nt 2021. aastal oli esimese kümne URL-i seas masintõlkelisi tekste kaheksa, esimese 200 seas 97 (48,5%) ning esimese 1000 seas 362 (36%).

Joonis 4. Masintõlkelised veebilehed *et.delachieve.com* (vasakul) ja *paikusemesimumm.ee* (paremal)

2.1.3. Veebitekstide URL-i põhine klassifitseerimine

2013. aastal tegime esmakordselt katse veebidomeene klassifitseerida, määrates neile võimalusel juurde tekstitüübi. Tollal eristasime kuut tekstitüüpi: ajakirjandustekstid, foorumid, blogid, teabetekstid, usutekstid ja ametlikud tekstid. ÜK 2017-s jäid tekstid klassifitseerimata, ÜK 2019-s kasutasime 13 tekstitüüpi (blogid, foorumid, haridus, ilukirjandus, toit, tervis, ajakirjad, uudised, religioon, teadus, seks, ühiskond, sport ja Vikipeedia). ÜK 2021-s järgisime kahetasandilist klassifikatsiooni (Suchomel 2020) ning jagasime tekstid žanridesse (laiem klass)[8] ja teemadesse (kitsam klass)[9]. Žanri määrab kirjutamisstiil, näiteks on ilukirjanduse ja foorumipostituste stiil täiesti erinev, kuid teemad võivad olla žanriülesed. Kahetasandiline klassifikatsioon tõi kaasa selle, et käsitleme nüüd osasid teemasid kohati detailsemalt (näiteks ÜK 2019-s kasutusel olnud tekstitüüp "ühiskond" hõlmas endas nii majandust kui ka õigust, mis on nüüd kumbki omaette teema) ning žanreid laiemalt (liitsime ÜK 2019 tekstitüübid "uudised" ja "ajakirjad" nüüd kokku üheks žanriks "perioodika"). Täiendavalt võtsime e-poodide laialdase leviku tõttu kasutusele täiesti uue žanri "veebikaubandus".

Veebilehtedele žanri ja teema määramine toimus paralleelselt masintõlkeliste tekstide tuvastamisega. Kui tekst oli kvaliteetne, määrasime talle võimalusel žanri, kuhu alla URL-ilt kogutud tekstid liigitusid (nt perioodika), ning teema(d), millest seal kirjutati (nt sport). Erinevaid žanreid koorus välja viis (vt tabel 2) ning teemasid 24 (vt tabel 3). Kokku määrasime 4002-st käsitsi kontrollitud URL-ist 528-le nii žanri kui ka teema. Näiteks liigitus Õpetajate Leht perioodika alla, mille teema on haridus ning Bauhof veebikaubanduse alla, mille teema on ehitus ja kinnisvara. Teatud juhtudel määrasime ühele URL-ile mitu teemat, näiteks on

[8] Suchomeli (2020) kogemuse põhjal on žanreid raskem määrata kui teemasid. Ka Sue Atkins ja Michael Rundell (2008) on öelnud, et tekste saab kategoriseerida mitmel viisil, kuid laiadel kategooriatel on piirid hägusad ning need ei ole alati üksteist välistavad.

[9] Tekstitüüp on žanri ja teema ülemmõiste (Suchomel 2020).

Haridusministeeriumi kodulehe teema haridus, aga ka poliitika ja valitsemine. Ainult žanri määrasime 746-le ning ainult teema 946-le URL-ile.

Veebitekstide klassifitseerimise[10] eesmärk oli mh ette valmistada treeningandmed ülejäänud ühendkorpuse tekstide automaatseks klassifitseerimiseks. Veebilehtede arv tabelis 2 ja 3 viitab treeningandmete mahule ehk sellele, kui mitu URL-i me neist käsitsi kontrollitud 4002-st lehest vastavasse žanrisse/teemasse määrasime, ning sõnede arv sellele, kui palju sõnesid neil veebilehtedel kokku esines. See tähendab seda, et 4002-st käsitsi kontrollitud URL-ist klassifitseerisime 2220 – see moodustas treeningmaterjali[11], mille abil Lexical Computing Ltd. tarkvaraarendajate treenitud klassifikaator ülejäänud URL-idelt kogutud tekstid automaatselt klassifitseeris. Tekstide automaatsel klassifitseerimisel kasutati kõrget lävendit – kui klassifikaator ei olnud žanris või teemas kindel, siis seda tekstile ei määratud (Suchomel 2020).

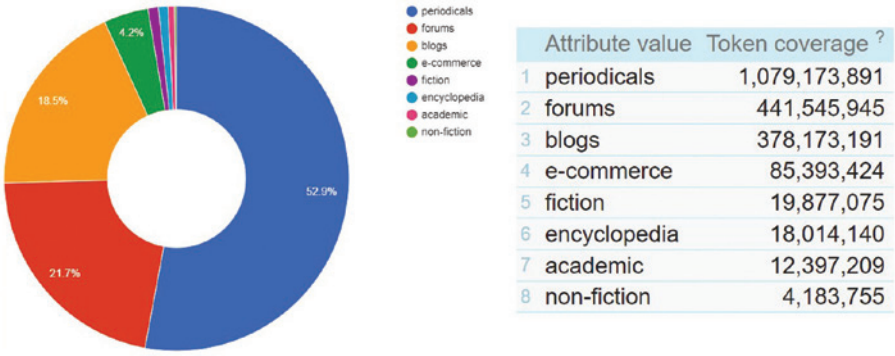
Nagu eelnevalt mainitud, koorus URL-ide kontrollimisel välja viis žanrit (tabel 2).

Tabel 2. Žanride klassifikatsioon URL-ide põhjal

Žanrid	Näited	Veebilehed	Sõned
blogid (*blogs*)	Mallukas, Marimell, Paljas Porgand, Päevakera blogi	761	174 642 782
entsüklopeedia* (*encyclopedia*)	KakuWiki	4	27 205 001
foorumid (*forums*)	Lapsemure, Soccernet, Rahafoorum	121	243 841 952
perioodika (*periodicals*)	Õhtuleht, Sirp, Horisont, Anne ja Stiil	208	297 687 046
veebikaubandus (*e-commerce*)	Photopoint, Loverte, Kaup24	180	36 183 138
Kokku		1274	794 081 388

* Entsüklopeedia alla liigitusid Vikipeedia-laadsed veebilehed.

Pärast tekstide automaatset klassifitseerimist on ÜK 2021 mahukaim žanr perioodika (52,9%), millele järgnevad foorumid (21,7%), blogid (18,5%) ja e-kaubandus (4,2%) (vt joonis 5).



Joonis 5. Žanride jaotus ÜK 2021-s [artikli veebiversioonis on diagramm värvilisena]

[10] Eestikeelsete veebitekstide automaatset liigitamist on katsetanud Kristiina Vaik ja Kadri Muischnek (2018).

[11] ÜK 2021 automaatseks klassifitseerimiseks kaardistasime ÜK 2019 tekstitüüpid, mille kattuvad osad omavahel ühendasime. Nendel andmetel treenitud klassifikaatori abil määrati uus klassifikatsioon ka varasematele, ÜK 2013–2019 tekstidele.

Joonisel 5 on lisaks välja toodud kolm žanrit – akadeemiline kirjutamine (*academic*), ilukirjandus (*fiction*) ja mitte-ilukirjandus (*non-fiction*) –, mis ei selgunud mitte URL-ide kontrollimise teel, vaid pärinevad konkreetsetest allkorpustest – avatud lähtekoodiga teadusartiklite korpusest (loe lähemalt ptk 2.4) ning kirjanduse korpusest (loe lähemalt ptk 2.6).

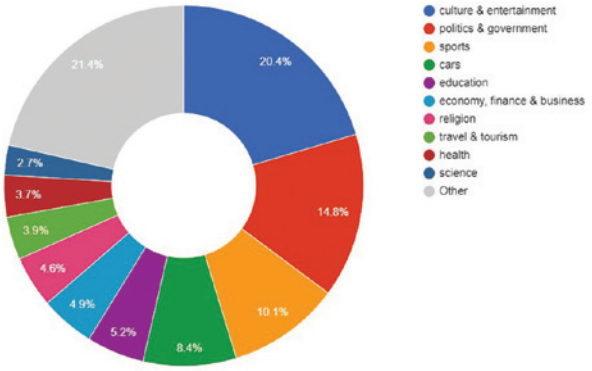
Teemasid koorus URL-e kontrollides välja 24 (vt tabel 3).

Tabel 3. Teemade klassifikatsioon URL-ide põhjal

Teemad	Näited	Veebilehed	Sõned
ajalugu (*history*)	ajalehtede ajalooteemalised alamlehed, ajakiri Imeline Ajalugu	10	2 077 467
autod (*cars*)	automarkide (Mazda, Volvo, BMW) foorumid	89	73 003 851
ehitus ja kinnisvara (*construction & real estate*)	ehitustarvete poed (K-Rauta, Bauhof), ehitusfoorumid, kinnisvaraportaalid	43	21 410 100
haridus (*education*)	Õpetajate Leht, koolide ja ülikoolide kodulehed	149	42 741 627
hasartmängud (*gambling & casinos*)	kõikvõimalikud loto- ja kasiinolehed ning kihlveokontorid	20	11 522 203
ilu (*beauty*)	ilukaubamajad ja -salongid	29	8 100 512
IT ja tehnoloogia (*technology & IT*)	meediaväljaannete tehnikateemalised alamlehed, elektroonikafirmade (Samsung, Olympus) või elektroonikat müüvate e-poodide (Euronics, Photopoint) veebilehed	64	28 296 009
kodu, pere ja lapsed (*home, family & children*)	laste- ja perefoorumid	23	8 086 156
kultuur ja meelelahutus (*culture & entertainment*)	muuseumite, teatrite, kunstisaalide, muusika-, tantsu-, kirjandus-, kunsti-, filmi- ja lugemisblogide ja samateemaliste lehtede ja ajakirjade kodulehed	241	114 453 884
majandus, rahandus ja äri (*economy, finance & business*)	majandusteemalised ajalehed (Äripäev, Ärileht), teiste meediaväljaannete (Postimees) majandusteemalised alamlehed, pankade kodulehed	45	26 915 932
loodus ja keskkond (*nature & environment*)	RMK koduleht, ilmaennustuse lehed, aiandusteemalised lehed	45	18 407 679
loomad (*pets & animals*)	kodutute loomade varjupaigad, erinevate loomaliikide (nt hobuste) ühingud, kalastusfoorumid	50	16 635 737
naised (*women*)	Delfi Naistekas, Buduaariturg, Ohmygossip	42	14 339 020
poliitika ja valitsemine (*politics & government*)	ministeeriumide ja erakondade veebilehed, erakondade meediakanalid (Kesknädal, Uued Uudised)	106	84 112 180

Teemad	Näited	Veebilehed	Sõned
põllumajandus (*agriculture*)	Maaeluministeeriumi koduleht, Eesti Põllumajandusloomade Jõudluskontrolli koduleht	13	1 772 201
reisimine ja turism (*travel & tourism*)	turismifirmade lehed, reisifoorumid ja -blogid	78	43 886 494
religioon (*religion*)	erinevate kirikute kodulehed, kristlikud meediakanalid	60	36 315 424
seks (*sex*)	tutvumisportaalid, sekslelude veebipoed	23	6 049 320
sport (*sports*)	meediaväljaannete sporditeemalised alamlehed, erinevate spordiklubide kodulehed	156	85 893 650
teadus (*science*)	meediaväljaannete teadusteemalised alamlehed (Novaator, Forte), ajakiri Horisont, Eesti Teadusagentuuri koduleht	15	6 827 925
tervis (*health*)	haiglate kodulehed, ajalehtede terviseteemalised alamlehed	63	27 342 274
toit ja joogid (*food & drinks*)	retseptikogud, kokandusblogid, toiduteemalised ajakirjad	71	13 789 566
videomängud (*video games*)	arvutimängude teemalised veebilehed ja foorumid	17	7 818 027
õigus (*law & justice*)	kohtute veebilehed, õigusaktide registrid	22	11 568 877
Kokku		1 474	712 192 346

Tekstide automaatse klassifitseerimise järel on ÜK 2021 mahukaimad teemad kultuur ja meelelahutus (20,4%), poliitika ja valitsemine (14,8%), sport (10,1%), autod (8,4%), haridus (5,2%), majandus, rahandus ja äri (4,9%), religioon (4,6%), reisimine ja turism (3,9%) ning tervis (3,7%) (vt joonis 6).



	Attribute value	Token coverage
1	culture & entertainment	279,730,261
2	politics & government	202,310,370
3	sports	138,761,332
4	cars	114,602,631
5	education	71,062,199
6	economy, finance & business	66,951,835
7	religion	63,546,395
8	travel & tourism	52,849,561
9	health	51,314,159
10	science	36,653,740
11	technology & IT	35,770,784
12	construction & real estate	33,983,523
13	women	31,058,272
14	pets and animals	30,932,007
15	law & justice	25,337,732
16	food & drinks	22,914,246
17	nature & environment	21,864,842
18	home, family & children	19,778,450
19	gambling & casinos	15,263,308
20	sex	13,108,994
21	beauty	12,923,899
22	video games	9,216,498
23	agriculture	4,632,784
24	history	3,182,480

Joonis 6. Teemade jaotus ÜK 2021-s [artikli veebiversioonis on diagramm värvilisena]

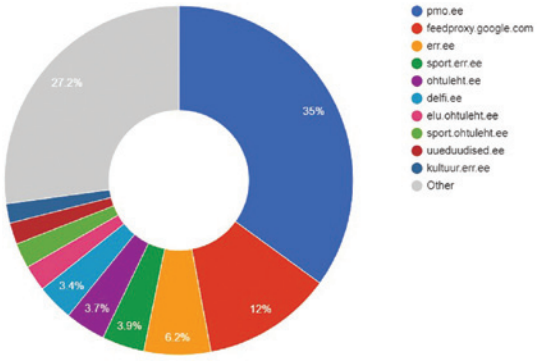
Lähtetekstide klassifitseerimine võimaldab nüüd Sketch Engine'i sõnavisandites analüüsida ka kollokatsioonide žanrilist ja temaatilist kuuluvust (vt ptk 3.2).

2.2. Uudisvood 2014–2021

Tavapäraselt on veebi kroolimine ühekordne tegevus, mille käigus kogutakse dokumente teatud kuupäeva seisuga. Nii valmib staatilise sisuga korpus, kust saab teha sama tulemusega päringu ka aastate pärast. Samas on olemas ka spetsiaalsed kroolijad (vt nt Cho, Garcia-Molina 1999, Fetterly jt 2009), mis valitud allikatest korpusi järk-järgult uuendavad. EKI on alates 2020. aastast kogunud kahte reaalajas täienevat korpust – eesti keele uudisvoo korpust (*Estonian RSS Feed Corpus*, edaspidi RSS uudisvoo korpus) ja JSI ajamärgistatud eesti keele uudisvoo korpust (*Timestamped JSI web corpus 2014-2020 Estonian*, edaspidi JSI uudisvoo korpus).[12] Korpuste tüpoloogia seisukohalt liigituvad RSS ja JSI uudisvoo korpused monitorkorpuste alla, kuna sinna lisatakse tekste reaalajas aina juurde ning selle tulemusena saab seirata muutusi keeles (Clear 1987, Cartier 2017). Monitorkorpused on head allikad just uudissõnade ja nende kasutusmustrite uurimiseks.

Selliste korpuste kogumiseks on vaja, et veebilehed kasutaksid RSS-funktsiooni, mis kogub automaatselt XML-kujul masinloetavas formaadis faile, mida regulaarselt uuendatakse. Interneti tavakasutajale on veebilehtede RSS-funktsioon ilmselt tuttav, kui ta kasutab uudisvoo lugejat (*feed reader*), mis võimaldab konkreetseid veebilehti külastamata nende viimaste uudistega kursis olla. Uudised saadetakse tellitud veebilehtedelt (nt meediaportaalist, blogidest) uudisvoo lugejasse reaalajas, postituse külge on märgitud pealkiri ja selle esimene lõik, postitamise aeg ning URL.

Need eesti veebilehed, mis RSS-funktsiooni kaudu uuendusi tellida lubavad ja millelt alla laaditud dokumentide abil me RSS uudisvoo korpust kogume, on tuvastatud käsitsi. 2020. aasta märtsis alustasime tekstide kogumist 336 URL-ilt, 2021. aasta novembris lisasime juurde ligikaudu 400 URL-i. 2022. aasta märtsikuu seisuga on neist aktiivsed 555 ning neilt on korpusesse kogutud ligikaudu 77 mln sõnet. Kõige rohkem dokumente on pärit *online*-meedia portaalidest (vt joonis 7), nagu Postimees, ERR, Õhtuleht ja Delfi. Kuigi tarkvarafirma kogub dokumente reaalajas, uuendatakse korpuse sisu Sketch Engine'is umbes kord kuus.



Joonis 7. Eesti keele uudisvoo korpuse suurimad allikad 2022. aasta märtsikuu seisuga [artikli veebiversioonis on diagramm värvilisena]

[12] Täname Lexical Computing Ltd. tarkvaraarendajat Ondřej Hermani RSS uudisvoo korpuse ja JSI uudisvoo korpuse haldamise ja täiendamise eest.

Teine sarnane uudisvoogudel põhinev korpus, mis on Sketch Engine'is alates 2021. aastast eesti keele jaoks avalik, on Jozef Stefani Instituudi (JSI) uudisvoo korpuste perekonda (Bušta jt 2017, Trampuš, Novak 2012) kuuluv JSI uudisvoo korpus. JSI perekonna uudisvoo korpused on alates 2014. aastast RSS-funktsiooni abil uudiseid kogunud u 80 000 veebilehe uudisvoogudest, mille kaudu avaldatakse erinevates keeltes 350 000–600 000 artiklit päevas.[13] Seitsme aasta jooksul (2014–2021) on eestikeelseid tekste alla laaditud 524 veebilehelt (märtsikuu seisuga 255 mln sõnet), mis osaliselt kattuvad eesti keele uudisvoo korpuse allikatega.

RSS ja JSI uudisvoo korpused on Sketch Engine'is kättesaadavad eraldiseisvate reaalajas täienevate korpustena, kuid nende 2014.–2021. a kogutud dokumendid on staatilise väljavõttena lisatud ka ÜK 2021 allkorpuseks (*Feeds 2014–2021*).

2.3. Koondkorpus ja tasakaalus korpus

Tartu Ülikoolis koostatud eesti keele koondkorpus sisaldab ajakirjandustekste aastatest 1995–2008, ilukirjandustekste alates 1990-ndatest, riigikogu stenogramme aastatest 1995–2001, Eesti ja Euroopa seaduseid, teadustekste (doktoritöid, teadusajakirju) ja uue meedia tekste (uudistekstid, jututoad, foorumid). Tartu Ülikoolis koostatud tasakaalus korpus sisaldab võrdses mahus ajalehti, ilu- ja teaduskirjandust.

2.4. Avatud lähtekoodiga teadusartiklite (DOAJ) allkorpus

Directory of Open Access Journal ehk DOAJ on veebikataloog, mis alates 2003. aastast koondab ligikaudu 18 000 avatud lähtekoodiga ajakirja. DOAJ andmebaasis on üle seitsme miljoni artikli, mis hõlmavad kõiki teaduse, tehnoloogia, meditsiini, sotsiaalteaduste ja humanitaarteaduste valdkondi. Eestikeelseid artikleid on seal näiteks ajakirjadest Eesti ja Soome-ugri Keeleteaduse Ajakiri, Eesti Rakenduslingvistika Ühingu aastaraamat, Lähivõrdlusi, Methis: Studia Humaniora Estonica, LingVaria, Folklore, Eesti Arst, Eesti Haridusteaduste Ajakiri, Ajalooline Ajakiri, Mäetagused, Estonian Journal of Earth Sciences, Eesti Majanduspoliitilised Väitlused ja Agraarteadus. ÜK 2021-s on DOAJ portaalist pärit tekstid koondatud allkorpusesse DOAJ, žanriks on määratud "akadeemiline kirjutamine" (*academic*) ning teemaks "teadus" (*science*).

2.5. Vikipeedia korpus

Eestikeelse Vikipeedia tekstidest oleme loonud kaks allkorpust – Vikipeedia (11 mln sõnet) ning Vikipeedia arutelud 2017 (*Wikipedia Talk 2017*) (7,6 mln sõnet). Vikipeedia arutelude korpus sisaldab 2017. aastal alla laaditud Vikipeedia artiklite toimetajate arutelusid.

[13] Ingliskeelseid sõnu koguneb ühe kuu jooksul uudisvoogude kaudu umbes 1 mld. JSI uudisvoo korpused on olemas veel araabia, katalaani, tšehhi, saksa, inglise, soome, prantsuse, horvaadi, ungari, itaalia, korea, hollandi, poola, vene, hispaania, serbia ja rootsi keele jaoks.

2.6. Kirjanduse korpus

Eesti keele korpustes on ilukirjandusest alati suur puudus olnud. Vanemad tekstid on autoriõiguste alt vabastatud[14], kuid uuemad mitte, mistõttu peab enne materjali korpusesse lisamist esmalt tegelema autoriõiguste ja intellektuaalse omandi õigusega. EKI alustas võimalike keeleressursside loovutamise üle kirjastustega läbirääkimisi 2021. aasta lõpus.[15] ÜK 2021 kokkupanemise ajaks olid instituudile teoseid loovutada jõudnud kirjastused Varrak, Koolibri ja Petrone Print – kolme peale kokku 228 raamatut. Nende seas on nii originaal- kui ka tõlketeoseid.

Kirjanduse allkorpuse (21 mln sõnet) moodustavad seega need 228 teost, aga ka koondkorpuses sisaldunud eesti algupärased ilukirjandusteosed (kokku 160 raamatut). Kõikide teoste metaandmetes on kirjas, kas tegemist on originaali (st eesti autori kirjutatud) või tõlkega, samuti on olemas originaali või tõlke pealkiri. Iga teose puhul on sarnaselt veebilehtede klassifikatsioonile määratud ka žanr, milleks võib olla kas ilukirjandus (*fiction*) või mitte-ilukirjandus (*non-fiction*). Teemade asemel kasutame kirjanduse korpuse dokumentide klassifitseerimisel hoopis tekstitüüpe – ilukirjanduslikud tekstitüübid on nt romaan (*novel*), novell (*short story*), lasteraamat (*children's book*), memuaar (*memoir*), lugude sari (*set of stories*); mitte-ilukirjanduslikud tekstitüübid on nt käsiraamat (*handbook*), reisikiri (*travel writing*), essee (*essay*), õpik (*textbook*), populaarteadus (*popular science*).[16]

3. Ühendkorpuste korpusleksikograafiline analüüs Sketch Engine'i funktsioonide näitel

Sketch Engine'i põhifunktsioone ja nende eesti keele mooduleid (sagedusloend (*Wordlist*), konkordants (*Concordance*), statistilised kollokaadid (*Collocates*), sõnavisand (*Word Sketch*), tesaurus (*Thesaurus*), heade näitelausete tuvastamine (GDEX)) on põhjalikult kirjeldatud (Kallas jt 2012, Kallas 2013, Kallas jt 2015, Kallas jt 2017, Koppel 2020).

ÜK 2021 tarbeks oleme uuendanud sõnavisandite grammatikat (*SketchGrammar*). Sõnavisandite grammatika versioonis 2.1 on kokku 113 grammatilist suhet.[17] Grammatika kirjutamisel on esmakordselt kasutatud makrokeelt m4. Makrodena on defineeritud nt kvantorid, alistavad sidesõnad, genitiivi nõudvad prepositsioonid, afiksaaladverbid, infiniitsed verbivormid. Grammatiliste suhete nimetused on muudetud eestikeelseks, nt suhte *ADJ_modifier* asemel kasutame nimetust *omadussõnaga*.

ÜK 2021 tulekuga lisandus neli uut funktsiooni, mida seni polnud Sketch Engine'is eesti keelele rakendatud: terminituvastus (*Multi-Word Terms*) (Baisa jt 2017), kollokatsioonide žanrilise ja temaatilise kuuluvuse tuvastus sõnavisandites

[14] Autoriõiguse seaduse järgi kehtib autoriõigus autori kogu eluaja jooksul ja 70 aastat pärast tema surma. 2021. aastal alustasime koostööd Tartu Ülikooli raamatukoguga, kes pakkus autoriõiguste alt vabanenud teoste digiteerimist. Samuti alustasime juba digiteeritud ja autoriõiguste alt vabade teoste kogumist avalikest andmebaasidest. Suur osa vabakasutuses olevatest teostest on olemas ainult PDF-ina, mille tekstituvastuse kvaliteet on aga väga varieeruv. Töö vanema ilukirjanduse kättesaadavamaks tegemiseks on käimas ning loodame seda lisada juba järgmisse, 2023. aasta ühendkorpuse versiooni.

[15] Täname Tartu Ülikooli intellektuaalse omandi õiguse professorit Aleksei Kellit koostöölepingute koostamise eest.

[16] Täname EKI leksikograafi Madis Jürvistet, kes suhtles kirjastustega ning klassifitseeris nii uued, 2021. a lõpus EKI-le saadetud teosed kui ka varasemalt koondkorpusesse kuulunud ilukirjandusteosed. Samuti täname EKI keeletehnoloogi Helen Kaljumäed kirjanduse korpuse kokkupanemise eest.

[17] Seni kasutusel olnud sõnavisandite grammatika versiooni 1.5 on kirjeldanud Kallas (2013).

(Suchomel 2020), diakrooniline analüüs ja trendid (*Trends*) (Herman 2019) ning sõnavektorid (*Word Embeddings*).

3.1. Terminituvastus

Terminituvastuseks vajaliku terminigrammatika (*TermGrammar*, loe lähemalt Baisa jt 2017) loomisel võeti aluseks IATE terminibaasi eestikeelsete terminite morfosüntaktilised struktuurid.[18] Terminigrammatika, milles on reeglitega kirjeldatud potentsiaalsete terminite pikkus ja struktuur, tuvastab 1–5-sõnalisi termineid, mille morfosüntaktiline struktuur on IATE-s vähemalt 55 eestikeelsel terminil (ehk vähemalt 0,15%-l kõikidest sealsetest eestikeelsetest terminitest). Joonisel 8 on näha esimesed 10 ühesõnalist, kahesõnalist ja kolme- või enamasõnalist terminit ÜK 2021 terviseteemalise allkorpuse põhjal.

	Word		Word		Word
1	arst	1	tartu ülikool	1	tartu ülikooli kliinikum
2	võima	2	kogu aeg	2	tervise arengu instituut
3	olema	3	ülikooli kliinikum	3	erakorralise meditsiini osakond
4	patsient	4	viimane aeg	4	väga hea arst
5	laps	5	vaimne tervis	5	päevaseks tarbimiseks soovitatav kogus
6	ravi	6	eelmine aasta	6	tartu ülikooli arstiteaduskond
7	inimene	7	tervise areng	7	läbi viidud uuring
8	haigus	8	pikk aeg	8	vaimse tervise probleem
9	tere	9	kuu aeg	9	vaimse tervise keskus
10	uuring	10	ida-tallinna keskhaigla	10	ülemiste hingamisteede viirusnakkus

Joonis 8. Terminituvastus Sketch Engine'is

Terminituvastuseks on lisaks suurele üldkeelekorpusele vaja ka valdkonnakorpuseid. Eestis tegeleb valdkonnakorpuste loomisega peamiselt Eesti Keele Instituudi keeletehnoloogia kompetentsikeskus. 2022. a juuni seisuga on loodud nt militaar-, õigus-, rahvatervise- ja kriisikorpused. Samuti võimaldab ÜK 2021 suur teemade hulk luua Sketch Engine'is alamkorpuseid, mille najal terminituvastust teostada.

3.2. Kollokatsioonide žanrilise ja temaatilise kuuluvuse tuvastus sõnavisandites

Žanreid ja teemasid kuvatakse Sketch Engine'i sõnavisandites kollokatsiooni juures siis, kui see teatud žanris ja/või teemas võrreldes teiste žanride ja/või teemadega esildub (vt joonis 9).

[18] Terminigrammatika autorid on EKI keeletehnoloog Eleri Aedmaa ja Lexical Computing Ltd. tarkvaraarendaja Marek Blahuš.

Joonis 9. Teemad ja žanrid Sketch Engine'i sõnavisandites

Joonisel 9 on näha, et kollokatsioonid *mune vahustama/kloppima/värvima* esilduvad toidu ja joogi temaatikas ning *mune munema* loomade ning looduse ja keskkonna temaatikas. Osundus *especially* 'eriti' viitab sellele, et selle konkreetse kollokatsiooni suhteline sagedus teatud teemavaldkonnas (või žanris) on vähemalt kaks korda kõrgem kui selle suhteline sagedus kogu korpuses. Osundus *usually* 'tavaliselt' (kasutusel võib olla ka *always* 'alati') viitab sellele, et 70–97% kollokatsiooni esinemise juhtudest esinevad konkreetses teemavaldkonnas (või žanris).

3.3. Eesti keele diakrooniline analüüs ja trendid

Sketch Engine'i funktsioon *Trends* analüüsib keelendi kasutust teatud ajaühiku (kuu, kvartal, aasta) jooksul, võrreldes selle kasutussagedust sama korpuse sees teiste samaväärsete ajavahemike lõikes. Tulemuseks on loend sõnadest, mille kasutussagedus on aja jooksul muutunud. Kasutuse suurenedes on trend positiivne (roheline nool suunaga üles); kasutuse langedes negatiivne (punane nool suunaga alla) (vt joonis 10). Diakroonilist analüüsi saab teha kahe eraldiseisva reaalajas täieneva korpuse (RSS uudisvoo korpus ja JSI uudisvoo korpus, vt ptk 2.2) peal. Lisaks saab seda teha ÜK 2021 uudisvoogude allkorpuse peal, kuid selle tulemus on staatiline ehk 2021. aasta lõpu seisuga.[19]

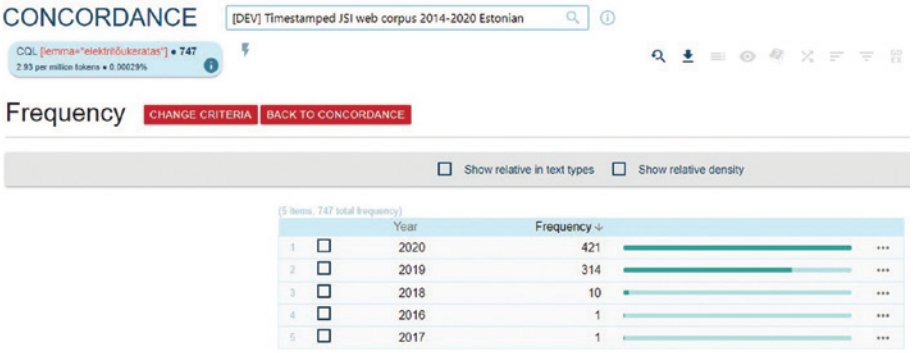
[19] Kuigi eesti keele koondkorpuse ajakirjandustekstide ilmumise aastad (1995–2008) on samuti teada, siis diakroonilise analüüsi funktsioon *Trends* sellele allkorpusele ei laiene. Ajamärgised on ÜK 2021-s küljes vaid RSS ja JSI uudisvoo korpuse tekstidel.

	Lemma	Trend ↓		Frequency	Sample
1	elektritõukeratas	↗	3.49	747	
2	podcas	↗	3.27	183	
3	eaklass	↗	3.27	86	
4	tootmiskvoot	↘	-3.08	83	
5	mõjuisik	↗	3.08	104	
6	ruumiloome	↗	3.08	83	
7	taskuhäälingusaade	↗	3.08	102	
8	metsasõda	↗	3.08	86	
9	podcasti	↗	2.90	611	
10	õngitsuskiri	↗	2.90	202	

	Lemma	Trend ↓		Frequency	Sample
11	leptospiroos	↗	2.90	90	
12	äpitakso	↗	2.90	108	
13	vallaarhitekt	↗	2.90	216	
14	läänerand	↗	2.90	366	
15	voogedastusplatvorm	↗	2.90	387	
16	roya	↗	2.90	90	
17	jalaväekompanii	↘	-2.75	225	
18	arengufond	↘	-2.75	280	
19	lisavaheaeg	↘	-2.75	94	
20	välisluureamet	↗	2.75	398	

Joonis 10. Positiivsed ja negatiivsed trendid aastate lõikes JSI uudisvoo korpuses

Jooniselt 10 selgub, et aastate 2014–2020 lõikes on JSI uudisvoo korpuses kõige suurema tõusva trendiga sõnad olnud *elektritõukeratas*, *podcast*[20], *eaklass*, *mõjuisik* ja *ruumiloome* ning kõige suurema langusega sõnad *tootmiskvoot*, *jalaväekompanii*, *arengufond* ja *lisavaheaeg*. Sõna järel asuvale kolmele punktile klikkides saab omakorda vaadata sõna kasutuse täpsemat jaotust aastate lõikes: jooniselt 11 on näha, et sõna *elektritõukeratas* hakati plahvatuslikult kasutama 2019. aastal. Selle funktsiooni abil saavad sõnavarauurijad ja keeleajaloo uurijad tuvastada uusi sõnu ehk neologisme ning analüüsida, millal teatud sõna kasutama hakati, millal selle kasutamine lõpetati või millisel ajaperioodil teatud sõna kasutus eriliselt tõusis või langes.

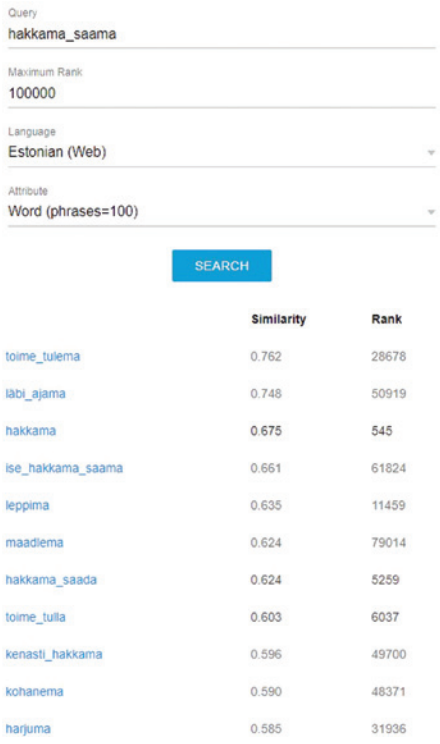


Joonis 11. Lemma *elektritõukeratas* kasutussageduse jaotus aastate lõikes JSI uudisvoo korpuses

3.4. Sõnavektorid

Leksikaalsemantiliste seoste (sünonüümid, antonüümid, hüpo- ja hüperonüümid) tuvastamiseks on lisaks tesaurusele eraldi tööriist Sõna vektoresitus (*Embedding Viewer*) (Bojanowski jt 2017, Herman 2021). Eesti keelt töötleb tööriist esialgu veebikorpuse 2017 andmete pealt. Alates 2021. aastast genereerib programm sõna vektoresitusi ehk sarnaste sõnade loendeid nii üksikute sõnade kui ka mitmesõnaliste üksuste jaoks. Mitmesõnalistele üksustele leksikaalsemantiliste seoste otsimiseks tuleb otsisõna kirjutada allkriipsuga, nt *hakkama_saama*, *must_auk* (vt joonis 12).

[20] Joonisel 7 on lemmatiseerimise viga (podcas *vs.* podcast).



Joonis 12. Verbi *hakkama saama* tesaurus

Jooniselt 12 nähtub, et *hakkama saama* kõrgeima sarnasusindeksiga (*similarity*) sõnad (st et esinevad sageli sarnases kontekstis) on *toime tulema* ja *läbi ajama*.

4. Kokkuvõte ja edasiarendused

Kirjeldasime artiklis eesti keele ühendkorpuste sarja tekkimisloo ja kogumise põhimõtteid. Nüüdseks on sarjas valminud neli versiooni: eesti keele ühendkorpus 2013, 2017, 2019 ja 2021.

EKI kogub veebikorpusi iga kahe aasta tagant ning koos sellega valmivad ka ühendkorpuse järgmised versioonid. Ühendkorpuste sarja seni viimase versiooni (2021) maht on 2,9 mld sõnet ja see sisaldab ühteteist allkorpust: veebikorpused 2013, 2017, 2019, 2021, uudisvood 2014–2021, Vikipeedia,Vikipeedia arutelud 2017, avatud lähtekoodiga teadusartiklid, kirjandus, koondkorpus ja tasakaalus korpus. Täiendavalt on kogu korpuse sisu klassifitseeritud žanridesse (blogid, foorumid, perioodika, veebikaubandus, entsüklopeedia, akadeemiline kirjutamine, kirjandus) ja teemadesse (kokku 24 teemat, nt loomad, teadus, sport, religioon). Eri tüüpi allkorpused ning nende klassifitseerimine võimaldavad eesti keelt mitmekülgselt analüüsida. Eesti keele ühendkorpus katab perioodi alates 1990-ndatest kuni tänapäevani, mis tähendab, et tänapäeva eesti keelt on võimalik uurida umbes 30 aasta lõikes.

Kuna veebikorpused moodustavad ühendkorpuste sarjas valdava osa (eesti keele ühendkorpuse 2021 mahust koguni 91%), kirjeldasime artiklis veebikorpuste kogumise, järeltöötluse ja puhastamise põhimõtteid ning nende protsessidega kaasnevaid probleeme. Kuigi veebikorpusi on kritiseeritud selle poolest, et neis pole tagatud lingvistiline variatiivsus (st et kaetud pole teatud žanrid, nt ilukirjandus ja suuline keel), kaalub nende suurus võimalikud puudused üle (Cvrček jt 2020): mahukast korpusest on võimalik kasutusnäiteid leida ka väga madala sagedusega keelenähtuste kohta (Pomikálek jt 2009), ning see on jätkuvalt odavaim ja lihtsaim viis suuri keeleandmeid koguda. Veebikorpuste kroolimise sagedad probleemid on keele identifitseerimine ja sarnaste keelte eristamine, tekstikodeeringu tuvastamine, tekstide (HTML-ist) puhastamine, duplikaatide eemaldamine, autorsuse ja loomevarguse tuvastamine ning masintõlkelised tekstid ja veebispämm. Eestikeelsete veebikorpuste kroolimisel on osutunud kõige tõsisemaks probleemiks masintõlkelised veebilehed, mille kõrvaldamine eeldas kroolitud veebikorpuse URL-ide käsitsi kontrollimist.

Koos eri tüüpi korpuste loomise meetodite arenguga on viimase aastakümne jooksul jõudsalt arenenud ka korpusleksikograafilised analüüsimeetodid. Korpusmaterjali kasutatakse märksõnastiku genereerimiseks, sõnade kollokatiivse ja süntaktilise käitumise uurimiseks, leksikaalsemantiliste suhete ja näitelausete (automaat)tuvastamiseks. Korpuspäringusüsteemi Sketch Engine näitel kirjeldasime artiklis uusi analüüsivõimalusi: terminituvastust, kollokatsioonide žanrilise ja temaatilise kuuluvuse tuvastust sõnavisandites, eesti keele diakroonilist analüüsi ja trende ning sõnavektoreid.

Korpusalane töö EKI-s jätkub. Ühendkorpuste sarja järgmise versiooniga loodame pakkuda keeleuurijatele ja leksikograafidele senisest veelgi suurema variatiivsusega keeleandmeid. Ühendkorpus esindab tänapäeva kirjalikku keelt, edaspidi oleks kindlasti vaja pöörata rohkem tähelepanu ka suulisele keelele. Oskuskeele seisukohalt on vaja kaasata rohkem valdkonnakorpusi ning diakroonilise uurimise seisukohalt möödunud sajandil kirjutatud tekste. Samuti jätkub EKI koostöö kirjastustega, mille tulemusena loodame tunduvalt suurendada kirjanduse allkorpuse mahtu.

Uued võimalused keelt uurida tekivad ka koos uute märgenduskihtide lisamisega (eelkõige semantiline märgendus). See võimaldab teha komplekssemaid päringuid ning analüüsida sõnade ja konstruktsioonide leksikaalsemantilisi omadusi. Eesti keele korpusleksikograafilise analüüsi seisukohalt on päevakorras jätkuvalt definitsioonide automaattuvastus, süntaktiliste mallide tuvastamine (*syntactic patterns*) ja vastete tuvastamine rööpkorpustest.

Viidatud kirjandus

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

Baisa, Vít; Michelfeit, Jan; Matuška, Ondřej 2017. Simplifying terminology extraction: OneClick Terms. – The 9th International Corpus Linguistics Conference. University of Birmingham.

Bojanowski, Piotr; Grave, Edouard; Joulin, Armand; Mikolov, Tomas 2017. Enriching word vectors with subword information. – Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

Bušta, Jan; Herman, Ondřej; Jakubíček, Miloš; Krek, Simon; Novak, Blaž 2017. JSI Newsfeed Corpus. – The 9th International Corpus Linguistics Conference. University of Birmingham.

Cartier, Emmanuel 2017. Neoveille, a web platform for neologism tracking. – Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia: Association for Computational Linguistics, 95–98. https://doi.org/10.18653/v1/E17-3024

Cho, Junghoo; Garcia-Molina, Hector 1999. The Evolution of the Web and Implications for an Incremental Crawler. Stanford InfoLab.

Clear, Jeremy 1987. Trawling the language: Monitor corpora. – ZURILEX Proceedings. Tübingen: Francke.

Cvrček, Václav; Komrsková, Zuzana; Lukeš, David; Poukarová, Petra; Řehořková, Anna; Zasina, Adrian Jan; Benko, Vladimír 2020. Comparing web-crawled and traditional corpora. – Language Resources and Evaluation, 54 (3), 713–745. https://doi.org/10.1007/s10579-020-09487-4

Fetterly, Dennis; Craswell, Nick; Vinay, Vishwa 2009. The Impact of Crawl Policy on Web Search Effectiveness. – Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 580–587. https://doi.org/10.1145/1571941.1572041

Gatto, Maristella 2014. Web as Corpus: Theory and Practice. A&C Black.

Herman, Ondřej 2019. Automatic Detection of Word Sense Shift. Ph.D. Thesis Proposal. Masaryk University Faculty of Informatics.

Herman, Ondřej 2021. Precomputed word embeddings for 15+ languages. – RASLAN 2021: Recent Advances in Slavonic Natural Language Processing, 41–46.

Jakubíček, Miloš; Kilgarriff, Adam; Kovář, Vojtěch; Rychlý, Pavel; Suchomel, Vít 2013. The TenTen Corpus Family. – CL 2013: 7th International Corpus Linguistics Conference. Lancaster, 125–127. http://ucrel.lancs.ac.uk/cl2013

Jakubíček, Miloš; Kovář, Vojtěch; Rychlý, Pavel; Suchomel, Vít 2020. Current challenges in web corpus building. – Proceedings of the 12th Web as Corpus Workshop. Marseille: European Language Resources Association, 1–4.

Kaalep, Heiki-Jaan; Vaino, Tarmo 2001. Complete morphological analysis in the linguist's toolbox. – Congressus Nonus Internationalis Fenno-Ugristarum. Pars V. Dissertationes sectionum: linguistica. II. Tartu, 9–16.

Kallas, Jelena 2013. Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias ['Syntagmatic Relations of Estonian Content Words in Corpus and Pedagogical Lexicography']. Dissertations on Humanities 32. Tallinn: Tallinna Ülikool.

Kallas, Jelena; Koppel, Kristina; Pool, Raili; Tsepelina, Katrin; Üksik, Tiiu; Alp, Pilvi; Epner, Anu 2021. Eesti keele kui teise keele õpetaja tööriistad Eesti Keele Instituudi keeleportaalis Sõnaveeb ['Estonian as a Second Language Teacher's Tools in the Institute of Estonian Language's Language Portal Sõnaveeb']. – Eesti Rakenduslingvistika Ühingu aastaraamat, 17, 61–80. https://doi.org/10.5128/ERYa17.04

Kallas, Jelena; Koppel, Kristina; Tuulik, Maria 2015. Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel ['New possibilities in corpus lexicography based on the example of the Estonian Collocations Dictionary']. – Eesti Rakenduslingvistika Ühingu aastaraamat, 11, 75–94. https://doi.org/10.5128/ERYa11.05

Kallas, Jelena; Suchomel, Vít; Khokholova, Maria 2017. Automated identification of domain preferences of collocations. – Electronic Lexicography in the 21st Century. Proceedings of eLex 2017 conference, 309–320.

Kallas, Jelena; Tuulik, Maria; Jürviste, Madis 2012. Leksikograafilise tarkvara Sketch Engine eesti keele moodul. – ESUKA/JEFUL, 3 (2), 57–77. https://doi.org/10.12697/jeful.2012.3.2.03

Kilgarriff, Adam 2012. Getting to know your corpus. – Text, Speech and Dialogue: Proceedings of the 15th International Conference, TSD 2012, Brno, Czech Republic, September 3–7. Springer, 3–15.
Kilgarriff, Adam; Rychlý, Pavel; Smrz, Pavel; Tugwell, David 2004. The Sketch Engine. – Proceedings of the 11th EURALEX International Congress, 105–115.
Kilgarriff, Adam; Baisa, Vít; Bušta, Jan; Jakubíček, Miloš; Kovář, Vojtěch; Michelfeit, Jan; Rychlý, Pavel; Suchomel, Vít 2014. The Sketch Engine: Ten years on. – Lexicography, 1, 7–36. https://doi.org/10.1007/s40607-014-0009-9
Koppel, Kristina 2020. Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele ['Corpus-based Automatic Detection of Example Sentences for Dictionaries for Estonian Learners']. Dissertationes linguisticae Universitatis Tartuensis 38. Tartu: Tartu Ülikooli Kirjastus. https://dspace.ut.ee/handle/10062/67138
Koppel, Kristina; Kallas, Jelena; Khokhlova, Maria; Suchomel, Vít; Baisa, Vít; Michelfeit, Jan 2019. SkELL Corpora as a part of the language portal Sõnaveeb: Problems and perspectives. – Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference, 1–3 October 2019. Sintra, Portugal, 763–782.
Laur, Sven; Orasmaa, Siim; Särg, Dage; Tammo, Paul 2020. EstNLTK 1.6: Remastered Estonian NLP Pipeline. – Proceedings of The 12th Language Resources and Evaluation Conference, 7152–7160.
Leech, Geoffrey 1992. 100 Million Words of English: The British National Corpus (BNC). https://www.semanticscholar.org/paper/100-Million-Words-of-English%3AThe-British-National-Leech/74795ed7f04541e6df82eab47158d5d3c88a1aad (17.1.2022).
Manning, Christopher D.; Schütze, Hinrich; Raghavan, Prabhakar 2008. Introduction to Information Retrieval. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511809071
Pomikálek, Jan 2011. Removing Boilerplate and Duplicate Content from Web Corpora. Masarykova univerzita, Fakulta informatiky. https://is.muni.cz/th/o6om2/ (18.11.2021).
Pomikálek, Jan; Jakubíček, Miloš; Rychlỳ, Pavel 2012. Building a 70 billion word corpus of English from ClueWeb. – LREC'12: Proceedings of the 8th International Conference on Language Resources and Evaluation, 502–506.
Pomikálek, Jan; Rychlỳ, Pavel; Kilgarriff, Adam 2009. Scaling to billion-plus word corpora. – Advances in Computational Linguistics, 41 (3), 3–13.
Suchomel, Vít 2020. Better Web Corpora for Corpus Linguistics and NLP. Masaryk University, Faculty of Informatics. https://is.muni.cz/th/u4rmz/?lang=en (17.1.2022).
Suchomel, Vít; Kraus, Jan 2021. Website properties in relation to the quality of text extracted for web corpora. – RASLAN 2021: Proceedings of Recent Advances in Slavonic Natural Language Processing, 167–175.
Suchomel, Vít; Pomikálek, Jan 2012. Efficient web crawling for large text corpora. – WAC7: Proceedings of the 7th Web as Corpus Workshop, 39–43.
Trampuš, Mitja; Novak, Blaž 2012. Internals of an aggregated web news feed. – Proceedings of 15th Multiconference on Information Society, 221–224.
Vaik, Kristiina; Muischnek, Kadri 2018. Eestikeelsete veebitekstide automaatne liigitamine ['Classifying Estonian web texts']. – Eesti Rakenduslingvistika Ühingu aastaraamat, 14, 215–227. https://doi.org/10.5128/ERYa14.13

Võrguviited

Autoriõiguse seadus. riigiteataja.ee/akt/128122011005?leiaKehtiv (20.3.2022).
DOAJ = Directory of Open Access Journal. doaj.org (20.3.2022).
Eesti keele koondkorpus. cl.ut.ee/korpused/segakorpus/index.php?lang=et (20.3.2022).
Eesti keele ühendkorpus 2017. doi.org/10.15155/3-00-0000-0000-0000-071E7L (20.3.2022).

Eesti keele ühendkorpus 2019. doi.org/10.15155/3-00-0000-0000-0000-08565 (20.3.2022).
Eesti keele ühendkorpus 2021. doi.org/10.15155/3-00-0000-0000-0000-08D17L (20.3.2022).
Eesti vana kirjakeele korpus. vakk.ut.ee (20.3.2022).
EKI piiblikonkordants. eki.ee/piibel (20.3.2022).
Embedding Viewer. embeddings.sketchengine.eu/ (20.3.2022).
Entu. entu.keeleressursid.ee (20.3.2022).
estNLTK teegid. nbviewer.org/github/estnltk/estnltk/tree/version_1.6/tutorials/nlp_pipeline/ (20.3.2022).
etTenTen13. doi.org/10.15155/1-00-0000-0000-0000-0012EL (20.3.2022).
IATE. iate.europa.eu/home (20.3.2022).
Jožef Stefan Institute. ijs.si/ijsw/JSI (20.3.2022).
OneClick Terms. terms.sketchengine.eu (20.3.2022).
Sketch Engine. sketchengine.eu (20.3.2022).
Stanza mudelid. entu.keeleressursid.ee/public-document/entity-9862/2021-05-29 (20.3.2022).
Sõltuvussüntaktiline märgendusskeem. github.com/EstSyntax/EstUD/blob/master/Eesti-UDdokumentatsioon.pdf (20.3.2022).
Tasakaalus korpus. cl.ut.ee/korpused/grammatikakorpus (20.3.2022).
TÜ ilukirjanduskorpus. cl.ut.ee/korpused/grammatikakorpus/ilukirjeldus (20.3.2022).
Vabamorf. github.com/Filosoft/vabamorf (20.3.2022).
Vikipeedia artikkel. et.wikipedia.org/wiki/Koer (20.3.2022).
Vikipeedia Talki artikkel. et.wikipedia.org/wiki/Arutelu:Koer (20.3.2022).

ESTONIAN NATIONAL CORPUS 2013–2021: THE LARGEST COLLECTION OF ESTONIAN LANGUAGE DATA

Kristina Koppel, Jelena Kallas
Institute of the Estonian Language

The paper describes the Estonian National Corpus 2021 (Estonian NC 2021), the latest and the largest edition in the Estonian National Corpora series. The entire series of Estonian NC consists of four corpora: Estonian NC 2013, 2017, 2019 and 2021. The series was compiled by the Institute of the Estonian Language in cooperation with the software company Lexical Computing Ltd. All corpora are accessible through the Sketch Engine interface, a corpus query system developed and maintained by Lexical Computing Ltd. The data are also stored in the repository Entu at Center of Estonian Language Resources.

The Estonian National Corpus 2021 contains eleven sub-corpora (i.e. Web 2013, Web 2017, Web 2019, Web 2021, Feeds 2014-2021, Wikipedia 2021, Wikipedia Talk 2017, the Open Access Journals (DOAJ), Literature, the Balanced Corpus, and the Reference Corpus) totalling 2.4 billion words. In addition, the corpus is divided into genres and topics.

The most extensive part of the Estonian NC 2021 is the Estonian Web Corpora, i.e. texts crawled from the web. In the paper, we outline the process of crawling the web, the process of cleaning and post-processing the crawled data, and the methodology for classifying web texts into genres and topics. We also introduce new tools for the analysis of corpus data in Sketch Engine, and suggest further perspectives and needs for corpus development.

Keywords: Estonian National Corpus, corpora, corpus lexicography, corpus query system, Estonian

Kristina Koppel (Eesti Keele Instituut) on EKI ühendsõnastiku töörühma liige, sünonüümide infokihi töörühma juht ja eesti-inglise toorsõnastiku projekti juht. Põhilised uurimisvaldkonnad: korpuslingvistika, korpusleksikograafia.
Roosikrantsi 6, 10119 Tallinn, Estonia
kristina.koppel@eki.ee

Jelena Kallas (Eesti Keele Instituut) on EKI ühendsõnastiku töörühma liige, vene keele infokihi töörühma juht, portaali Keeleõppija Sõnaveeb projekti juht ja projekti "Õpetaja tööriistad" juht. Põhilised uurimisvaldkonnad: korpuslingvistika, korpusleksikograafia, õppeleksikograafia.
Roosikrantsi 6, 10119 Tallinn, Estonia
jelena.kallas@eki.ee