Eesti keeletehnoloogia valdkonna arendamise ning keeletehnoloogia vahendite laialdase kasutuselevõtu tegevuskava 2022–2024

SISSEJUHATUS

Keeletehnoloogia on infotehnoloogia ja keeleteaduse osa, mis tegeleb inimkeele töötlemisega, hõlmates nii kirjutatud kui ka suulist keelt. Keeletehnoloogia arendamine on Euroopa keelelise ja kultuurilise mitmekesisuse säilitamisel prioriteet. See tähendab, et iga rahvuskeele jaoks tuleb luua keeletehnoloogilised andmestikud ja vahendid, mille tõhus kasutamine tagab ka väikeste keelte püsimise avatud maailmas. Eesti keeletehnoloogia arengut on alates 2006. aastast riiklikult toetatud keeletehnoloogia teadusprogrammide kaudu. Praegu on käimas kolmas järjestikune programm „Eesti keeletehnoloogia 2018–2027“[1]. Riiklike programmide ajal on käivitunud 90 projekti, mille tulemusel on peamiselt loodud vajalikke baastehnoloogiaid ja keeleandmestikke, kuid fookus on olnud ka keeletehnoloogia kasutusele võtmisel nii avalikus kui ka erasektoris.

Tegevuskava eesmärgid, ülesehitus ja valdkonnad

Tegevuskava annab ülevaate nii valdkonna hetkeseisust kui ka eesti keeletehnoloogia arendamise tegevustest aastatel 2022–2024. Tegevuskava kestus lähtub eeldusest, et infotehnoloogia arengud on kiired ning keeletehnoloogia valdkonna projekte viiakse ellu etapiti. Tegevuskava alusel edendatakse mitme riikliku arengukava – „Eesti keele arengukava 2021–2035“, „Digiühiskonna arengukava 2030“, „Eesti teadus- ja arendustegevuse, innovatsiooni ning ettevõtluse arengukava 2021–2035“ ning „Kratikava 2022–2023“ – tegevussuundi, et arendada eesti keeletehnoloogia võimalikult kõrgele tasemele ja võtta eesti keeletehnoloogia vahendid kasutusele võimalikult paljudes avaliku sektori teenustes ja infosüsteemides, mis kokkuvõttes tagab eesti keele elujõu digiühiskonnas.

Tegevuskava jaotub neljaks suuremaks omavahel seotud osaks: k**eeletehnoloogia arendamine, keeletehnoloogia kasutuselevõtt ja populariseerimine, keeleandmestikud ning koostöö ja seire**.

[1] https://www.hm.ee/et/tegevused/teadus/teadusprogrammid

Tegevuskava elluviimine ja rahastamine

Peamine tegevuskava elluviija on Eesti Keele Instituudi (EKI) keeletehnoloogia kompetentsikeskus kui keeletehnoloogia arendamise tugikeskus, mis teeb koostööd Riigi Infosüsteemi Ametiga (RIA), et eesti keeletehnoloogilised vahendid jõuaksid kasutusele võimalikult paljudesse avaliku ja erasektori loodud lahendustesse. Kuna eesmärkide ja tegevuste poolest on see ministeeriume ja asutusi ühendav kava, oleneb elluviimine kõigi keeletehnoloogia arendamisega tegelejate panusest.

Tegevuskava rakendamist rahastatakse Haridus- ja Teadusministeeriumi, Majandus- ja Kommunikatsiooniministeeriumi ning Euroopa Liidu struktuurivahendite eelarvetest. Haridus- ja Teadusministeeriumi keeletehnoloogia TA-programmi (EKT) aastane eelarve on ligikaudu 1,5 miljonit eurot aastas ning Euroopa Liidu struktuurivahendite (SF ja RRF) keeletehnoloogia valdkonda käsitlevate vahendite umbkaudne eelarve on 1 miljon eurot aastas.
Tegevuskava elluviimise, juhtimise ning seiramisega tegeleb Haridus- ja Teadusministeeriumi juhitav riigiasutuste ja võtmepartnerite esindajatest koosnev kogu, mis muu hulgas arutab ja kavandab vajaduse korral täiendavaid tegevusi. Seatud tegevuste täitmist analüüsitakse jooksvalt ning tegevuskava periood lõpeb aruandega.

I KEELETEHNOLOOGIA ARENDAMINE

Hetkeolukord

Rahvusvahelises võrdluses on olemas keele töötlemiseks vajalikud baaskomponendid[2], kuid puudu on mitu võtmekomponenti[3], mille arendamine on kriitilise tähtsusega.

Eestikeelse **kõnetuvastuse** (ehk kõne-tekstiks) üldine tase on väga hea ning võimaldab juba praegu genereerida kvaliteetseid reaalajalisi subtiitreid või näiteks radioloogia uuringuaruandeid. Madalam on kõnetuvastuse kvaliteet mürarohkete salvestuste ja variatiivsema kõne (aktsendid, murded, seenioride või laste kõne jms) puhul.

Eestikeelse **kõnesünteesi** (ehk tekst-kõneks) valdkonnas on lisaks statistilistele häältele arendatud viimaste aastate jooksul ka tehisnärvivõrkudel põhinevaid loomuliku kõlaga hääli. Praegused hääled sobivad hästi lugema ette erinevaid tekste, kuid vaja on arendada ka spontaanse kõne sünteesi.

Eesti keele jaoks on loodud olulise tähtsusega **tekstianalüüsi** baaskomponendid, millele toetub suuresti järgmiste keeletasandite – eelkõige semantilise ja pragmaatilise tasandi – automaatne analüüs. Mõnes valdkonnas vajab tekstianalüüsi vahendite tehnoloogiline tase uuendamist, näiteks tuleb liikuda reeglipõhisest ning statistilisi meetodeid kasutavast tehnoloogiast järjest enam masinõppelise tehnoloogia poole. Süvitsi on tegeletud morfoloogia ja süntaksiga, vähem on tähelepanu pööratud semantilisele ja pragmaatilisele analüüsile.

[2] http://www.meta-net.eu/whitepapers/overview
[3] https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf

Tegevus	2022	2023	2024	Elluviija	Rahastus
1.1 Arendatud on vähemalt viis seni puuduvat tekstianalüüsivahendit, sh on täiustatud eesti keele eri tasandite automaatse analüüsi vahendeid ning loodud lõppkasutajale mõeldud üldraamistikke	Grammatikakontrollija on arendatud 1. etapi tasemeni			EKI ja TA-asutused	EKT
	Eesti tekstide kokkuvõtja on arendatud 1. etapi tasemeni				
	Eesti tekstide lihtsustaja on arendatud 1. etapi tasemeni				
	Välja on arendatud anonümiseerija/pseudonümiseerija				
		Valeuudiste (*fake news*) tuvastaja arendusega on alustatud			
	Vihakõne tuvastamise arendusega on alustatud				
		Semantikavahendite arendamisega on alustatud			
	Dialoogimudelite analüüs on läbi viidud				
	Täiustatud ja edasi on arendatud olemasolevaid vajalikke tekstitöötlusvahendeid				
1.2 Arendatud on kõnetehnoloogiate baastehnoloogiaid, tõstetud on nende tehnoloogiate kvaliteeti	Kõnetuvastuse baastehnoloogia on arendatud TVT 4. tasemele võimekusega tuvastada spontaanset kõnet piiratud kontekstides, erineva helikvaliteediga keskkondades ning eri kõnetüüpide puhul (aktsendid, pealerääkimine jms). Arendatakse edasi eestikeelset kõnesünteesi			EKI ja TA-asutused	EKT
1.3 Arendatud on masintõlke baastehnoloogiat, tõstetud on selle kvaliteeti	Baastehnoloogia on arendatud TVT 1.-3. tasemele, mis tagab täisautomaatse kõrge kvaliteediga masintõlke piiratud/kontrollitud valdkondades (ilma järeltoimetamiseta)			EKI ja TA-asutused	EKT
1.4 Välja on arendatud vajalik ja innovaatiline keeletaristu keeletehnoloogia jaoks	Edasi on arendatud valitud korpusepäringusüsteemi, sh korpuste märgendamise vahendeid			EKI	EKT

II KEELETEHNOLOOGIA KASUTUSELEVÕTT JA POPULARISEERIMINE

Hetkeolukord

Seni on keeletehnoloogia komponente kasutatud ja integreeritud juhtumipõhiselt. Keeletehnoloogia vahendid on küll loodud ja vabalt kättesaadavad, kuid nende leidmine ja kasutamine on olnud pigem tagasihoidlik. Järjest valmivad uued eesti keeletehnoloogia olulised komponendid ning on tähtis, et need saaksid integreeritud vastavalt nende funktsioonidele võimalikult paljudesse teenustesse ja toodetesse.

Paljude suurte tehnoloogiafirmadega (nt Google, Apple, Facebook), kelle ärimudelid põhinevad samuti keeletehnoloogial, koostöö puudub. Olemas on regulaarsed konverentsid, kuid puuduvad erinevatele sihtgruppidele suunatud koolitused, töötoad ja seminarid. Puudu on kasutuselevõttu toetav tugisüsteem.

<table>
<tr><th>Tegevus</th><th>2022</th><th>2023</th><th>2024</th><th>Elluviija</th><th>Rahastus</th></tr>
<tr><td rowspan="6">2.1 Edendatud on keeletehnoloogia kasutuselevõttu avaliku sektori infosüsteemides, keeletehnoloogia vahendeid on rakendatud vähemalt seitsme avaliku sektori teenuses ja infosüsteemis</td><td colspan="3">Tõlkevärava arendusetapid 1-3 on läbi viidud</td><td rowspan="6">EKI ja RIA</td><td rowspan="6">EKT, SF</td></tr>
<tr><td colspan="2" rowspan="2"></td><td>Grammatikakontrollija on liidestatud (Microsoft, Google, avaliku sektori infosüsteemid jms)</td></tr>
<tr><td colspan="2">Subtitreerimistarkvara on liidestatud või on loodud mikroteenus otseedastuse jaoks</td><td></td></tr>
<tr><td colspan="3">#bürokratt on liidestatud vajalike keeletehnoloogia vahenditega</td></tr>
<tr><td>Loodud on ärinimede automaatne kontrollija</td><td colspan="2"></td></tr>
<tr><td>ERRi otsesaadete jaoks (nt pressikonverentsid, jutusaated, uudised) töötab automaatne eestikeelne subtitreerimine</td><td colspan="2"></td></tr>
</table>

2.2 Toetatud on keeletehnoloogia kasutuselevõtmist erasektoris		Koostatud on keeletehnoloogia vahendite ja kasutusvõimaluste kataloog		EKI	EKT, SF
	Läbi on viidud ettevõtete vajadustest lähtuvad konsultatsioonid				
	Huvitatud osapoolte (ettevõtlus ja keeletehnoloogia arendajad) kokkuviimine toimub pidevalt				
	Korraldatud on (süva)töötube				
2.3 Hinnatud ja testitud on kõne- ja tekstianalüüsivahendite kasutamist keeleõppes, sh eesti keele kui teise keele õppeks		Analüüsitud on kõnetehnoloogiate kasutusvõimalusi		EKI ja TA-asutused	SF
		Grammatikakontrollija on kohandatud keeleõppe vajadusest lähtuvalt			
		Tekstianalüüsivahendid on kohandatud keeleõppe vajadustest lähtuvalt			
2.4 Läbi on viidud vähemalt 35 avalikule ja erasektorile mõeldud koolitust ning koostatud on juhendid, mille eesmärk on tõsta teadlikkust keeletehnoloogia vahendite rakendamisest	Korraldatud on keeletehnoloogiat tutvustavaid koolitusi			EKI	EKT, SF
	Planeeritakse ja viiakse läbi iga-aastane keeletehnoloogia konverents				
	Korraldatakse ideekorje kampaania				
	Esinetakse laiemale avalikkusele, sh üldhariduskoolidele				
	Luuakse keeletehnoloogia brändiraamat				

III KEELEANDMESTIKUD

Hetkeolukord

Keeleandmestike rohkus ja kvaliteet on keeletehnoloogia arendamisel olulise tähtsusega. Seni on keeleandmestikke peamiselt kogutud keeletehnoloogiliste projektide vajadusest tulenevalt ning nende avaldamiseks mõeldud töövood pole tugevalt juurdunud. Jätkuvalt on puudu erinevate keeletehnoloogiate arendamiseks vajalikust variatiivsest keelematerjalist, nt kõneandmed, tõlkematerjalid, viipekeele andmestikud, aga ka keeleuurimiseks ja keelearenduseks vajalikud ükskeelsed andmestikud, nagu näiteks erialatekstid. Järjest enam on fookuses keeleandmestike kogumine ning nende taaskasutamise võimaldamine, kuid siiani on suur murekoht keeleandmestike õiguslik regulatsioon ja nende kasutamine.

<table>
<tr><th>Tegevus</th><th>2022</th><th>2023</th><th>2024</th><th>Elluviija</th><th>Rahastus</th></tr>
<tr><td rowspan="2">3.1 Seadusandlusest tulenevad piirangud ja võimalused keeleandmete (sh tekstiliste ja kõneandmete) kogumiseks ja taaskasutamiseks on selgelt kirjeldatud</td><td>Läbi on viidud keeleandmestike kasutamis e õiguslik analüüs</td><td colspan="2"></td><td rowspan="2">EKI ja HTM</td><td rowspan="2">EKT</td></tr>
<tr><td>Koostatud on keeleandmestike parimate kasutusviiside juhend</td><td colspan="2"></td></tr>
<tr><td rowspan="8">3.2 Keeletehnoloogia ja tehisintellekti arendamiseks on kogutud igal aastal juurde vähemalt 10 mitmekesist keeleandmestikku sisaldavat korpust, mille õiguslik staatus on selge/määratud ning mis vastavad FAIR-printsiibile. Kogutud on vähemalt 4000 tundi kõnematerjali kõnetehnoloogiate arendamiseks. Kogutud ja märgendatud on vähemalt 100 000 sotsiaalmeediateksti (eelkõige vihakõne tuvastamise jaoks)</td><td colspan="2">Kogutud on viipekeele materjale</td><td></td><td rowspan="8">EKI ja RIA</td><td rowspan="8">RRF, SF, EKT</td></tr>
<tr><td colspan="3">Kogutud on valdkondlikke tõlkematerjale masintõlke arendamiseks</td></tr>
<tr><td colspan="3">Kogutud on kõnematerjali kõnetehnoloogiate arendamiseks</td></tr>
<tr><td colspan="3">Kogutud on NER-keeleandmestikke</td></tr>
<tr><td colspan="3">Kogutud on ilukirjandustekste</td></tr>
<tr><td colspan="3">Kogutud on sotsiaalmeediatekste</td></tr>
<tr><td>Algatatud on „Anneta kõnet“ kampaania kõnematerjali kogumiseks</td><td colspan="2"></td></tr>
<tr><td colspan="3">Kogutud variatiivsed keelematerjalid on korpustes märgendatud, sh transkribeeritud, joondatud, vajaduse korral ajamärgistatud</td></tr>
</table>

	Korpused on puhastatud, korrastatud, vajaduse korral anonümiseeritud ja pseudonümiseeritud			
	Loodud on eri keelemudeleid			
3.3 Keeletehnoloogia ja keeleandmestikud on süsteemselt korrastatud, töödeldud ning kättesaadavad kõigile kasutajatele Eesti teabevärava kaudu	Avaandmete portaal/teabevärav on kohandatud keeleandmestike jaoks		EKI ja RIA	EKT, SF
	Olemasolevad keeleandmestikud on üle viidud/peegeldatud avaandmete portaali			
	Korraldatud on koolitusi teabevaldajatele avaandmete portaali kasutamise kohta ja keeleandmestike kirjeldamise kohta			
3.4 Tagatakse efektiivne keeleandmestike kogumise protsess, mille jaoks on loodud keeletehnoloogia jt seotud uurimisvaldkondade vajadustest lähtuvad vahendid	Läbitud on eesti keele teksti- ja käekirjatuvastuse 1. etapp		EKI	EKT
		Alustatud on metaandmete automaatse genereerija arendamisega		
3.5 Haridus- ja Teadusministri määrus „Keeleandmestike nimekiri, avaldamise ja taaskasutamise tingimused ning kord“	Eelnõu on ette valmistatud ja määrus kinnitatud		HTM	

IV SEIRE JA KOOSTÖÖ

Hetkeolukord

Aastal 2012 viidi läbi seni viimane rahvusvaheline keeletehnoloogia seire (META-NET *white paper*[4]) , mis hindas eesti keeletehnoloogia baaskomponentide taseme heaks. Eesti keeletehnoloogiat arendades on seire tulemustest ka osaliselt juhindutud, kuid süstemaatiliselt pole tegeletud keeletehnoloogia komponentide, tehnoloogiate ja vajalike keeleandmestike kaardistamise ja seiramisega. Samuti ei ole terviklikult välja selgitatud avaliku ja erasektori kriitilisi vajadusi, mille täitmisel oleks keeletehnoloogia rakendamisest otsene kasu.

Eesti on esindatud rahvusvahelistes keeletehnoloogia taristutes ja võrgustikes (nt CLARIN, ELRC, ELG), kuid seni pole nende võimalusi maksimaalselt ära kasutatud. Kõik nimetatud võrgustikud pakuvad mitmeid võimalusi laienemiseks ja koostööks, sh rahastust, koolitusi ning keeletehnoloogia vahendite ja keeleandmestike avaldamis- ja turundamisvõimalusi rahvusvahelisel tasandil.

<table>
<tr><th>Tegevus</th><th>2022</th><th>2023</th><th>2024</th><th>Elluviija</th><th>Rahastus</th></tr>
<tr><td rowspan="3">4.1 Erinevates avaliku ja erasektori sihtgruppides on läbi viidud uuringud, mille alusel saab kaardistada ning arendada keeletehnoloogia vahendeid vajaduspõhiselt</td><td>Korraldatud on keeletehnoloogiavahendite kasutusuuring (1. etapp)</td><td></td><td>Korraldatud on keeletehnoloogia vahendite kasutusuuring (2. etapp)</td><td rowspan="3">EKI</td><td rowspan="3">EKT, SF</td></tr>
<tr><td>Läbi on viidud keeletehnoloogiliste vajaduste uuring (1. etapp)</td><td></td><td>Läbi on viidud keeletehnoloogiliste vajaduste uuring (2. etapp)</td></tr>
<tr><td>Välja on töötatud pideva ideekorje põhimõtted</td><td colspan="2"></td></tr>
<tr><td>4.2 Käivitatud on keeletehnoloogia arendamiseks vajalike TA-tegevuste, (sidus)valdkondade ja</td><td colspan="3">Tehtud on keeletehnoloogia ja keeleandmestike seiret nii Eesti piires kui ka rahvusvaheliselt, raportid on esitatud kord aastas</td><td>EKI</td><td>EKT</td></tr>
</table>

[4] http://www.meta-net.eu/whitepapers/overview

<table>
<tr><td rowspan="4">tugiprotsesside pidev seiramine. Eesti keeletehnoloogiate taset on regulaarselt hinnatud nii Eesti piires kui ka võrrelduna rahvusvaheliselt tunnustatud tasemega</td><td>Prioriteetsetest keeletehnoloogia valdkondadest on saadud ülevaade (1. etapp)</td><td></td><td>Prioriteetsetest keeletehnoloogia valdkondadest on saadud ülevaade (2. etapp)</td><td rowspan="4"></td><td rowspan="4"></td></tr>
<tr><td>Välja on selgitatud keeleliste erivajadustega inimeste vajadused</td><td colspan="2"></td></tr>
<tr><td colspan="2">Kindlaks on määratud keeletehnoloogia jt sidusvaldkondade (nt terminoloogia, humanitaaria, sotsiaalteadused) vajadused</td><td></td></tr>
<tr><td>Valminud on ELE Euroopa keelte keeletehnoloogiliste tasemete võrdlus</td><td colspan="2"></td></tr>
<tr><td rowspan="3">4.3 Suurenenud on keeletehnoloogia-alane koostöö rahvusvahelisel tasemel, maksimaalselt on ära kasutatud võrgustike rahastusvõimalusi, pidevalt on aja- ja asjakohastatud Eesti keeletehnoloogia vahendite ja keeleandmestike info rahvusvahelistes keeletehnoloogia võrgustikes ning võrgustike võimalusi on tutvustatud Eesti teadlastele ja keeletehnoloogia arendajatele. Eesti teadusasutustel, ettevõtetel ja avaliku sektori asutustel on juurdepääs olulistele rahvusvahelistele keeletehnoloogia võrgustike hüvedele</td><td colspan="3">Osaletakse ELG võrgustikus[5]</td><td rowspan="3">HTM, EKI ja TA-asutused</td><td rowspan="3">EKT</td></tr>
<tr><td colspan="3">Osaletakse ELRC võrgustikus[6]</td></tr>
<tr><td colspan="3">Osaletakse CLARINi taristus[7]</td></tr>
</table>

[5] https://www.european-language-grid.eu/
[6] https://lr-coordination.eu/
[7] https://www.clarin.eu/