Margit Langemets

eesti keele instituudi sõnaraamatute peatoimetaja, juhtinud mitmeid eesti keele instituudi sõnaraamatutöid, mh suure seletava sõnaraamatu väljaandmist, tegelnud elektroonilise leksikograafiaga ja uurinud eesti nimisõnade tähendusvaheldusi

Jelena Kallas

eesti keele instituudi teadur ja vanemleksikograaf. Teaduslikeks huvialadeks on korpus- ja õppeleksikograafia ning eesti keele kui teise keele õpetamise metoodika

Sõnaraamatud arvutis ehk elektrooniline leksikograafia

Teatavasti tuleb sõna *leksikograafia* kreeka keelest: *lexikon* 'sõnaraamat' + *graphō* 'kirjutan'. Vanimad sõnaraamatud pärinevad juba antiikajast, eeskätt idamaadest, nt Sumerist (praegusest Iraagist), Indiast ja Hiinast, kus ühelt poolt pandi kirja vähetuntud (luulelisi) sõnu, teiselt poolt praktilisi näpunäiteid nt naabritega kaubandussuhete ajamiseks.

Ühiskonna areng ja keeles toimuvad muutused on tinginud vajaduse aina uute sõnaraamatute järele. Samuti on üheskoos tehnoloogia arenguga muutunud sõnaraamatute kuju: sõnaraamatud on ilmunud saviplaatidel, pärgamendil, papüürusel, viimase 500 aasta jooksul paberil ja nüüd – umbes viimase 50 aasta jooksul – arvutis.

XX ja XXI sajandi piirimail sündis leksikograafia uus haru – *elektrooniline* ehk *e-leksikograafia*, mis on kõige noorem ja kõige kiiremini arenev suund. Teadusharu uurib veebisõnaraamatute koostamise ja kasutamise küsimusi. Peale leksikograafide tegeleb selle valdkonnaga suur hulk keeletehnolooge (tarkvaraarendajaid, korpuslingviste, veebimeistreid jt).

Tänapäeval on e-leksikograafia muutunud rahvusvaheliselt tunnustatud tegevusvaldkonnaks. Iga kahe aasta tagant toimuvad Euroopas elektroonilise leksikograafia konverentsid, viimati möödunud sügisel Tallinnas. 2013. aasta konverents kandis pealkirja „Elektrooniline leksikograafia 21. sajandil: mõeldes väljaspool paberit" ning selle korraldasid eesti keele instituut ja Sloveenia rakenduslingvistika instituut Trojina.[1] Konverentsil tõdeti esiteks seda, et sõnastike koostajad peaksid oma kasutajat paremini tundma õppima, temaga ühendust hoidma, et infot paremini esitada. Teiseks tuleks täpsemini määratleda sõnaraamatute roll ja funktsioonid laiemas internetikeskkonnas, eelkõige otsingumootorite ja teatmeportaalide kõrval. Kolmandaks ei tuleks e-sõnaraamatut käsitleda valmis, lõpetatud tootena, vaid pigem abivahendina üha uute rakenduste loomiseks. Elektrooniline sõnaraamat on nagu Tallinna linn, mis kunagi valmis ei saa.

Elektrooniline sõnaraamat kui innovatsioon

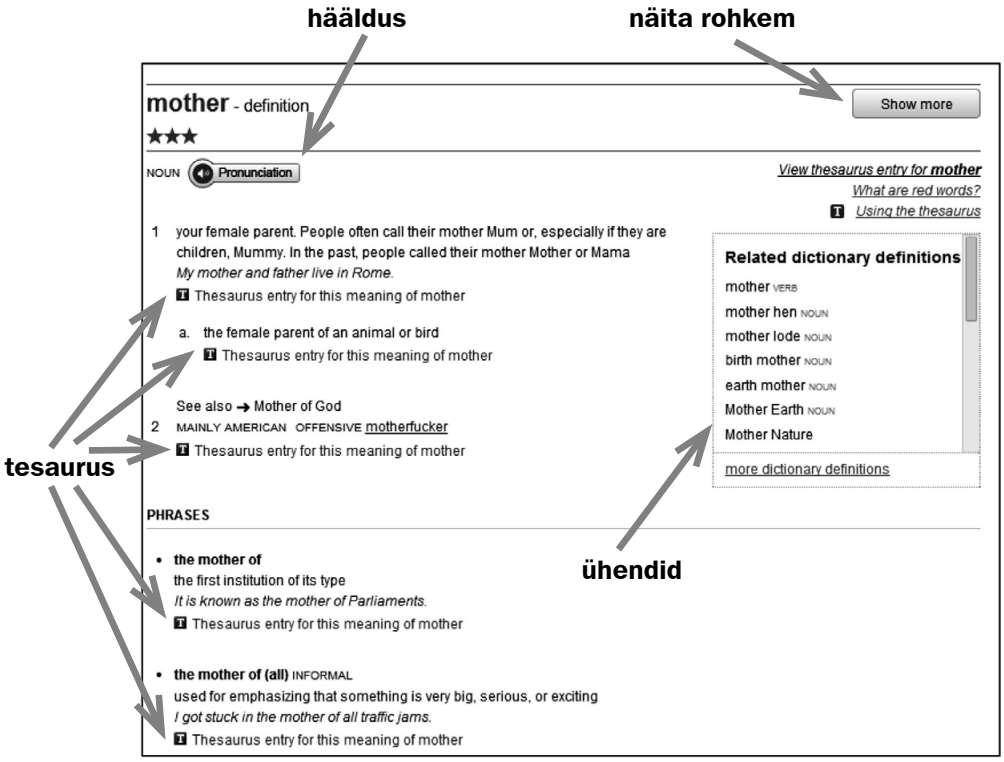
Meie lugejate seas ei ole ehk ühtegi inimest, kes pole pabersõnaraamatut käes hoidnud. „Mille poolest on siis elektrooniline sõnaraamat erinev?" küsite teie. Ikka on seal info, kuidas sõna õigesti kirjutada, mida see tähendab jne. Jah, tõepoolest, suures osas langeb pabersõnastikes ja internetisõnastikes esitatud info kokku. Ja mõned veebisõnastikud näevadki väliselt välja nagu pabersõnastike digitrükid. Selle põhjuseks on sageli see, et ollakse traditsioonilise esituse raamides kinni, ei rakendata multimeedia võimalusi, sõnaraamatut ollakse harjunud nägema kui staatilist tihedas kirjas teost, mitte kui paindlikku teenust, mille funktsionaalsust tuleb pidevalt arendada.

Millised on elektroonilise sõnastiku eelised

Kõigepealt on see ruumiga seotud piirangute kadumine. Koostaja ei pea enam muretsema sõnastiku mahu ega trükikulude pärast. Pealegi võimaldab tehnoloogia esitada infot sihtotstarbeliselt osade kaupa: keeleinfo on justkui sahtlitesse pakitud ja kui kasutaja teab, mida otsib, avab ta vaid selle sahtli, kus on teda huvitav info, mitte kõiki sahtleid korraga. Allpool toome näiteks Macmillani inglise keele sõnaraamatu artikli *mother* 'ema'. Linkide taga on peidus info häälduse, sõnavormide, ühendite kohta. Eraldi

[1] Konverentsi eLex 2013 koduleht http://eki.ee/elex2013/ (31.1.2014)

aknas on kuvatud kõik teised märksõnad, kus see sõna esineb. Kui kogu see info oleks kasutajale korraga kuvatud, upuks ta selle sisse ära.



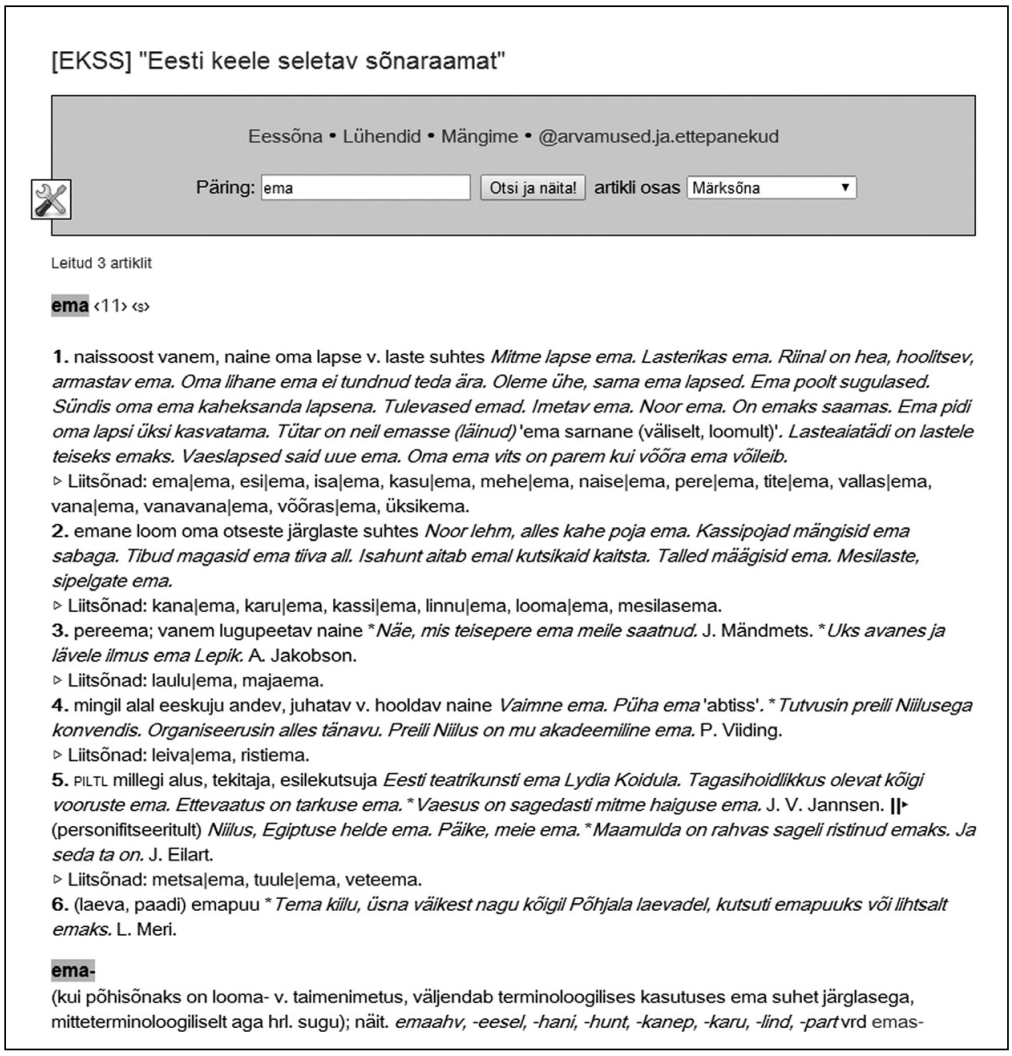
Joonis 1. Artikkel *mother* 'ema' Macmillan Dictionary2 e-sõnaraamatus. Eri tüüpi sõnastiku leksikograafiline info on esitatud linkide abil.

Seega muutub sõnaraamat hüpertekstiks, millel on sõnastikusisesed, aga tihti ka sõnastikuvälised lingid. Sõnastikusiseselt viidatakse sõnadele, mis on sama sõnastiku märksõnad. Sõnastikuvälised lingid suunavad kasutajat aga teiste ressursside, nt Wikipedia, rahva-Wiktionary (kust leiab 170 keelt) või eri tüüpi lingvistiliste andmebaaside (näiteks WordNet) juurde. Tulemuseks on teatmeteos, kus kasutaja saab vajaduse korral surfata eri allikates, kuni leiab vajaliku informatsiooni.

Võrdluseks võiks vaadata eesti keele seletava sõnaraamatu[3] artiklit *ema*, nii nagu see on hetkel veebis nähtav. Kahjuks näeb meie veebisõnaraamat veel üsna raamatu moodi välja.

[2] Internetis aadressil http://www.macmillandictionary.com/ (31.1.2014)

[3] Internetis aadressil www.eki.ee (31.1.2014)


[EKSS] "Eesti keele seletav sõnaraamat"

Eessõna • Lühendid • Mängime • @arvamused.ja.ettepanekud

Päring: ema Otsi ja näita! artikli osas Märksõna

Leitud 3 artiklit

ema ‹11› ‹s›

1. naissoost vanem, naine oma lapse v. laste suhtes *Mitme lapse ema. Lasterikas ema. Riinal on hea, hoolitsev, armastav ema. Oma lihane ema ei tundnud teda ära. Oleme ühe, sama ema lapsed. Ema poolt sugulased. Sündis oma ema kaheksanda lapsena. Tulevased emad. Imetav ema. Noor ema. On emaks saamas. Ema pidi oma lapsi üksi kasvatama. Tütar on neil emasse (läinud)* 'ema sarnane (väliselt, loomult)'. *Lasteaiatädi on lastele teiseks emaks. Vaeslapsed said uue ema. Oma ema vits on parem kui võõra ema võileib.*
▹ Liitsõnad: ema|ema, esi|ema, isa|ema, kasu|ema, mehe|ema, naise|ema, pere|ema, tite|ema, vallas|ema, vana|ema, vanavana|ema, võõras|ema, üksikema.
2. emane loom oma otseste järglaste suhtes *Noor lehm, alles kahe poja ema. Kassipojad mängisid ema sabaga. Tibud magasid ema tiiva all. Isahunt aitab emal kutsikaid kaitsta. Talled määgisid ema. Mesilaste, sipelgate ema.*
▹ Liitsõnad: kana|ema, karu|ema, kassi|ema, linnu|ema, looma|ema, mesilasema.
3. pereema; vanem lugupeetav naine **Näe, mis teisepere ema meile saatnud.* J. Mändmets. **Uks avanes ja lävele ilmus ema Lepik.* A. Jakobson.
▹ Liitsõnad: laulu|ema, majaema.
4. mingil alal eeskuju andev, juhatav v. hooldav naine *Vaimne ema. Püha ema* 'abtiss'. **Tutvusin preili Niilusega konvendis. Organiseerusin alles tänavu. Preili Niilus on mu akadeemiline ema.* P. Viiding.
▹ Liitsõnad: leiva|ema, ristiema.
5. PILTL millegi alus, tekitaja, esilekutsuja *Eesti teatrikunsti ema Lydia Koidula. Tagasihoidlikkus olevat kõigi vooruste ema. Ettevaatus on tarkuse ema. *Vaesus on sagedasti mitme haiguse ema.* J. V. Jannsen. ‖▸ (personifitseeritult) *Niilus, Egiptuse helde ema. Päike, meie ema. *Maamulda on rahvas sageli ristinud emaks. Ja seda ta on.* J. Eilart.
▹ Liitsõnad: metsa|ema, tuule|ema, veteema.
6. (laeva, paadi) emapuu **Tema kiilu, üsna väikest nagu kõigil Põhjala laevadel, kutsuti emapuuks või lihtsalt emaks.* L. Meri.

ema-
(kui põhisõnaks on looma- v. taimenimetus, väljendab terminoloogilises kasutuses ema suhet järglasega, mitteterminoloogiliselt aga hrl. sugu); näit. *emaahv, -eesel, -hani, -hunt, -kanep, -karu, -lind, -part* vrd emas-

Joonis 2. Artikkel *ema* eesti keele seletavas sõnaraamatus.

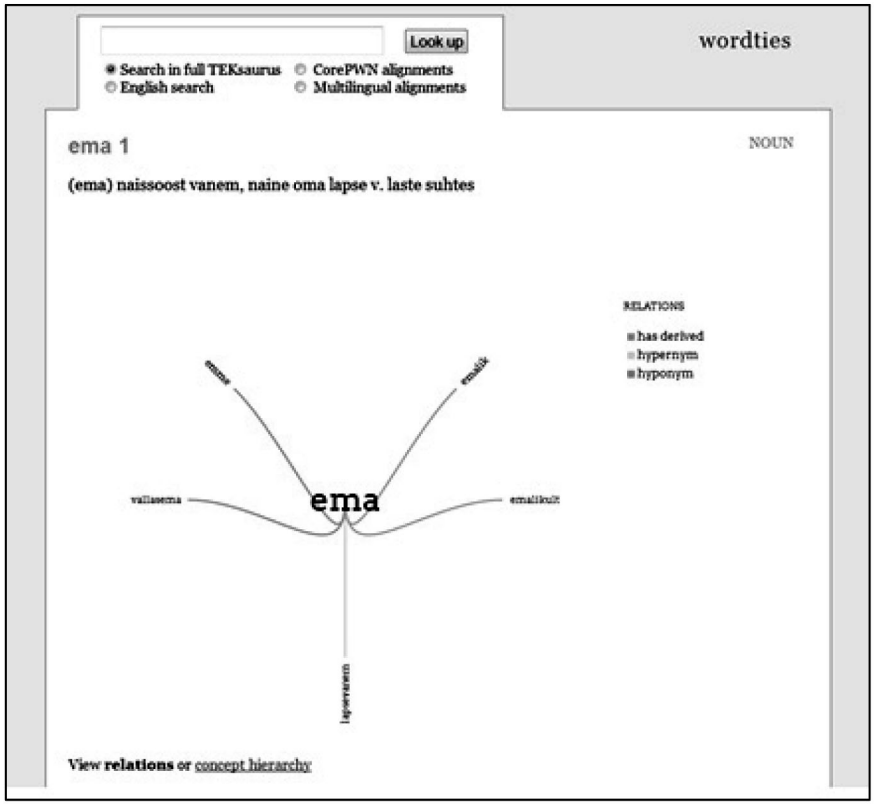
Teiseks suureks e-sõnastiku eeliseks on mugav otsing. Kasutajal on vaja vaid sõna sisse trükkida ja ta saab otsitud info kätte (ilma tähestikku meelde tuletamata). Lisaks aitavad paljud elektroonilised sõnaraamatut vigase vormi sisestamisel nii, et pakuvad valiku sõnavormidest, mida kasutaja võiks otsida.

Elektrooniliste sõnastike lahutumatu osa on multimeedia rakendused, eri tüüpi heli-, video- ja pildifailid. Sageli saab sõnastike abil kuulata sõna ja selle vormide häädlust. Siin kasutatakse kahte meetodit. Üks võimalus on see, et helisalvestatakse inimhääle faile ning need lingitakse sõnastikku. Teine võimalus on kasutada kõnesünteesi. Eesti keele jaoks on selliseid sõnastikke proovitud teha: üks seda tüüpi eksperimentaalne sõnastik asub aadressil http://www.eki.ee/dict/psh. Mõnikord kasutatakse helifaile ka

sõna tähenduse avamisel. Nii on kasutatud näiteks lindude ja loomade häälitsusi sisaldavaid faile.

Aina populaarsemaks muutuvad sõnastike videotutvustused. Kõige sagedamini kasutatakse aga videoid siiski erisõnastike, näiteks viipekeelesõnastike koostamisel. EKI ja Eesti viipekeeletõlkide ühingu koostööna on hiljuti valminud esimene eesti keele – eesti viipekeele sõnastik[4].

Sõnastike interaktiivsemaks muutmiseks kasutatakse sõnastikes palju ka fotosid ja pilte. Viimasel ajal on muutunud populaarseks eri tüüpi graafikute ja sõnapilvede (ingl *word cloud*) kasutamine, need võimaldavad sõnadevahelisi seoseid (nt sünonüüme, ülamõisteid ja alamõisteid) visualiseerida. Selline võimalus on olemas ka eesti keele jaoks: Tartu ülikoolis loodud Eesti Wordnet[5] näitab oma sisu ka graafiliselt:



Joonis 3. Sõna *ema* tähenduse „naissoost vanem, naine oma lapse v. laste suhtes" seoste graafiline esitus Eesti Wordnetis.

[4] Internetis aadressil http://eelex.eki.ee/run/ViipeKeel/ (31.1.2014)

[5] Eesti Wordneti graafiline päring http://wordties.cst.dk/wordties-estwn/ (8.01.2014)

Inglise sõnu koos oma seostega saab vaadata näiteks aadressil http://www.visuwords.com/ või http://www.visualthesaurus.com/.

Elektroonilised sõnastikud portaalide osana

Üsna levinud lähenemine on tänapäeval see, et sõnastikke ei lingita omavahel, vaid need koondatakse ühte portaali. See võimaldab vaadata sama sõna esitust eri sõnaraamatutes korraga.

Eesti keele sõnastikest saab sellise komplekspäringu teha näiteks portaalis www.keeleveeb.ee. Keeleveebi on koondatud eesti keele üldkeele sõnastikud (nt õigekeelsussõnaraamat, kõnekäändude ja fraseologismide andmebaas), erialasõnastikud (nt ehitaja sõnastik, kunstileksikon), tõlkesõnastikud, koolisõnastikud ja võõrkeelsed sõnastikud.

Eesti keele instituut on loonud ka keeleprofessionaalidele (toimetajatele, tõlkijatele) mõeldud e-keelenõu portaali http://keelenou.tk/. Portaali on valitud sõnaraamatud ja teatmeteosed, millega spetsialistid oma igapäevatöös kokku puutuvad. Antakse ka soovitusi ühe või teise sõna kasutamise kohta.

Taaskasutus – tänapäeva oluline võlusõna

Asi, millele elektrooniliste sõnaraamatute koostajad hoolega mõtlema peavad, on see, mismoodi saab sõnastikes esitatud materjali taaskasutada. Siin on abiks veebipõhised leksikograafi töökeskkonnad ja sõnastikusüsteemid (ingl Dictionary Writing System). Sõnastikusüsteemide ülesehitus on selline, et igal sõnaraamatu üksusel on oma tunnus või kood: näiteks on omaette tunnusega tähistatud sõnade häältdus, sõnaliik, muutevormid, näitelaused, ühendid, sünonüümid jne. Vajaduse korral saab leksikograaf ühe ja sama tunnusega üksused korraga välja võtta ja nii saab ühe sõnaraamatu põhjal teha täiesti uue sõnaraamatu, näiteks sünonüümide sõnaraamatu.

Eestis on enim kasutatud eesti keele instituudis loodud sõnastikusüsteem EELex (http://eelex.eki.ee/), kus praegu on kokku ligi 50 sõnastikku. Kõik need sõnastikud on standardse märgendusega taaskasutatavad keelekogud, mida saavad kasutada nii leksikograafid ja keeletehnoloogid kui ka tavakasutajad.

Elektroonilise sõnaraamatu sisu: kust info tuleb

Tänapäeva leksikograafide üks olulisemaid töövahendeid on tekstikorpused, mis kujutavad endast mingil viisil süstematiseeritud tekstikogusid. Neid täiendatakse pidevalt, samuti arendatakse nende töötlemise vahendeid. Sõnaraamatu koostaja võib üsna hõlpsasti läbi lugeda ja analüüsida 50 lauset, kus mingit sõna on kasutatud, võib-olla ka 500 lauset, aga kui kasutusi on 5000 – sageli veel palju rohkem –, siis seda käsitsi läbi töötada pole enam võimalik. Sestap on loodud spetsiaalsed korpusepäringusüsteemid (nt Sketch Engine firmalt Lexical Computing Ltd.), mis toovad statistika põhjal välja erinevad leksikaalsed ja grammatilised kombinatsioonid. Sketch Engine'i süsteemis on praegu 400 korpust, mis esindavad 70 keelt, sh ka eesti keelt.

Eesti keele uurimise jaoks on koostatud eri tüüpi korpuseid, mis võimaldavad jälgida eesti keele arengut alates 16. sajandist tänapäevani. Korpuste kogud on olemas Tartu ülikooli arvutilingvistika uurimisrühma kodulehel http://www.cl.ut.ee/, Keeleveebi portaalis www. keeleveeb.ee ja eesti keele instituudi kodulehel www.eki.ee.

Üks uusimaid eesti keele korpusi on möödunud aastal loodud veebikorpus etTenTen. Spetsiaalse tarkvara abil koguti internetist kokku eestikeelsed tekstid ja saadi tekstikogu, mille suurus on 250 miljonit tekstisõna. Sama meetodit on varem rakendatud ka mitme teise keele jaoks. Eelkõige sisaldab korpus ajakirjanduskeelt, lisaks on ka palju uue meedia tekste. Need on foorumid, netikommentaarid ja blogid. Internetikeele osa on eriti huvitav, kuna võimaldab analüüsida tänapäeva eesti internetikeele eripäraseid jooni, milleks on lühendid, emotikonid, toorlaenud, täpitähtede asendused (*ä*, *ö*, *õ* ja *ü* esitus on vastavalt *2*, *8*, *6* ja *Y*), nt *l2hen ylikooli* pro *lähen ülikooli*, ja slängi rohkus. Veebikorpus kannab tulevikku ulatuvat nime etTenTen, tähendades 10^{10} ehk kümmet miljardit eesti sõna, kui arvestada, et seni on meie kasutuses „üksnes" 250 miljonit ehk neljandik miljardit sõnakasutust. Võrdluseks: maailma suurim veebist kokku kogutud korpus English ClueWeb sisaldab 80 miljardit (!) sõna. Siit järeldus: mida rohkem eesti tekste, uusi ja vanu eesti sõnu veebis ringleb, seda jõulisem on eesti keel ja seda rohkem on leksikograafidel uurimismaterjali.

Kes koostab e-sõnaraamatu: leksikograaf, arvuti või meie kõik?

Elektroonilise leksikograafia konverentsidel arutatakse selle üle, kuidas sõnastike koostamist kiirendada. Kui palju saab arvuti abi kasutada, mida oleks võimalik teha automaatselt? Arvuti oskab juba ise genereerida sõna muutevorme, tuvastada sagedasemaid leksikaalseid ja grammatilisi kombinatsioone, otsib näitelauseid jne. Mõned katsed on üsna edukad, kuid selles valdkonnas on veel palju teha. Praegune seis on ikkagi selline, et arvuti on suuteline genereerima mitte terviklikke sõnaartikleid, vaid nende toorikuid, mis lähevad seejärel toimetaja kätte. Sealjuures saavad automaatselt koostatud sõnaartikleid aidata n-ö puhastada kõik inimesed, kellel meeldib seda talgute korras teha (ingl *crowd sourcing*). Leksikograafi ülesanne oleks vaid artiklile viimane lihv anda. Arvuti abi on eriti edukalt kasutatud uute sõnade tuvastamisel ja kirjeldamisel ning siin on inimesed saanud abiks olla, klõpsates sobiva esituse poolt või vastu. Üldiselt ongi nii, et sõnastike tavakasutajaist on saamas nõustajad ja miks mitte ka kaasautorid. Sõnaraamatute juures on enamasti foorumid, jututoad ja blogid, kus kasutajad saavad oma arvamust väljendada või näiteks omalt poolt seletuse või tõlkevaste pakkuda. Nii et tänapäeva elektrooniline sõnaraamat ei ole enam range akadeemiline väljaanne, vaid avatud keskkond, kus saab kasulikult aega veeta ja teiste kasutajatega keeleküsimusi arutada.