Väärtuslike keeleandmestike avaldamisest ja taaskasutamisest

Sirli Zupping
Haridus- ja Teadusministeeriumi keelepoliitika osakonna nõunik

Helen Kaljumäe
Eesti Keele Instituudi keeletehnoloogia kompetentsikeskuse keeletehnoloog

Kvaliteetsetel andmetel on tänapäeva innovatiivses infoühiskonnas täita oluline roll. Enamasti vajatakse andmeid selleks, et luua uusi ja parandada olemasolevaid e-teenuseid, hinnata ja täiustada protsesside efektiivsust ja teha andmetele tuginevaid otsuseid. Andmepõhise lähenemise laialdasem rakendamine on toonud kaasa vajaduse erisuguste andmete, sh andmekogude järele. Seetõttu räägitakse andmete avalikustamisest ja nende taaskasutamise võimalustest üha enam.

Andmepõhisus puudutab ka keelevaldkonda, sest viimasel ajal on kasvanud vajadus keeleandmete (eelkõige masinloetavate) järele. Keeleandmed hõlmavad nii teksti- kui ka kõneandmeid. Tekstiandmetena kogutakse ja kasutatakse näiteks tõlkeid, dokumente, raamatuid ja artikleid. Kõneandmed on inimese kõnet sisaldavad helisalvestised, nagu telefonikõned, loengusalvestised, raadiote vestlussaated, taskuhäälingud jmt.

Selliste keeleandmestike kasutusala on võrdlemisi lai, ulatudes keele- ja kultuuriloo uurimisest või keeleõppe toetamisest suurte masinõppeliste mudelite treenimiseni keeletehnoloogias. Keeleandmestikud on asendamatud allikad näiteks Bürokrati[1] ja Tõlkevärava[2] loomisel, isikuandmete automaatse anonüümija[3] väljatöötamisel, aga ka keeleõppevahendite koostamisel.

Keeleandmestikud kui väärtuslikud andmed

Euroopa Liidu avaandmete direktiivi[4] ülevõtmisega Eesti õigusse jõustati muu hulgas nõuded avaandmetele. Eesmärk on suurendada avaandmete kättesaadavust ja taaskasutatavust, et edendada innovatsiooni ja majandust ning infoühiskonda.[5] Avaandmete direktiivi ülevõtmisega loodi ka võimalus määratleda keeleandmed väärtuslike andmetena.[6] Väärtuslike keeleandmestike nimekirja koostamisel tuleb lähtuda avaandmete direktiivist tulenevast kontseptsioonist – väärtuslikud andmestikud, mille taaskasutamist seostatakse ühiskonnale, keskkonnale ja majandusele tekkivate oluliste hüvedega, sest need sobivad lisaväärtusteenuste ja -rakenduste loomiseks. Lähtudes avaandmete definitsioonist ja võttes arvesse keeleandmete olemust, on väärtuslikud keeleandmestikud nii avaandmetena kui ka juurdepääsupiiranguga

[1] www.kratid.ee.
[2] **M. Peetris**. Riigi loodav keskne tõlkekeskkond ja selle roll õiguskeeles. – Õiguskeel 2021/4, lk 1–5.
[3] https://github.com/buerokratt/Data-Anonymizer.
[4] Euroopa Parlamendi ja nõukogu direktiiv (EL) 2019/1024 avaandmete ja avaliku sektori valduses oleva teabe taaskasutamise kohta.
[5] Avaliku teabe seaduse muutmise seaduse avaldamine Riigi Teatajas.
[6] Avaliku teabe seaduse (RT I, 06.08.2022, 20) § 4[1] lõike 3 alusel.

andmetena kättesaadav keelematerjal, sh kõnematerjal (avaliku sektori andmed, teadusandmed, erasektori andmed). Praegu kaalutakse, kas keeleandmestike senisest laiemat avaldamist ja taaskasutamist on vaja reguleerida määrusega või piisab ka ametlikest, selgetest juhistest ja teavitusest, mis julgustaks teabevaldajaid oma keeleandmestikke avaldama.

Võimalikult paljude kvaliteetsete eestikeelsete tekstide kättesaadavus avaandmetena hõlbustab koostada asjakohaseid ja mahukaid keelekorpuseid, mis on vajalikud teadusuuringuteks ning tänapäevaste sõnastike koostamiseks. Keeletehnoloogias on väärtuslike keeleandmestike abil võimalik parandada eesti keele masintõlke, kõnetuvastuse ja kõnesünteesi kvaliteeti. See omakorda parandab era- ja avaliku sektori teenuste hõlpsamat kättesaadavust. Juba praegu on kõikide oluliste keeletehnoloogiliste lahenduste puhul aluseks avaandmed, mis tagavad võrdsed võimalused ka väikese ja keskmise suurusega ettevõtetele. Avaandmete taaskasutamine aitab kokku hoida avaliku sektori kuludelt ning soodustab ettevõtluse arengut.

Väärtuslike keeleandmestike liigid

Avaliku teabe seaduse § 3 lg 1 järgi on avalik teave „mis tahes viisil ja mis tahes teabekandjale jäädvustatud ja dokumenteeritud teave, mis on saadud või loodud seaduses või selle alusel antud õigusaktides sätestatud avalikke ülesandeid täites“. Seega kuuluvad väärtuslike keeleandmestike alla kõik avaliku sektori valduses olevad, avaliku sektori loodud või avaliku sektori rahastatud ning juurdepääsupiiranguta, sh üks-, kaks- ja mitmekeelsed keeleandmestikud. Keeleandmeid sisaldavad väga paljud andmekandjad, kuid lähtudes Euroopa keeleressursside koordineerimise konsortsiumi (ELRC) võrgustikus välja töötatud soovitustest[7], eristuvad väärtuslikena laias laastus neli liiki keeleandmestikke.

- **Nimede loendid**, sh kohanimede, isikunimede, kaubamärkide, organisatsioonide ja ettevõtete nimede loendid. Nimeloendeid kasutatakse automaatsetes info eraldamise ülesannetes. Edukas infoeraldus on oluline loomuliku keele mõistmise vahendite, nt küsimus-vastus- ja otsingusüsteemide arendamiseks, andmete automaatseks anonüümimiseks.
- **Terminibaasid**, sh erialased sõnakogud, sõnastikud, sõnaloendid, ontoloogiad ja tesaurused. Eestikeelse terminoloogia avaldamine ja levitamine toetab eestikeelset kõrgharidust, aitab arendada eesti keelt, sh hoida ringluses erialast terminivara, aitab hoida erialavaldkondi omakeelsena ning aitab eestikeelse masintõlke arendamist, sh võimaldab ühtset terminite kasutust masintõlkesüsteemides. Lisaks on vaja luua erialakeelekorpusi ehk suuri tekstikogumeid, millest saab sobivate tööriistade abil oskussõnad automaatselt välja sõeluda. Nii saab luua senisest palju tõhusamini eri valdkondade terminikogusid.
- **Tõlked** (lähte- ja sihtkeele tekstid), sh järeltoimetatud masintõlked, ja tõlkemälud. Siia alla kuuluvad näiteks pressiteadete, uudiste jt ametlike tekstide kvaliteetsed tõlked ja järeltoimetatud masintõlked (vandetõlkide tõlgitud ametlikud tekstid, Riigi Teataja tõlked). Tõlkemälud, mis on asutuste enda loodud ja/või teenusepakkuja loodud ning

[7] **European Language Resource Coordination**. ELRC white paper. Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe. Why language data matters. 2019.

mida on tõlkimisel kasutatud, kuuluvad kahtlemata väärtuslike keeleandmestike hulka. Tõlkematerjale kasutatakse masintõlkesüsteemide treenimisel ning seeläbi saavutatakse masintõlke täpsem kvaliteet.

- **Kõneandmestikud**, näiteks Riigikogu ja kohalike omavalitsuste istungite stenogrammid võimaluse korral koos helifailidega ning protokollidega. Helifaile, stenogramme ning nende põhjal loodud protokolle kasutatakse automaatse eestikeelse kõnetuvastuse treenimisel ning automaatsete protokollide loomise süsteemi treenimisel.

Kus saab keeleandmestikke avaldada?

ELi avaandmete direktiivi kohaselt peavad väärtuslikud andmestikud olema tasuta kättesaadavad, masinloetavad, rakendusliideste kaudu esitatavad ja asjakohasel juhul hulgi allalaaditavad.[8] Need tingimused peaks kohalduma ka keeleandmestike avaldamisele.

Keeleandmestikke on Eestis võimalik avaandmetena kättesaadavaks teha mitut moodi, alustades üldisemast avaandmete portaalist ja lõpetades spetsiifilisema andmete avaldamise platvormiga.

- **Teabevärav**[9] on Eesti riigi avaandmete portaal, kus tehakse kättesaadavaks teabevaldajate avaandmed, tagatakse avaandmete tasuta allalaadimine ja taaskasutamine mis tahes eesmärkidel. Teabevärava kaudu on igaühel ligipääs nii avaliku sektori juurdepääsupiiranguteta kui ka era- ja kolmanda sektori jagatud litsentsitud andmetele. Teabevaldajate määratud litsentsi alusel on võimaldatud andmete taaskasutamine nii ärilistel kui ka mitteärilistel eesmärkidel.
- **META-SHARE**[10] on keeleressursside register, mis koondab infot keeleandmestike ja keeletarkvara kohta. Registris on kirjeldatud keeleressursside metaandmed ning hallatakse ka ressursside allalaadimisvõimalust. Keeleressursside metaandmeid saavad vaadata ja vabakasutuses keeleressursse alla laadida kõik huvilised sisse logimata.
- **Tõlkevärav** (praegu arendustööd käivad)[11] on keskne tõlkeplatvorm, mis koondab endas masintõlke ja tõlkeabivahendite kasutamise keskkonda ning võimaldab süsteemselt tegeleda tõlketööde korraldamise ja tõlkimise protsessidega. Keskkond arvestab lähtetekstide, tõlgete jm andmete konfidentsiaalsusnõudeid, kasutades vajaduse korral anonüümimistehnikaid. Tõlkevärav on tasuta kasutamiseks avalikkusele ja avaliku sektori asutustele ning seda on võimalik kohandada asutuste spetsiifikale.
- **Ekilex**[12] on sõnastiku- ja terminibaasisüsteem, mis loodud sõnastike ja terminibaaside koostamiseks ja ajakohastamiseks, sinna koondatakse infot sõnade ja terminite kohta erinevatest sõnakogudest ja terminibaasidest. Ekilex aitab vähendada sõnakogude andmete dubleerimist ning lihtsustab tehtu avalikustamist. Süsteemi saavad tasuta kasutada kõik, kes soovivad oma terminibaasi koostada.

[8] Avaandmete direktiivi art 14 lg 1.
[9] Eesti avaandmete teabevärav.
[10] https://metashare.ut.ee/.
[11] https://riigihanked.riik.ee/rhr-web/#/procurement/4728940/general-info/.
[12] Eesti Keele Instituudi sõnastiku- ja terminibaasisüsteem Ekilex.

Kuidas tagada keeleandmestike õiguspärane avaldamine ja kasutamine?
Keeleandmestike õiguspärase kasutamise osas on jätkuvalt ebaselgust nii andmevaldajatel kui ka andmete kasutajatel. Keeleandmestike loomise, avaldamise ja taaskasutamisega seotud nüansse on teadlased ja erialaspetsialistid varemgi käsitlenud ning olenevalt rõhuasetusest on leitud ka lahendused või vähemalt edasimõtlemist vajavad küsimused.[13, 14]

2022. a detsembris korraldas Eesti Keele Instituut koostöös Haridus- ja Teadusministeeriumiga keeleandmestike teemal laiema kaasamisürituse, et selgitada välja andmete valdajate ja kasutajate põhilised probleemid seoses keeleandmestikega. Aruteludelt kogutud kasutusnäited ja küsimused annavad alust analüüsida kehtivat õigust, et tuvastada kehtivas regulatsioonis probleemid, mis on seotud keeleandmestike avaldamise ja taaskasutamisega.

Läbiviidud aruteludes ilmnes, et väärtuslike keeleandmestike hulgas on kõige selgem olukord nimede loendite avaldamise ja taaskasutamisega. Nimede loendite valdajad ega kasutajad pole oma töös õiguslikke takistusi täheldanud, pigem toodi välja vajadus mugavamate tehniliste lahenduste järele. Olulisemad nimede loendid on kättesaadavad riiklikest registritest (nt juriidilised nimed äriregistris, kohanimed riiklikus kohanimede registris) ning need on sisult kõrge kvaliteediga.

Terminibaaside kättesaadavaks tegemisel tõstatuvad litsentside valikuga seotud küsimused. Terminibaasides sisalduvat materjali peaks olema võimalik esitada mis tahes vajalikul moel ja töödelda eri viisidel, nt õppematerjaliks, rakenduse lähtematerjaliks, ent oluline on kaitsta ka autorite huve. Näiteks on Eesti Keele Instituudi sõnastiku- ja terminibaasisüsteemis Ekilex, kuhu koondatakse eesti sõnavara, ka terminivara, senini teabevaldajad ehk terminibaaside autorid andnud nõusoleku oma sõnakogud avaldada tüüplitsentsiga CC BY 4.0.

Litsentsidele lisaks on ebaselgus valdav ka terminibaaside sisu ja sellele kohalduda võivate autoriõiguste osas. Näiteks sisaldavad terminibaasid andmeid väiksematest terminisõnastikest (mõistekirje, definitsioon), ent koostajate vaatest pole selge, kas ja millised osised on kaitstud autoriõigusega ning mida seetõttu vajaduse korral teistmoodi tuleks teha. Osa vanemaid terminisõnastikke sisaldab ka refereeriva sisuga, ent allikaviiteta definitsioone.

Tõlketekste on erinevat laadi, nii sisu kui juurdepääsetavuse vaatest, ning sellest tuleneb ka nende keeleandmestike problemaatika. Tõlgete avaldamist ja taaskasutamist raskendavad neis sisalduda võivad isikuandmed ja muu konfidentsiaalne info. Selliste andmete avaldamisel tekivad mitmed küsimused: kes peab tagama tundliku info eemaldamise; kas sellist anonüümimise kohustust on võimalik kokkuleppel määrata eelistatud poolele (ja mil määral jääb teine pool siiski veel milleski vastutavaks); millega peab arvestama, kui kogu äriliselt tundlikku infot ei ole võimalik eemaldada; kas isikuandmete automaatne anonüümimine on õiguslikult võrdne inimese tehtuga (ja milline on protsess vigade hilisemal avastamisel). Andmete avaldaja vaatest on oluline esmalt hõlmata, missuguseid tõlketekste, kellega ja

[13] **A. Kelli, A. Tavast, K. Lindén**. Vestlusrobotid ja autoriõigus. – Juridica 2020/5, lk 345–354.
[14] **A. Kelli jt**. Isikuandmeid sisaldavate keeleandmete jagamisega seonduv õiguslik raamistik: teadlase ja teadusasutuse kohustused ning vastutus. – Eesti Rakenduslingvistika Ühingu Aastaraamat 2021/17, lk 99–121.

milliste kasutustingimustega avaldada üldse võib – on need võib-olla mõeldud vaid asutusesiseseks või kindla haldusala piires kasutamiseks; tõlkemäludega võivad kaasneda ka metaandmed, mis on väärtuslik infoallikas, ent samuti võivad sisaldada isikuandmeid, mida on vaja kaitsta.

Kõneandmestikud on eri liiki isikuandmeid sisaldavad andmestikud, mistõttu oodatakse võimalikke abistavaid suuniseid nii selliste andmete kogumise ja töötlemise kui ka avaldamise kohta. Et kõneandmestikud võivad sisaldada isikuandmeid ka kõne sees, on siingi oluline leida sobivad litsentsid avaldamiseks ja anonüümimise lahendused. Teadustöö tarbeks kogutakse palju kõneandmeid andmesubjekti nõusolekule tuginedes. Ometi jääb teadlaste aruteludest kõlama palju ebaselgust nõusolekuvormide ja neis esitatava informatsiooni kohta, liiati on praegused lähenemised töögrupiti erinevad, need on kujunenud ja kohandatud oma töö käigus ning vastavad seetõttu küll suuresti teadlaste vajadustele, kuid mitte tingimata täielikult õiguslikule regulatsioonile. Näiteks on ebaselge, mis vormis peab nõusolek olema esitatud ning kui konkreetne ja informatiivne sisult olema (arvestades, et alati pole nõusolekut küsides täpselt teada, kuhu teadustöö tulemusel jõutakse).

Eraldi küsimus on, kuidas saaks taaskasutada olemasolevate avalike teenuste kõneandmeid (nt erinevad infoliinid, töötukassa ja häirekeskuse telefonikõned), mida tekib arvestataval hulgal ja iga päev töö käigus, ning mis oleks keeletehnoloogia arendamisel väärtuslikud. Selliste andmete taaskasutamiseks võimaluste leidmine, esmalt teaduskasutusekski, on äärmiselt oluline.

Kokkuvõte

Keeleandmestikud on tänapäeva infoühiskonnas olulised andmed, mis on tugevalt seotud eesti keele ja keeletehnoloogia arenguga. Tänu Euroopa Liidu avaandmete direktiivi ülevõtmisele Eesti õigusse on võimalik määratleda keeleandmestikud väärtuslike andmetena ning teha senisest suurem hulk keeleandmestikke kättesaadavaks. Keeleandmestike avaldamise ja taaskasutamisega seoses on olenevalt keeleandmestike liigist ja kasutusotstarbest siiski veel üksjagu küsimusi lahendamata, mistõttu on oluline enne regulatsioonide väljatöötamist kaasata arutellu võimalikult palju keeleandmestike valdajaid ja kasutajaid, et selgitada välja sobivaim viis nende andmete avaldamise ja taaskasutamise jaoks.

Võimalikult suure hulga kvaliteetsete keeleandmestike avaldamine ja kättesaadavus toetab eesti keele jätkusuutlikkust ja keeletehnoloogia arengut, need avardavad teadusuuringute võimalusi nii kitsamalt keeleteaduses kui ka sellega külgnevatel uurimisaladel. Lisaks on kvaliteetsed keeleandmestikud vajalikud keelt kasutavate rakenduste arenduses, näiteks viisaka ja korrektse keelekasutusega juturoboti disainimisel. Pikemas perspektiivis on keeleandmestikel kanda oma roll ka tehisintellekti arengu soodustamisel.