Teema: **Anonümiseerimise tehnikate rakendamine kakskeelsetele andmetele eesti-inglise või eesti-vene keelepaari näitel**

Juhendajad: Eesti Keele Instituudi keeletehnoloogia Kompetentsikeskuse töötajad Silver Vapper, BSs, Kadri Vare, MA ja Martin Luts, M.Sc.

Sissejuhatus valdkonda. Tarbetekstide (sh tarkvara kasutajaliidesed, abitekstid, kasutusjuhendid) lokaliseerimisel annab olulist ajalist ja rahalist võitu varasemalt tõlgitud tekstidest sarnaste tekstiosade uuskasutus. Selleks koondatakse inimese poolt tõlgitud või masintõlgitud ja inimese poolt järeltoimetatud tekstid tõlkemäludesse – andmebaasidesse, mis sisaldavad tõlkeüksuseid, näiteks:

```
<tu>
      <tuv xml:lang="et">
        <seg>Andrus Veerpalu poeg Andreas jäi 2019. aasta reidil veredopinguga vahele.</seg>
      </tuv>
      <tuv xml:lang="en">
        <seg> Andrus Veerpalu's son Andreas was caught with blood doping during the 2019
raid.</seg>
      </tuv>
    </tu>
```

Eesti keele jaoks on loodud anonümiseerimise vahendid ntTEXTA Toolkit Anonymizer https://docs.texta.ee/et/anonymizer.html

Probleemipüstitus. Eesti Keele Instituut koondab Tõlkevärava projektis kogu eesti avaliku sektori tõlkemälud ja teeb need avaandmetena kättesaadavaks. Isikuandmete kaitse nõuab andmete anonümiseerimist, tehniline lahendus mitmekeelsete andmete anonümiseerimiseks – mis on kohandatud eesti keelele – puudub.

Ülesanne ehk töö eesmärk. Uurida võimalikke probleemi lahendusteid, koostada algoritm ja luua tarkvaralise lahenduse prototüüp kakskeelsetes (millest üks on eesti keel) tekstides isikut identifitseerivate andmete tuvastamiseks ja isiku identiteedi kustutamiseks või üldistamiseks. Testida lahendust reaalelulistel andmetel.

Juhendajad pakuvad töö teostajale tuge reeglite defineerimiseks (mida tuvastada, millega asendada), testandmeid (nt EOPC korpus, meditsiinialased või kohtute dokumendid) ja võimalust töö tulemusi rakendada Tõlkevärava arendusprojektis.

Viited

- Sissejuhatus riigi avaandmete andmete anonümiseerimisse https://www.ria.ee/sites/default/files/content-editors/publikatsioonid/avaandmete_loomise_juhend.pdf
- Eesti Avatud Rööpkorpus https://metashare.tilde.com/repository/search/?q=estonian+open+parallel+corpus
- EKI keelekool: kodanik XXXX XXXX isikukoodiga XXXXXXXXXXX soovib taotleda dokumenti... https://leht.postimees.ee/7484652/ak-eki-keelekool-kodanik-xxxx-xxxx-isikukoodiga-xxxxxxxxxxx-soovib-taotleda-dokumenti#_ga=2.168653093.509453857.1665129767-903145450.1647614642
- Tõlkevärava äri-, arhitektuuri- ja mõjuanalüüs, kasutajaliidese prototüüp https://wiki.rik.ee/pages/viewpage.action?pageId=82480517

- Tõlkemäludest https://phrase.com/blog/posts/translation-memory/
- https://mapa-demo.pangeamt.com/mapa/1.0/anonymization/model_showcase