https://doi.org/10.54013/kk764a6

Sõnastikukogust keeleportaaliks

MARGIT LANGEMETS, KRISTINA KOPPEL, JELENA KALLAS,
ARVI TAVAST

XXI sajandil on traditsiooniline sõnaraamatutöö oluliselt muutunud, tuues kaasa nii hõlbustusi kui ka katsumusi sõnaraamatu koostajale ja kasutajalegi. Siin artiklis keskendume digipöördele, mis sai Eesti Keele Instituudis (EKI) alguse 2017. aastal uue sõnastikusüsteemi Ekilex loomisega. Ent seda ei saa sugugi pidada sõnaraamatutöö ainsaks digipöördeks. Esimeseks digipöördeks võiks paradoksaalsel moel lugeda õnnelikku juhust, et 1960-ndatel alustatud „Eesti kirjakeele seletussõnaraamat" (EKSS) ei jõudnud ilmuma hakata varem kui 1980-ndate lõpul – ajal, mil instituudis olid arvutid juba kasutusel. Nõnda sattusime kohe n-ö arenenud leksikograafiamaade hulka, kel juba 1990/2000-ndatel oli elektroonilisel kujul olemas oma keele tähtsaim ja mahukaim sõnakogu (enam-vähem struktureeritud tekstina). Lihtsama ülesehitusega 1976. aasta õigekeelsussõnaraamat oli samuti arvutis olemas. Digipöördeks võib pidada ka oma sõnastikusüsteemi EELex (Langemets jt 2010; Jürviste jt 2011) ehitamist alates 1990-ndate lõpust. Kasutuskõlblikuks sai EELex 2003. aastal esimesena eesti-vene sõnaraamatu (EVS) jaoks, misjärel süsteemi laiendati ja arendati paljude sõnastike tarvis kuni 2015. aastani. Oma hiilgeaegadel oli EELexis ligi 70 valminud või töös olevat sõnastikubaasi. Ühtlasi oli tolleks ajaks saanud selgeks, et samal viisil jätkamine ei ole jätkusuutlik: internet oli plahvatuslikult levinud, tarkvara ja andmeformaadid olid muutunud, leksikograafia kogu Euroopas ja maailmas oli tugevasti oma suunda muutmas.

Sarnaste probleemidega seisid silmitsi ka teiste Euroopa riikide leksikograafid: suurem osa leksikaalsest teadmusest oli talletatud vaid paberile, sõnastikud olid suures osas struktureerimata, nad ei olnud masinloetavad ega sobinud seetõttu ei loomuliku keele töötluse arendamiseks ega muude keeleandmetel põhinevate rakenduste jaoks. Nagu Eestis, nii ka mujal Euroopas alustati hoogsalt järgmise põlvkonna sõnastikusüsteemide ehitamist, millega kaasnes laiaulatuslik pabersõnastike digiteerimine. Algul kasutas enamik sõnastikubaase XML-vormingut, kuid varsti sai selgeks, et hädavajalik on välja töötada andmeformaat, mis võimaldaks keeleinfot kirjeldada ja andmeid vahetada (nii eri keelte kui ka andmebaaside vahel). Muutunud olud tõukasid taas tagant uut digipööret, mille tunnussõna on andmete ühendamine ja linkimine, mis sõnaraamatumaailmas tähendab leksikaalse info koondamist.

Värskelt ilmunud elektroonilise leksikograafia valge raamatu „The Future of Academic Lexicography" (Steurs jt 2020) soovitus riiklikele leksikograafiaasutustele on mitte jätkata üksikute sõnastike koostamist, vaid ehitada üleriigiline taristu – ühtne andmebaas, mis koondab seni koostatud ressursse ning mida saab linkida teistega nii oma riigi sees kui ka rahvusvaheliselt. 35 Euroopa riiki ühendav Horisont 2020 projekti „Euroopa leksikograafiataristu" (ELEXIS) eesmärk on ühendada linkandmete

formaadis eri keelte sõnastikud ja keeleandmed ning pakkuda kõigile sõnaraamatutööga seotud asutustele uusi tööriistu ja teenuseid.[1] Sellele on pühendunud mitmed Euroopa koostööprojektid, kus osaleb ka Eesti Keele Instituut. Seni on loodud pilvepõhine PDF- ja XML-vormingute konverteerimisteenus Elexifier, RDF-põhine linkimistööriist NAISC ja korpuspäringusüsteemiga Sketch Engine (vt ka Kilgarriff jt 2004; Kilgarriff jt 2014) ühendatud sõnastikusüsteem Lexonomy (vt ka Měchura 2017).

Järgnevas puudutame samu võtmeküsimusi, mida eespool mainisime: liikumist ühtse andmebaasi poole, andmete linkimist ja üldist sõnaraamatutöö muutumist. EKI-s loodud sõnastikusüsteemi Ekilex ja keeleportaali Sõnaveeb oleme tutvustanud Euroopa konverentsidel (Tavast jt 2018; Koppel jt 2019a; Tavast jt 2020), siin kirjeldame ka uusimat EKI sõnakogu – EKI ühendsõnastikku – ja selle koostamisel kasutatud Ekilexi andmemudelit. Sõnaraamatutöö on pöördeliselt muutunud mitmes mõttes: tehniliselt (ühtne sõnastikusüsteem ja andmebaas, arenenud korpuspäringusüsteemid), leksikograafiliselt (sõnastike ühendamine ja koostamispõhimõtete ühtlustamine, töö infokihtidega, andmete dünaamilisus) ja töökorralduslikult (kollektiivsuse kasv, avatuse suurenemine). Arutame nende asjaolude üle lähemalt.

1. Kolm sõnastikusüsteemi, sadakond sõnakogu

Ajalooliselt esimene sõnastikusüsteem oli Eesti Keele Instituudis EELex (kasutusel alates 2003. aastast), selle kõrvale lisandusid teised ning ühel hetkel kasutasime üht sõnastikusüsteemi ja kaht terminibaasi, igal oma head ja vead. EELex oli mõeldud üldkeele sõnaraamatute jaoks, terminoloogid kasutasid terminibaasisüsteeme Termeki (pole enam kasutusel) ja MultiTerm.

EELexi andmemudel (Langemets jt 2010; Jürviste jt 2011) on üles ehitatud sõnast lähtuvalt, esindades nii semasioloogilist mudelit. Töörühmad olid (ja on) autonoomsed, igaüks töötas oma soovide järgi kujundatud sõnastikubaasis. Tulemus oli, et kõik baasid (kokku umbes 70) on „natuke struktureeritud", igaüks ise viisil, kõigis on mingis ulatuses korduvat infot (nt märksõnad, morfoloogiline info, tähenduste seletused), mis samuti on esitatud mitmel eri moel; uued sõnad lisandusid igasse üksikusse baasi eraldi. Algusaegade XML-andmebaas kasvas teatavaks XML-i ja relatsioonilise andmebaasi segamudeliks. Hiljem lisandus ka mõistepõhine ehk onomasioloogiline andmemudel, mis siiski väga jõudsat kasutust ei leidnud. Iga valminud sõnaraamat avaldati omaette veebilehel.

Termeki süsteemi arendas Werkdata Ltd aastatel 2007–2015, aastal 2012 omandas EKI õiguse seda Eesti terminitöö jaoks kasutada. Termekis töötasid eeskätt instituudivälised terminikomisjonid ja üksiktegijad. Termeki on relatsiooniline andmebaas osaliselt modifitseeritava onomasioloogilise andmemudeliga. Selles loodi umbes 40 terminibaasi ja üks kakskeelne üldkeelesõnastik (läti-eesti). Iga valminud sõnaraamat tehti kasutajale elektrooniliselt kättesaadavaks Termeki suhteliselt tehnilises keskkonnas ning omaette baasina.

[1] https://elex.is/tools-and-services (10. II 2021).

MultiTermi süsteemi kasutas kaks suurt terminibaasi: mitmekeelne Esterm ning sõjanduse ja riigikaitse termineid koondav Militerm[2]. MultiTerm on samuti relatsiooniline andmebaas osaliselt modifitseeritava onomasioloogilise andmemudeliga.

EELexi tüüpi süsteemi piiratust kogesime kõige teravamalt 2016. aastal, kui lõime poolautomaatselt (kolme olemasoleva baasi põhjal) eesti-soome sõnaraamatu baasi (vt Langemets jt 2017), milles eesti osa aluseks sai samal ajal töös olnud uus eesti keele sõnaraamat (EKS 2019, mõlemad sõnaraamatud ilmusid 2019). Rööbitine töö nende sõnaraamatutega sisaldas väga palju dubleerimist: kui eesti poolel mingi otsus tehti ja esitust muudeti, tuli sedasama käsitsi korrata ka eesti-soome baasis. Oli vaja kardinaalselt teistsugust, andmete vahetust ja taaskasutust võimaldavat, ühtset sõnastikusüsteemi.

Terminitöös seati ühtne üldkasutatav terminibaas eesmärgiks juba 2008. aastal algatatud riiklikus terminoloogiaprogrammis (Valge 2013). Kümme aastat hiljem polnud me päriselt sinna veel jõudnud: olid olemas „natuke ühtne" Termeki ja eraldiseisev Multiterm, aga ühtset baasi ei olnud. Füüsikud küsisid otse: „Kas on olemas plaan, kuidas lõpuks tekib üks päringuaken, kuhu sõna sisestades pöördutakse kõigi erialasõnastike poole?" (Kuusk, Reivelt 2019) Füüsikute arupärimise ajal oli keeleinfo ühendamine tegelikult EKI-s juba alanud. Üks päringuaken tekkis esimesena 2019. aastal üldkeeleinfo jaoks, aasta hiljem ka erialasõnastike jaoks. Sõnastikuinfo koduks sai uus keeleportaal Sõnaveeb.

2. Kogu keeleinfo ühes kohas

Sõnastikuinfo ühendamine algas EKI-s 2017. aastal, kui koostöös tarkvarafirmaga TripleDev OÜ alustasime EKI uue andmebaasi Ekilex loomist (Tavast jt 2018). Ühtse baasi selgrooks ehk tuumikuks kujundasime „Eesti keele sõnaraamatu 2019" (EKS 2019; Tavast jt 2018). Oleme järginud digimaailma põhimõtet, et andmed ja andmete esituskuju seisavad üksteisest lahus (vt ka Gorjanc jt 2017: 11; Langemets jt 2018; Tavast jt 2018).

Sõnastiku- ja terminibaasisüsteem Ekilex on leksikograafide ja terminoloogide profitööriist, mida kasutavad nii EKI enda kui ka majavälised sõnakogude koostajad ja toimetajad (kokku üle 300). Baasis tegutsejate arv kasvab pidevalt koos uute keelte lisamise ja uute terminibaaside loomisega. Eesti keelendeid on Ekilexi andmebaasis kokku ligi 300 000 (seisuga aprill 2021) ehk kaks korda rohkem kui omaaegses suurimas, eesti keele seletavas sõnaraamatus (EKSS 2009). Nii üldkeele- kui ka oskussõnastikes esinevaid keelendeid on 60 000. Üksnes terminibaasides esinevaid termineid on 120 000 (kokku on üle 80 terminibaasi).

Keeleportaal Sõnaveeb on koht, kus kasutajad saavad ühe päringuakna kaudu kogu infot uurida. Keeleõppijale on mõeldud vähema sõnavalikuga (umbes 6000 sõna) ja lihtsama ülesehitusega Keeleõppija Sõnaveeb. Sõnaveebi kasutab kuus üle 32 000 inimese (Google Analyticsi andmed, aprill 2021). 2019. ja 2020. aasta võrd-

[2] Alates märtsist 2021 jätkub Militermi koostamine Ekilexis.

luses paistab kasv püsiv ja ühtlane, mitte hüppeline: aasta jooksul on kasutajate arv (ühe kuu kohta) kasvanud 10 000 võrra.

EKI ühendsõnastik (Sõnaveebis) on eesti keele mahukaim sõnakogu, kuhu oleme koondanud mitmest EKI tänapäeva keele sõnastikust pärit info. Ajapikku hakkab seda täiendama ka õigekeelsussõnaraamat. Ühendsõnastiku tuumik on eesti keele sõnaraamat (EKS 2019), koos teiste sõnastikega nimetame seda alates 2020. aastast EKI ühendsõnastikuks. Sõnastike ühendamine üheks ühiseks andmebaasiks on keeruline protseduur, mida täielikult automatiseerida ei saa. Andmete ühendamisel oleme kasutanud spetsiaalselt loodud tähenduste ühendamise tööriista. Ühendamise rõõmustav pool on, et eri sõnakogudest korjub kokku väärtuslikke andmeid, näiteks sõnu ja tähendusi, mida andmebaasis veel pole registreeritud, samuti mitmekesist infot sünonüümide, rektsioonide jm kohta. Tülikam pool on, et leksikograafid peavad kõvasti vaeva nägema, et andmeid korrastada ning puhastada (eri sõnakogudest tulnud) liiasest infost, kordustest ja ebajärjekindlustest. Näiteks eesti-vene sõnaraamatus (EVS) on eesti keele tähendusjaotust osaliselt painutatud sihtkeele (vene keele) järgi. Kui eesti *nägu* tähistab nii inimese kui ka looma kehaosa, siis vene keeles jaguneb mõiste kaheks: inimesel on *лицо* ja loomal *морда*; sama on sõnaga *ämm*: eesti keeles on ühendmõiste 'mehe- või naiseema', aga vene keeles eraldi mõisted 'meheema' (*свекровь*) ja 'naiseema' (*тёща*). Nüüdset Ekilexi selgroogu hoiame kindlalt paigas eesti keele järgi (sealhulgas tähendused võivad muutuda), ent see nõuab vene keele osas ulatuslikku toimetamis- ja korrastustööd. Oleme väitnud enesekriitiliselt, et alahindasime impordi võimekust andmete harmoniseerimisel (Tavast jt 2018). Õppetunni saanuna ei impordi me edaspidi tervet järjekordset sõnakogu, vaid analüüsime põhjalikult, millist uut infot sellest kogust Ekilexi juurde tuua. Nii toimime näiteks alates 2020. aastast töös oleva õigekeelsussõnaraamatu ÕS info puhul, kus oleme alustanud ÕS-i soovituste ülevaatamist ja ühendsõnastikus esitamist (Langemets, Päll 2020).

Võib öelda, et oleme pikkade sammudega liikunud nn tõelise e-sõnaraamatu põhipunktide poole, mis on välja toodud ligi kümme aastat tagasi (Fuertes-Olivera, Bergenholtz 2011; Langemets 2012). Nüüd kinnitame ka meie, et moodne sõnaraamat 1) on info juurde pääsemise abivahend, infotehnoloogiline tööriist (info on päritav ka API ehk rakendusliidese kaudu); 2) on (relatsiooniline) andmebaas, kus iga andmeelement on ainukordne; 3) on allikas, millest erinevate linkide ja seoste abil luuakse (virtuaalne) sõnaraamat või valikuline väljavõte soovitud infost (vt ka Bajčetić, Declerck 2020); 4) arvestab andmete esitamisel kasutajate eri tasemete ja vajadustega. Kui enamik kasutajaid on ilma suuremate nõudmisteta tavainimesed, siis eksperdid soovivad võimalikult detailset infot. Näiteks saab Sõnaveebis morfoloogilist infot vaadelda kolmel tasemel: põhivormide ulatuses, täisparadigmana ja morfofonoloogilises transkriptsioonis (koos välte, rõhu jt märkidega) esitatud täisparadigmana.

Sõnaveebi tagasiside kaudu teame, et paljud kasutajad soovivad mugavamat võimalust kasutajaliidest enda käe järgi kohandada, infot detailsemalt otsida ning enda jaoks paremini organiseerida. See töö seisab meil alles ees.

3. Muutuv sõnaraamatutöö

Moodne aeg on leksikograafiasse ja leksikograafide jaoks kaasa toonud uusi vaatenurki, millest nii mõnigi on nõudnud harjumist, et muutus omaks võtta. Uut moodi on korraldatud sisuline töö (infokihid, korpuspõhisus, andmete dünaamilisus) ja omavaheline koostöö. Kõik on pidanud kohanema uue ühtse sõnastikusüsteemiga, aga ka aru saama andmemudelist, nii selle võimalustest kui ka piiridest.

3.1. Töö infokihiga, mitte sõnastikuga

Kui varem töötas iga töörühm oma sõnastikuga, siis nüüdse aja leksikograafid töötavad ühtses baasis, aga enamik neist keskendub teatavale andmekihile (vt lähemalt vt Tavast jt 2020), näiteks sünonüümidele, teiste keelte vastetele, ÕS-i soovitustele, sõnaliikidele (vt lähemalt Paulsen jt 2020), kohanimedele, teatava valdkonna terminitele jne. Samal ajal näevad leksikograafid sõna või mõiste kohta kogu infot ja vajadusel täiendavad seda. Uusi andmekihte on võimalik luua jooksvalt, vastavalt vajadusele. Leksikograafid on pidanud piltlikult öeldes hakkama kohe sõitma vahendil (Ekilex), mida samal ajal ise ehitame ja arendame.

EKI ühendsõnastiku ja ühtse baasi sünniga on leksikograafid pääsenud suurest hulgast korduvast tööst, mis varem oleks kuulunud vältimatult iga uue sõnastikuprojekti juurde: näiteks märksõnaloendi loomine, häälduse ja morfoloogilise info esitus, eesti keele tähendusjaotus.

Iga andmekiht toob mõistagi kaasa probleemid, millega tuleb eraldi tegelda. Nii oleme näiteks sünonüümikihi koostamisel muutnud andmemudelit mitu korda. Kui algul koostasime automaatselt genereeritud valiku najal sünonüüme ühepoolselt (a = b), siis alates 2020. aasta lõpust teeme seda kahesuunaliselt (kui a = b, siis b = a). Teisisõnu: kui algul sai *koer* käsitsi koostamise ajal sünonüümiks *peni*, siis vastupidi seost automaatselt ei tekkinud, aga nüüd tekib (Tavast jt 2020). Teiste keelte lisandudes tuleb lahendada keeltevaheliste tähenduserinevuste esitamise probleem: näiteks vaste võib olla eesti sõnast/terminist laiem või kitsam (nt eesti *ämm* jaguneb vene keeles kaheks mõisteks: *meheemaks* ja *naiseemaks*). Andmemudel võimaldab küll tähistada säärast tähendusseost, ent kasutajale me mõistlikul viisil tulemust veel esitada ei saa. See on seotud sellega, et tänapäeva leksikograafias seisavad andmebaas ja info esituskuju (kasutajaliides) üksteisest lahus, on tehniliselt eri asjad. Keeltevahelised tähendusseosed on plaanis lahendada lähitulevikus.

ÕS-i soovituste andmekihi ühendame välislinkide abil teiste õigekeelsusallikatega, eeskätt (samuti töös oleva) EKI teatmikuga ehk õigekeelsuskäsiraamatuga. Nii on võimalik kohe tutvuda põhjalikuma taustamaterjaliga.

3.2. Tekstikorpused ja deskriptiivne keelekirjeldus

Leksikograafilise kirjelduse põhiallikad on tänapäeva leksikograafias mahukad elektroonilised tekstikogud ehk tekstikorpused. Korpuste analüüsimiseks kasutatakse korpuspäringusüsteeme (nt Sketch Engine, KORP), mis võimaldavad arvandmetele

toetudes sõnade kasutust ja kasutuse muutumist mitmekülgselt analüüsida. Traditsiooniliselt on korpusi loodud hoolikalt valitud, kindla päritolu ja kvaliteediga allikatest: nii on loodud nt eesti keele koondkorpus.[3]

Alates 2000. aastatest on järjepidevalt kasvanud veebitekstide maht ning neid on hakatud kasutama veebikorpuste loomiseks (Grefenstette, Nioche 2000; Cavagliá, Kilgarriff 2001; Pomikálek jt 2012; Suchomel, Pomikálek 2012). Veebitekstid kogutakse kokku spetsiaalse programmiga (ingl *crawler*) SpiderLing (vt lähemalt Suchomel, Pomikálek 2012; Suchomel 2020), mis mööda veebilinke liikudes laeb alla sealse tekstilise materjali. Tekst kodeeritakse UTF-8 vormingusse ning puhastatakse (nt muukeelsetest tekstidest), samuti eemaldatakse identsed või väga sarnased dokumendid. Seejärel tekstid lemmatiseeritakse, märgendatakse morfoloogiliselt[4] ning laetakse üles korpuspäringusüsteemi (EKI-s Sketch Engine'isse)[5]. Veebikorpuste puhul kerkivad esile uued probleemid: näiteks on keeruline tuvastada või eristada üht keelt teisest sarnasest keelest (nt vene keelt ukraina või valgevene keelest, vt Koppel jt 2019b: 776), tuvastada duplikaate, veebispämmi ja masintõlkelisi tekste (vt lähemalt Suchomel 2020). Peale korpuse sisu (allikate valiku) ja mahu probleemide on sagedad ka lemmatiseerimise, morfoloogilise märgenduse ja (osa)lausestamise ning mitmesõnaliste üksuste tuvastamise vead, (vormi)homonüümia ja polüseemia (Koppel 2020: 56–63).

Kuigi veebikorpusi on kritiseeritud selle poolest, et neis pole tagatud lingvistiline variatiivsus (st kaetud pole teatud žanrid, nt ilukirjandus ja suuline keel), kaalub nende suurus puudused üle (Cvrček jt 2020): mahukast korpusest on võimalik kasutusnäiteid leida ka väga madala sagedusega keelenähtuste kohta (Pomikálek jt 2009).

Eesti keele jaoks pole kunagi varem olnud nii suuri (ja reaalajas täienevaid) tekstikorpusi, kui on viimase kümne aasta jooksul sündinud EKI ja tarkvarafirma Lexical Computing Ltd. koostöös. Suurim korpus – eesti keele ühendkorpus 2019 (1,5 mld sõna) – koosneb umbes 88% ulatuses just veebist kogutud tekstidest. Ühendkorpus on žanriliselt mitmekesiseim allikas tänapäeva eesti keele uurimiseks, mille allkorpustest saab omakorda eraldi päringuid teha.

Tekstikorpustest avaneb leksikograafile pilt tegelikust keelekasutusest. Mida automaatsemalt korpust analüüsida, seda deskriptiivsem (objektiivsem?) on leksikograafiline keelekirjeldus. Juba on võimalik genereerida märksõnaloendeid, tuvastada definitsioone, kollokaate, tähendusseoseid, näitelauseid ja muid üksusi (vt lähemalt

[3] Koondkorpus on loodud Tartu Ülikoolis ning see sisaldab ajakirjandus-, ilukirjandus- ja teadustekste, riigikogu stenogramme ning Eesti ja Euroopa seadustekste aastatest 1990–2008. https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et (31. I 2021).

[4] Eesti keele korpuste lemmatiseerimiseks ja märgendamiseks kasutatakse tekstianalüsaatorit EstNLTK. Uusimad (alates 2019. aastast loodud) korpused on märgendatud versiooniga 1.6, mille teek on kättesaadav aadressil https://www.keeletehnoloogia.ee/et/estnltk-1-6-teegi-stabiilne-versioon (31. I 2021).

[5] Märgendamata kujul on korpus allalaadimiseks kättesaadav CC-BY-SA litsentsi alusel Eesti Keeleressursside Keskuse repositooriumis Entu.

Koppel 2020: 11–14).[6] Korpusleksikograafia alusepanija John Sinclair (1991: 39, 60–61) on rõhutanud, et puhtobjektiivne keelekirjeldus *per se* pole piisavalt üldistatud: andmed luuakse küll automaatselt, ent korpusandmeid hindavad ja kokkuvõtte teevad ikka leksikograafid.

Järgmist õigekeelsussõnaraamatut koostades puutume kokku deskriptiivse (korpuspõhise) ja preskriptiivse (ÕS 2018-st pärit) keelekirjelduse vastuoludega. Seni paiknesid eri tüüpi sõnaraamatud (EKSS 2009 ja ÕS 2013; EKS 2019 ja ÕS 2018) üksteisest lahus, omaette veebilehel. Nüüd otsime lahendusi, kuidas kasutajale arusaadaval moel esitada ühendsõnastikus mõlemat: nii tegeliku keele kirjeldust kui ka keelekorralduslikke ÕS-i soovitusi. Oleme alustanud ÕS-i soovituste ülevaatamist teemavaldkondade kaupa. Ka Sinclair (1991: 60–61) on hoiatanud, et preskriptiivsed väited ei tohiks keeleandmeid ignoreerida või neist liiga eemalduda, muidu võivad ettekirjutused kaotada tõsiselt võetavuse.

3.3. Märgatav kollektiivsuse kasv, info dünaamilisus

Varem eraldi töötanud väiksemad eri sõnastike töörühmad (nt ÕS-i koostajad, seletava sõnaraamatu koostajad, naabersõnastiku koostajad) on nüüd koondunud üheks suureks EKI meeskonnaks, kes kõik tegutsevad alates 2020. aastast ühes andmebaasis. Iga sõnaraamat on enamasti olnud kollektiivse töö tulemus, mis toetub projekti jaoks loodud ja aeg-ajalt täienevale stiiliraamatule või koostamisjuhendile. Kõik leksikograafid teavad, kui keeruline on eri inimestel sarnast joont hoida, et subjektiivsust ei õnnestu täielikult vältida: igas kollektiivselt sündinud keelekogus on annus ebaühtlust. Ühendsõnastiku baasis, kus töötab üheskoos üle 40 inimese (neist suur hulk instituudiväliseid kaastöötajaid) ning mille maht kasvab pidevalt, on ühtsuse hoidmine tunduvalt keerulisem. Arvatavasti peame – eeskätt leksikograafid ise – edaspidi harjuma n-ö alatise teelolemisega, mõnikord ka asjaolude muutumisega teekonnal ideaalse (saavutamatu) lõpp-punkti suunas. Aga suur muutus on see ka kasutajate jaoks: polda harjutud, et töö käib kõigi silme all. Kui varem ilmus sõnastik siis, kui ta oli lõpuni viimistletud, käsikiri korrektuurini üle loetud, siis nüüd ilmutame uut infot jooksvalt: EKI ühendsõnastik uueneb iga päev, samuti kõik terminibaasid, kus parasjagu töö käib. Kasutaja kaaluda on, mis on parem: kas lõpliku kuju saanud staatiline sõnastik (mis hakkab vananema alates ilmumisele järgnevast päevast) või pidevalt uuenev, ühiskonnas aktuaalset keelekasutust registreeriv või keelemurele lahendust otsiv keeleportaal.

Koos rahvahanke algatustega muutub kogu sõnaraamatutöö aina laiapõhjalisemaks, selles osaleb üha rohkem inimesi: näiteks võib vabatahtlikke rakendada eri keelte vastete lisamisel, etteantud vastete või sünonüümide sobivuse hindamisel. Ses osas on Eestil arenguruumi küllaga. Seda tüüpi kaasamise pea ainus näide on 2019. aastal ilmunud eesti assotsiatsioonisõnastik (Vainik 2019), kus esitatavad seosed on kogutud testimise teel tavalistelt keelekasutajatelt.

[6] 2018. aasta üleeuroopalisest küsitlusest selgus, et Euroopas luuakse poolautomaatselt ligikaudu 31% ja täisautomaatselt umbes 7,5% leksikaalsetest andmebaasidest (Kallas jt 2019).

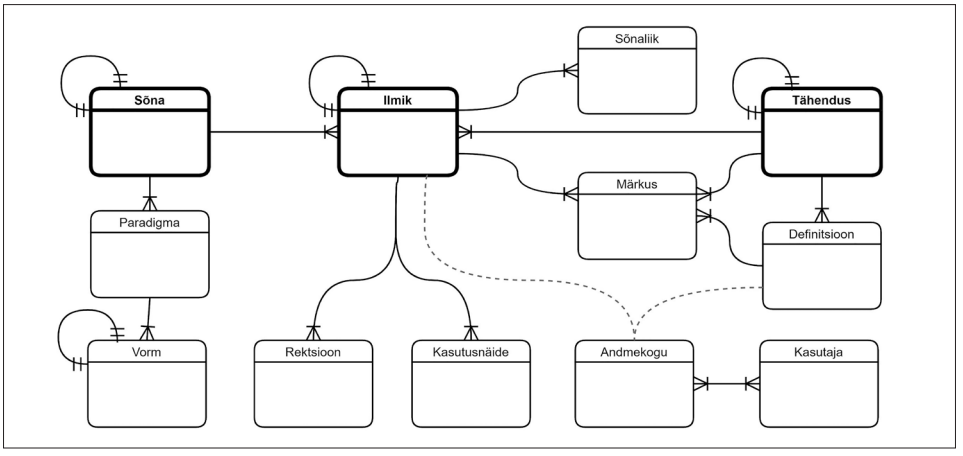
Digiajastu heade külgede ja vigade üle arutlenud Jaan Undusk ja Marek Tamm (2021) on osaluskultuuri ja vabatahtlikku koosloomet hinnanud digiajastu positiivseks küljeks – need on uued vormid ühistes platvormides, mis tihti aitavad asju kiiremini edendada. Spetsialistidele, sh leksikograafidele jääb järelkontrolli roll.

3.4. Infotehnoloogilist: andmemudel ja API

Leksikograafiliste andmete kirjeldamiseks on pakutud mitu erinevat mudelit, näiteks OntoLex-Lemon (McCrae jt 2017) ja selle leksikograafiline moodul Lexicog, TEI Lex-0 (Bański jt 2017; Tasovac jt 2018), LMF ISO-24613 (Francopoulo jt 2006), Eestis Ekilex (Tavast jt 2018).

Ekilexi andmemudel erineb varasematest sõnastikusüsteemidest (EELex, Termeki, Multiterm) seose tüübi poolest. Kolmes varasemas andmemudelis on vormi (sõna/keelendi) ja tähenduse vahel üks mitmele seos 1:n, st ühel sõnal on mitu tähendust (semasioloogilises mudelis) ja ühel mõistel on mitu terminit (onomasioloogilises mudelis). Nende andmemudelite puuduste kohta vt Tavast jt 2018. Ekilexi andmemudelis kasutame vormi (sõna/keelendi) ja tähenduse vahel mitu mitmele seost m:n, st ühel sõnal võib olla mitu tähendust ja üks tähendus (mõiste) võib olla seotud mitme sõnaga, ehk mõlemat pidi on lubatud palju seoseid. Kuna mitu mitmele seose puhul ei saa luua otseseost, siis seob neid vahetabel (vt joonist 1, vahetabeli rollis on ilmik, selle kohta allpool). Sellist seost võib kirjeldada ka kui sõna ja tähenduse andmete pidevat taaskasutamist: relatsioonilistes andmebaasides ühendavad olemeid mitmesed seosed (tabelid), m:n iga binaarse seose jaoks luuakse eraldi seos (tabel).

Andmemudelis nimetame seda mitu mitmele seosega olemit *lekseemiks*, mis lahtiseletatult tähendab 'see sõna selles tähenduses' (seos sõna ja tähenduse vahel; joonisel 1 tähistab seda seos sõnast tähenduseni). Kui algul oli see pelgalt tehniline



Joonis 1. Ekilexi andmemudeli skeem 2021.

seos, siis arenduse käigus on sellest seosest saanud andmemudeli võtmeüksus, mis omakorda seob teisi olemeid: enamik andmeüksusi (nt sõnaliik, rektsioon, õppekommentaarid, kasutusnäited) on lekseemi parameetrid (vt joonist 1).[7]

Ekilexi andmemudel koondab ühtsesse baasi kogu keelelise info, mille algkodu on olnud paljudes eraldiseisvates sõnakogudes. Sõna ja tähenduse mitu mitmele seose vahetabeli rolli täidab olem, mida andmemudelis nimetame *ilmikuks* (vt joonist 1), mis lahtiseletatult tähendab 'see sõna selles tähenduses selles sõnakogus'. Teoreetilise panusena leksikograafiasse saame lekseemi toel kirjeldada sõna ja tähendust kui keeleinfot, mitte kui sõnastikuinfot (Tavast jt 2018). Ehk teisisõnu, püüame ühe sõnatähenduse kirjeldamiseks pakkuda kasutajale variante, nt detailset seletust, vähemat tähendusvihjet, lihtsustatud keeles seletust. Eri sõnastikud on seni kasutanud kõiki variante läbisegi.

Kolme aasta jooksul oleme Ekilexi mudelit mitu korda muutnud, ka nüüd on juba näha, et muutusi tuleb veel. Algne mudel ei olnud seevõrra rekursiivne nagu nüüdne: erinevus seisneb selles, et nüüd on ühtsustatud sõnaks/keelendiks kõik vormiga seotud olemid (sõna/keelend, kollokatsioon (naabersõnad), kasutusnäide, võib-olla edaspidi isegi definitsioon). Toona kaalusime sama, aga otsustasime teisiti: kuigi neil üksustel on mõni sarnane omadus (nt eriti tähendus), aga siiski paljud omadused erinevad, siis jäi iga nimetatu omaette olemiks (vt joonist Tavast jt 2018: 753). Tol hetkel tundunuks see samm ka liiga radikaalne. Kolmeaastase arenduse tulemusel oleme otsustanud ikkagi mudelit muuta rekursiivsuse suunas: kui enne olid kollokatsioonid eraldi olemid, mis olid lekseemiga seotud, siis nüüd on kollokatsioon ise sõna/keelend, mille juures saab (sõnakoguspetsiifiliste) ilmikuseostega esitada, millistest komponentidest ja kuidas see koosneb (vt joonisel 1 omaseost ilmiku kasti küljes). Kollokatsiooni komponendid on sõnad oma mingis tähenduses. Ilmikuseosega on andmemudelis seotud näiteks allmärksõna (vt EKI ühendsõnastikus *aadress* : *aadressil*) ja õppekommentaar (vt Keeleõppija Sõnaveebis *ükssada*). Tavalise (mitmesõnalise) märksõna ja kollokatsiooni vahe seisneb selles, et viimasel esineb (peale muude ilmikuseoste) ka kollokatsioonitüüpi ilmikuseoseid (nt info kollokaadi sõnaliigi ja grammatilise suhte kohta). Oleme jõudnud samale järeldusele, mida 2019. aasta e-leksikograafia konverentsil väljendas üks OntoLex-Lemoni arendajaid Thierry Declerck: „I don't like the idea of lexicographic entry" ('mulle ei meeldi mõte sõnaartiklist'). Ekilexis oleme teel ühtse universaalse struktuuri poole, mis sobiks kõigile keeleandmetele.

Kahtlemata on leksikograafide jaoks Ekilexi andmemudel (vt joonist 1) keeruline. Samal ajal on sellel vaieldamatult suur mõju leksikograafilistele tööprotsessidele (Paulsen jt 2020: 180–181), näiteks sõnaliigi määramisel (lekseemi parameetrina) on mõistlik olla kooskõlas toimivate tekstitöötlusvahendite (nt EstNLTK) liigitusega. Kui keeleteaduses on näiteks sünonüümiat, antonüümiat ja kaashüponüümiat tõlgendatud ühtmoodi kui leksikaalseid suhteid, siis on raske mõista, kuidas siin mudelis käituvad nad erinevalt. Leksikaalsed suhted on mudelis defineeritud kahel viisil: 1) täissünonüümid ja täisvasted on sama tähendusega sõnad (ühes keeles või eri

[7] OntoLex-Lemoni mudelis (McCrae jt 2017) vastab lekseemile leksikaalne tähendus („Lexical Sense"), mis töötab põhimõtteliselt samal viisil („mapping from a word to a concept").

keeltes), nad on seotud ühe tähendusega, st nende jaoks pole eraldi lisaseost; 2) kõik teised leksikaalsed suhted (nt kaashüponüümid, vastandid, osasünonüümid[8]) on mõistelised seosed eri tähenduste vahel (mitte sõnade vahel).

Ekilexi programmiliides (API) on mõeldud arendajatele, kes soovivad kasutada Ekilexi andmeid eri tüüpi rakenduste loomiseks, näiteks liitsõnaraamatuks, õppematerjaliks, äpiks jne. API võimaldab andmeid küsida JSON-vormingus. Seni on EKI API-t kasutatud eelkõige Horizon 2020 projekti ELEXIS raames.

4. Kokkuvõtteks

Läinud 20 aastat on näidanud, et digimaailma normaalne olek ongi pidevalt pöördeid teha. Ühelt poolt kogeme kustumatut uudishimu leiutada uusi võimalusi keele uurimiseks ja kirjeldamiseks, teiselt poolt peame pidevalt tegelema olemasolevate süsteemide järeleaitamisega. Sõnaraamatutöö on muutunud väga interdistsiplinaarseks: leksikograafid peavad tegema koostööd arvutilingvistide ja keeletehnoloogidega, linkandmete, semantilise veebi ja tehisintellekti spetsialistidega (Steurs jt 2020). Samuti ollakse suuremaks kasvanud oma riigist: hulk töövahendeid on kõigile vabalt saadaval üle Euroopa. Kindlasti tehakse koostööd veebiliideste spetsialistidega. Tänapäeva leksikograafi pädevus ulatub lingvistilistest teadmistest väljapoole: tal on vaja oma igapäevatöö jaoks tehnilist, n-ö andmebaasilist mõtlemist ja tehnilisi oskusi. Leksikograafi töö muutub ühelt poolt üha rohkem järeltoimetamiseks, teiselt poolt üha laiapõhjalisemaks ja demokraatlikumaks, hõlmates vabatahtlikke kaastöölisi ja uusi töövorme uutes ühisplatvormides. See on veel paljuski avastamata maa.

Eestis oleme (Ekilexis ja Sõnaveebis) saanud juba kasutada mitmeid keeletehnoloogilisi lahendusi ja rakendusi, muuhulgas kõnesünteesi (näitelausete ettelugemiseks), kõnetuvastust ja hääljuhtimist (sõnaotsingul), automaatset kollokatsioonide ja veebilausete tuvastamist ning kuvamist, sünonüümi- või vastekandidaatide pakkumist.

Kuhu leksikograafia edasi areneb? Võib tunduda ootamatu, aga ... nähtamatuse suunas. Gilles-Maurice de Schryver (2019) ja teised keeletehnoloogid (nt Tasovac 2010) on kirjeldanud leksikaalseid andmeid kui üksusi, mis on märkamatult integreeritud mitmesugustesse rakendustesse, näiteks keeleõpperakendustesse, masintõlkeplatvomidesse, tekstitöötlusprogrammidesse. Ehk teisisõnu, tuleb aeg, kui tekst sisaldab sõnastikku samavõrra kui praegu sõnastik sisaldab teksti.

Artikkel on valminud Haridus- ja Teadusministeeriumi projekti EKKD64 „Eesti keele sõnavara ja korraldus: deskriptiivne ja preskriptiivne vaatenurk", Euroopa Liidu programmi Horisont 2020 projekti INFRAIA-02-2017 „European Lexicographic Infrastructure" ja Eesti Keele Instituudi baasfinantseerimise toel.

[8] Osasünonüüme ehk peaaegu sarnase tähendusega sõnu on nimetatud ka lähisünonüümideks. Siin kasutame terminit osasünonüümid.

VEEBIVARAD

EELex. Eesti Keele Instituudi sõnastikusüsteem. https://eelex.eki.ee
Eesti keele ühendkorpus 2019. https://dx.doi.org/10.15155/3-00-0000-0000-0000-08565L
Ekilex. Eesti Keele Instituudi sõnastiku- ja terminibaasisüsteem. https://ekilex.eki.ee
Ekilex API. https://github.com/tripledev/ekilex/wiki/Ekilex-API
EKI ühendsõnastik 2021. Eesti Keele Instituut. Sõnaveeb, 2021. https://sonaveeb.ee/collections
EKS 2019 = Eesti keele sõnaraamat 2019. Eesti Keele Instituut. Sõnaveeb, 2019. www.eki.ee/dict/eks; https://doi.org/10.15155/3-00-0000-0000-0000-08240L
Elexifier. https://elexifier.elex.is
ELEXIS = European Lexicographic Infrastructure. https://elex.is
Entu. https://entu.keeleressursid.ee
Esterm. https://termin.eki.ee/esterm
EstNLTK = Estonian Natural Language ToolKit. Kogumik teeke eestikeelsete tekstide töötluseks. Versioon 1.6.2. https://estnltk.github.io
EVS = Eesti-vene sõnaraamat 2019. 2., täiendatud ja kohandatud veebiväljaanne. Eesti Keele Instituut. Sõnaveeb, 2019. https://www.eki.ee/dict/evs
Keeleõppija Sõnaveeb. Eesti Keele Instituudi keeleportaal. Versioon 1.21. https://sonaveeb.ee/lite
KORP. Korpuspäringusüsteem. https://korp.keeleressursid.ee
Lexicog. https://www.w3.org/2019/09/lexicog
Lexonomy. https://www.lexonomy.eu/
MultiTerm. https://www.trados.com/products/multiterm-desktop/
NAISC. https://github.com/insight-centre/naisc
Sketch Engine. Korpuspäringussüteem. https://www.sketchengine.eu
Sõnaveeb. Eesti Keele Instituudi keeleportaal. Versioon 1.21. https://sonaveeb.ee

KIRJANDUS

Bajčetić, Lenka; Declerck, Thierry 2020. Interlinking Slovene language datasets. – Proceedings of XIX EURALEX Congress: Lexicography for Inclusion. Kd I. Toim Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras. Democritus University of Thrace, lk 73–80.

Bański, Piotr; Bowers, Jack; Erjavec, Tomaz 2017. TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. – Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of eLex 2017 Conference, Leiden, September 2017. Toim Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, Vít Baisa. Brno: Lexical Computing CZ s.r.o., lk 485–494.

Cavagliá, Gabriela; Kilgarriff, Adam 2001. Corpora from the Web. – Proceeding of the Fourth Annual CLUCK Colloquium, Sheffield, UK, lk 120–124.

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, kd 54, nr 3, lk 713–745. https://doi.org/10.1007/s10579-020-09487-4

de Schryver, Gilles-Maurice; Chishman, Rove; da Silva, Bruna 2019. An overview of digital lexicography and directions for its future: An interview with Gilles-Maurice de Schryver. – Calidoscópio, kd 17, nr 3, lk 659–683. https://doi.org/10.4013/ld.2019.173.13

EKSS = Eesti kirjakeele seletussõnaraamat. Kd I–VII. Tallinn: Eesti Keele Instituut, 1991–2007.

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. http://www.eki.ee/dict/ekss (5. I 2021).

Francopoulo, Gil; George, Monte; Calzolari, Nicoletta; Monachini, Monica; Bel, Nuria; Pet, Mandy; Soria, Claudia 2006. Lexical markup framework (LMF). – Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). Genoa: European Language Resources Association, lk 233–236.

Fuertes-Olivera, Pedro; Bergenholtz, Henning (toim) 2011. e-Lexicography: The Internet, Digital Initiatives and Lexicography. Great Britain: Continuum.

Gorjanc, Vojko; Gantar, Polona; Kosem, Iztok; Krek, Simon (toim) 2017. Dictionary of Modern Slovene: Problems and Solutions. Ljubljana: University of Ljubljana, Faculty of Arts. https://doi.org/10.4312/9789612379131 https://doi.org/10.4312/9789612379131

Grefenstette, Gregory; Nioche, Julien 2000. Estimation of English and non-English language use on the WWW. – Recherche d'Information Assistée par Ordinateur (RIAO). 6th International Conference, College de France, France, April 12–14. Proceedings. Toim Joseph-Jean Mariani, Donna Harman. Paris, lk 237–246.

Jürviste, Madis; Kallas, Jelena; Langemets, Margit; Tuulik, Maria; Viks, Ülle 2011. Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary. – Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011, Bled, 10–12 November. Toim Iztok Kosem, Karmen Kosem. Ljubljana: Trojina, Institute for Applied Slovenian Studies, lk 106–112.

Kallas, Jelena; Koeva, Svetla; Langemets, Margit; Tiberius, Carole; Kosem, Iztok 2019. Lexicographic practices in Europe: Results of the ELEX survey on user needs. – Proceedings of the eLex 2019 conference. 1–3 October, Sintra, Portugal. Toim I. Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Isabel Pereira, J. Kallas, Miloš Jakubíček, Simon Krek, Carole Tiberius. Brno: Lexical Computing CZ, s.r.o, lk 519–536.

Kilgarriff, Adam; Rychlý, Pavel; Smr, Pavel; Tugwell, David 2004. The Sketch Engine. – Proceedings of the XI EURALEX International Congress. Toim Geoffrey Williams, Sandra Vessier. Lorient, France: Université de Bretagne Sud, lk 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: Journal of ASIALEX, kd 1, nr 1, lk 7–36. https://doi.org/10.1007/s40607-014-0009-9

Koppel, Kristina 2020. Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele. (Dissertationes linguisticae Universitatis Tartuensis 38.) Tartu: Tartu Ülikooli Kirjastus.

Koppel, Kristina; Tavast, Arvi; Langemets, Margit; Kallas, Jelena 2019a. Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. – Pro-

ceedings of the eLex 2019 conference. 1–3 October, Sintra, Portugal. Toim Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Toim Isabel Pereira, J. Kallas, Miloš Jakubíček, Simon Krek, Carole Tiberius. Brno: Lexical Computing CZ, s.r.o, lk 434–452.

Koppel, Kristina; Kallas, Jelena; Khokhlova, Maria; Suchomel, Vít; Baisa, Vít; Michelfeit, Jan 2019b. SkELL corpora as a part of the language portal Sõnaveeb: Problems and perspectives. – Proceedings of the eLex 2019 conference. 1–3 October, Sintra, Portugal. Toim Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Isabel Pereira, J. Kallas, Miloš Jakubíček, Simon Krek, Carole Tiberius. Brno: Lexical Computing CZ, s.r.o, lk 763–782.

Kuusk, Piret; Reivelt, Kaido 2019. Füüsika eestikeelsest terminivarast riiklikul oskuskeelekuul. – Sirp 31. V.

Langemets, Margit 2012. Lingvistilisest kolonialismist sõnaraamatus. – Keel ja Kirjandus, nr 8–9, lk 598–613.

Langemets, Margit; Loopmann, Andres; Viks, Ülle 2010. Dictionary management system for bilingual dictionaries. – eLexicography in the 21st Century: New Challenges, New Applications. Toim Sylviane Granger, Magali Paquot. Louvain-la-Neuve: Presses universitaires de Louvain, Cahiers du CENTAL, lk 425–429.

Langemets, Margit; Hein, Indrek; Heinonen, Tarja; Koppel, Kristina; Viks, Ülle 2017. From monolingual to bilingual dictionary: The case of semi-automated lexicography on the example of Estonian–Finnish Dictionary. – Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of eLex 2017 Conference, Leiden, September 2017. Toim Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, Vít Baisa. Brno: Lexical Computing CZ s.r.o., lk 155–171.

Langemets, Margit; Uibo, Udo; Tiits, Mai; Valdre, Tiia; Voll, Piret 2018. Eesti keel uues kuues. Eesti keele sõnaraamat 2018. – Keel ja Kirjandus, nr 12, lk 942–958.

Langemets, Margit; Päll, Peeter 2020. Kust vaadata kirjakeele normi? EKI keelekool. – Postimees. Arvamus ja Kultuur 19. XII.

McCrae, John P.; Bosque-Gil, Julia; Gracia, Jordi; Buitelaar, Paul; Cimiano, Philipp 2017. The OntoLex-Lemon Model: Development and applications. – Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of eLex 2017 Conference, Leiden, September 2017. Toim Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, Vít Baisa. Brno: Lexical Computing CZ s.r.o., lk 587–597.

Měchura, Michal B. 2017 Introducing Lexonomy: An open-source dictionary writing and publishing system. – Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of eLex 2017 Conference, Leiden, September 2017. Toim Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, Vít Baisa. Brno: Lexical Computing CZ s.r.o., lk 662–679.

Paulsen, Geda; Vainik, Ene; Tuulik, Maria 2020. Sõnaliik leksikograafi töölaual: sõnaliikide roll tänapäeva leksikograafias. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 16, lk 177–202. https://doi.org/10.5128/ERYa16.11

Pomikálek, Jan; Rychlý, Pavel; Kilgarriff, Adam 2009. Scaling to billion-plus word corpora. – Advances in Computational Linguistics, nr 41, lk 3–13.

Pomikálek, Jan; Jakubíček, Miloš; Rychlý, Pavel 2012. Building a 70 billion word corpus of English from ClueWeb. – Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC'12). Toim Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis. European Language Resources Association, lk 502–506.

Sinclair, John 1991. Corpus Concordance and Collocation. Oxford: Oxford University Press.

Steurs, Frieda; Schoonheim, Tanneke; Heylen, Kris; Vandeghinste, Vincent (toim) 2020. The Future of Academic Lexicography – A White Paper. Version 1.2 https://ivdnt.org/wp-content/uploads/2021/02/The-Future-of-Academic-Lexicography-A-White-Paper.pdf (25. IV 2021).

Suchomel, Vít 2020. Better Web Corpora For Corpus Linguistics And NLP. Doctoral Theses. Brno: Masaryk University, Faculty of Informatics.

Suchomel, Vít; Pomikálek, Jan 2012. Efficient web crawling for large text corpora. – Proceedings of the 7th Web-as-Corpus Workshop (WAC7). 17 April, Lyon, France. Toim Adam Kilgarriff, Serge Sharoff. Lyon, lk 39–43.

Tasovac, Toma 2010. Reimagining the dictionary, or why lexicography needs digital humanities. – Digital Humanities. King's College London, 7th–10th July. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-883.html (5. II 2021).

Tasovac, Toma; Romary, Laurent; Banski, Piotr; Bowers, Jack; de Does, Jesse; Depuydt, Katrien; Erjavec, Tomaž; Geyken, Alexander; Herold, Axel; Hildenbrandt, Vera; Khemakhem, Mohamed; Petrović, Snežana; Salgado, Ana; Witt, Andreas 2018. TEI Lex-0: A baseline encoding for lexicographic data. Version 0.8.6. DARIAH Working Group on Lexical Resources. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html (5. II 2021).

Tavast, Arvi; Langemets, Margit; Kallas, Jelena; Koppel, Kristina 2018. Unified data modelling for presenting lexical data: The case of EKILEX. – Proceedings of the XVIII EURALEX International Congress. EURALEX: Lexicography in Global Contexts, Ljubljana, 17–21 July. Toim Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek. Ljubljana: Ljubljana University Press, Faculty of Arts, lk 749–761.

Tavast, Arvi; Koppel, Kristina; Langemets, Margit; Kallas, Jelena 2020. Towards the super-dictionary: Layers, tools and unidirectional meaning relations. – Proceedings of XIX EURALEX Congress: Lexicography for Inclusion. Kd I. Toim Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras. Alexandroupolis: Democritus University of Thrace, lk 215–223.

Undusk, Jaan; Tamm, Marek 2021. Digimaailm ja mõnda. – ERR, Ööülikool 16. I. https://podcastid.ee/ooulikool/ooulikool-jaan-undusk-ja-marek-tamm-digimaailm-ja-monda/ (1. II 2021).

Vainik, Ene 2019. Eesti keele assotsiatsioonisõnastik. Eesti Keele Instituut. https://www.eki.ee/dict/assotsiatsioonid (25. IV 2021).

Valge, Jüri 2013. Kaks terminoloogiaprogrammi: 2008–2012 ja 2013–2017. – Õiguskeel, nr 4. https://www.just.ee/sites/www.just.ee/files/juri_valge._kaks_terminoloogiaprogrammi_2008-2012_ja_2013-2017.pdf (5. II 2021).

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. http://www.eki.ee/dict/qs2013 (26. VIII 2021).

ÕS 2018 = Eesti õigekeelsussõnaraamat ÕS 2018. Toim Maire Raadik. Koost Tiiu Erelt, Tiina Leemets, Sirje Mäearu, M. Raadik. Eesti Keele Instituut. Tallinn: EKSA. https://www.eki.ee/dict/qs2018 (26. VIII 2021).

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

Jelena Kallas (snd 1976), PhD, Eesti Keele Instituudi vanemarvutileksikograaf (Roosikrantsi 6, 10119 Tallinn), jelena.kallas@eki.ee

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

From a collection of dictionaries to a language portal

Keywords: e-lexicography, corpora, dictionary writing system, language portal, data model, API

The article aims to describe some major changes that have taken place in e-lexicography in recent decades in Europe generally and in Estonia in particular. Digital changes have permeated not only the dictionary compilation process but also whole workflow from lexicographic content creation to publication. The focus has shifted from building specific dictionaries to building a central database and infrastructure that can be adapted for further user and NLP applications.

We describe methods and technologies used to better integrate lexicographic data (several tools have been developed within the Horizon 2020 project European Lexicographic Infrastructure), and to better access lexicographic information.

As a turning point for digital change in Estonian lexicography, we consider the start of the development of the new Dictionary Writing System Ekilex and its user interface Sõnaveeb in 2017. The long-term goal is to have a single data source to provide consistent information about the Estonian language. In connection with Ekilex and Sõnaveeb, we discuss several issues: the theoretical foundations of the Ekilex biggest lexicographic dataset, the EKI Combined Dictionary, improvements in lexicographic workflow, and the Ekilex data model and API. The EKI Combined Dictionary contains information layers imported from several monolingual explanatory dictionaries, bilingual dictionaries, a collocations dictionary, and an etymology and

morphology database. The improvements in lexicographic workflow include working in one general database, more cooperation between research groups in the institute and more active involvement of external users.

The Ekilex data model meets the requirements for treating both words and meanings as independent entities and for representing both semasiological and onomasiological data. Created data are stored in Ekilex's PostgreSQL database and comply with all current standards of data exchange. As of April 2021, Ekilex contains approx. 300,000 headwords from general-language dictionaries and more than 90 terminological databases.

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

Jelena Kallas (b. 1976), PhD, Institute of the Estonian Language, Senior Computational Lexicographer (Roosikrantsi 6, 10119 Tallinn), jelena.kallas@eki.ee

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