

Tehisaru¹ toel mitmekeelselt: Euroopa Komisjoni teenused teadlikule kasutajale

Kristiina Suviste²

Euroopa Komisjoni kirjaliku tõlke peadirektoraadi tehnoloogiaosakond

Tehisintellekt muudab keelekasutust

Viimastel aastatel on tehisintellekti areng toonud kaasa murrangulisi muutusi mitmes valdkonnas, sealhulgas keeletehnoloogias. Masintõlkesüsteemid, mis varem piirdusid lihtsate tekstide automatiseeritud tõlkimisega, on nüüdseks arenenud üsna usaldusväärseks töövahendiks ka valdkondlike ja erialatekstide, sealhulgas haldus-, meditsiini- ja õigustekstide tõlkimisel. Üha rohkem spetsialiste, kes ei ole elukutselised tõlkijad, kasutavad tehisintellektipõhiseid tööriistu oma igapäevases töös.

Selle arengusuunaga on kaasa läinud ka Euroopa Komisjon, mis pakub mitmekesisist valikut tehisarupõhiseid keele- ja tõlketööriistu, et toetada mitmekeelset suhtlust nii Euroopa Liidu institutsioonides kui ka liikmesriikides. Euroopa Komisjoni arendatud süsteemid, nagu eTranslation ja teised teenused, põhinevad usaldusväärsel tehnoloogial ning arvestavad Euroopa Liidu õiguskeskkonna, terminoloogilise järjepidevuse ja kvaliteedistandarditega. Erinevalt paljudest kommertsteenustest töötavad need tööriistad turvalises Euroopa Komisjoni keskkonnas, kus kasutajaandmeid ei talletata, jagata ega taaskasutata. See tagab, et ka tundlikke andmeid sisaldavaid tekste saab tõlkida masina abil, ilma et ohustataks turvalisust või andmekaitset. Need tööriistad ei ole mõeldud vaid tõlgete loomiseks, vaid ka keelelise ligipääsetavuse ja kommunikatsiooni tõhustamiseks eri keeleoskuse tasemetel.

Ajalooline areng: ECMT-st sai eTranslation

Euroopa Komisjon on keeletehnoloogia valdkonnas olnud teerajaja juba 1970. aastatest alates. Esimene reeglipõhine masintõlkesüsteem ECMT töötas kuni 2010. aastani, mil see asendus statistilisel meetodil põhineva MT@EC süsteemiga. 2013. aastal käivitatud MT@EC toetas kohe kõiki EL-i ametlikke keeli – 552 keelepaari. Järgnevate aastate jooksul tegi tehnoloogia tohutu arenguhüppe: 2017. aastaks jõuti neuromasintõlkeni, mis hõlmas tohutul hulgal tõlkecorpuseid ja Euroopa Liidu institutsioonide tõlkijate keeleoskust.

¹ Siin artiklis kasutatakse *tehisaru* ja *tehisintellekt* sünonüümidenä.

² Artiklis väljendatud seisukohad on autori isiklikud ega pruugi kajastada Euroopa Komisjoni ametlikku seisukohta.

Süsteemi uueks nimeks sai eTranslation ja selle kasutajaskond laienes programmi “Digitaalne Euroopa”³ rahastamise tõttu – lisaks Euroopa Liidu institutsioonidele said sellele juurdepääsu ka liikmesriikide ametiasutused, ülikoolid ja hiljem ka väikeettevõtted ja mittetulundusühingud.

Kuidas neuromasintõlge töötab?

Neuromasintõlge põhineb neuronitest koosneval struktuuril, mida on treenitud ulatuslike paralleelkorpuste andmetega, st tekstide kogumitega, kus sama sisu esineb mitmes keeles. Treeningu käigus õpivad mudelid tuvastama keelemustreid, süntaktilisi struktuure ja semantilisi seoseid. See võimaldab mudelitel teha keeleliselt loomulikke ja kontekstitundlikke tõlkeotsuseid, mida reeglipõhised või statistilised süsteemid ei suuda sama usaldusväärselt tagada. Olulist rolli mängib seejuures konteksti arvestamine lause- või isegi lõigutasandil, see aitab vähendada tüüpilisi masintõlkevigu, nagu mitmetähenduslike sõnade vale tõlgendamise või terminite sobimatu kasutus. Euroopa Liidu tõlkekorpused ning professionaalsete tõlkijate panus mudelite treeningusse tagavad, et tõlked jäävad stiililt ja terminoloogiliselt vastavaks ametlikule keelekasutusele. Tänu sellele sobib neuromasintõlge üha rohkem kasutamiseks ka õigustekstides, haldusdokumentides ja muudes ametlikes dokumentides.

Lisaks pööravad Euroopa Liidu arendatud tööriistad üha enam tähelepanu väikekeelte, nagu eesti, leedu või malta, paremale toetamisele, võimaldades kvaliteetsemat tõlget ka nende keelte puhul, millel on varem olnud vähem keeleandmeid ja tehnoloogilist tuge.

Praktilised tööriistad mitmekeelseks suhtluseks

eTranslation võimaldab tekste tõlkida 32 keele vahel – Euroopa Liidu kõigi 24 ametliku keele vahel ja lisaks veel mitmesse keelde, näiteks araabia, jaapani, hiina, norra, türki, vene ja ukraina. Süsteemi saab kasutada nii veebiliidese kui ka rakendusliidese kaudu. Unikaalse lisavõimalusena on võimalik valida tõlkemootoreid vastavalt tõlgitava teksti valdkonnale – olgu see õigustekst, Euroopa Liidu teematika, finantsvaldkonna tekst või üldisemad tekstid, nagu artiklid või veebilehed. Näiteks õigustekstide tõlkimiseks mõeldud tõlkemootor Court of Justice Case Law põhineb Euroopa Liidu Kohtu praktilal. Finantsvaldkonna tõlkemootorit on treenitud Euroopa Keskpanga ja liikmesriikide pankade tekstidel. Üldisemate tekstide tõlkimiseks on välja töötatud tõlkemootor General Text, et toetada mittespetsiifiliste ja laiemale avalikkusele mõeldud tekstide tõlkimist.

Terminoloogia ja kohandatavus

eTranslationi üks uusi võimalusi on lisada tõlkepäringule oma sõnastikke. See aitab täpsemalt tagada terminoloogilist järjepidevust väga spetsiifilise valdkonna tekstide masintõlkimisel. Näiteks on see eriti oluline õigustekstides. Kasutajad saavad lisada kuni

³ [Digitaalse Euroopa programm](#).

250 terminit keelepaari kohta ning tuleb meeles pidada, et lühikesed ja täpsed terminid annavad parima tulemuse.

Näiteks on võimalik luua asutusepõhiseid või isegi valdkonnapõhiseid sõnastikke, mis peegeldavad organisatsiooni spetsiifikat ja keelekasutust. See aitab vältida terminite varieeruvust, mis muidu võiks segadust tekitada.

Tõlkekvaliteedi tagamine

Neuromasintõlge on spetsialiseeritud tehisintellekti vorm (transformermudel), kuid uued tehnikad, näiteks suured keelemudelid, saavad seda veelgi täiustada. Nüüd on võimalik Euroopa Liidu tõlkijate kahe aastakümne jooksul tehtud varasemaid tõlkeid otseselt ära kasutada, et muuta masintõlke kvaliteet veelgi paremaks.

Kui kasutaja valib suvandi „Kasuta tehisintellekti abi“, võrdleb eTranslation tõlgitava dokumendi iga segmenti (lause, pealkiri jne) varasemate EL-i tõlgetega, et püüda leida sarnaseid tõlkeid või mittetäielikke vasteid (*fuzzy matches*). Kui mõni neist leitakse, töötleb tehisintellekt tõlgitavat lauset, et saada parim võimalik masintõlge. Kui vasteid ei leita, tõlgitakse lause tavalisel viisil.

Masintõlke kvaliteeti hinnatakse Euroopa Komisjonis testide ja kasutajate tagasiside põhjal. Komisjoni keeleteadlased teevad katseid, kus võrreldakse masintõlget inimtõlkega, mõõtes näiteks terminite õigsust, stiili ja grammatika täpsust. TER (*Translation Edit Rate*) on masintõlke kvaliteedi hindamiseks kasutatav mõõdik, mis näitab, kui palju muudatusi tuleb teha automaatselt tõlgitud tekstis, et see vastaks professionaalse tõlke tasemele. Mida väiksem on TER, seda parem on masintõlke tulemus – see tähendab, et vaja on vähem toimetamist. TER arvestab sisestamisi, kustutamisi, asendamisi ja ümberpaigutusi ning annab kvantitatiivse ülevaate vajadusest tõlke redigeerimise järele. Euroopa Komisjon kasutab TER-i ühe võrdlusnäitajana, et hinnata eTranslationi tõlkemootorite arengut ja sobivust ametlike tekstide jaoks. Euroopa Komisjonis kasutatakse tõlkemootorite kvaliteedi hindamiseks ka BLEU skoori (*Bilingual Evaluation Understudy*), mis põhineb masintõlke ja ühe või mitme inimtõlke kattuvusel ning annab tulemuseks skoori vahemikus 0 kuni 100 – mida suurem skoor, seda enam sarnaneb masintõlge inimtõlkega. Kvaliteedikontroll ei ole ühekordne tegevus, vaid seda tehakse pidevalt. Suuremaid tõlkemootoreid uuendatakse vähemalt kaks korda aastas ning enne ja pärast treeningut tehakse loomulikult ka kvaliteedikontroll, et näha, kas uuendatud tõlkemootorite pakutava masintõlke kvaliteet on parem.

Tekstide lihtsustamine: WebText ja Accessible Text

Masintõlke kõrval on Euroopa Komisjoni kirjaliku tõlke peadirektoraadi tehnoloogiaosakond välja töötanud tekstitööriistad WebText ja Accessible Text. WebTexti

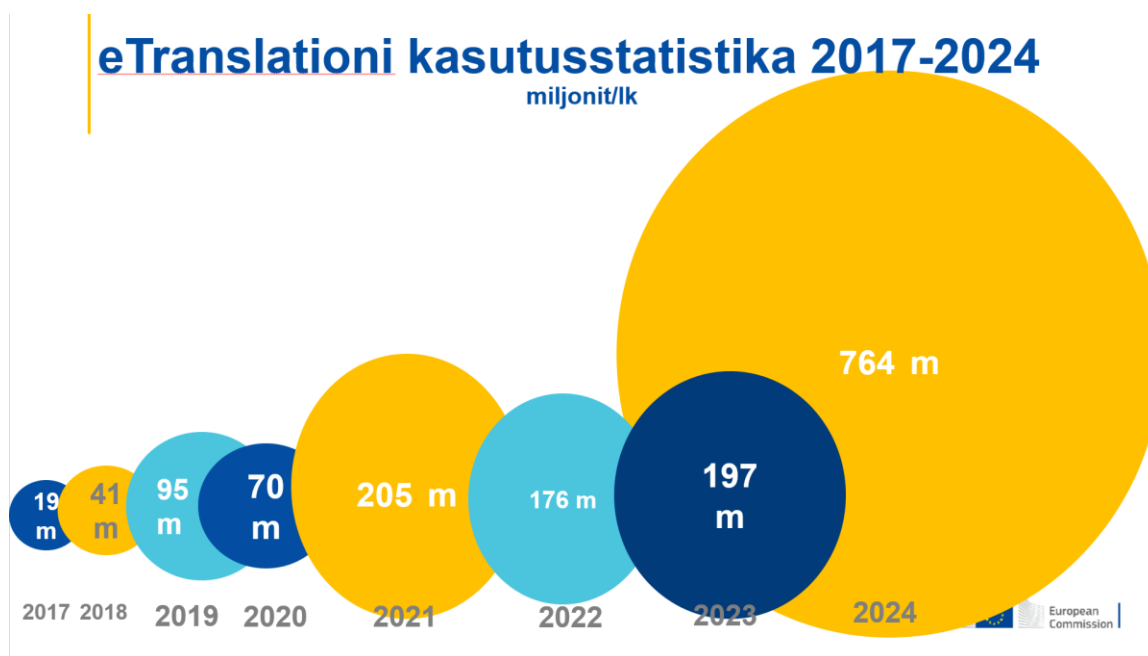
eesmärk on optimeerida tekste veebis kasutamiseks, et need oleksid arusaadavamad ja kaasavamad. Accessible Text järgib lihtsustatud keele standardeid, et muuta tekst arusaadavamaks inimestele, kellel on lugemisraskusi või kelle emakeel erineb teksti keelest. Mõlema tööriista üks eesmärkidest on edendada keelelist kaasatust.

Kasutajatugi

Kasutajad saavad valida eTranslationi veebiliidese ja rakendusliidese vahel. Veebiliides sobib individuaalseks tööks, samas kui rakendusliides võimaldab integreerimist olemasolevate infosüsteemidega. Kasutajatugi nii veebiliidese kui ka rakendusliidese kasutajale on kättesaadav e-posti teel ning pakub tehnilist tuge, koolitusmaterjale ja kasutusjuhendeid. Euroopa Komisjon korraldab regulaarselt ka töötubasid ja (veebi)seminare, et toetada teenuste oskuslikku kasutamist.

Statistika ja kasutusmahud

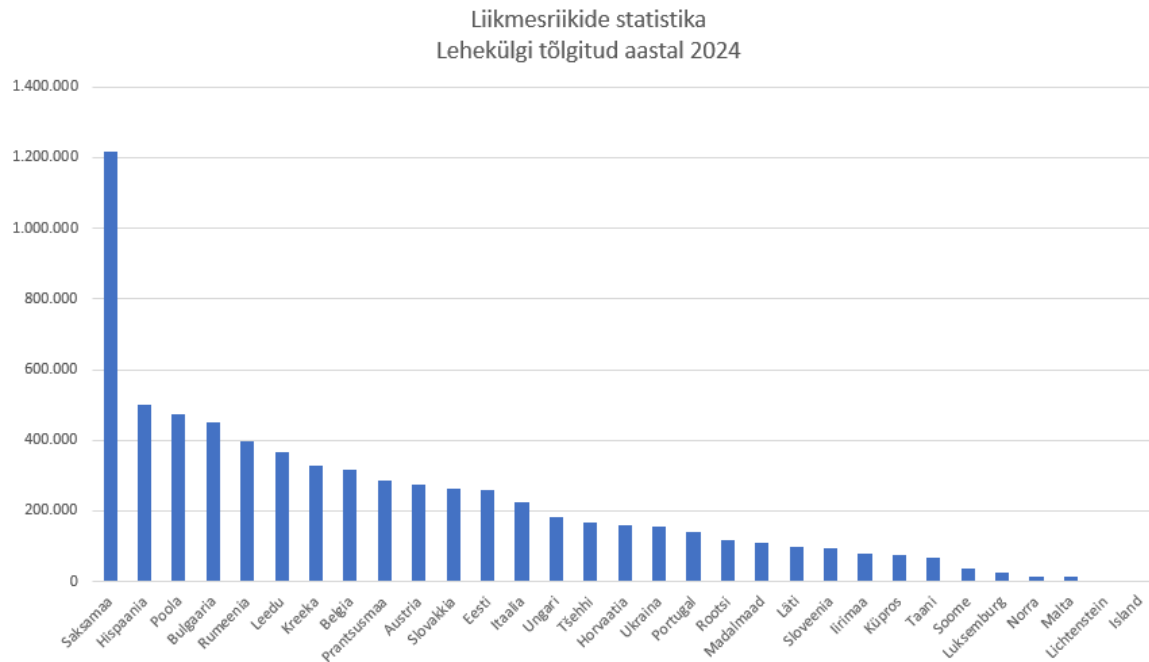
eTranslationi kasutamine on algusaegadest 2017. aastal praeguseks jõudsalt kasvanud: 19 miljonilt leheküljelt aastas 764 miljoni leheküljeni 2024. aastal. Aktiivseimad kasutajad on EL-i institutsioonid ja liikmesriikide haldusasutused.



Joonis 1. eTranslationi kasutusstatistika aastate jooksul

Kõige enam tõlgitakse dokumente inglise, prantsuse ja saksa keelde, kuid märkimisväärselt on kasvanud ka väiksemate keelte osakaal, näiteks leedu, soome ja eesti keel moodustavad üha suurema osa päringutest.

Allolevalt diagrammilt on näha, et Saksamaa, Hispaania ja Poola olid 2024. aastal tõlgitud lehekülgede arvu poolest eTranslationi kõige aktiivsemad kasutajamaad. Balti riikidest paistab silma Leedu ja Eesti stabiilselt suur kasutus – Leedu kuendal ja Eesti 12. kohal. Samas jääb Läti pigem tagasihoidlikumasse kasutajate gruppi. See viitab sellele, et Balti riigid integreerivad masintõlget oma haldusasutuste töösse järjest enam, eriti Eesti, kus digitaalne avalik sektor on hästi arenenud.



Joonis 2. Tõlgitud lehekülgede arv 2024. aastal riikide kaupa

Kasutusstatistika aitab Euroopa Komisjonil suunata arendustöid ning kaardistada kasutajate vajadusi ja uusi funktsioone.

Masintõlke piirangud ja riskid

Kuigi masintõlke kvaliteeti arendatakse pidevalt ja tulemused on üha paremad, tuleb siiski arvestada teatud piirangutega. Masintõlge võib näiteks eksida konteksti, idiomaatika või kultuuriliste vihjete tõlgendamisel. Näiteks võib sõna *strong* tähendada erinevates kontekstides *tugev*, *autoritaarne* või *pädev*.

Samuti võivad tekkida probleemid juriidiliste mõistete täpse edasiandmisega. Masintõlge ei pruugi alati tabada juriidiliste terminite täpset tähendust ega nende kasutust eri õigussüsteemides, mistõttu on elukutselise tõlkija või jurist-lingvisti kaasamine mõnes olukorras hädavajalik.

Näiteks on ingliskeelne termin *discovery*, mida tavakasutuses võib tõlkida kui *avastus*. Juriidilises kontekstis viitab see aga menetlustoimingule, mille käigus pooled jagavad üksteisega tõendeid enne kohtumenetlust.

Nende piirangute tõttu ei asenda masintõlge elukutselist tõlkijat, vaid täidab pigem abivahendi rolli, toetades tõlkijat, aga ka tavakasutajaid näiteks dokumentide esialgsete tööversioonide loomisel, korduvate tekstiosade kiiremaks töötlemiseks või mitmekeelses suhtluses taustmaterjalina. Lõplik keeleline ja sisuline vastutus jääb alati inimesele, kes hindab konteksti, valib sobiva stiili ja tagab terminoloogilise täpsuse, eriti valdkondades, kus ebatäpsused võivad põhjustada olulisi tagajärgi, nagu juriidikas, meditsiinis või avalikus halduses.

Turvalisus ja andmekaitse

Erinevalt kommertsteenustest töötab eTranslation Euroopa Liidu turvaliste kaitsemüüride taga, kus andmeid ei salvestata ega kasutata kommertseesmärkidel ega ka tõlkemootorite treenimisel. Andmekaitse ja konfidentsiaalsus on tagatud vastavalt Euroopa Liidu normidele, mis teeb sellest sobiva ja turvalise valiku avalikule sektorile ja teistele tundlikke andmeid sisaldavate dokumentide käsitlejaile. Tõlgitud tekstid kustutatakse süsteemist automaatselt 24 tunni jooksul või kohe, kui kasutaja sellise valiku teeb. Andmeid ei talletata, taaskasutata ega jagata.

Masintõlke praktilised kasutusvõimalused mittetõlkijatele avalikus ja erasektoris

Masintõlget ei kasuta ammu enam üksnes tõlkijad. Tänapäeval kasutavad eTranslationit ja teisi Euroopa Komisjoni tehisintellektil põhinevaid keeleteenuseid paljud eri valdkondade töötajad ametnikest ja teadlastest kuni juristide, ettevõtjate ja kommunikatsioonispetsialistideni. Alljärgnevad näited toovad esile, kuidas Euroopa Komisjoni loodud tööriistad toetavad praktiliselt ja turvaliselt igapäevast mitmekeelset suhtlust erinevates tööolukordades.

Õigusosakonna ametnik ministeeriumis kasutab eTranslation masintõlkesüsteemi seaduseelnõude ja määruste eeltõlkimiseks inglise ja prantsuse keelde. Tänu spetsiaalsele õigusvaldkonna tõlkemootorile ja võimalusele lisada ministeeriumi enda terminiloendeid on tõlked täpsemad ja järjepidevamad. Masintõlge kiirendab oluliselt tööprotsessi, eriti olukorras, kus esialgseid tööversioone on vaja kiiresti teiste liikmesriikide ekspertidele jagada.

Kohalikus omavalitsuses kasutab arendusosakonna töötaja eTranslationit selleks, et tõlkida infomaterjale mitmesse Euroopa Liidu ja kolmandate riikide keelde. See võimaldab pakkuda võrdset juurdepääsu teabele ka välismaalastele ja mitmekeelsele kogukonnale. Tõlked luuakse kiiresti, säilitades samas olulise teabe sisu ja arusaadavus.

Rahandusministeeriumi ametnik vajab iga nädal inglise-, prantsuse- ja saksa keelsete dokumentide eestikeelset tõlget. Kasutades eTranslationit, saab ta paari minutiga tõlgitud versiooni, mida toimetab, enne kui saadab kolleegidele.

Väikeettevõtja, kes soovib pakkuda oma klientidele infot oma kodulehel viies keeles, on integreerinud eTranslationi rakendusliidese või WebT sisuhaldussüsteemide pistikprogrammi kaudu automaatse tõlkefunktsiooni. Tulemused on piisavalt kvaliteetsed, et pakkuda küllastajatele arusaadavat sisu, samas säilitades kontrolli kohandatud terminoloogia üle.

Oluline on rõhutada, et hoolimata tööriistade üsna heast kvaliteedist tuleb inimesel endal kõik masintõlke abil loodud tekstid enne lõplikku kasutamist alati üle vaadata, et tagada keeleline täpsus, sobiv terminikasutus ja vastavus kontekstile. Sama kehtib kõikide generatiivsete tehisintellektil põhinevate tööriistade puhul – lõplik vastutus sisu õigsuse, sobivuse ja kasutamise eest lasub alati inimesel.

Tulevikusuunad ja arendused

Masintõlke kvaliteedi parandamine on pidev protsess, mis hõlmab nii tehnoloogiliste lahenduste täiustamist kui ka kasutajate tagasiside süstemaatilist arvestamist. Et masintõlketeenus eTranslation vastaks kasutajate vajadustele ja tehnoloogia arengule, on Euroopa Komisjonil plaanis mitmeid olulisi täiustusi. Oluline osa arengust on seotud Euroopa Liidu oma suure keelemudeli loomise, arendamise ja katsetamisega. Masintõlke ja suurte keelemudelite arendamiseks vajalike treeningandmete mahud on äärmiselt suured, mistõttu on nende töötlemiseks ja mudelite treenimiseks vaja kasutada tiiptasemel superarvuteid ja nende suurt arvutusvõimsust. On olemas spetsiaalne Euroopa suure jõudlusega andmetöötluse ühissetevõte EuroHPC JU⁴, mille kaudu arendatakse ja hallatakse Euroopas mitmeid maailmatasemel superarvuteid. Neuromasintõlke ja suure keelemudeli koos kasutamine võimaldab muuta masintõlget valdkonnateadlikumaks ning parandada üleüldist kvaliteeti. Kõige rohkem võivad sellest võita väiksemad keeled, kuna Euroopa Komisjonil on olemas ka masinate treenimiseks vajalikud juba eespool mainitud väga kvaliteetsed andmekorpused.

Kokkuvõte

Tehisintellektil põhinevad keele- ja tõlketööriistad, nagu masintõlkesüsteem eTranslation, on kujunenud oluliseks tugisambaks Euroopa Liidu mitmekeelse suhtluse ja halduse ligipääsetavuse tagamisel. Euroopa Komisjoni arendatud lahendused ühendavad keelelise täpsuse, turvalisuse ja kasutusmugavuse, võimaldades tõhusamat teabevahetust kõikides ametlikes keeltes ning toetades väiksemate keelte arengut.

⁴ Vt [The European High Performance Computing Joint Undertaking](#).

Lisaks tõlkijatele on eTranslationist ja teistest tehisintellektil põhinevatest teenustest kasu ka näiteks juristidele ja avaliku sektori töötajatele, kelle igapäevatöös on oluline täpne, kontekstitundlik ja turvaline mitmekeelne suhtlus. Õigustekstide, määruste, kohtupraktika ja haldusdokumentide tõlkimisel pakuvad valdkonnaspetsiifilised tõlkemootorid ja isikupärastatavad sõnastikud olulist tuge, aidates tagada terminoloogilist järjepidevust ning vähendades tõlkevigade riski. Kuna sageli esineb tekstides tundlikku teavet, on Euroopa Komisjoni turvaline IT-taristu ja andmekaitsemeetmed eriti olulised. Masintõlge ei asenda küll juriidilist ekspertiisi ega inimtõlget, kuid toetab oluliselt tööprotsessi tõhusust ja mitmekeelse dokumentatsiooni kättesaadavust kogu EL-i haldussüsteemis. Samas ei tohi unustada, et masintõlge on abivahend, mitte asendus elukutselisele tõlkijale – lõplik vastutus jääb alati inimesele.

Tulevikus, koostöös tipptasemel teadusvõrgustike ja superarvutitaristutega, nagu EuroHPC, liigub Euroopa Liit edasi omaenda suurte keelemudelite arendamise suunas. See võimaldab pakkuda veelgi kvaliteetsemaid, valdkonnaspetsiifilisemaid ja kaasavamaid keelelahendusi, mis toetavad nii haldust, õigust, haridust kui ka ettevõtlust kogu Euroopas.

Lisainfo: DGT-AI-Language-Services-Advisory@ec.europa.eu ja <https://language-tools.ec.europa.eu/>