JUSTIITS- JA DIGIMINISTEERIUM

Tehisaru võimekuse arendamisest eesti keele ja kultuuri alal

Liisa Pakosta

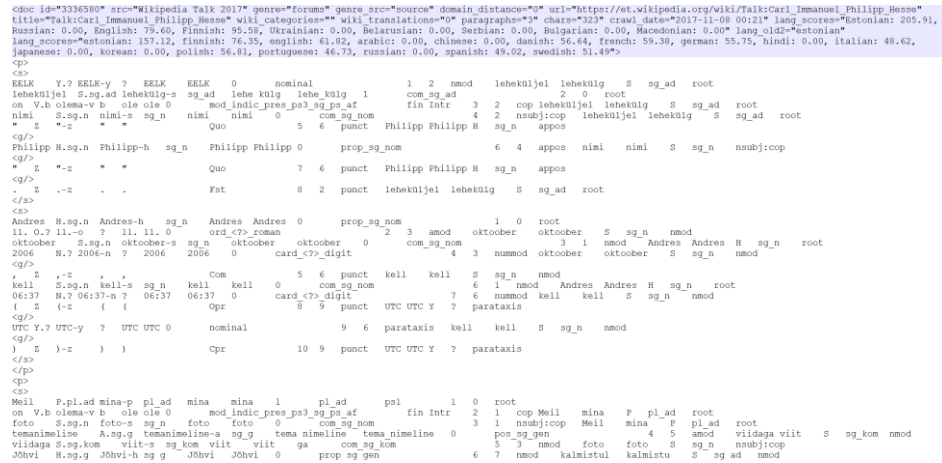
Justiits- ja digiminister

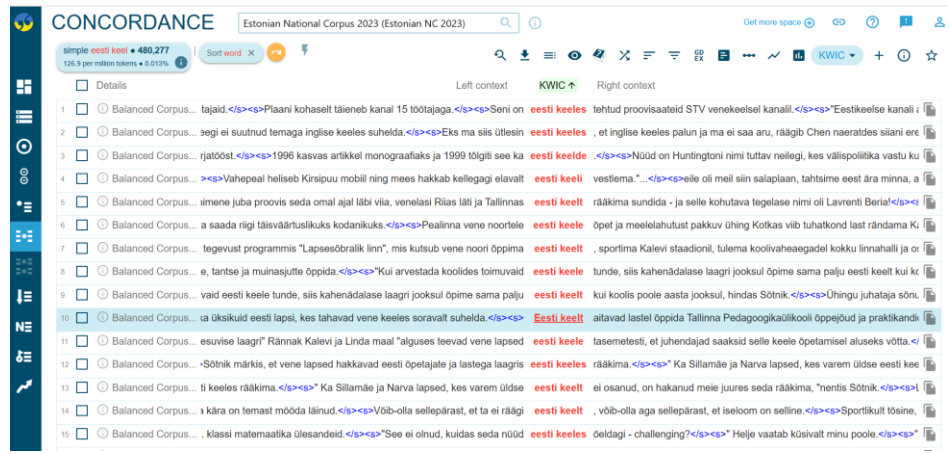
Probleem

- Eesti keele kestlikkus eeldab tänapäeval lisaks muule ka selle esindatust tehisaru suurtes keelemudelites
- Keelemudelid juba on laialdaselt kasutusel, ka hariduses, ja mõjutavad sealtkaudu keelt
- Mudeli eesti keele oskus ja eesti kultuuri tundmine sõltub tema treenimiseks kasutatud eestikeelsete tekstide hulgast ja kvaliteedist
- Mudeli treenimine ise on kulukas protsess (eelarved miljardites), meie ainus võimalus on koostöö
- Eesti keele jõudmiseks mudelitesse tuleb tagada keeleandmete laialdane kättesaadavus ning nende kasutus ettevõtete hulgas
- Selle käigus tuleb järgida autoriõiguse ja isikuandmete kaitse regulatsioone, et mitte kahjustada kellegi huve

Eesti ühendkorpus ja selle levitamine

- Eesti keele korpuseid on seni levitatud maksimaalselt avatud standardlitsentsiga, nt ühendkorpus CC-BY 4.0
  - Litsentsitingimused määrab teabevaldaja lähtudes konkreetsetest asjaoludest (nt viitamise kohustus algallikale, ärilise kasutuse piirang)
- Korpust on levitatud näiteks:
  - rahvusvahelistele projektidele (nt HPLT);
  - teadusasutustele (TÜ, TalTech);
  - ettevõtetele (Meta, Tilde)
- Meta sai ligipääsu kogu EKI ühendkorpusele, sh teaduseesmärgil ilukirjanduse alamkorpusele





Andmekaitseõigus

- Korpus sisaldab isikuandmeid;
- Andmete töötlemiseks on vajalik õiguslik alus ning järgida tuleb töötlemise eesmärgipärasust, minimaalsust, õigsust, ajakohasust, piiratud säilitamist ja turvalisust (IKÜM art 5);
- Teadusuuringute eesmärgil lubab IKÜM isikuandmeid töödelda, järgides Eesti isikuandmete kaitse seaduse tingimusi (eriti § 6);
- Keelekorpuse arendamine võib kuuluda IKÜMi kohaselt teadusuuringute, tehnoloogiaarenduse ja tutvustustegevuste alla;
- Vaja on analüüsida andmetöötluse õiguslikke aluseid ning alustatud on vajalike tehniliste ja organisatoorsete meetmete rakendamisega (sealhulgas andmekaitsespetsialist määramisega).

Autoriõigused

- Koondab eri allikatest pärit materjali, sh. sisaldab autoriõigusteta, aegunud õigustega või litsentsiga kasutatavaid tekste (nt Vikipeedia);
- Osa korpuse materjalist on autoriõigusega kaitstud (nt artiklid);
- Siseriiklik õigus (AutÕS):
  - lubab teadusasutustel vabalt kasutada veebist kraabitud avalikke andmeid teksti- ja andmekaeveks;
  - võimaldab ka ärilisel eesmärgil teksti- ja andmekaeve erandi alusel, kuid tingimused erinevad teaduseesmärgist;
- Ärilisel eesmärgil:
  - andmekaeve puhul võivad õiguste omajad selle kasutamise keelata (nt robots.txt);
  - Seega praegusel juhul on oluline küsimus, kas, millal ja millisel moel on õiguste omajad oma materjali juures andmekaeve välistuse (nn *opt-out*) teinud;
  - Tänane valdkonna kaardistus on näidanud, et ettevõtted pole välistanud andmekaeve erandi alusel andmete kogumist ning töötlemist tehniliste meetmetega (lähenetud on väga konkreetsete ettevõtete välistamise läbi, näiteks Postimees Grupp ei luba Yandexil andmekaevega tegeleda);

Edasised sammud

Tegevus	Vastutaja	Tähtaeg
Koostatud on meetme tingimused, et pakkuda rahastust investeeringuteks, mis on vajalikud avaliku sektori keeleandmete kättesaadavaks muutmisel.	JDM, EKI	Q2 2025
Koostatakse avaliku sektori keeleandmete kaardistus.	EKI	Q2 2025
Töötatakse välja keeletehnoloogia juhtimismudel, mis tagaks keeleressursside efektiivse haldamise ja tehnoloogilise kasutuse.	EKI, JDM, HTM	31.05.2025
Tõstetakse keeleandmetega seotud õiguskompetentsi.	JDM, EKI	31.08.2025
Koostatakse nõuded, millele riigi infosüsteemide tekstiandmed vastama peavad, ning plaan andmete vastavusse viimiseks, sh tagades andmete sobivuse generatiivse tehisintellekti mudelite treenimiseks ja kasutamiseks.	JDM, HTM, EKI; kaasatakse teisi osapooli	31.12.2025
Viiakse läbi analüüs tekstiandmete kasutamise lihtsustamiseks suurte keelemudelite treenimisel ning esitada see hiljemalt 31.12.2025 valitsuse kabinetiistungile koos ettepanekutega suurte keelemudelite tõhusamaks kasutuseks Eesti majanduse konkurentsivõime toetamisel.	EKI, kaasatakse JDM, HTM, KUM jt osapooli	31.12.2025
Arendatakse edasi eesti keelt ja kultuuri arvestavat keelemudelit.	JDM, HTM	31.12.2025
Viiakse läbi analüüs riigi ressursside koondamisest tehisaru rakendamiseks ja andmehalduse tõhusamaks korraldamiseks.	JDM, RaM	31.12.2025
Õigusakti(de) muudatus(ed) on jõustunud.	HTM	-
Viiakse läbi keele andmeruumi analüüs ja katseprojekt.	EKI, JDM	31.03.2026

JUSTIITS- JA DIGIMINISTEERIUM

Suur tänu