Riigi loodav keskne tõlkekeskkond ja selle roll õiguskeeles

Mari Peetris
Justiitsministeeriumi Riigi Teataja talituse nõunik

2021. aasta juunis kiitis Vabariigi Valitsus heaks riigile keskse tõlkekeskkonna loomise eesmärgiga koondada ühele platvormile seni väljaarendatud keeletehnoloogilised lahendused ja tõlketööriistad, et võimaldada avalikul sektoril tulevikus tõhusamalt ametitekste tõlkida.

2019. aastal valmis avaliku sektori tõlkekorralduses kaks olulist analüüsi. Haridus- ja Teadusministeerium ning Majandus- ja Kommunikatsiooniministeerium kaardistasid avaliku sektori tõlkevajadusi.[1] Uuringus osales kokku 58 asutust, sealhulgas ministeeriumid ja nende hallatavad asutused. Vastajate sõnul on tõlkekulud suured, tõlkemälusid suuremalt jaolt ei koguta või neid ei olegi ning terminibaase töösse ei kaasata. Ka masintõlget ei kasutata või kui, siis ei ole see alati lahendatud kõige turvalisemalt, näiteks kopeeritakse tundlikud andmed Google Translate'i. Vajadus tõhusama tõlkekorralduse järele on suur. Koguni 76% vastanud asutustest pidas vajalikuks mingit keskset platvormi, küll aga sellist, mis arvestaks nende eripäraga, näiteks terminoloogiliste, keeleliste ja korralduslike vajadustega.

Samal aastal tegi tõlketehnoloogia ettevõte Tilde analüüsi[2], mille kohaselt kulutas avalik sektor, sealhulgas kohaliku omavalitsuse üksused, aastatel 2015–2018 tõlkimisele 6,5 miljonit eurot. Andmete kogumiseks korraldati masinkorje 30 000-eurosest hanke piirmäärast suuremate hangete kohta, mistõttu väiksemahulised tellimused jäid arvestusest välja. Samuti ei ole kogusummas kajastatud vandetõlkide poolt avaliku ülesandena tehtud seaduste ja välislepingute tõlgete kulu, vabade kutsete tõlkekulu ega asutuste koosseisus olevate tõlkijate kulu. Kokkuvõtvalt ja hinnanguliselt võib öelda, et kogu avaliku sektori tõlkekulu on üle kahe miljoni euro aastas.

Ülalmainitud tõlkevajaduste kaardistuse põhjal on tõlkekorraldus avalikus sektoris killustunud, kasutusel olevad lahendused on ühtlustamata või puuduvad. Juba 2016. aasta Riigi Tugiteenuste Keskuse koostatud memorandumis „Riigi tsentraliseeritud keeleressursi haldamine ja tõlgete hanked" on märgitud, et suur osa tõlgetest on kordustõlked ja pelgalt tõlkemälu kasutamisega kaasnev rahaline kokkuhoid oleks ligikaudu 20%. Paraku avalik sektor suuremalt jaolt tõlkemälusid ei kogu ega halda. Memorandumi koostamise ajal veel masintõlkest juttu ei tehtud, kuid viimastel aastatel masintõlke kiiret arengut arvesse võttes suureneks kokkuhoid kordades.

Naabrite poole vaadates näeme, et avalikus sektoris on masintõlge edukalt kasutusele võetud. Lätis on juba avalikuks kasutamiseks välja arendatud masintõlget ja uusi keeletehnoloogilisi

[1] **Haridus- ja Teadusministeerium, Majandus- ja Kommunikatsiooniministeerium**. Avaliku sektori tõlkevajaduste kaardistus aastal 2018. Kokkuvõte. Tallinn, 2018.
[2] **Tilde**. Report on National Translation Contracts in Estonia. Public Procurement Market Research: Process and Findings.

võimalusi sisaldav portaal Hugo.lv.[3] Samasugune portaal Vertimas.vu.lt on valminud ka Leedus.[4] Oma tõhusust on tõestanud Euroopa Liidu Nõukogu eesistujate jaoks kohandatud masintõlkesüsteemide sari EU Council Presidency translator[5], mille esmased tulemused olid Soome puhul[6] nii paljulubavad, et see kaasati 2019. aasta sügisel Soome eesistumise keeleteenuste töösse. Keskkond on soome kui väikese ja struktuurilt keerulise keele puhul oma kasulikkust tõlkeabitööriistana tõestanud, andes tulemuseks arvestatava kvaliteediga tõlke, võimaldades Soomel püstitada eesistujana lausa uue tõlkemahu rekordi – 12,7 miljonit sõna kuue kuuga.[7] Masintõlkesüsteem jäi Soome avalikus sektoris kasutusse ka pärast eesistumist.

Tõlkimisvajadus aga kasvab laialdase rahvusvahelise suhtluse tõttu aina enam – Euroopa andmeportaali hinnangul on Euroopa tõlketeenuste turg, mis võtab enda alla poole ülemaailmsest tõlketurust, viimase kümne aasta jooksul kahekordistunud.[8] Kahtlemata on see avaliku sektori jaoks suur kuluallikas ja vajadus ressursse kokku hoida kasvavate tõlkemahtude pealt on juba omaette ülesanne. Samas on keeletehnoloogilised võimalused tõlkimise tõhusamaks korraldamiseks oluliselt avardunud ja mitmekesistunud.

Eestis on erinevaid keeletehnoloogilisi lahendusi viimastel aastatel jõudsalt arendatud. Näidetena võib esile tuua Riigikogu istungite stenografeerimise süsteemi Hans, Ametlike Teadaannete ning äriregistri teabesüsteemi ja kinnistusraamatu otsinguportaali masintõlget. Eestis on tasuta kasutatavaid keeletehnoloogiaid juba arvukalt: Tartu Ülikooli keeletehnoloogia uurimisrühma arendatud tõlkemootor TartuNLP[9], Tartu Ülikooli EstNLTK[10] teekide kogumik eestikeelsete tekstide töötluseks, Eesti Keele Instituudi tekstitöötlusvahendid ja kõnerobotid, terminibaasid ning sõnastiku- ja terminibaasisüsteem Ekilex[11] ning mitmed teised. Mõistlik oleks kvaliteetsed üksikkomponendid, mida on arvestatav hulk, koondada ühtseks mugavaks süsteemiks.

Kuna avaliku sektori asutused kasutavad või kavandavad kasutusele võtta samadel eesmärkidel tõlketehnoloogilisi lahendusi, kaasneb dubleerivate tegevuste ja topelt kulu oht. Väikese riigina ei ole aga mõistlik igaühel oma lahendust luua. Teada on seegi, et iseseisvalt on masintõlke harjutamiseks piisava materjali leidmine keeruline. Masin vajab õppimiseks ohtralt keeleandmestikku. Seda enam on kasu kesksest tõlkekeskkonnast, mis võimaldab koondada ühiste andmestike põhjal aina iseõppivaid keeletehnoloogiaid ja tõlkekorraldust.

Riigi Tugiteenuste Keskuse algatus jäi pärast 2016. aastat soiku, kuid teema tõstatus taas, kui Vabariigi Valitsuse tegevusprogrammis aastateks 2019–2023 tehti Justiitsministeeriumi ülesandeks koostada analüüs ja ettepanekud eesti keele automaatfunktsioonide

[3] Läti keeletehnoloogia veebileht Hugo.lv.
[4] Leedukeelse masintõlke portaal.
[5] Euroopa Liidu Nõukogu eesistujate masintõlkesüsteem.
[6] Euroopa Liidu Nõukogu eesistujate masintõlkesüsteem. Soome keel.
[7] CEF Digital. Finland sets record with the EU Presidency Translator. 2020.
[8] European Data Portal. The Economic Impact of Open Data. Opportunities for value creation in Europe. 2020.
[9] Tartu Ülikooli keeletehnoloogia uurimisrühma arendatud tõlkemootor TartuNLP.
[10] Tartu Ülikooli teekide kogumik EstNLTK eestikeelsete tekstide töötluseks.
[11] Eesti Keele Instituudi sõnastiku- ja terminibaasisüsteem Ekilex.

kasutuselevõtuks – ennekõike küll õigusdokumente silmas pidades. Kuna Riigi Teataja talitus korraldab ka seaduste tõlkimist, ning, püüab tagada paremat tõlkekvaliteeti, pakkus see võimaluse teemasse põhjalikumalt süüvida. Arvestades sedagi, et valmistasime koostöös Registrite ja Infosüsteemide Keskusega ette Ametlike Teadaannete masintõlke kasutusele võtmist.[12]

Olles analüüsinud andmeid ja keeletehnoloogia väljavaateid, esitasime enim tõlkivate ministeeriumitega ühisettepaneku luua keskne tõlkemälu ja masintõlke keskkond. Eesmärk oli minna kaugemale üksnes keele automaatfunktsioonidest ja kaasata ka töövoo haldus, vahendid statistika kogumiseks, kvaliteedikontrolli meetmed ja palju muud. Avalikus sektoris tekste pärast tõlkimist üldjuhul ei toimetata, kuid ka sellele murekohale püütakse keskses tõlkekeskkonnas leida lahendus. Kindlasti peab platvorm võimaldama ühildumist asutustel juba hangitud tõlkeabi tarkvaradega Trados ja memoQ. Peale selle luuakse võimalus kasutada terminibaase, tõlkemälusid ja masintõlkemootoreid, mis on samuti mainitud tõlkeabitarkvaradega liidestatavad. Keskkonna enda tõlkeabitööriist jääb vabalt kasutatavaks kõigile, kes vähegi soovivad.

Pärast seda, kui Vabariigi Valitsus tehtud ettepaneku heaks kiitis, koostati lähteanalüüs ja loodi prototüüp. Need valmisid 2021. aasta juunis. Analüüsi põhjal saavutatakse keskse tõlkekeskkonna kasutuselevõtul avalikus sektoris kõige suurem ajaline kokkuhoid tõlkimiselt, mis moodustab umbes 50%. Unustada ei tasu ka halduskoormuse vähenemist tõlkimise korraldamisel. Samuti on märkimisväärne keskkonna majanduslik tasuvus, millega on võimalik kokku hoida umbes 1,3 miljonit eurot aastas. See sääst ei teki kohe, kuid kui keskkonnaga ühineb võimalikult palju avaliku sektori asutusi, on see saavutatav. Kulude kokkuhoiust ehk veelgi olulisem on tõlgete kvaliteedi parandamine. Liidestatud komponentide abil on võimalik saavutada näiteks kasutatavate terminite ühtsus terminibaaside põhja, ühtsustatumad tõlkelaused tänu tõlkemälule ja kvaliteedikontrolli funktsionaalsusele, õigekirjakontroll tagab vigadeta teksti jne. Liidestatud komponendid on mõeldud aitama ka neid, kes iga päev ei tõlgi.

Koostatud rakendusplaani järgi luuakse tõlkekeskkond 20 kuu jooksul kokku kolmes arendusetapis, eeldatav arenduskulu on kokku 1,48 miljonit eurot. Rakendusplaani elluviimine on antud keelevaldkonda juhtivale Haridus- ja Teadusministeeriumile ning nende hallatavale Eesti Keele Instituudile. Justiitsministeerium jääb projektiga seotuks ka edaspidi ja esimeste tulemusteni püütakse koostöös jõuda 2022. aasta lõpus.

Tänapäeval on digitaalne võimekus arenenud keele tunnus. Esmaseid eesmärke keskse tõlkekeskkonna loomisel on tagada eesti keele püsimine ja kasutatavus maailmas. Arvestades eesti keele vähest kasutajaskonda ei ole me keeletehnoloogia suurarendajate jaoks prioriteetne keelesuund, mistõttu tuleb riigil siin oma õlg alla panna, näiteks selleks, et saaksime tulevikus oma seadmetega siiski eesti keeles suhelda. Kvaliteetse süsteemi loomine aitab tagada eesti keele digiteerituse ja jätkusuutlikkuse mitmekeelses maailmas.

[12] Justiitsministeerium. Ametlikes Teadaannetes võeti kasutusele masintõlge.

Õigus on paratamatult seotud kõikide muude valdkondadega ja puudutab neist igaüht. Ametitekstides tuleb lähtuda õiguskeele ja -tõlke põhimõtetest. Tähelepanu tuleb pöörata terminoloogilisele ühtsusele: oluline on, et kasutatakse samu termineid, mis seadustes. Isegi avalikus sektoris koostatud kiri peab kasutama seaduse termineid seaduses antud tähenduses. Ühtluse tagamiseks vajavad avaliku sektori ameti- ja tarbetekstid kvaliteetset õigusvaldkonna keeletehnoloogiat. Keskse tõlkekeskkonna valdkondlike masintõlkemootorite hulgas on õigusvaldkonnal kindlasti oma koht ja sellise mootori loomine juba käib. Keelesuundadest tehakse see kõigepealt kasutatavaks tõlkimisel eesti keelest inglise, saksa ja vene keelde ning vastupidi. Esmased keelesuunad valiti just tõlkemahu järgi ning neisse keeltesse tõlgitakse praegu enim, mis tähendab, et nendes keeltes on ka kõige rohkem harjutusmaterjali. Masin õpib näiteks Riigi Teataja seadusetõlgetest, erialastest tõlkemäludest, mis loodud väljaspool avalikku sektorit, ja Euroopa Parlamendi digitaalsest keeleandmestikust, kuid ka Estermi sõnastike ja Ekilexi oskussõnastike väljavõtetest.

Praegu ei arvesta keeletehnoloogilised lahendused õiguskeele terminoloogilise iseseisvuse ja trafaretsuse põhimõttega. Seetõttu ei anna üldkeele tõlkemälud või -mootorid piisavalt täpseid tulemusi, rääkimata üldkeele sõnastikest. Tähendusväljad erinevad ka keelesuuniti. Näiteks ei ole vasted alati ühesed Euroopa Liidus inglise keelest eesti keelde tõlgitud õigusaktide ja meil siin eesti keelest inglise keelde tõlgitud seadusetekstide vahel. Neil võib olla hoopis erinev õiguslik tähendus.

Meil on põhjust tunda uhkust, et kõik Eesti seadused on meie õiguskorra tutvustamiseks ning rahvusvahelise ameti- ja ärialase suhtluse lihtsustamiseks tõlgitud inglise keelde ning on Riigi Teatajast ajakohasena kättesaadavad – midagi, mille poolest me Euroopas silma paistame.

Seega eelistatavalt nopitakse määratletud õigustermineid asjakohastest teabesüsteemidest ja seadustest, kuid mugavat vahendit selleks veel olemas ei ole. Praegu tuleb näiteks seadusetõlgetes kasutatud vasteid n-ö käsitsi otsida. Kohati aitaksid seda murekohta lahendada tõlkemälud, kuid ka neist tuleb teatud tingimustel vasteid otsida või saadud otsingutulemustest sobivaim välja sõeluda. Keskses tõlkekeskkonnas on see lahendatud terminibaaside liidestamisega. Eesti Keele Instituudi sõnastiku- ja terminibaasisüsteem Ekilex on mõeldud teabe koondamiseks paljudest sõnakogudest ja terminibaasidest, mida on kokku juba üle 50. Tõlkijad kasutavad neid hoolega, mistõttu on nende kaasamine igati õigustatud. Samas on kavas hõlmata ka ELi terminibaasid, näiteks IATE. Kuid terminite väljaotsimine ja võrdlemine on suures osas käsitsitöö. Terminikogude sidumine võimaldab ajasäästu tõlkijatele terminite otsimisel ja paremat kvaliteeti, sest tõlked on ühtlustatumad ja tõlkija saab terminite jahtimise asemel keskenduda rohkem sisule ja loetavusele.

Paistab, et vastete leidmine tühjalt lehelt tõlkimist alustades on seega suuremalt jaolt lahendatud? Keskses tõlkekeskkonnas enamasti aga ei alustata nullist ja aega aitab säästa aina täiustuv ja juurde õppiv valdkondlik masintõlkemootor. Masintõlkega on senini aga olnud mure, et see töötab lausepõhiselt ja kasutab lausest lausesse erinevaid vasteid. Seega võib ühes lõigus olla ühel mõistel nii palju erinevaid vasteid, kui tal on esinemiskordi. Siin ilmneb vastuolu masintõlke iseõppimise ja olemasolevate terminite kasutamise kohustuslikkuse vahel.

Õigustõlkes ei kasutata sünonüüme ning juba kasutusel olevad terminid ja definitsioonid peavad olema kasutuses trafaretselt, sünonüümide ja eranditeta. Seetõttu ei ole õigustekstis võimalik alati vältida samade terminite kordamist ühes ja samas lauses. Nii peaks ka õigusvaldkonna masintõlge eelistama juba seaduste tõlgetes kasutatud ja terminibaasidesse kantud tähendusi ja termineid. See aga on mõnevõrra vastuolus masintõlkemootorite iseõppimise põhimõttega, mis võib termineid asendada ja muudel alustel eelistada. See ei käi ühte jalga ühtsuse, läbivuse ja trafaretsusega – kolme õigusterminoloogia alustalaga. Õiguskeeles peab olema selge, millest räägitakse, ning olema kindel, et seda tehakse läbivalt ja ühtemoodi. Ainult nii võib sedavõrd keerukas valdkonnas selgust luua. Sünonüümidel ja ilustatud keelel siin ruumi ei ole. Õnneks kuulub masintõlketehnoloogia suuremalt jaolt nende lahenduste hulka, mis kunagi lõplikult valmis ei saa. Kuigi masintõlke juured ulatuvad 1950ndatesse, on see kiiresti arenema hakanud võrdlemisi hiljuti, alates 2014. aastast tehisnärvivõrkudel põhinevate neurotõlkemootoritega.

Järjest on hakatud maadlema ka nende probleemidega, mis seni on masintõlke kasutust piiranud või selle vastu ebausku tekitanud – seda, et masin läheneb tekstile lausehaaval, mistõttu loogiliselt omavahel lauseid ei seo, ja kasutab vasteid meelevaldselt. Nüüd suudavad uuemad mootorid juba vaadata ka eelnevaid ja järgnevaid lauseid ning aina enam on hakatud uut lähenemist kombineerima muude lahendustega, sealhulgas vanade masintõlkelahendustega, näiteks töödeldes teksti enne või pärast neuromasintõlget.

Kokkuvõttes tähendab see, et õigeid lahendusi pakkuv süsteem on aina keerukam ja annab häid tulemusi kitsastes valdkondades, kus on kindel terminoloogia. Kui valdkond on ette määratud ja terminoloogia on olemas, saab seda masintõlkele nii-öelda peale suruda ja pakutud tõlked on tõlkijale meelepärasemad. Kuid samas tasub olla ettevaatlik mugavustunde tekkimise ja masintõlke liigse usaldamisega, sest mida parem on masintõlge, seda ohtlikum see on – lihtsamalt jäävad märkamata vead, mis suurema süvenemiseta tunduvad siiski piisavalt lihtsad ja loogilised, et neid tähelepanuta jätta. Tõlkemälu kasutuselevõtuga võivad sellised vead kanduda tõlkest tõlkesse.

Seega ei ole võimalik aina mahukamate andmete ja lahenduste kuhjamisega täielikult inimest asendada, tõlkimine on selleks piisavalt keeruline, autori arvates eriti õigusvaldkonnas. Mida usaldusväärsem on masin, seda valvsam peab olema tõlkija või toimetaja.

Keskse tõlkekeskkonna lähteanalüüsi, prototüübi ja palju muuga on võimalik tutvuda projekti veebilehel.[13]

[13] Keskse tõlkemooduli ärianalüüs ning prototüübi loomine.