Eesti keele digitaalne päästeoperatsioon: kuidas tehisaru meie emakeele selgeks õpib?

NELE NISU
Eesti Keele Instituudi andmeõiguse ja -poliitika juht, Tartu Ülikooli nooremteadur

KADRI VARE
Eesti Keele Instituudi keeletehnoloogia osakonna juhataja

Tehnoloogia areng on mitmekesine. Ühiskonnas on näiteks ootus, et ükskord leitakse vähiravim või saavutatakse läbimurre mõne harvikhaiguse ravis. Selleks tuleb võimaldada teha teadust ja uurida tekkinud andmeid. Soov alusandmeid uurida on igas valdkonnas – see puudutab ka eesti keelt.

Eestis ollakse üldiselt „krattide“ usku. Nii naudime Soome sõites automaatkontrolli, kus „masinlugeja“ võimaldab meil piirikontrolli läbida 15 sekundiga, metsa kaugseire infosüsteem aitab tuvastada lageraieid ning meditsiinis kasutatakse seadmeid, mis võimaldavad arstil lihtsa vaevaga hinnata peaaju pöördumatult kahjustunud ja päästetava koe mahtu (Kratid, i.a). Ammugi ei ole uudiseks mugavad, reaalaja automaatsubtiitrid, mis parandavad teabele ligipääsu. Selline arengusuund on ülemaailmne.

Stanfordi Ülikooli 2025. aasta tehisaru indeksi raportis tuuakse näiteks, et tehisaru liigub kiiresti laboriseinte vahelt igapäevaellu – tervishoiust transpordini –, kus üks USA suurimaid operaatoreid pakub nädalas üle 150 000 autonoomse sõidu (Stanford University, 2025). Meie igapäevaelu on muutunud väga mugavaks. Tarbijana on meil ligipääs kogu maailma e-kaubandusele, veebilehti saab otse veebilehitsejas eesti keelde tõlkida ning pärast pikka tööpäeva aitab silmi säästa ekraanilugeja, mis loeb soovitud teksti meile ise ette. Reisil olles on tõlkerakendused saanud asendamatuteks abimeesteks, kuid näiteks murdekeelte laialdase digitaalse toeni ei ole me veel jõudnud. Samas on see areng pidev protsess – keeleandmestik täieneb ning selle maht ja kvaliteet kasvab pidevalt. Tegu on väärtusliku ressursiga, mis võimaldab keelt põhjalikumalt uurida, luua uusi seoseid, leida uudissõnu ning

arendada eestikeelsele õppele üleminekuks vajalikke rakendusi, näiteks hääldusharjutusi (Eesti Keele Instituut, i.a-a).

Eesti keele arendamine tehisaru ajastul ja tingimustes saab toimuda üksnes tänu andmete kättesaadavusele. Selleni jõudmine on tihti nagu läbi kibuvitsapõõsa minek – kriibib ja kraabib küll, kuid ühel hetkel saab läbi. Üle maailma vaieldakse, kas ja mida võib autoriõigustega kaitstud sisuga teha ja mida mitte (Euroopa Liidu Intellektuaalomandi Amet, 2025). Seega on paljudel tekkinud küsimus, mida see tehnoloogia siis õigupoolest teeb ja miks käivad keeleandmete ümber vägagi laetud diskussioonid.

Kui tööd on digiteeritud, muutuvad need arvutusanalüüsi tooraineks.

TEHISARU KÜTUS ON ANDMED

Paljude valdkondade teadlaste ja teadustöö tegemise protsessi jaoks on probleem infoküllus – iga päev avaldatakse potentsiaalselt asjakohaseid tööks vajalikke materjale, millest teadlane peaks leidma sobiva osa, seda lugema ja analüüsima (Carroll, 2019). Sellises mahus materjali käsitsi läbitöötamine tänapäeva meetodite juures ei ole mõistlik.

Teksti- ja andmekaeve meetodid võimaldavad automaatselt töödelda ja analüüsida suurt hulka materjali, näiteks leida erinevaid mustreid ja seoseid. Teadlased teevad teksti- ja andmekaeve protsessi käigus andmetest koopiaid, kogudes ja koondades andmeid, vormindades neid andmetöötluseks ning säilitades neid tulemuste valideerimiseks.

Kui tööd on digiteeritud, muutuvad need arvutusanalüüsi tooraineks (Carroll, 2019). Suured keelemudelid ammutavad teadmisi tohututest treeningandmetest, et luua algoritmilisi protsesse, mis suudavad genereerida ja väljastada uut, sarnaste omadustega sisu (Euroopa Liidu Intellektuaalomandi Amet, 2025). Teksti ja andmekaeve mustrite ja seoste alusel avastatud leiud võivad avada täiesti uusi uurimissuundi ja seda mitte ainult keeleuurimisel. Näiteks kui geenide ja lihasfunktsiooni vahelise seose leidmine ootamatus kehaosas tuvastaks ravimite uusi rakendusvõimalusi, aga annaks ka uusi teadmisi teatud tüüpi geenide ja lihaste vahelisest vastasmõjust. Need uuringud võivad aidata konkreetset uurimissuunda edasi arendada. Veel enam – keskenduda ei tohiks ainult tulemustele, näiteks teatud haiguse eest vastutava geeni leidmisele, vaid ka uute uurimissuundade avastamisele. Teisisõnu, teksti- ja andmekaeve ei aita teadlastel leida üksnes õigeid vastuseid, vaid võib aidata leida ka õigeid küsimusi (Carroll, 2019). Sisuliselt tugineb tehisaru arendamine masinõppe tehnikatele, mis annavad algoritmile võime parandada oma tulemuslikkust kogemuste kaudu (Meys, 2020).

Tehisaru ei pruugi tähendada üksnes suuri keelemudeleid, vaid ka neile tuginevaid rakendusi, millega inimesed iga päev kokku puutuvad. Kui tahame, et keeleõppe tehnoloogiad, tekstitöötlusvahendid või häälassistendid kasutaks head ja mitmekesist eesti keelt või isegi meie murdeid, peab arendajal olema ligipääs võimalikult laiale valikule alusmaterjalile, olgu selleks kõneandmed, viipekeelevideod või raamatud.

Just keeleuurimise tulemusel tekivad uued sõnad sõnaveebi (Eesti Keele Instituut, i.a-b). Keele uurimine, hoidmine ja arendamine on pidev protsess, mis ei alga ega lõppe ühe konkreetse ajaga. Selle protsessi aluseks on keeleandmed, kus ilma mitmekesiste ning ajas uuenevate keeleandmeteta pole võimalik keelt uurida, kirjeldada ega põhjendatud keelenõu anda.

Keele uurimisel ja keele kohta järelduste tegemisel ei saa lähtuda üksnes ametlikest

tekstidest, nagu seadustest ja kohtulahenditest Riigi Teatajas, mis ei kuulu autoriõigusega kaitstava teabe hulka (Autoriõiguse seadus, 2025). Kui keeleandmed pärineks ainult sellistest allikatest, oleks ka keelekirjeldus üsna ühekülgne ega kajastaks tegelikku keelt. Sellisel juhul hakkaksid ka tehnoloogilised lahendused kasutama üsna kantseliitlikku ja ühekülgset keelekasutust. Ka sõnaraamatuid (õigekeelsussõnaraamatut, seletavat sõnaraamatut) ei saa koostada ainult nende tekstide pinnalt, sest tervikliku pildi keelest sellisena, nagu see tegelikult on, annavad ka ilukirjandus, ajakirjandus, argikeel ja släng. Seetõttu ongi hädavajalik keeleandmeid pidevalt koguda ja uuendada. Ainult nii on võimalik mõista keele arengut, teha põhjendatud keeleotsuseid ning tagada, et keelekirjeldus vastab elavale keelele, mitte ei ole mineviku peegeldus.

Seega on keeleandmed vältimatu alus nii klassikalisele keeleteadusele kui ka laiemalt humanitaar- ja sotsiaalteadustele ning rahvusteadustele, võimaldades keelt uurida, kirjeldada ja mõtestada. Selle arendamine on riigi vastutada, seda ei tee keegi meie eest. Samal ajal on keeleandmetel kujunemas ka uus roll – need on aluseks võrdlusandmete loomisele, mille abil hinnata teaduspõhiselt suurte keelemudelite keelekasutust, selle õigsust ja kvaliteeti.

RIIGI KOHUSTUS ON TAGADA KEELE PÜSIMAJÄÄMINE

Tehisaru areng on olnud viimastel aastatel avalikkuse tähelepanu ja arutelude keskpunktis. Kiire areng ja laialdane kasutus on tekitanud küsimusi just autoriõiguse vallas (Euroopa Liidu Intellektuaalomandi Amet, 2025). Ühelt poolt tuleb arvestada autorite õigustega, teisalt avaliku huviga eesti keele säilimise ja uurimise vastu.

Eesti põhiseaduse preambul rõhutab, et riik peab tagama eesti rahvuse, keele ja kultuuri säilimise läbi aegade (Põhiseadus, 2025). Keele mainimine preambuli teiste aluspõhimõtete hulgas tähendab eesti keele tunnistamist rahvuse südamikuna ning on eesti rahvuse säilimise põhiseaduslik garantii. Mida suurem on preambuli sätete üldistusaste, seda laiem on selle roll põhiseaduse ülejäänud sätete üle. Ka Eesti Riigikohus on lähtunud preambuli normatiivsest mõjust (Eesti Vabariigi põhiseaduse kommentaarid, 2020). Keele hoidmisel ja arendamisel ei piisa sellest, et eesti keel on riigikeel, vaid vaja on ka toetavaid tegevusi, et erinevad tehnoloogilised lahendused kasutaks korrektset eesti keelt ja keelekasutus ajas ei hääbuks.

Inimestele peavad olema eesti keeles kättesaadavad riigi ja kohaliku omavalitsuse asutuste poolt tagatud meditsiin, haridus, õigusabi jms avalikud teenused, samuti on ühiskonna toimimiseks vältimatu toimiv suhtlus ja sotsiaalne sidusus selle liikmete vahel. Seetõttu on põhjendatud, et riik edendab eesti keele kasutamist ka neis ühiskonnaelu valdkondades, mis ei ole otseselt seotud avaliku võimu teostamisega, näiteks äris, kultuuris, hariduses ja ajakirjanduses. Põhiseaduse kommentaarides rõhutatakse sedagi, et koostoimes preambuliga tuleneb §-st 6 riigi kohustus tagada eesti keele igakülgne arenemine teadus- ja kultuurkeelena (Eesti Vabariigi põhiseaduse kommentaarid, 2020).

Nagu artikli alguses selgitatud, eeldab igasugune tehnoloogia arendamine andmeid, olenemata sellest, kas need sisaldavad isikuandmeid või autoriõigusega kaitstud sisu. Küsimus ei seisne niivõrd õiguste olemasolus, vaid nende selges ja üheselt mõistetavas tõlgendamises, mis võimaldaks tehnoloogiat – sealhulgas tehisaru – arendada.

Vähemalt isikuandmetega on teatud selgus loodud – andmekaitse üldmäärus rõhutatakse, et kui isikuandmeid töödeldakse teadusuuringute eesmärgil, tuleks seda määrust kohaldada ka sellise töötlemise suhtes, samuti hõlmab teadusuuring tehnoloogia arendust (Isikuandmete kaitse üldmäärus, 2016). Sõltumata sellest, milleks andmeid algselt koguti, ei loeta isikuandmete töötlemist

esialgse eesmärgiga vastuolus olevaks, kui andmeid töödeldakse teaduseesmärgil (Isikuandmete kaitse üldmäärus, 2016). Küll tuleks riigil ette näha sobivad kaitsemeetmed ja need on Eestis sätestatud isikuandmete kaitse seaduse §-s 6 (Isikuandmete kaitse seadus, 2026). Euroopa Andmekaitsenõukogu kinnitas samuti, et eristada tuleb treenimisfaasi ja nn toote kasutust, samuti selgitas nõukogu töötlemise aluseks olevaid sobivaid õiguslikke aluseid ning võimalikke riske (Euroopa Andmekaitsenõukogu, 2024). Selle pinnalt on selge, et isikuandmeid saab töödelda treenimiseks (tehnoloogia arendamiseks). Paraku ei ole autoriõigustega olukord sama selge.

Vaja on ka toetavaid tegevusi, et tehnoloogilised lahendused kasutaks korrektset eesti keelt ja keelekasutus ajas ei hääbuks.

Sarnaselt isikuandmete töötlemisega on ka autoriõigusega kaitstud teoste kasutamisel eristatud õiguskirjanduses erinevaid etappe. Autoriõiguse seaduse kohaselt on teadus- ja kultuuripärandiasutustel õigus autori nõusolekuta ja autoritasu maksmiseta reprodutseerida teadusuuringute eesmärgil teksti- ja andmekaeveks teoseid, millele neil on seaduslik juurdepääs (Autoriõiguse seadus, 2025).

Sätte tõlgendamiseks puudub kohtupraktika ning selle aluseks oleva direktiivi põhjendused ei anna praktika ühtlustamiseks ammendavat raamistikku (Euroopa Parlament ja Euroopa Liidu Nõukogu, 2019).

Teksti- ja andmekaeve tähendab automatiseeritud analüüsimeetodit, millega analüüsitakse digitaalkujul tekste ja andmeid, et saada teavet muu hulgas mustrite, suundumuste ja korrelatsioonide kohta (Autoriõiguse seadus, 2025). Tegu ei ole ammendava, vaid näidisloeteluga.

Erand on tehtud üksnes reprodutseerimiseks ehk koopia tegemiseks ning koopiaid võivad teha selle erandi alusel ainult teadus- ja kultuuripärandi asutused. Nagu eelnevalt selgitasime, siis töötlevad teadlased suurt infomahtu tänapäevaste meediavahenditega. Siinjuures on õiguskirjanduses diskussioone, kas on asjakohane tuua paralleel inimesega ja võrrelda andmete kaevandamist „õigusega lugeda". Kirjanduses on toodud, et teose lugemine või lihtsalt nautimine ei riku ühtegi autoriõigusega kaitstud õigust ja seega ei kujuta endast autoriõiguse seadusega seotud tegu, samas kui teksti- ja andmete kaevandamine toimub tingimata arvuti abil, mis teostab tehniliselt vajalikke sisemisi kopeerimistoiminguid ja mida tuleb seetõttu käsitada reprodutseerimisena (Konertz, 2025).

Praktikas on tekkinud küsimus, kes täpselt võib ja saab seda tänapäevast meetodit suurte andmehulkade jaoks teaduseesmärgil kasutada. On selge, et kogu innovatsiooni ei taga üksnes klassikalised teadusasutused, vaid tihti hoopis erasektor. Isikute ring, kes sellele erandile võib tugineda, on ka erinevates seadustes reguleeritud erinevalt, mis võib praktikas innovatsiooni takistada.

Teadusasutuseks peetakse autoriõiguse seaduse tähenduses teadus- ja arendustegevuse ning innovatsiooni korralduse seaduse § 16 lõikes 1 sätestatud tingimustele vastavat juriidilist isikut või asutust, sealhulgas ülikooli ja selle raamatukogu, samuti teadusinstituuti või muud asutust, mille peamine eesmärk on teha teadusuuringuid või tegeleda õppetööga, mis hõlmab ka teadusuuringuid, ja ta teeb seda mittetulunduslikul alusel või reinvesteerides kogu kasumi oma teadusuuringutesse või täites avalikes huvides olevaid ülesandeid nii, et teadusuuringute tulemused ei ole soodustingimustel kättesaadavad

ettevõtjale, kellel on otsustav mõju sellise asutuse üle (Autoriõiguse seadus, 2025).

Teadus- ja arendustegevuse ning innovatsiooni korralduse seaduse alusel loetakse teadus- ja arendustegevuse osalisteks teadlased, teadus- ja arendusasutused, ülikoolid, rakenduskõrgkoolid ning teised osalised, kes kavandavad, viivad läbi, korraldavad, rahastavad, toetavad või hindavad teadus- ja arendustegevust või avaldavad teadus- ja arendustegevuse tulemusi (Teadus- ja arendustegevuse ning innovatsiooni korralduse seadus, 2025). Kahe seaduse loetelu võimaldab erinevaid lähenemisi, kus esimene on veidi kitsam, teine laiem. Samas ei peaks määrav olema üksnes tegija vorm (teadusasutus või erasektor), vaid tegevuse eesmärk ja selle ühiskondlik kasu.

Praktikas võib juhtuda, et innovatsiooni tekitamiseks saab küll isikuandmeid töödelda, kuid kuna andmestikud sisaldavad ka autoriõigustega kaitstud teoseid, takerdub innovatsioon selle taha. See on eriti oluline keeleandmestike ehk keeletehnoloogiate, tarkvarade arendamisel, kus treeningandmed sisaldavadki enamasti autoriõigustega kaitstud sisu (tekste vms).

Kuigi sätete aluseks olev direktiiv võimaldab teha teadusuuringute raames erasektoriga koostööd, on sellegi piirid hägused. Näiteks mainitakse direktiivis selgitustes eraldi ära üks näide, kuid mitte teisi isikuid – näiteks tuuakse, et mõiste peaks hõlmama ka selliseid üksusi nagu teadusuuringutega tegelevad haiglad (Euroopa Parlament ja Euroopa Liidu Nõukogu, 2019).

Euroopas on vähemalt üks esimese astme kohtulahend, kus kohus leidis, et mittetulundusühing, kes tegeles keelemudeli arendamisega, võis seda teha viidatud teksti- ja andmekaeve erandi alusel. LAION, kes on Saksa mittetulundusühing ja tegutseb tehisaru turul, sai n-ö kraapimise teel avalikest allikatest andmeid, mis puhastati ja mugandati pildisüsteemide treenimiseks. Saksa kohus pidas töötlust autoriõiguse erandi alusel õiguspäraseks, sest tegevusel olid teaduslikud eesmärgid ja see oli suunatud uute teadmiste saamisele. Peale selle sedastas kohus, et tegevus ei olnud äriline, mida tõendas asjaolu, et saadud andmestik tehti avalikult tasuta kättesaadavaks, seda olenemata organisatsiooni rahastusest või töötajaskonnast (Euroopa Liidu Intellektuaalomandi Amet, 2025). Seega on kohtud tõlgendanud erandi sätteid lähtuvalt eesmärgist, mitte subjektist.

Kuigi nimetatud teksti- ja andmekaeve erand tehti teadlikult suurte andmehulkade töötluseks (Euroopa Parlament ja Euroopa Liidu Nõukogu, 2019), puudub taas ühine arusaam, kust jookseb lubatu piir – kas teadlane võib teksti- ja andmekaeve meetodil teha teadustööd suure hulga ilukirjanduse alusel või on see lubatav kahe raamatu või mõne lehekülje puhul.

Nagu eelnevalt selgitatud, ei sõltu teadustöö mitte lehekülgede arvust, vaid teadusuuringu eesmärgist ning seisneb enamasti keele laialdases uurimises. Lisaks lähtuvad teadus- ja arendustegevuse osalised akadeemilisest vabadusest, mille kohaselt on teadus- ja arendustegevuse tegijal õigus otsustada teadus- ja arendustegevuse sisu ja meetodite üle (Teadus- ja arendustegevuse ning innovatsiooni korralduse seadus, 2025). Vabadus peab tagama, et teadlane lähtuks oma töös teaduslikust metoodikast, mitte ühiskondlikust, majanduslikust või muust välisest survest. Teadusega tegelemiseks on ka andmete kogumine ja analüüs (Eesti Vabariigi põhiseaduse kommentaarid, 2020).

Miks siis jõutakse praktikas selle mahu küsimuseni? Autoriõiguse seaduse järgi tuleb erandite puhul kaaluda, ega need ei ole vastuolus teose tavapärase kasutamisega ega kahjusta põhjendamatult autori seaduslikke huve (Autoriõiguse seadus, 2025). Mitmes riigis on algatatud kohtuvaidlusi nii selle üle, miks autoriõigustega kaitstud teoseid kasutati treenimisel (arendamine), kui ka selle üle, et sarnased tulemid esinevad ka väljundandmetes (kasutus) (Euroopa Liidu Intellektuaalomandi Amet, 2025).

Õiguskirjanduses on treenimise puhul tõdetud, et erand ei ole vastuolus

tavapärase kasutusega ega kahjusta põhjendamatult autori seaduslikke huve. Näiteks kirjeldab Senftleben, et olukorras, kus tehisaru väljund peegeldab üksnes üldisi ideid, kontseptsioone ja stiile, on keeruline tuvastada konflikti teose tavapärase kasutamisega. Kuigi see võib avaldada kirjandus- ja kunstiteoste turule häirivat mõju, siis sellist ideede, kontseptsioonide ja stiilide abstraktsel tasandil toimuvat mittespetsiifilist konkurentsi ei saa pigem pidada piisavaks, et järeldada konflikti. Seega, kui tehisaru treenimiseks kasutatud teoste töötlemine viib väljundini, mis kajastab üksnes kaitsmata ideid, kontseptsioone ja stiile, ei ole alust rääkida konfliktist teose tavapärase kasutamisega kolmeastmelise testi mõttes (Senftleben, 2025).

Ent eristada tuleb treenimisfaasi ja tehnoloogia kasutust. Kui LAION-i kaasuses leidis kohus, et treenimine oli lubatud, siis teistes lahendites on vaidlus selle üle, kas avalikustatud väljundandmed on liialt sarnased algteostega, mis juhul oleks tegu lubamatu kopeerimise ja avalikustamisega (näiteks SNE *vs.* Meta ja New York Times *vs.* OpenAI) (Euroopa Liidu Intellektuaalomandi Amet, 2025). Küsimus on seega, kas treeningandmed esinevad ka väljundandmetes. Näiteks lahendis GEMA *vs.* Open AI leidis kohus, et tegu on lubamatu kopeerimise ning üldsusele avalikustamisega (Bird&Bird, 2025). Milliste järeldusteni jõutakse Euroopas teistes kohtuvaidlustes, näitab aeg.

Vähemalt „lubatud“ mahu küsimust ei ole seni eraldi tõstatatud ja erand näib olevat suunatud just suurte andmehulkade töötluseks. Näiteks LAION-i tegevus seisneb selles, et ta pakub enam kui viiest miljardist pildi ja teksti paarist koosnevat andmebaasi, mis viib kokku internetis avalikult kättesaadavate piltide hüperlingid ja pildi sisu kirjeldava tekstilise teabe. See andmestik uueneb iga kuu. Sisuliselt on LAION-i andmestik filtreeritud, puhastatud ja poolstruktureeritud alamhulk, mis on optimeeritud generatiivsete pildisüsteemide treenimiseks (Euroopa Liidu Intellektuaalomandi Amet, 2025).

DIGITAALNE KEELEARHIIV KUI VÕIMALUS UURIDA KEELEPÄRANDIT

Eelnevast tulenevalt tekib küsimus, kuidas käsitleda sellist keeleandmestikku pikemas perspektiivis. Keel on kultuuri kandja ning mõlemad on ajas muutuvad ja arenevad – kui tehisaru eesti keelt ei kasutaks, tähendaks see eesti keele mahajäämust. Tehisaru ajastul väljendub keele kui kultuuri kandja roll üha enam andmetes, mille kaudu keel säilib, areneb ja kandub edasi ka tehnoloogiasse. Nagu eelnevalt selgitatud, siis toimub keeleteaduses uurimine ja uuendamine pidevalt, see ei alga ega lõppe. Seetõttu tuleks kaaluda, kuidas suurendada selgust ja säilitada keeleandmestik, mis moodustab omaette väärtusliku andmekihi, mida teadlased saavad keele uurimiseks kasutada. Kas väärindatud keeleandmestik võiks olla uus digitaalne keelearhiiv ja teatud kontekstis kultuuripärand?

Kultuuripärandit käsitleti rahvusvahelises õiguses esimest korda 1957. aastal ning alates 1950. aastatest on UNESCO ja teised valitsustevahelised organisatsioonid välja töötanud hulga rahvusvahelisi lepinguid ja tekste selle kaitseks. Kultuuripärandi kaitset käsitlev rahvusvaheline õigus algas suhteliselt kitsaste eesmärkidega, nimelt kultuuriväärtuste kaitsmisega sõja ajal (Blake, 2008). Võib tõdeda, et raske on tõlgendada isegi põhimõisteid, nagu kultuuripärand, kultuuriväärtus ja inimkonna kultuuripärand, mis võivad sisaldada väga erinevat sisu, näiteks „materiaalne, rituaalne ja sümboolne kultuur“ või „keel kui kultuur, väärtused ja uskumused“ ning mõnel juhul võivad sinna kuuluda ka „ideed, ideoloogiad ja tähendused“ (Blake, 2008).

Sõltumata sellest, kuidas pärandit sisustada, defineerib autoriõiguse seadus igal juhul kultuuripärandi asutuse kui sellise, milleks on seaduse kohaselt avalik raamatukogu, muuseum, **arhiiv** või filmi- või audiopärandi **säilitamisega tegelev**

asutus (Autoriõiguse seadus, 2025). Arhiiv võib sisaldada erineva valdkonna ja vormiga teavet, sh andmeid. Näiteks on arhiiviseaduse mõttes dokument mis tahes teabekandjale jäädvustatud teave (Arhiiviseadus, 2019) ning arhiiviväärtuslikeks andmeteks on mitmed andmekogude andmed (Rahvusarhiiv, i.a-a). Andmed eksporditakse teatud aja järel kokkulepitud mahus (arhiivi)vormingusse ja edastatakse Rahvusarhiivile pikaajaliseks säilitamiseks (Rahvusarhiiv, i.a-b).

Arhiivi kasutus lähtub seaduses sätestatud kasutuspiirangutest, näiteks isikuandmete kaitsest või autoriõigusest ja sellega kaasnevatest piirangutest (Avaliku teabe seadus, 2026). Eeltoodut arvestades on ka keeleandmestik arhiiv, mida on vaja eesti keele hoidmiseks ja arendamiseks, et hoida kultuuri ja tagada keele kestlikkus. Digitaalne keeleandmestik ei ole pelgalt tehniline andmehalduse keskkond, vaid see aitab tagada kultuurilist järjepidevust keele säilimise kaudu (Eesti Keele Instituut, 2026). Digitaalne vorm ei välista kogu käsitlemist samaväärselt füüsilise andmekoguga. Kultuuripärandi asutuse definitsiooni aluseks olevas direktiivis ei ole samuti arhiivile ette nähtud kindlat vormi, mistõttu võib arhiiv olla digitaalse keeleandmeruumi vormis (Euroopa Parlament ja Euroopa Liidu Nõukogu, 2019).

Märgendatud ja töödeldud keelandmestik võimaldab luua mustreid ja teha järeldusi sõnade kooskasutusest, muutumisest, tekkest või sootuks hääbumisest. Digitaalne andmestik võimaldab leida uusi uurimisküsimusi ja arendada keeletehnoloogiaid nii õppijatele, õpetajatele kui ka ligipääsetavuse parandamiseks laiemalt. Andmestikku saavad kasutada teadlased, arvestades seadustest tulenevaid piiranguid (Avaliku teabe seadus, 2026). Digitaalsed arhiivid on digiajastul kriitilise tähtsusega ja nende tänapäevane lahtimõtestamine oleks kahtlemata oluline samm edasi.

Keeleandmestikku kui märgendatud ja digitaalset varamut tuleks käsitleda strateegilise ressursina, mille säilitamine ja kättesaadavus ei ole üksnes tehniline, vaid ka keele- ja kultuuripoliitiline küsimus, mis eeldab riigilt selget raamistikku, et andmeid nii hoida kui ka teadus- ja arendustegevuses kasutada.

Erandite puhul tuleb kaaluda, ega need ei ole vastuolus teose tavapärase kasutamisega ega kahjusta põhjendamatult autori huve.

KOKKUVÕTE

Tehisaru areng on loonud olukorra, kus keele püsimajäämine ei sõltu enam üksnes selle kasutamisest igapäevaelus, vaid ka sellest, kas keel on esindatud ja kasutatav digilahendustes. Eesti keele puhul tähendab see, et keele uurimine, arendamine ja tehnoloogiline rakendamine eeldab ulatuslikku, mitmekesist ja ajakohast keeleandmestikku. Ilma väärtusliku keeleandmestikuta ei ole võimalik luua kvaliteetseid keeletehnoloogilisi lahendusi ega tagada, et eesti keel toimiks täisväärtusliku teadus-, haridus- ja suhtluskeelena.

Samas on keeleandmete kasutamine tihedalt seotud õiguslike piirangutega, eelkõige autoriõiguse ja andmekaitsega, mille tõlgendused ei ole praktikas üheselt selged. See tekitab olukorra, kus tehnoloogiline areng võib takerduda õigusliku ebakindluse taha. Eriti oluline on eristada tehisaru arendamise etappe – treenimist ja kasutust – ning tagada, et õigusruum toetaks teadus- ja arendustegevust, kahjustamata samal ajal kellegi õigusi.

Eesti riigil lasub põhiseaduslik kohustus tagada eesti keele säilimine ja areng, mis

tehisaru ajastul tähendab muu hulgas aktiivset rolli keeleandmete kättesaadavuse, kasutamise ja säilitamise korraldamisel. Seetõttu on vaja kujundada selge, tasakaalustatud ja tulevikku suunatud raamistik, mis võimaldab keeleandmeid teaduses ja tehnoloogiaarenduses laiapõhjaliselt kasutada, tagades samal ajal isikute või õiguste omajate õiguste kaitse. Ainult nii on võimalik kindlustada, et eesti keel püsib elujõuline, arenemisvõimeline ja konkurentsivõimeline ka kiiresti muutuvas digitaalses maailmas.

KASUTATUD ALLIKAD

ARHIIVISEADUS (2019). RT I, 13.03.2019, 33. – https://www.riigiteataja.ee/akt/113032019033.

AUTORIÕIGUSE SEADUS (2025). RT I, 12.07.2025, 9. – https://www.riigiteataja.ee/akt/112072025009.

AVALIKU TEABE SEADUS (2026). RT I, 06.03.2026, 3. – https://www.riigiteataja.ee/akt/106032026003.

BIRD&BIRD (2025). Landmark ruling of the Munich Regional Court on copyright and AI training. – https://www.twobirds.com/en/insights/2025/landmark-ruling-of-the-munich-regional-court-(gema-v-openai)-on-copyright-and-ai-training.

BLAKE, J. (2008). On defining the cultural heritage, lk 61, 62, 68. Cambridge University Press. – https://www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/abs/on-defining-the-cultural-heritage/93909A5DCDC2A7F6A65E08897A9C9155.

CARROLL, M. W. (2019). Copyright and the progress of science: why text and data mining is lawful, lk 895, 900, 901, 903. UC Davis Law Review. – https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3531231.

EESTI KEELE INSTITUUT (i.a-a). Häälduslharjutused. – https://sonaveeb.ee/pronunciation-exercises/.

EESTI KEELE INSTITUUT (i.a-b). Uued sõnad EKI ühendsõnastikus – https://sonaveeb.ee/newwords.

EESTI KEELE INSTITUUT (2026). Tellitud analüüs. Kultuuripärandi asutuse määratlemine keeleandmeruumi kontekstis.

EESTI VABARIIGI PÕHISEADUSE KOMMENTAARID (2020). – https://pohiseadus.ee/sisu/3467.

EUROOPA ANDMEKAITSENÕUKOGU (2024). Arvamus 28/2024 tehisintellektimudelite kontekstis isikuandmete töötlemisega seotud teatavate andmekaitseaspektide kohta. – https://www.edpb.europa.eu/news/news/2024/edpb-opinion-ai-models-gdpr-principles-support-responsible-ai_et.

EUROOPA LIIDU INTELLEKTUAALOMANDI AMET (2025). Development of generative artificial intelligence from a copyright perspective, lk 20, 69 jj, 70, 72, 77 jj. – https://www.euipo.europa.eu/en/publications/genai-from-a-copyright-perspective-2025.

EUROOPA PARLAMENT JA EUROOPA LIIDU NÕUKOGU (2019). Direktiiv (EL) 2019/790, 17.04.2019, mis käsitleb autoriõigust ja autoriõigusega kaasnevaid õigusi digitaalsel ühtsel turul ning millega muudetakse direktiive 96/9/EÜ ja 2001/29/EÜ, PE/51/2019/REV/1, ELT L 130. – https://eur-lex.europa.eu/legal-content/ET/ALL/?uri=CELEX:32019L0790.

ISIKUANDMETE KAITSE SEADUS (2026). RT I, 06.03.2026, 10. – https://www.riigiteataja.ee/akt/106032026010.

ISIKUANDMETE KAITSE ÜLDMÄÄRUS (2016). Euroopa Parlamendi ja nõukogu määrus (EL) 2016/679 füüsiliste isikute kaitse kohta isikuandmete töötlemisel ja selliste andmete vaba liikumise ning direktiivi 95/46/EÜ kehtetuks tunnistamise kohta (isikuandmete kaitse üldmäärus), ELT L 119. – https://eur-lex.europa.eu/eli/reg/2016/679/oj/est.

KONERTZ, R. (2025). Künstliche Intelligenz und § 44b UrhG, lk 1260. – https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00002226.

KRATID (i.a). Politsei- ja Piirivalveamet, Keskkonnaagentuur, Regionaalhaigla, Eesti Rahvusringhääling, kasutuslood. – https://www.kratid.ee/kasutuslood-kratid.

MEYS, R. (2020). Data Mining Under the Directive on Copyright and Related Rights in the Digital Single Market: Are European Database Protection Rules Still Threatening the Development of Artificial Intelligence? Lk 457. GRUR International. – https://academic.oup.com/grurint/article/69/5/457/5827596.

PÕHISEADUS (2025). RT I, 11.04.2025, 3. – https://www.riigiteataja.ee/akt/111042025003.

RAHVUSARHIIV (i.a-a). Andmekogude arhiveerimine ja andmekogud. – https://www.ra.ee/arhiivihaldus/digitaalarhiivindus/andmekogude-arhiveerimine/.

RAHVUSARHIIV (i.a-b). Väärtuslikud andmekogud. – https://www.ra.ee/astra/site/databases.

SENFTLEBEN, M. (2025). TDM, GenAI and the Copyright Three-Step Test, International Review of Intellectual Property and Competition Law. – https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5373903.

STANFORD UNIVERSITY (2025). Artificial Intelligence, Index Report. – https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf.

TEADUS- JA ARENDUSTEGEVUSE NING INNOVATSIOONI KORRALDUSE SEADUS (2025). RT I, 12.07.2025, 1. – https://www.riigiteataja.ee/akt/112072025001.