doi:10.5128/ERYa16.04

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 16, 59–76

KEELEANDMETE ÕIGUSLIKU REŽIIMI MÕJU NENDE ABIL LOODUD KEELEMUDELITELE

Aleksei Kelli, Kadri Vider, Arvi Tavast, Krister Lindén, Ramūnas Birštonas, Penny Labropoulou, Age Värv, Irene Kull, Gaabriel Tavits, Carri Ginter

Ülevaade. Artikli eesmärgiks on selgitada, millises ulatuses mõjutab keeleandmetele kohalduv õiguslik režiim keelemudelite arendamist ja kasutamist. Autorid lähtuvad oma käsitluses protsessiskeemist, alustades algandmetest ning lõpetades keeletehnoloogiat sisaldavate valmistoodetega (nt kõneliidesega külmik). Keeletehnoloogias kasutatavad algandmed sisaldavad tihti autoriõiguslikult kaitstavaid teoseid, autoriõigusega kaasnevate õiguste objekte (esitus, salvestus) ja isikuandmeid (isiku hääl, isiku kohta käiv muu info), mida säilitatakse annoteerimata ja annoteeritud andmekogudes. Keelandmete õiguslikke küsimusi on juba varem uuritud. Õiguslikult on läbi uurimata aga keelemudelite õiguslikud aspektid. Autorid on seisukohal, et reeglina ei mõjuta keelemudelite edasist õiguslikku staatust kasutatud algandmete õiguslik režiim, sest autoriõigusega kaitstavad teosed mudelis pigem ei säili. Küll aga võib õiguslikke probleeme tekitada isiku hääle kasutamine keelemudelis. Autorid analüüsivad võimalikke lahendusvariante nende probleemide ületamiseks. Artiklis vaadeldakse ka uue autoriõiguse direktiiviga kehtestatavat andmekaeve regulatsiooni ja selle rakendamist keelemudelite loomise kontekstis.*

Võtmesõnad: autoriõigus, isikuandmete kaitse, keelemudel, keeletehnoloogia, teksti- ja andmekaeve

1. Sissejuhatus

Keeletehnoloogia arendamine põhineb keeleandmete kasutamisel. Keeleandmed on sageli hõlmatud erinevate õigustega (autoriõigus, autoriõigusega kaasnevad õigused, isikuandmed). Sedalaadi andmete kasutamine võib põhineda õiguse omaja nõusolekul või tema õiguste seadusest tuleneval piirangul, mis andmete kasutamist võimaldab. Kuna keeleandmete ja nende kasutamisega seonduvat temaatikat on

* Artikli valmimist on toetanud ERF-i projekt 2014-2020.4.01.16-0134 "Eesti Keeleressursside Keskuse (EKRK) ühendatud sisuotsing" tegevusest "Riikliku tähtsusega teaduse infrastruktuuri toetamine teekaardi alusel" ning Haridus ja Teadusministeeriumi keeletehnoloogia teadus- ja arendustegevuse programm "Eesti keeletehnoloogia 2018–2027"

eelnevalt põhjalikult uuritud (lähemalt vt Kelli jt 2018, Klavan jt 2018, Tavast jt 2013), siis kajastatakse neid küsimusi käesolevas artiklis üksnes niivõrd, kui see on vajalik tõstatatud probleemide analüüsimiseks.

Artikli põhifookuses on küsimus, kas ja millises ulatuses laienevad keeleandmete suhtes kehtivad õiguslikud piirangud nende andmete abil arendatud keeletehnoloogiale (KT). Autorite eesmärk on selgitada, mil määral tuleb keeletehnoloogia loomise protsessis arvestada autoriõiguse ja isikuandmete kaitse nõuetega. Kui võtame näiteks salvestatud telefonikõne, siis on ilmne, et selle salvestise koopiale laienevad autoriõigused ja andmekaitse normid. Sama ilmne on, et autoriõigus ja andmekaitse ei kohaldu külmiku hääleliidesele (kasutajaliidesele), ehkki viimast treeniti autoriõigusega kaitstud teoseid ja isikuandmeid sisaldava keelekorpuse abil. Piir algsete õiguste kehtivuse ja nende mõju lõppemise vahel peaks asuma kusagil eelpool toodud näidetes kirjeldatud olukordade vahepeal. Seaduses sellist piiri kehtestatud ei ole, mistõttu on vaja keeletehnoloogia arendajate jaoks õiguste piir võimalikult selgelt defineerida. Nii ongi autorid seadnud eesmärgiks vähendada õiguslikku ebaselgust, pakkudes välja, millistel tingimustel võiksid keeleandmeid katvad õigused lõppeda.

Kuivõrd artiklis tegeletakse keeletehnoloogia seisukohalt põhimõtteliste küsimustega, mis seonduvad keeletehnoloogia arendamis- ja kasutusvabadusega, kaasati artikli kirjutamisse nii keeletehnoloogе kui õigusekspertе, et tagada käsitluse laiapõhjalisus. Analüüs tugineb autorite ulatuslikel praktilistel keeletehnoloogia arendamise ja sellega seotud õiguslike probleemide lahendamise siseriiklikel ja rahvusvahelistel kogemustel, mille tõttu on artiklil praktiline suunitlus. Artikkel arendab edasi autorite eelnevat selleteemalist uurimust (vt Kelli jt 2019a).

2. Keeletehnoloogia arendamise etapid

Artiklis analüüsitud probleemide mõistmiseks on vaja eristada andmepõhise keeletehnoloogia arendamise etappe.

1. Algandmete kogumine (kirjalikud tekstid, kõnesalvestised, videod jne). Algandmed võivad sisaldada autoriõigustega kaitstud materjali ja isikuandmeid. Algandmete kogumine ei hõlma tavaliselt muid tegevusi peale andmete salvestamise, esmase puhastamise ja mõistlikkuse kontrolli. Andmete autoriõiguse ja neis sisalduvate isikuandmetega võivad olla seotud õiguslikud küsimused nagu nt autoriõigusega kaitstud teose kasutamise ja üldsusele kättesaadavaks tegemise lubatavus, isikuandmete kaitsenõuete järgimine. Probleemid tulenevad peamiselt asjaolust, et üldjuhul ei ole võimalik muuta algandmeid anonüümseks selliselt, et isikute tuvastamine või oluliste autoriõigusega kaitstud teoste osade reprodutseerimine oleks matemaatiliselt võimatu.

2. Andmestike või andmekogude koostamine (töötlemata tekstikorpused nagu Google News, Common Crawl[1] või OpenSubtitles[2], eesti veebikorpus 2013[3], kõnekorpused nagu eesti keele spontaanse kõne foneetiline korpus[4] jne). Tegemist on algandmetega, mis on kogutud ja korrastatud kindlat kriteeriumi silmas pidades (nt teatud kindlal teemal konkreetse piirkonna elanike kõnesalvestuste koondamine, et jäädvustada piirkonna aktsent). Kirjeldatud andmestikud või andmekogud on

[1] https://commoncrawl.org/
[2] https://www.opensubtitles.org/
[3] https://doi.org/10.15155/1-00-0000-0000-0000-0011FL
[4] https://doi.org/10.15155/1-00-0000-0000-0000-001A3L

tavaliselt nii suured, et üksikud teose osad või isikuandmed moodustavad vaid tühise osa tervikust.

Töötlemata andmetest erinevad andmestikud selle poolest, et andmete suur maht vähendab oluliselt võimalust leida teavet algandmete kohta ning seega piirata autoriõiguste ja isikuandmete kaitset.

Autoriõiguse ja isikuandmete kaitse vaatenurgast ei erine andmestikud algandmetest. Peamine praktiline erinevus seisneb selles, et andmete suure mahu tõttu on inimesel tehniliselt keeruline saada tevet selle kohta, et tema isikuandmed või teosed on lisatud andmekogusse. Samuti saab isiku andmeid andmekogumist selliselt eemaldada, et see ei mõjuta andmekogumi kasutatavust.

Andmekogu loomine hõlmab sageli panust andmete kogumisel, korraldamisel, indekseerimisel, esitamisel, majutamisel jms. Seetõttu on andmekogud reeglina autoriõiguslikult kaitstavad *sui generis* andmebaasidena.

3. Annoteeritud (märgendatud) andmekogude loomine (vormianalüüsiga tekstikorpus nagu Estonian National Corpus 2017[5], süntaktiliselt analüüsitud korpused (ingl *syntactically parsed corpora*) nagu eesti keele puudepank[6] jne). Annoteeritud andmekogu tähendab, et andmestikku (vt eelmine punkt) on analüüsitud ja andmetele on lisatud märgendus vastavalt uurimiseesmärgile. Annoteeritud andmekogude autoriõiguse ja isikuandmete kaitse ei erine töötlemata andmete autoriõiguse ja isikuandmete kaitsest – ehkki töötlemata andmete ja märgenduste puhul võivad autoriõiguse omajad olla erinevad. Märgenduskihte võib säilitada eraldi ja need võivad andmetest ka lahus olla. Tavapärane on töödelda originaalandmete koopiaid koos annotatsioonikihtidega nii, et saadud andmestik sisaldab kõiki algseid andmeid. Annotatsiooniga andmestiku loomine hõlmab andmete käsitsi, poolautomaatset või automaatset analüüsi.

4. Keelemudelid – sõnastikud, sõnaloendid, sagedusjaotused, n-grammid nagu Google'i n-grammid, eeltreenitud sõnavektorid (vt Grave jt 2018), eeltreenitud keelemudelid (vt Devlin jt. 2018). Need on andmepõhised keeletehnoloogilised tulemused, mis on välja töötatud teatud kindlal viisil eelnevalt kirjeldatud andmete või kogumite töötlemisel, kuid ei pruugi neid sisaldada. Keelemudelid püüavad keelekasutust modelleerida, st seda esitada või kirjeldada.

Mudeli loomine nõuab märkimisväärselt enam tööd, teadmisi ja (arvutuslikke) ressursse, kui on vajalik eelnevate etappide läbimiseks. Mudelite loomine hõlmab vähemalt algoritmi loomist või valimist, algoritmi rakendamist tarkvaras, riistvara seadistamist (võib sisaldada isegi kohandatud riistvara arendamist), hüperparameetri optimeerimist, mudeli valideerimist.

Harvadel juhtudel võivad mõned mudelitüübid olla mõeldud vahetult kasutamiseks (nt sõnaraamatud). Peamiselt kasutatakse mudeleid aga järgnevates etappides toodete loomiseks.

5. Pooltooted (kõnesünteesi mootor või visuaalsete objektide tuvastaja) ja **valmistooted** (kõneliidesega külmik). Käesolevas artiklis pooltooteid ja valmistooteid ei analüüsita, kuna keeleandmete õiguslik režiim ei oma neile mingit mõju.

Keelemudelite õiguslikke küsimusi peetakse erialakirjanduses reeglina kõige keerukamaks (vt nt De Castilho 2018: 1267). Käesoleva artikli autorid on seisukohal, et keeleandmetele kohalduvad õigused ei ulatu enamasti keelemudeliteni,

[5] https://doi.org/10.15155/3-00-0000-0000-0000-071E7L
[6] https://doi.org/10.15155/1-00-0000-0000-0000-00089L

mis tähendab, et autoriõigusest ja isikuandmete kaitsest tulenevad algandmete kasutamise piirangud ei laiene keelemudelitele. See tähendab, et mudeli levitamis- ja kasutuspiirangud ei ole algandmete õiguste omajate määrata.[7] Eelöeldu ei laiene tingimata kõikidele keelemudelitele. Mõnedes keelemudelites võivad sisalduda ka autoriõiguslikult kaitstavad teosed, autoriõigusega kaasnevate õiguste objektid (esitus, salvestis) ja isikuandmed.

Järgnevas uuritakse keelemudelite õiguslikku režiimi, eelkõige autoriõigusi, võttes arvesse Euroopa Parlamendi ja nõukogu direktiivi (EL) 2019/790 autoriõiguste kohta digitaalsel ühisel turul (edaspidi digiühiskonna direktiiv, DÜD), milles esmakordselt reguleeritakse *expressi verbis* teksti- ja andmekaevet Euroopa Liidu (EL) tasemel. Nimetatud direktiiv tuleb EL liikmesriikide õigusesse sisse viia 7. juuniks 2021 (DÜD art 29(1)), mistõttu analüüsitakse ka direktiivi ülevõtmisega seotud õigusteoreetilisi ja normitehnilisi probleeme. Lisaks digiühiskonna direktiivile on asjakohane ka 2001. aastal vastu võetud infoühiskonna direktiiv 2001/29/EÜ (IÜD) ja teised Eesti autoriõiguse seaduse aluseks olevad direktiivid. Isikuandmete kaitse seisukohalt on aktuaalsed 2016. aastast pärinev isikuandmete kaitse üldmäärus (EL) 2016/679 (edaspidi ÜM) ja 2019. aastal jõustunud uus isikuandmete kaitse seadus (IKS).

3. Keelemudelis sisalduva keeleandmestiku autoriõiguslik kaitse

Selleks, et vastata küsimusele, kuidas autoriõigus mõjutab keelemudeleid, tuleb esmalt lühidalt välja tuua kriteeriumid, millele autoriõigusega kaitstav teos peab vastama. Eesti autoriõiguse seaduse kohaselt loetakse teoseks

> "mis tahes originaalset tulemust kirjanduse, kunsti või teaduse valdkonnas, mis on väljendatud mingisuguses objektiivses vormis ja on selle vormi kaudu tajutav ning reprodutseeritav kas vahetult või mingi tehnilise vahendi abil" (AutÕS § 4(1)).

Peamine nõue autoriõiguse tekkimiseks on teose **originaalsus**, st teos on autoriõigusega kaitstav üksnes siis, kui see on originaalne. Originaalsuse kriteerium määrab ühtlasi, kas mudeli loomisel sisendina kasutatud andmetele või mudelis sisalduvatele andmetele kohaldub autoriõigus. Huvitaval kombel pole seda peamist kriteeriumi rahvusvahelistes lepingutes või Euroopa Liidu õiguses üheselt defineeritud.[8] Peamiselt on teose originaalsuse kriteeriumi sisustamisega tegelenud Euroopa Liidu Kohus (EK), lahendades teose autoriõigusliku kaitsega seotud vaidlusi. EK ühes olulisemas otsuses "Infopaq …" (C-5/08) selgitas kohus, et originaalsus tähendab autori enda intellektuaalset loomingut.[9] Euroopa Kohus on samas lahendis väljendanud järgmist seisukohta:

> "Mis puutub kaitstavate teoste osadesse, siis tuleb märkida, et need koosnevad sõnadest, mis eraldi võetuna ei moodusta neid kasutanud autori intellektuaalset loomingut. Alles nende sõnade valik, kasutus ja kombineerimine võimaldas autoril väljendada oma loomingulist meelelaadi algupäraselt ja

[7] Oluline on, et mudeli looja võib seada mudeli kasutamisele piiranguid.

[8] Ent seda on defineeritud mõnes Euroopa Liidu direktiivis seoses spetsiifiliste teostega, nagu arvutiprogrammid. Tarkvaradirektiivi 2009/24/EÜ art 1(3) kohaselt määratletakse algupärasust (originaalsust) autori enda intellektuaalse loominguga.

[9] Ka Eesti autoriõiguse seadus sisustab originaalsust autori enda intellektuaalse loominguga. Vt AutÕS § 4(1).

> aitas tal jõuda tulemuseni, mis kujutab endast intellektuaalset loomingut" (C-5/08, punkt 45).

Käesoleva artikli kontekstis on teose originaalsuse nõue oluline järgmistel põhjustel. Esiteks, kui originaalsus puudub, ei ole andmekogumis sisalduv tekst kaitstud ja selle kasutamiseks ei ole tarvis luba küsida. Seega, isegi kui andmekogu teksti osad on mudelis kopeeritud, pole need autoriõigusega kaitstud. Teiseks, isegi kui tekst tervikuna on originaalne ja järelikult autoriõigusega kaitstud, jääb üles küsimus, kas mudelis kasutatud teksti osad on iseenesest originaalsed ja seega kaitstavad. Kuid kui mudelis kasutatavad teksti osad originaalsed ei ole (väga väikesed fragmendid), võib neid kasutada ilma, et selleks peaks luba küsima. Seega tuleb kindlaks teha mitte ainult kogu teose originaalsus, vaid kõikide selles kasutatud osade originaalsus. Euroopa Kohus on selgitanud, et

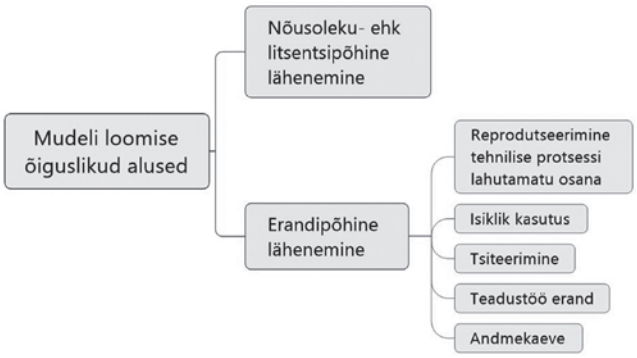
> "ei saa välistada, et teatavad eraldiseisvad laused või isegi lauseliikmed asjaomases tekstis võivad anda lugejale edasi niisuguse avaldatud teksti nagu ajaleheartikli algupärasust, edastades talle niisuguse tekstiosa, mis iseenesest on selle artikli autori enese intellektuaalse loomingu väljendus" (C-5/08, punkt 47).

Samas lahendis tõi kohus välja, et 11 üksteisele järgnevat sõna võib kujutada endast autoriõiguslikult kaitstavat teost (C-5/08, punkt 48). Sisuliselt tähendab see seda, et kohtu arvates peab juba 11 järjestikuse sõna kasutamisel tuginema autori nõusolekule või siis autoriõiguslikule piirangule (vt ptk 4).

4. Keeleandmete kasutamise õiguslikud alused keelemudeli loomisel

Autoriõiguslikult kaitstavate keeleandmete kasutamine mudeli loomiseks võib tugineda erinevatele õiguslikele alustele.

Lihtsustatult saab jagada mudeli loomisel andmete kasutamise õiguslikud alused kaheks: 1) teoseid sisaldavaid andmeid kasutatakse õiguste omaja nõusoleku (litsentsi) alusel; 2) mudeli loomisel tuginetakse autoriõiguslikele eranditele. Võimalikud õiguslikud alused on visualiseeritud lisatud joonisel 1.



Joonis 1. Mudeli loomise õiguslikud alused

Keelemudeli loomiseks viiakse läbi andmekaeve. Digiühiskonna direktiiv määratleb andmekaeve ning sätestab selle läbiviimise regulatsiooni. Direktiivis on defineeritud teksti- ja andmekaeve (edaspidi andmekaeve) järgmiselt:

> "automatiseeritud analüüsimeetod, millega analüüsitakse digivormingus tekste ja andmeid, et saada teavet muu hulgas mustrite, suundumuste ja korrelatsioonide kohta" (DÜD, art 2(2)).

Põhiline piiratav õigus mudeli loomiseks (andmekaeveks) on reprodutseerimisõigus (kopeerimisõigus), st mudeli loomiseks vajalikud andmed tuleb kõigepealt kopeerida.

Analüüsime järgnevalt muid keelemudeli loomise õiguslikuks aluseks olla võivaid aluseid peale nõusoleku (litsentsi).

4.1. Reprodutseerimine tehnilise protsessi lahutamatu osana

Infoühiskonna direktiivi art 5(1) kohaselt on EL liikmesriigid kohustatud jätma reprodutseerimisõiguse alt välja tehnilise reprodutseerimise. Eesti autoriõiguse seaduses on see sätestatud järgnevalt:

> "Ilma autori nõusolekuta ja tasu maksmata on lubatud teoste ajutine või juhuslik reprodutseerimine, mis toimub tehnilise protsessi lahutamatu ja olulise osana ning mille eesmärk on vahendada teose edastamist võrgus kolmandate isikute vahel või teha võimalikuks teose või autoriõigusega kaasnevate õiguste objekti seaduspärane kasutamine ning millel puudub iseseisev majanduslik eesmärk" (AutÕS § 18^1(1)).

Õiguskirjanduses on leitud, et tehnilise reprodutseerimise erand on teatud tingimustel kasutatav ka keelemudeli loomiseks (De Castilho jt 2018: 1272-1273). Seda võimalust on rõhutatud ka digiühiskonna direktiivis.[10]

4.2. Isiklik kasutus

Piiratud teadustöö erandiga riikides tuginevad teadlased andmekaeve teostamisel ja mudeli loomisel isikliku kasutamise erandile. Nimetatud erandi aluseks on infoühiskonna direktiivi artikli 5(2) punkt b. Autoriõiguse seaduse sõnastuse kohaselt saab isikliku kasutamise erandile tugineda üksnes füüsiline isik ning tema tegevus ei tohi taotleda ärilisi eesmärke (AutÕS § 18(1)).

4.3. Tsiteerimine

Keelemudeli loomisel saab andmete kasutamisel tugineda ka tsiteerimise erandile. Berni konventsiooni kohaselt on lubatud

[10] Digiühiskonna direktiivi selgituse kohaselt "võib esineda teksti- ja andmekaeve juhtusid, millega ei kaasne reprodutseerimine või mille puhul reproduktsioonid kuuluvad ajutise reprodutseerimise suhtes kohaldatava kohustusliku erandi alla, mis on sätestatud direktiivi 2001/29/EÜ artikli 5 lõikes 1, mille kohaldamist tuleks jätkata teksti- ja andmekaeve meetodite puhul, millega ei kaasne koopiate valmistamist väljaspool kõnealuse erandi kohaldamisala" (põhjenduspunkt 9).

> "tsiteerida teost, mis on juba õiguspäraselt üldsusele kättesaadavaks tehtud, kuid tingimusel, et tsiteerimine vastab ausale praktikale ja selle ulatus ei ületa eesmärgiga põhjendatud mahtu" (art 10(1)).

Sarnased piirangud tsitaatide kasutamisele sätestab ka infoühiskonna direktiiv, kus tsiteerimist piirab kooskõla mõistlike tavadega ning konkreetse eesmärgi jaoks kasutamise ulatuse vajalikkus.[11] Eeltoodut arvesse võttes tuleb küsida, millises ulatuses on keeletehnoloogia arendamisel tsitaatide kasutamine lubatud. Eesti autoriõiguse seadus nõuab tsiteerimisel autorile ja allikale viitamist, tsiteeritav objekt peab olema eelnevalt õiguspäraselt avaldatud, tsitaadi maht peab olema motiveeritud ning tsiteerimisel ei tohi moonutada tsiteeritava objekti mõtet. Kaasnevate õiguste regulatsioon eeldab kaasnevate õiguste objekti (nt esitus) tsiteerimisel informatsiooni andmise eesmärki (AutÕS § 19 p 1, § 75(1) p 4). Tsiteerimine on lubatud ka ärilisel eesmärgil. Euroopa Kohtu sämplit[12] puudutava lahendi kohaselt

> "peab kaitstud teost kasutava ja tsiteerimise erandile tugineda sooviva isiku eesmärk olema selle teosega dialoogi astumine" (C-476/17, punkt 71).

Tsiteeritud Euroopa Kohtu seisukoht muudab keeletehnoloogia arendamisel tsiteerimisõigusele tuginemise probleemseks. Sämpli (nagu ka tsitaadi) eesmärk on ikkagi sama, mis tsiteeritaval teosel, mõlemad kuuluvad samasse kunstivaldkonda ja mõlema puhul on oluline muusikaline kvaliteet, teksti sisu, väljendusviis vms. Keeletehnoloogia arendamise juures ükski neist eesmärkidest huvi ei paku, vaid uuritakse nendega mitteseotud teemat: missugune on eesti keel. Tekstifragment keeletehnoloogia protsessis reeglina muu keeleandmestikuga "dialoogi ei astu" (ei hakka muus andmestikus toodud seisukohti vaidlustama, kinnitama, muul viisil analüüsima). Seega pole küsimus dialoogi astumisest põhimõtteliselt rakendatav.

4.4. Teadustöö erand

Enne andmekaeve erandi kehtestamist tugines keeletehnoloogia arendamine teadustöö erandile, millel on oma õiguslik tähendus ka käesoleval ajal. Teadustöö erand on sätestatud infoühiskonna direktiivis ja erandi kohaselt piiratakse teadustöö eesmärgil reprodutseerimisõigust ja õigust üldsusele edastamisele ja kättesaadavaks tegemisele. Infoühiskonna direktiivi kohaselt on lubatud teose või sellega kaasnevate õiguste objekti kasutamine illustreeriva materjalina õppetöös või teadusuuringutes vajalikus ulatuses mitteärilisel eesmärgil, tuues ära allika ja autori nime, kui see on võimalik (IÜD art 5(3)a)). Autoriõiguse seadus sätestab teadustöö erandi mõnevõrra laiemalt. Autoriõiguse seaduse kohaselt on lubatud (autori ja allika äranäitamisel):

1) õiguspäraselt avaldatud teose kasutamine illustreeriva materjalina õppe- ja teaduslikel eesmärkidel motiveeritud mahus ja tingimusel, et selline kasutamine ei taotle ärilisi eesmärke ning

---

[11] Infoühiskonna direktiivi kohaselt "Liikmesriigid võivad artiklites 2 [reprodutseerimisõigus] ja 3 [õigus üldsusele edastamisele ja kättesaadavakstegemisele] sätestatud õiguste puhul näha ette erandeid või piiranguid järgmistel juhtudel:

> d) tsiteerimine kriitikas või ülevaadetes, kui tsitaadid on seotud teose või muu objektiga, mis on juba seaduslikult üldsusele kättesaadavaks tehtud, tingimusel, et märgitakse ka allikas, sh autori nimi, kui see ei ole võimatu, ja et tsitaatide kasutamine on kooskõlas mõistlike tavadega ning neid kasutatakse konkreetse eesmärgi jaoks vajalikus ulatuses" (IÜD art 5 (2)d)). Tsiteerimisele piirangute kehtestamine on aga jäetud liikmesriigi otsustada.

[12] "Eesti keele seletav sõnaraamat" määratleb sämplit järgmiselt: "helilõik, mis on salvestatud arvutisse, andmekandjale, süntesaatorisse v. samplerisse eesmärgiga seda helilõiku mingi töö loomisel kasutada" (EKSS: sämpel).

2) õiguspäraselt avaldatud teose reprodutseerimine õppe- ja teaduslikel eesmärkidel motiveeritud mahus haridus- ja teadusasutustes, mille tegevus ei taotle ärilisi eesmärke (AutÕS § 19 p-d 2 ja 3).

Digiühiskonna direktiivi põhjenduspunkti 10 kohaselt võidakse andmekaevet läbi viia teadustöö erandile tuginedes, viidates samas asjaolule, et infoühiskonna direktiivis sisalduva teadustöö erandi sisseviimine siseriiklikku õigusesse on vabatahtlik. Lisaks ei ole teadustöö erand piisavalt kohandatud tehnoloogia kasutamisele teadusuuringutes.

4.5. Andmekaeve

Autoriõiguse seaduse kohaselt on autori nõusolekuta ja autoritasu maksmiseta lubatud õiguste objekti töötlemine teksti- ja andmekaeve eesmärkidel. Sellisel tegevusel ei tohi olla ärilisi eesmärke ning autorile ja avaldamisallikale tuleb viidata (AutÕS § 19(3^1)). Positiivseks võib pidada, et andmekaeve erand laieneb nii autoriõiguslikult kaitstud teostele kui autoriõigusega kaasnevate õiguste objektidele (esitus, fonogramm, jmt). Mõnevõrra probleemne on asjaolu, et piiratava õigusena nimetatakse andmekaeve erandis töötlemisõigust. Töötlemisõigus on autori varaline õigus ning tähendab õigust "teha teosest kohandusi (adaptsioone), töötlusi (arranžeeringuid) ja teisi töötlusi (õigus teose töötlemisele) " (AutÕS § 13(1) p 5). Samas kattub töötlemisõigus õigusega teose puutumatusele (AutÕS §12(1) p 3), mis aga autori isiklik õigus.

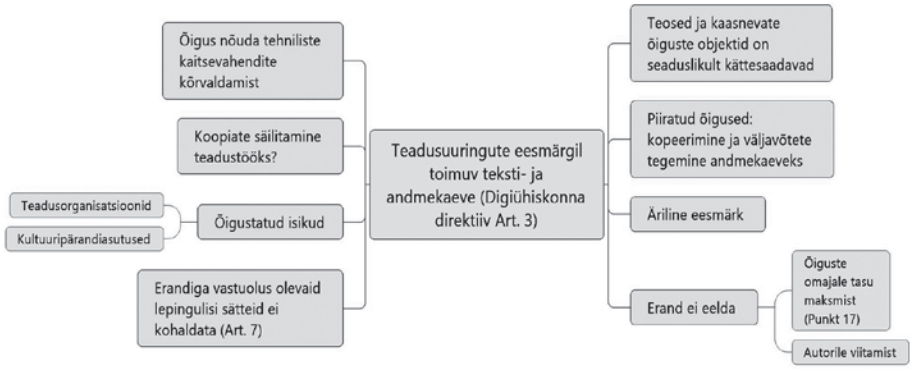
Isegi kui keeleandmete annoteerimist pidada teose töötlemiseks, on andmekaeve puhul põhiliseks piiratavaks õiguseks reprodutseerimisõigus.[13] Sellest loogikast lähtuti ka Justiitsministeeriumi egiidi all välja töötatud autoriõiguse ja autoriõigusega kaasnevate õiguste seaduse eelnõus (autoriõiguse seaduse eelnõu). Nimelt nägi eelnõu ette järgmise regulatsiooni:

> "Lubatud on õiguste eset kasutada järgmisel viisil, kuid tingimusel, et on ära märgitud õiguste omaja isik, õiguste eseme nimetus ning avaldamisallikas, välja arvatud juhul, kui taoline viitamine on võimatu:
> – õiguste eseme kopeerimine ja töötlemine teksti- ja andmekaeve eesmärkidel ning tingimusel, et selline kasutamine ei taotle ärilisi eesmärke" (autoriõiguse seaduse eelnõu § 43 (1) p 3).

Kokkuvõtvalt saab siiski asuda seisukohale, et andmekaeve regulatsioon kehtivas autoriõiguse seaduses on oma olemuselt positiivne, sest peegeldab seadusandja tahet andmekaevet lubada. Seega saab andmekaeve regulatsiooni tõlgendada selliselt, et ka teoste ja autoriõigusega kaasnevate õiguste objekti kopeerimine andmekaeve jaoks on lubatud.

Analüüsitud autoriõiguse seaduses olevat andmekaeve erandit tuleb peagi muuta tulenevalt digiühiskonna direktiivist, mis kehtestab kaks andmekaeve erandit. Esiteks erand, mis on avaram ja mis laieneb teadusasutustele (DÜD art 3) ning teiseks piiratum erand, millele saavad tugineda kõik (DÜD art 4). Käesolevas artiklis analüüsitakse esimest, mis on kokku võetud joonisel 2.

[13] Tegelikult tuleneb see ka digiühiskonna direktiivist, mis annab järgmise selgituse: "Teatavatel juhtudel võib teksti- ja andmekaeve hõlmata materjali, mis on kaitstud autoriõigusega või *sui generis* andmebaasi tegija õigusega või mõlemaga, eelkõige seoses teoste või muu materjali reprodutseerimisega või andmebaasist sisu väljavõtete tegemisega või mõlemaga, mis juhtub näiteks siis, kui andmed teksti- ja andmekaeve käigus normaliseeritakse" (põhjenduspunkt 8).



Joonis 2. Teadusuuringu eesmärgil toimuv andmekaeve

Digiühiskonna direktiivi kohaselt on õigustatud isikuks, kes saavad andmekaeve erandile tugineda, teadusorganisatsioonid ja kultuuripärandiasutused (DÜD art 3(1)). Digiühiskonna direktiiv määratleb teadusorganisatsiooni järgnevalt:

> "ülikool, kaasa arvatud selle raamatukogud, teadusinstituut või muu üksus, mille peamine eesmärk on teha teadusuuringuid või tegeleda õppetööga, mis hõlmab ka teadusuuringuid
>
> a) mittetulunduslikul alusel või reinvesteerides kogu kasumi oma teadusuuringutesse või
> b) täites liikmesriigi poolt tunnustatud avalikes huvides olevaid ülesandeid nii, et teadusuuringute tulemused ei ole soodustingimustel kättesaadavad ettevõtjale, kellel on otsustav mõju sellise organisatsiooni üle" (DÜD art 2(1)).
>
> Kultuuripärandi asutus digiühiskonna direktiivi kohaselt on "avalik raamatukogu, muuseum, arhiiv või filmi- või audiopärandi säilitamisega tegelev asutus" (DÜD art 2(3)).

Kuna digiühiskonna direktiivi siseriiklikku õigusesse sisseviimine seisab alles ees, siis tuleb otsustada, millises õigusaktis andmekaeveks õigustatud isikud defineerida. Loogiline tundub, et kultuuripärandiasutus määratletakse autoriõiguse seaduses. Teadusorganisatsiooni puhul tuleb arvestada, et Eestis kehtib teadus- ja arendustegevuse korralduse seadus (TAKS), mis määratleb teadus- ja arendusasutused (TAKS § 3). Asjakohane oleks säilitada teadusorganisatsiooni TAKS-i määratlus seda vajadusel digiühiskonna direktiivi valguses täiendades.

Digiühiskonna direktiiv lubab teadusuuringute eesmärgil toimuva teksti- ja andmekaeve jaoks reprodutseerida ja teha väljavõtteid autoriõiguslikult kaitstavatest ja *sui generis*[14] andmebaasidest, teostest ning ajakirjandusväljaannetest[15] (DÜD art. 3 (1)). Sel eesmärgil piiratakse infoühiskonna direktiivis sätestatud reprodutseerimisõigust (art. 2) ning andmebaasi direktiivist tulenevat autoriõiguslikult kaitstava andmebaasi autori reprodutseerimisõigust (andmebaasi direktiiv 96/9/EÜ art 5(a))[16] ja *sui generis* andmebaasi looja õigust keelata andmebaasi

[14] Autoriõiguse seaduse kohaselt on *sui generis* andmebaas teoste, andmete või muu materjali süstemaatiliselt või metoodiliselt korraldatud kogu, mis on individuaalselt kasutatav elektrooniliste või muude vahendite abil. Andmebaasi tegija on isik, kes on teinud kas laadilt, väärtuselt või suuruselt olulise investeeringu selle andmebaasi sisuks olevate andmete kogumiseks, omandamiseks, kontrollimiseks, süstematiseerimiseks või kättesaadavaks tegemiseks (AutÕS § 75^2, 75^3).

[15] Tegemist on digiühiskonna direktiiviga siseriiklikusse õigusesse sisse viidava uue vastuolulise kaasneva õigusega, mida antud artiklis lähemalt ei analüüsita (vt DÜD art. 15).

[16] Vt AutÕS § 13(1) p 1.

kogu sisust või kvantiteedilt või kvaliteedilt olulisest osast väljavõtte tegemine ja/või selle taaskasutamine (andmebaasi direktiiv 96/9/EÜ art 7(1)).[17] Samuti piiratakse digiühiskonna direktiivist tulenevat ajakirjandusväljaannete reprodutseerimisõigust (DÜD art 15(1)).

Andmekaeve erand ei piira andmekaeve teostamist üksnes mitteärilise eesmärgiga. Digiühiskonna direktiivist tuleneb, et teadusorganisatsioonid peaks saama andmekaeve erandile tugineda ka avaliku ja erasektori partnerluse raames (DÜD põhjenduspunkt 11). Direktiiv sätestab samuti, et andmekaevega tekkiv kahju õiguste omajatele on minimaalne, mistõttu seda ei pea neile kompenseerima (DÜD põhjenduspunkt 17).

Keskseks küsimuseks andmekaeve erandi rakendamisel saab suure tõenäosusega olema kasutatavale andmematerjalile seadusliku juurdepääsu nõue (DÜD art 3(1)).[18] Piirangud võivad siin olla nii lepingulised kui ka tehnilised. Lepinguliste piirangute osas näeb digiühiskonna direktiiv ette regulatsiooni, mille kohaselt andmekaeve erandiga vastuolus olevaid lepingu sätteid ei kohaldata (DÜD art 7(1)).

Tehniliste piirangutega on olukord mõnevõrra keerulisem. Ühelt poolt digiühiskonna direktiiv lubab õiguste omajal kohaldada tehnilisi kaitsemeetmeid.[19] Teiselt poolt on need küllalt piiratud. Direktiiv üritab tasakaalu saavutada järgmise sõnastusega:

> "Õiguste omajatel on õigus kohaldada meetmeid, et tagada nende võrkude ja andmebaaside turvalisus ja terviklikkus, kus majutatakse nende teoseid või muud materjali. Kõnealused meetmed ei tohi minna kaugemale sellest, mis on vajalik nimetatud eesmärgi saavutamiseks" (DÜD art 3(3)).

Tekib küsimus, mis saab siis, kui õiguste omaja läheb oma meetmetega kaugemale, kui viidatud artiklis lubatud (kasutab tehnilisi kaitsemeetmeid kopeerimise takistamiseks). Vastuse annab tehniliste kaitsemeetmete regulatsioon. Autoriõiguse seaduse kohaselt

> "teose või muu kaitstud objekti vaba kasutamise juhtudel peab õiguste omaja kohaldama oma teosele või muule õiguste objektile selliseid meetmeid, mis võimaldavad õigustatud isikul kasutada seda ulatuses, mis seadusega on vabaks kasutamiseks ette nähtud, tingimusel et vabaks kasutamiseks õigustatud isikul on seaduslik juurdepääs teosele või muule õiguste objektile" (AutÕS § 80^3(4)).

Ühtne praktika puudub aga juhul, kui õiguste omaja ei järgi seda kohustust ning piirab tehniliste meetmetega vabakasutust (nt andmekaevet).

Keeleteadlaste ja keeletehnoloogia arendajate jaoks on üheks olulisemaks küsimuseks, mida tohib teha andmekaeveks kasutatava andmekoguga. Digiühiskonna direktiivi kohaselt võib andmekaeveks tehtud koopiaid säilitada nõuetekohase turvalisuse tasemega "teadusuuringute, sealhulgas nende uuringute tulemuste kontrollimise tarvis" (DÜD art 3(2)). Kuna digiühiskonna direktiiv piirab

---

[17] Vt AutÕS § 75^4(2).

[18] Digiühiskonna direktiiv annab seoses seadusliku juurdepääsuga järgmise juhise: "Seaduslikku juurdepääsu tuleks mõista nii, et see hõlmab juurdepääsu sisule avatud juurdepääsu poliitika või õiguste omajate ja teadusorganisatsioonide või kultuuripärandiasutuste vaheliste lepinguliste kokkulepete alusel, nagu abonemendid, või muudel seaduslikel viisidel. Näiteks teadusorganisatsioonide või kultuuripärandiasutuste abonementide puhul tuleks eeldada, et nendega seotud isikutel, kellele asjaomane abonement laieneb, on seaduslik juurdepääs. Seaduslik juurdepääs peaks samuti hõlmama juurdepääsu sisule, mis on veebis tasuta kättesaadav" (DÜD põhjenduspunkt 14).

[19] Infoühiskonna direktiiv määratleb tehnilise kaitsemeetme mis tahes tehnoloogia, seadme või komponenendina, mille eesmärk tavapärase toimimise puhul on takistada või piirata teoste või muude objektidega seotud toiminguid, milleks autoriõiguse või sellega kaasnevate õiguste valdaja ei ole luba andnud (IÜD art 6(3)).

reprodutseerimisõigust, siis autoriõiguse ja kaasnevate õigustega kaitstavaid objekte sisaldavat andmekogu ei ole lubatud avalikult kättesaadavaks teha. Selleks on vaja õiguste omaja nõusolekut. Digiühiskonna direktiivist ei selgu üheselt, kas teadlased tohivad omavahel andmekogusid jagada teadustöö tegemiseks. Seda teemat peaks kindlasti käsitlema eelnõus ja selle seletuskirjas, millega direktiiv võetakse üle riigisisesesse õigusesse. Loodetavasti valitakse teadustööd soosiv lähenemine, mis lubab andmekogusid teadlaste vahel jagada.

5. Keelemudelis sisalduvate isikuandmete kaitse

Isikuandmete teemat keeletehnoloogia valdkonnas (rõhuasetusega häälel) on eelnevalt käsitletud (vt Kelli jt 2018, Ilin, Kelli 2019, Kelli jt 2019b), mistõttu siinkohal piirdutakse käsitlusega, mis on vajalik üksnes käesoleva artikli jaoks. Isikuandmete kaitse üldmääruse kohaselt on isikuandmeteks

> "igasugune teave tuvastatud või tuvastatava füüsilise isiku ("andmesubjekti") kohta" (ÜM art 4(1)). ÜM art 4 p 14 kohaselt on biomeetrilised andmed "konkreetse tehnilise töötlemise abil saadavad isikuandmed isiku füüsiliste, füsioloogiliste ja käitumuslike omaduste kohta, mis võimaldavad kõnealust füüsilist isikut kordumatult tuvastada või kinnitavad selle füüsilise isiku tuvastamist, näiteks näokujutis ja sõrmejälgede andmed".

Ka inimese häält saab lugeda biomeetrilisteks andmeteks, sest see kujutab endast teavet konkreetse isiku füsioloogiliste omaduste kohta, mille alusel saab seda isikut tuvastada, st isikute grupist eristada. ÜM art 4 p 1 viitab mitte üksnes "tuvastatud", vaid ka "tuvastatavale" isikule, seega piisab hääle isikuandmeteks lugemiseks juba sellest, kui inimest on hääle järgi põhimõtteliselt *võimalik* teistest eristada.

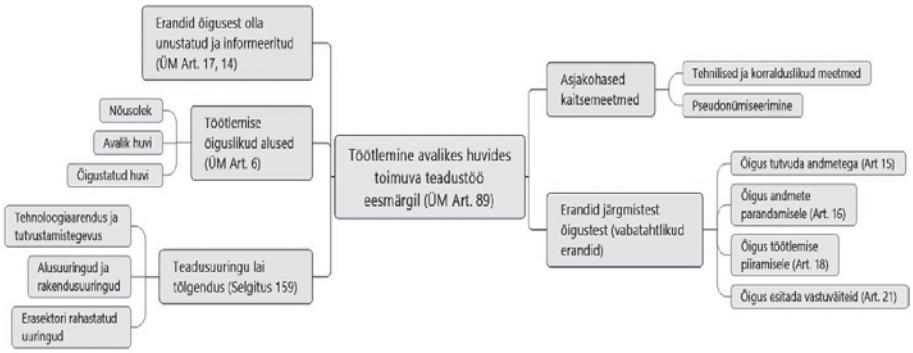
Isikuandmete kaitse alla ei kuulu anonüümsed andmed (vt ÜM põhjenduspunkt 26). Inimhääle salvestust on võimalik tehniliste vahenditega moonutada selliselt, et hääle põhjal ei ole enam võimalik ära tunda, kes konkreetselt salvestises kõneleb.[20] ÜM selgituste kohaselt tuleb füüsilise isiku tuvastatavuse kindlakstegemisel arvesse võtta kõiki vahendeid, mida vastutav töötleja või keegi muu võib füüsilise isiku otseseks või kaudseks tuvastamiseks mõistliku tõenäosusega kasutada (ÜM põhjenduspunkt 26), sealjuures tuleks "mõistliku tõenäosuse" all silmas pidada nt isiku tuvastamiseks ette võetavate meetmete maksumust ja selleks kuluvat aega, võttes arvesse ka tehnoloogilisi arenguid. Seetõttu võib asuda seisukohale, et inimkõne salvestisele ei kohaldu isikuandmete kaitse nõuded juhul, kui salvestist on tehniliselt töödeldud sellisel viisil, et konkreetse kõneleja häält ei ole enam võimalik ära tunda ning puuduvad tehnilised võimalused algsel salvestisel olnud hääle taastamiseks.

Kuna autoriõiguse alapunkti juures käsitleti teadustöö eesmärgil toimuvat andmekaevet, siis siinkohal esitatakse teadustöö eesmärgil toimuva isikuandmete töötlemise[21] joonis 3, mis aitab järgnevat selgitust paremini mõista.

Isikuandmete puhul on teoreetiliselt võimalik, et väikesed, kuid identifitseeritavad infoühikud jäävad keelemudelisse alles. Näiteks võib sõnaloend sisaldada nime või e-posti aadressi. Identifitseerivate infoühikute säilimist keelemudelis on võimalik vältida anonümiseerimise või pseudonümiseerimise abil. Seejuures tuleks pidada silmas, et isikuandmete seisukohalt on kaitstavad ka väga väikesed ühikud.

[20] Euroopa andmekaitseasutuste töörühm on seisukohal, et anonümiseerimine kui isikuandmete edasine töötlemine on kooskõlas isikuandmete algse töötlemise eesmärgiga eeldusel, et isikuandmed ka tegelikult anonümiseeritakse (WP29 2014: 7).

[21] Täiendavad nõuded isikauandete teaduseesmärgil töötlemiseks tulenevad isikuandmete kaitse seaduse §-st 6.



Joonis 3. Teadusuuringu eesmärgil toimuv isikuandmete töötlemine

Isikuandmete kasutamiseks keelemudelis peab oleama õiguslik alus, st isikuandmete töötlejal peab olema õigus isikuandmeid töödelda. Üldiselt võib mudelite loomiseks kasutatud isikuandmeid sisaldavate andmekogumite koostamine põhineda 1) isiku nõusolekul (vt ÜM art 4 p 11, art 7, WP29 2018) või olla lubatud 2) avaliku huvi (vt ÜM art 6 (1) p e) või 3) andmete töötleja või kolmanda isiku õigustatud huvi korral (vt ÜM art 6 lg 1 p f, WP29 2014a). Kui on olemas isiku nõusolek tema andmete töötlemiseks teaduseesmärkidel või kui töötlemine tugineb avalikele huvidele ja isikuandmeid sisaldavat saadud mudelit kasutatakse ka teaduseesmärkidel (seda ei tehta avalikkusele kättesaadavaks ega kasutata ärilistel eesmärkidel), on isikuandmete töötlemine õigustatud. Samuti ei teki õiguslikke probleeme juhul, kui isiku nõusolek hõlmab ärilist kasutamist ja avalikku levitamist (nõusoleku temaatikaga seoses vt WP29 2018).

Olukord muutub aga keeruliseks juhul, kui isikuandmeid sisaldavat andmekogumit töödeldakse uuringuks küsitud nõusoleku või avaliku huviga seotud erandi alusel, kuid saadud keelemudelit (kuhu võivad jääda isikuandmed) kavatsetakse kasutada ärilisel eesmärgil või teha üldsusele kättesaadavaks. Kirjeldatud olukorra lahendamiseks on järgmised võimalused:

1) töödelda isiku häält selliselt, et see ei ole kokku viidav konkreetse isikuga (anonümiseerida andmed);
2) küsida isiku nõusolekut tema andmete äriliseks kasutamiseks ja üldsusele kättesaadavaks tegemiseks.

6. Kokkuvõtteks

Keelemudelid on keeletehnoloogia kui teadus- ja arendustegevuse valdkonna üks peamisi väljundeid, millega üritatakse kirjeldada inimkeele toimimist, samuti nagu näiteks füüsika üritab oma mudelitega kirjeldada füüsilise maailma toimimist. Mudelite loomisel on mitmeid tegevusi, mis hõlmavad inimese keerulist intellektuaalset tegevust, näiteks andmekogude valimine, märgendussüsteemi loomine ja andmete märgendamine, algoritmide loomine või valimine, nende realiseerimine tarkvaras ja algoritmi parameetrite kohandamine. Nagu modelleerimine muudes teadusvaldkondades, ei ole ka keelemudelite loomine tänapäeval realistlik ilma

mahuka andmetöötluseta. Sageli võib ühe või mitme programmi käivitamine olla ka viimane etapp mudeli loomisel.

On selge, et töötlemata andmeid, andmekogusid ja märgendatud andmekogusid mõjutavad autoriõiguse ja isikuandmete regulatsioonid. Mingil määral tuginevad andmekogudele ka keelemudelid. Autoriõigusega kaitstud sisu mudelid tavaliselt ei sisalda, kuna algsete teoste fragmendid mudelis on selleks liiga lühikesed, ja algse teose originaalsete osade taastamine mudeli põhjal ei ole võimalik.

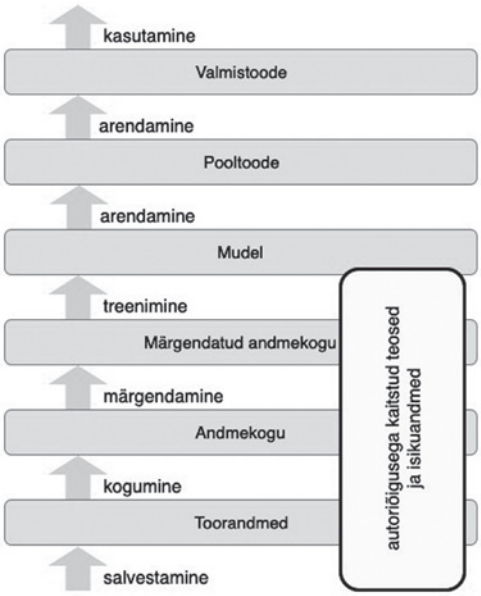
Seega saame artikli uurimisküsimusele toorandmete õiguste kehtivuse ulatuse kohta vastata, et need kehtivad nii kaua, kui vastavad andmed on veel äratuntavad. Tänapäeval levinud arendusprotsessis on keelemudel esimene vahetulemus, kus toorandmed äratuntavad enam ei ole.

Isikuandmed on võimalik mudelitest kaotada anonümiseerimise teel, välja arvatud kõnemudelitest, mis kõne kui isikuandmete sisaldamise või taasloomise võimaluse tõttu peavad siiski käsitlema isikuandmetega seotud probleeme.

Kui keelemudel ei sisalda autoriõigustega kaitstud teoste originaalseid osi, on selle edasise kasutuse tingimused mudeli looja määrata. Muuhulgas on mudeli loojal võimalik mudel avalikkusele kättesaadavaks teha ja seda kas ise ärieesmärgil kasutada või lubada teistel kasutada mudelit keeletehnoloogiliste toodete loomiseks ärieesmärgil.

Avalikkusele kättesaadavaks tegemine on oluline avatud teaduse seisukohalt, kiirendades keeletehnoloogiliste toodete arendamist ja soodustades konkurentsi mudelite loomise alal. Keelemudelite kasutamine ärieesmärkidel on vältimatu, kuna lõppkasutajatele mõeldud toodete (nt kõneliidesega külmikute) valmistamine ei ole teadusasutuste ülesannete ega pädevuste hulgas. Nende toodete olemasolu aga parandab üldist elukvaliteeti ning konkreetselt eesti keele puhul aitab täita põhiseaduslikku eesmärki tagada eesti keele kestmine läbi aegade.

Autorite peamised järeldused võtab kokku joonis 4.



Joonis 4. Keeletehnoloogilise toote loomise protsess

Viidatud kirjandus

Andmebaasi direktiiv = Euroopa Parlamendi ja nõukogu direktiiv 96/9/EÜ, 11. märts 1996, andmebaaside õiguskaitse kohta ['Directive 96191EC of the European Parliament and of the Council on the legal protection of databases']. EÜT L 77, 27.3.1996, 20-28. https://eur-lex.europa.eu/legal-content/ET/TXT/?qid=1571897258524&uri=CELEX:31996L0009 (24.10.2019).

AutÕS = Autoriõiguse seadus ['Estonian Copyright Act']. RT I 1992, 49, 615 ... RT I, 19.03.2019, 13. https://www.riigiteataja.ee/akt/119032019055 (15.10.2019).

Autoriõiguse seaduse eelnõu = Autoriõiguse ja autoriõigusega kaasnevate õiguste seaduse eelnõu ['Estonian Copyright Act project']. Versioon: 19-7-2014. https://www.just.ee/sites/www.just.ee/files/autos_en_19-7-2014.pdf (24.10.2019).

Berni konventsioon = Berni kirjandus- ja kunstiteoste kaitse konventsioon ['Berne Convention for the Protection of Literary and Artistic Works']. RT II 1994, 16, 49. https://www.riigiteataja.ee/akt/13101723 (18.10.2019).

Birštonas, Ramunas; Usonienė, Jurate 2013. Derivative works: Some comparative remarks from the European Copyright Law. – UWM Law Review, 5, 65–80.

C-476/17 = Kohtuasi C-476/17. Pelham GmbH jt *vs.* Ralf Hütter jt. (29. juuli 2019). https://eur-lex.europa.eu/legal-content/ET/TXT/?qid=1576587562212&uri=CELEX:62017CJ0476 (17.12.2019).

C-5/08 = Kohtuasi C-5/08. Infopaq International A/S *vs.* Danske Dagblades Forening (16. juuli 2009). https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:62008CJ0005&from=EN (15.10.2019).

De Castilho, Richard Eckart; Dore, Giulia; Margoni, Thomas; Labropoulou, Penny; Gurevych, Iryna 2018. A legal perspective on training models for natural language processing. – Nicoletta Calzolari et al. (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). European Language Resources Association, 1267–1274. http://www.lrec-conf.org/proceedings/lrec2018/pdf/1006.pdf (16.10.2019).

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805 [Cs].

Digiühiskonna direktiiv = Euroopa Parlamendi ja nõukogu direktiiv (EL) 2019/790, 17. aprill 2019, mis käsitleb autoriõigust ja autoriõigusega kaasnevaid õigusi digitaalsel ühtsel turul ning millega muudetakse direktiive 96/9/EÜ ja 2001/29/EÜ ['Directive (EU) 2019/790 of the European Parliament and of the Council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC']. https://eur-lex.europa.eu/legal-content/ET/TXT/?qid=1571145936304&uri=CELEX:32019L0790 (15.10.2019).

EKSS = Eesti keele seletav sõnaraamat 2009 ['The Explanatory Dictionary of Estonian 2009']. Margit Langemets, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks, Piret Voll (Toim.). Eesti keele instituut. http://www.eki.ee/dict/ekss/ (18.12.2019).

Grave, Edouard; Bojanowski, Piotr; Gupta, Prakhar; Joulin, Armand; Mikolov, Tomas 2018. Learning word vectors for 157 languages. https://arxiv.org/abs/1802.06893 [Cs].

IKS = Isikuandmete kaitse seadus ['Personal Data Protection Act']. RT I, 04.01.2019, 11. https://www.riigiteataja.ee/akt/104012019011 (24.10.2019).

Ilya, Ilin; Aleksei, Kelli 2019. The Use of Human Voice and Speech in Language Technologies: The EU and Russian Intellectual Property Law Perspectives. – Juridica International, 28, 17–27. https://doi.org/10.12697/JI.2019.28.03

Infoühiskonna direktiiv = Euroopa Parlamendi ja nõukogu direktiiv 2001/29/EÜ, 22. mai 2001, autoriõiguse ja sellega kaasnevate õiguste teatavate aspektide ühtlustamise kohta infoühiskonnas ['Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and

related rights in the information society']. https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:32001L0029&from=EN (15.10.2019).

Kelli, Aleksei; Tavast, Arvi; Lindén, Krister; Vider, Kadri; Birštonas, Ramunas; Labropoulou, Penny; Kull, Irene; Tavits, Gaabriel; Värv, Age 2019a. The extent of legal control over language data: The case of language technologies. – Kiril Simov, Maria Eskevich (Eds.), Proceedings of CLARIN Annual Conference 2019. Leipzig, Germany: CLARIN, 69–74. https://office.clarin.eu/v/CE-2019-1512_CLARIN2019_ConferenceProceedings.pdf (18.3.2020).

Kelli, Aleksei; Lindén, Krister; Vider, Kadri; Kamocki, Pawel; Birštonas, Ramunas; Calamai, Silvia; Labropoulou, Penny; Gavrilidou, Maria; Pavel Straňák 2019b. Processing personal data without the consent of the data subject for the development and use of language resources. – Inguna Skadina, Maria Eskevich (Ed.), Selected Papers from the CLARIN Annual Conference 2018, Pisa, 8–10 October 2018. Linköping: Linköping University Electronic Press, 72–82. http://www.ep.liu.se/ecp/article.asp?issue=159&article=008&volume= (25.10.2019).

Kelli, Aleksei; Vider, Kadri; Kull, Irene; Siil, Triin; Lindén, Krister; Tavast, Arvi; Värv, Age; Ginter, Carri; Meister, Einar 2018. Keeleressursside loomise ja kasutamisega seonduvaid isikuandmete kaitse küsimusi ['Data protection issues relating to the development and utilisation of language resources']. – Eesti Rakenduslingvistika Ühingu aastaraamat, 14, 77–94. https://doi.org/10.5128/ERYa14.05

Klavan, Jane; Tavast, Arvi; Kelli, Aleksei 2018. The legal aspects of using data from linguistic experiments for creating language resources. – Frontiers in Artificial Intelligence and Applications, 307, 71–78. https://doi.org/10.3233/978-1-61499-912-6-71

TAKS = Teadus- ja arendustegevuse korralduse seadus ['Organisation of Research and Development Act']. RT I 1997, 30, 471 ... RT I, 19.03.2019, 12. https://www.riigiteataja.ee/akt/119032019092 (24.10.2019).

Tarkvaradirektiiv = Euroopa Parlamendi ja nõukogu direktiiv 2009/24/EÜ, 23. aprill 2009 , arvutiprogrammide õiguskaitse kohta (kodifitseeritud versioon) (EMPs kohaldatav tekst) ['Directive 2009/24/ec of the European Parliament and of the Council on the legal protection of computer programs (Codified version) (Text with EEA relevance)']. ELT L 111, 5.5.2009, 16-22. https://eur-lex.europa.eu/legal-content/ET/TXT/?qid=1572194035371&uri=CELEX:32009L0024 (27.10.2019).

Tavast, Arvi; Pisuke, Heiki; Kelli, Aleksei 2013. Õiguslikud väljakutsed ja võimalikud lahendused keeleressursside arendamisel ['Legal challenges and possible solutions in developing language resources']. – Eesti Rakenduslingvistika Ühingu aastaraamat, 9, 317–332. https://doi.org/10.5128/ERYa9.20

ÜM = Euroopa Parlamendi ja nõukogu määrus (EL) 2016/679, 27. aprill 2016, füüsiliste isikute kaitse kohta isikuandmete töötlemisel ja selliste andmete vaba liikumise ning direktiivi 95/46/EÜ kehtetuks tunnistamise kohta (isikuandmete kaitse üldmäärus) ['Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)']. ELT L 119, 4.5.2016, 1-88. https://eur-lex.europa.eu/legal-content/ET/TXT/?qid=1555312258399&uri=CELEX%3A32016R0679 (17.10.2019).

WP29 2018 = Artikli 29 töörühm ['Article 29 Working Party']. Suunised määruse (EL) 2016/679 kohase nõusoleku kohta Vastu võetud 28. novembril 2017. Viimati muudetud ja muudatused vastu võetud 10. aprillil 2018. https://www.aki.ee/sites/default/files/inspektsioon/rahvusvaheline/juhised/suunised_nousoleku_kohta_wp259_rev_0.1_et.pdf (17.12.2019).

WP29 2014 = Article 29 Working Party (WP29). Opinion 05/2014 on Anonymisation Techniques. https://iapp.org/media/pdf/resource_center/wp216_Anonymisation-Techniques_04-2014.pdf (17.12.2019).

WP29 2014a = Article 29 Working Party (WP29). Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf (17.12.2019).

Võrguviited

Common Crawl. https://commoncrawl.org (18.3.2020).
Eesti keele puudepank. https://doi.org/10.15155/1-00-0000-0000-0000-00089L
Eesti keele spontaanse kõne foneetiline korpus. https://doi.org/10.15155/1-00-0000-0000-0000-001A3L
Eesti veebikorpus 2013. https://doi.org/10.15155/1-00-0000-0000-0000-0011FL
Estonian National Corpus 2017. https://doi.org/10.15155/3-00-0000-0000-0000-071E7L
Google News. https://news.google.com (18.3.2020).
OpenSubtitles. https://www.opensubtitles.org (18.3.2020).

INFLUENCE OF LEGAL REGIME OF LANGUAGE DATA ON LANGUAGE MODELS

Aleksei Kelli[1], Kadri Vider[1,2], Arvi Tavast[1,3,4], Krister Lindén[5], Ramūnas Birštonas[6,7], Penny Labropoulou[8], Age Värv[1], Irene Kull[1], Gaabriel Tavits[1], Carri Ginter[1]

University of Tartu[1], Center of Estonian Language Resources[2], Institute of the Estonian Language[3], Qlaara Labs[4], University of Helsinki[5], Vilnius University[6], Mykolas Romeris University[7], Institute for Language and Speech Processing (Athena)[8]

This article aims to explain the extent to which the legal regime applicable to language data affects the development and use of language models. In their approach, the authors follow a process chart, starting from raw data to finished products containing language technology (eg a refrigerator with a speech interface). The raw data used in language technologies often include copyrighted works, objects of related rights (performances, sound recordings) and personal data (voice, other information about the person) stored in non-annotated and annotated databases. The legal issues of language data have already been studied. However, the legal aspects of language models have not been throughly explored. The authors are of the opinion that, as a rule, the legal status of the language models is not affect by the legal status of the used raw language data, since copyrighted works usually do not remain in the model. However, the use of a person's voice in a language model can create legal problems. The authors analyze possible solutions to overcome these problems. The article also outlines the regulation of data mining introduced by the new copyright directive and its implementation in the context of development of language models.

Keywords: copyright, personal data, language model, language technology, text and data mining

Aleksei Kelli on intellektuaalse omandi õiguse professor Tartu Ülikoolis. Ta tegeleb ka digitaalsete keeleressursside õiguslike küsimustega TÜ-s ning Eesti Keele Instituudis ning on CLARIN ERIC õigus- ja eetikakomitee esimees. Teadustöös on tema peamised huvid avatud teadus, teadmussiire ja keeletehnoloogia õiguslikud küsimused.
Kaarli pst 3, 10119 Tallinn, Estonia
aleksei.kelli@ut.ee

Kadri Vider on Eesti Keeleressursside Keskuse (EKRK) tegevjuht, CLARIN-i riiklik koordinaator Eestis ja Tartu Ülikooli arvutiteaduse instituudi keeletehnoloogia teadur. EKRK ülesandeid on mh riiklike keeletehnoloogia teadus- ja arendusprogrammide tulemusena loodud ressursside ja tarkvara haldamine, kättesaadavaks tegemine ning litsentsimine või sellealane konsulteerimine.
Tartu Ülikool, Arvutiteaduse instituut, Narva mnt 18, Tartu, Estonia
kadri.vider@ut.ee

Arvi Tavast on EKI uue sõnastikusüsteemi Ekilex arendaja, alates 1.9.2020 EKI direktor. Tegelenud sõnastike ja keelemudelite koostamisega tundlike keeleandmestike põhjal nii era- kui ka akadeemilises sektoris (qlaara, EKI, TÜ).
Roosikrantsi 6, 10119 Tallinn, Estonia
arvi@tavast.ee

Krister Lindén on on Helsinki Ülikooli keeletehnoloogia uurimisrühma juht, Kielipankki teadusdirektor ning FIN-CLARIN-i juht, CLARIN-i riiklik koordinaator Soomes ja CLARIN ERIC õigus- ja eetikakomitee liige. 2008–2011 juhatas ta õiguslike küsimustega tegelemist CLARIN-it ettevalmistavas ESFRI projektis.
University of Helsinki, Unioninkatu 40, office 310, FIN-00014, Finland
krister.linden@helsinki.fi

Ramūnas Birštonas on Vilniuse ja Mykolas Romerise ülikoolide (Leedu) professor. Ta on samuti autori- ja kaasnevate õiguste komisjoni esimees Leedu Kultuuriministeeriumis. Ta on osalenud mitmetes teadusprojektides intellektuaalse omandi põhieksperdina ning publitseerinud intellektuaalse omandi ja sellega seonduvatel teemadel.
Vilnius University, Faculty of Law, Sauletekio av. 9 - I block, 10222 Vilnius, Lithuania
birstonas@mruni.eu

Penny Labropoulou on Ateena Teaduskeskuse Keele ja Kõne Töötlemise Instituudi juhtivteadur ja CLARIN ERIC õigus- ja eetikakomitee liige. Tema uurimisvaldkonnaks on keeleressursside metaandmete modelleerimine, keeletehnoloogiliste ressursside jagamine ja kasutamine, arvutileksikoloogia ja arvutiterminoloogia. Koostöös õigusekspertidega on ta ulatuslikult töötanud litsentseerimise teemadega ning keeleressursside õiguslike meta-andmetega suuremate teadustaristute kontekstis.
ILSP/Athena RC, Epidavrou & Artemidos 6, Athens, Greece
penny@athenarc.gr

Age Värv on Tartu Ülikooli võlaõiguse dotsent. Ta tegeleb lepinguväliste võlasuhete ning võlaõiguse üldosa ning võrdleva õiguse küsimustega. Teadustöös keskendub ta deliktiõigusele ja alusetu rikastumise õigusele ning intellektuaalomandi õiguse ja võlaõiguse kokkupuutepunktidele.
Kaarli pst 3, 10119 Tallinn, Estonia
age.varv@ut.ee

Irene Kull on Tartu Ülikooli tsiviilõiguse professor. Ta tegeleb Eesti ja EL lepinguõigusega. Teadustöö peamised huvid on EL lepinguõiguse ühtlustamine, tsiviilõiguse üldprintsiipide kohaldamine ning lepinguliste kohustuste tasakaalu õiguslikud probleemid.
Näituse 13 a, 50409 Tartu, Estonia
irene.kull@ut.ee

Gaabriel Tavits on Tartu Ülikooli sotsiaalõiguse professor. Ta tegeleb töösuhete ja sotsiaalse kaitse õiguslike probleemidega digitaalse majanduse ja digitaalse ettevõtluse tingimustes. Samuti on tema teadustöö huviks tehisintellekti kasutamine töösuhetes ja sellega seotud õiguslikud probleemid.
Näituse 20, 50409 Tartu, Estonia
gaabriel.tavits@ut.ee

Carri Ginter on Tartu Ülikooli Euroopa õiguse dotsent, kes tegeleb EL õiguse rakendamise küsimustega. Teadustöös on tema peamised huvid Euroopa Liidu ja liikmesriikide õigustike omavahelised kokkupuutepunktid.
Kaarli pst 3, 10119 Tallinn, Estonia
carri.ginter@ut.ee