Kadri Vare
eesti keele instituudi keele- ja kõnetehnoloogia osakonna juhataja
foto: Jake Farra

Keeletehnoloogia Eestis – uutest ja huvitavatest arendusprojektidest

Keeletehnoloogia valdkond on nii Eestis kui ka mujal maailmas viimastel aastatel teinud märkimisväärseid edusamme, muutudes keele õpetamisel ja õppimisel, uurimisel ja säilitamisel üha olulisemaks. Keeletehnoloogilised vahendid on saavutanud juba väga hea kvaliteedi, mis võimaldab neid igapäevaelus kasutada paljude protsesside hõlbustamiseks ning rakendada mitmesugustes teenustes ja toodetes. Suuremad keeletehnoloogia väljakutsed on praegusel ajal aga seotud eelkõige tehisintellekti, nagu näiteks kõigile tuttava ChatGPT arendamisega, sh väikseid keeli arvestades. Ülevaateartiklis tulebki juttu sellest, mis see keeletehnoloogia õigupoolest tähendab, millised on eesti keele keeletehnoloogilised vahendid, millised keeletehnoloogiaalased innovatsiooniprojektid käsil on ja miks üldse selle valdkonnaga peaks Eestis tegelema.

Keeletehnoloogia vahendid igapäevaelus

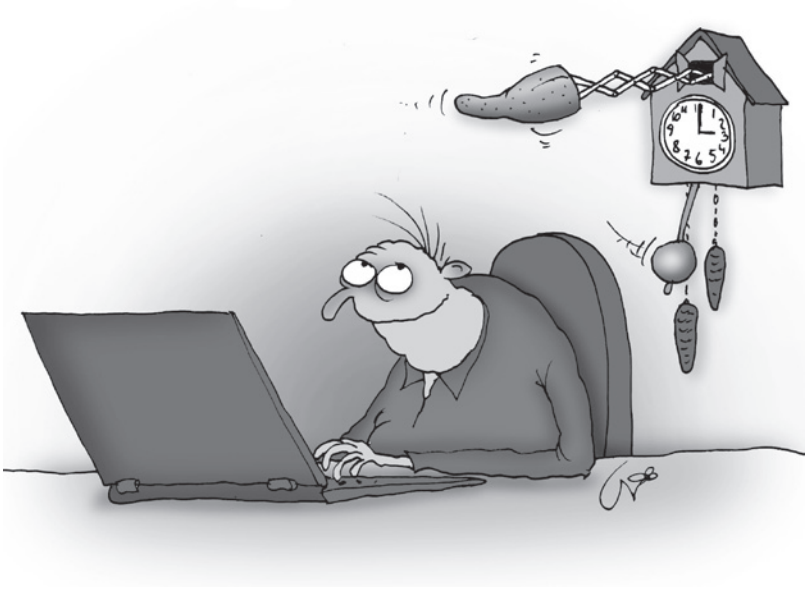
Keeletehnoloogia on vägagi mitmest teadusharust koosnev interdistsiplinaarne valdkond, mis seob endas matemaatikat, statistikat, keerulisi masinõppelisi ja närvivõrkudel põhinevaid tehnoloogiaid, aga muidugi ka teadmisi keelest ja selle toimimisest. Masinõppe meetodite rakendamine keeletehnoloogias on kiirelt kasvav suund, mis võimaldab süsteemidel keelt õppida ning seda kiiremini ja tõhusamalt töödelda. Tehisintellekti kasutuselevõtt keele mõistmise ja genereerimise protsessides avab uusi võimalusi nii tõlkimise, teksti automaatse koostamise kui

ka keeleõppe jaoks. Siiski on eesti keele töötlemisel oluline arvestada ka keele komponenti ja keelelisi eripärasid, nagu morfoloogiline rikkus ja süntaktiline keerukus.

Küllap on kõige tuntumad ja kõige kasutatavamad keeletehnoloogiad masintõlge, teksti kõneks ja kõnet tekstiks automaatselt genereerivad rakendused. **Kõnetuvastuse tehnoloogia** on viimastel aastatel teinud suuri edusamme, muutes selle mitmekülgselt kasutatavaks eri rakendustes ja teenustes. Näiteks kasutavad nimetatud tehnoloogiat virtuaalassistendid Siri, Google Assistant, Alexa ja Cortana, et kasutajad saaksid suhelda seadmetega loomuliku keele abil. Seeläbi saavad inimesed küsida ilmaprognoosi, mängida muusikat, seada meeldetuletusi, juhtida nutikodu seadmeid ja palju muud, kasutades lihtsalt oma häält. Sama tehnoloogia võimaldab luua (reaalajalisi) automaatseid subtiitreid, et pakkuda vaegkuuljatele võimalust jälgida otsesaateid, uudiseid, valimisdebatte jne. Aga kõnetuvastust saab ära kasutada ka selle jaoks, et salvestatud suulisi intervjuusid kergema vaevaga teksti kujule viia või oma dikteeritud mõtteid lihtsasti kirjalikult vormistada. Kõnetuvastuse tehnoloogia abil saab järjest enam arendada keeleõppe rakendusi, mis suudavad hinnata ja parandada kasutaja hääldust reaalajas. See tehnoloogia võib tulevikus oluliselt toetada ka (erivajadustega) õppijaid, võimaldades neil suhelda õppeprogrammidega kõne abil. Teisipidine tehnoloogia, kus **tekst kõneks** muudetakse, pakub võimalusi just vaegnägijatele ja pimedatele, kes saavad nõnda tekste endale ette lugeda lasta. Juba praegu võib seda tehnoloogiat kasutada igaüks, kes tunneb, et silmad on lugemisest väsinud või kes viibib parasjagu näiteks autoroolis, kus lugemine võimatu. **Masintõlge** on saanud igapäevaosaks mitte ainult tõlkebüroodele, vaid kõigile huvilistele. Nagu näha, on keeletehnoloogia märkamatult, aga edukalt toimetamas mitmel pool ja paljudes eluvaldkondades.

Suured keelemudelid ja eesti keel

Praegusel ajal ei ole keeletehnoloogias võimalik mööda vaadata suurte keelemudelite eesti keele ja kultuuri tundlikuks arendamise vajadusest. Suured keelemudelid, nagu ChatGPT, Llama, Gemini jt, on tehisintellekti evolutsiooni olulised verstapostid, võrreldavad mineviku märkimisväärsete leiutiste trükipressi või aurumasinaga. Eesti keele toetamine suurtes keelemudelites on keerukas, kuna praegu pakuvad eesti keele kasutamise



võimalust ainult mõned suured USA firmad oma suletud baasmudelite kaudu. Avatud lähtekoodiga mudelid, mis võimaldaksid laiemat ja paindlikumat ehk läbipaistvamat kasutust, ei sisalda veel eesti keelt. Selle peamiseks põhjuseks on suurte andmehulkade kättesaadavuse ja seaduste piirangud. Siin on väga oluline kaasnev teema ka kohaliku teadus- ja arendusvõimekuse suurendamine: tuleb tagada, et Eesti teadussüsteemis oleks piisav võimekus suurte keelemudelite loomiseks ja nendega töötamiseks. Selleks investeeritaksegi praegu ressursse teadusasutuste suutlikkuse tõstmisele ja tehakse koostööd rahvusvaheliste partneritega. Aktiivselt peab osalema avatud mudelite arendamises ja olemasolevate mudelite eesti keelele kohandamises, loomaks eestikeelseid stiimulõppe andmestikke ja parandades suletud mudelite kvaliteeti. Üks tähtis suund on selliste suurte mudelite hindamine ja valideerimine: hoida ülevaadet valdkonna arengusuundadest ja aidata nii era- kui avalikul sektoril teha informeeritud otsuseid tehisintellekti tehnoloogiate kasutuselevõtuks. Hindamisel keskendutakse keeleoskusele, kvaliteedile, ohutusele ja kultuurilisele sobivusele. Ei taha ju keegi, et tuleviku seadmed meile ropendades vastaksid, oleksid diskrimineerivad või levitaksid suisa ohtlikku informatsiooni.

Miks on eestlaste osalus suurte keelemudelite arendamises nii oluline? Sest peame tagama, et eesti keel ja kultuur ei jääks tehisintellekti arengus tagaplaanile. Peame looma võimalusi, et eesti keelt kõnelevad inimesed

saaksid kasutada uusimaid tehnoloogiaid oma emakeeles, säilitades samal ajal kultuurilise identiteedi ja suurendades eesti keele kasutusvõimalusi digiajastul. Samuti on oluline rõhutada, et koostöö eri sektorite vahel ning avatud suhtumine innovatsiooni ja tehnoloogiliste lahenduste arendamisse on võtmeks, et saavutada seatud eesmärgid ja kohandada suuri keelemudeleid eesti keele vajadustele.

Viipekeele tehnoloogia

Huvitav tegevussuund on käivitunud eesti viipekeele uurimise ja arendamise alal. Viipekeel kui unikaalne kurtide inimeste kommunikatsioonivorm vajab erilist tähelepanu ja ressursse, et tagada selle kasutajaskonna täielik kaasamine ühiskonda. Eesti keele instituut ja riigikantselei on käivitanud viiplemisroboti loomise projekti, mille eesmärk on mitte ainult eesti viipekeele säilitamine ja arendamine, vaid ka viipekeelega seotud keele- ja kõnetehnoloogiate väljatöötamine. Projekti idee on arendada välja viiplemisrobot, mis suudaks tõlkida eesti keelest eesti viipekeelde. Roboti eesmärk on muuta avalikud teenused ja informatsioon kuulmispuudega inimestele eri infoedastuskanalites kättesaadavamaks. Selleks arendatakse masinõppe jaoks kasutatavat eesti viipekeele korpust ja mitut spetsiifilise sõnavaraga viipekeele masinõppemudelit.

Projekt on vajalik ELi ligipääsetavusnõuete kohustuse täitmiseks, sest Eesti on viipekeelse kogukonna kaasamise suhtes ülejäänud Euroopast maha jäänud. Lisaks aitab projekt tõsta üldsuse teadlikkust kuulmispuude olemusest ja viipekeelse kogukonna erivajadustest. Eesti keele instituut teeb viiplemisroboti projekti jaoks ja viipekeele arendamise valdkonnas koostööd mitme asutusega, sealhulgas Tallinna ülikooli viipekeele uurimisrühmaga, majandus- ja kommunikatsiooniministeeriumi, riigi infosüsteemi ameti, häirekeskuse, Eesti viipekeele seltsi ja Eesti kurtide liiduga. Viipekeele tehnoloogiate arendamine on eesti keeletehnoloogia valdkonnas suur samm edasi, pakkudes uusi võimalusi keeleliste barjääride ületamiseks ja kogu ühiskonna kaasamiseks. Kuigi viipekeele tõlkeroboti loomine on keerukas ülesanne, on sellel märkimisväärne potentsiaal muuta info kuulmispuudega ja kurtidele inimestele kättesaadavamaks. Projekt näitab Eesti pühendumust innovatsioonile ja ligipääsetavuse parandamisele, luues samal ajal aluse edasisteks uurimistöödeks ja arendusteks viipekeele tehnoloogia vallas.

Eestikeelse teksti automaatkorrektuur

Väga olulist rolli kannavad keeletehnoloogia valdkonnas ka paljud tekstianalüüsi ja teksti mõistmise vahendid teksti ja õigekirja kontrollijast kuni automaatse märksõnastamise tehnoloogiateni. Siinkohal väärib märkimist projekt, mille käigus arendati eestikeelse teksti automaatkorrektuuri vahendeid. Projektis osalesid teadlased Tartu ja Tallinna ülikoolist ning see sai rahastust aastatel 2021–2023 eesti keeletehnoloogia riiklikust programmist. Projekti eesmärk oli arendada ja parandada olemasolevaid eesti keele õigekirja- ja grammatikakorrektuuri vahendeid.

Projekti uudsus peitus andmete täiendamises, siirdeõppe kasutamises ning suurte keelemudelite, näiteks GPT4 võimekusega võrdlemises. Eesmärk oli luua süsteeme, mis suudaksid automaatselt tuvastada eesti keele vigu, pakkudes samal ajal kasutajatele usaldusväärseid ja täpseid parandusi. Projektis rakendatud meetodid ja mudelid näitasid, et on võimalik saavutada paremaid tulemusi kui suurte keelemudelite põhjal, keskendudes vabavaralistele lahendustele, mis muudavad tulemused ja töötavad teksti automaatkorrektuuri rakendused kättesaadavaks laiemale kasutajaskonnale.

Projekti käigus loodi kvaliteetseid eesti keele andmekogusid, mis hõlmavad nii emakeelsete kõnelejate kui ka eesti keele kui teise keele õppijate tekste. Need andmestikud võimaldavad masinõppe mudelitel õppida ja tuvastada vigu mitmekesises tekstikogumis, parandades seeläbi automaatkorrektuuri täpsust ja usaldusväärsust. Eesti keele grammatika- ja õigekirjakontrollijate arendamiseks kasutatud metodoloogiad ja andmestikud on nüüd avalikult kättesaadavad, luues aluse edasisteks uurimusteks ja arendusteks selles valdkonnas.

Automaatkorrektuuri arendajad rõhutavad vajadust aina rohkemate andmete järele, et veelgi parandada automaatkorrektuuri tulemusi. On pakutud, et suurte keelemudelite genereeritud tehisvead võiksid oluliselt panustada treeningandmetesse. Lisaks on projekti tulemustes viidatud võimalustele keelemudelite efektiivsemaks kasutamiseks automaatkorrektuuris, sealhulgas vabavaraliste mudelite peenhäälestamise ja uute, spetsiifilisemate lahenduste arendamise näol.

Lõpetuseks

Kokkuvõtteks tuleb tänada Eesti teadlasi, poliitikakujundajaid, ettevõtteid ja huvilisi selle eest, et juba mitukümmend aastat on mõistetud keeletehnoloogia arendamise vajadust ning et seda on sihipäraselt sama kaua riiklikult toetatud. Need tegevused tagavad, et saaksime ka tulevikus oma igapäevaseadmetega rääkida just oma emakeeles, eesti keeles.