Nimeüksuste tuvastaja loomine puudepanga korpuse põhjal
Juhendaja: Siim Orasmaa (*siim . orasmaa @ ut . ee)*

Nimeüksuste tuvastamise eesmärgiks on automaatselt leida ja märgendada tekstis olulised infoüksused, nt isiku-, organisatsiooni- ja kohanimed. Eesti keele jaoks on välja töötatud mitmeid nimeüksuste tuvastamise mudeleid, nt EstBert-il põhinevad neuromudelid (Sirts, 2023) ning EstRoberta-l põhinev vana kirjakeele nimeüksuste tuvastamise mudel (Orasmaa jt 2022). Ka Eesti keele puudepangale on lisatud käsitsi nimeüksuste märgendused (Muischnek ja Müürisep 2023), aga teadaolevalt ei ole sellele korpusel veel automaatseid nimeüksuste tuvastamise katseid tehtud ega mudeleid loodud. Käesoleva töö eesmärgiks ongi täita see lünk ning eksperimenteerida nimeüksuste tuvastamisega puudepanga korpustel (nii kirjakeele kui veebikeele puudepankadel), leida parim nimeüksuste tuvastamise mudel ning võrrelda selle märgenduskvaliteeti ka olemasolevate EstBert-il põhinevad neuromudelite (Sirts, 2023) kvaliteediga.

Töö eeldab huvi nimeüksuste tuvastamise temaatika ja masinõppe vastu ning head Pythoni programmeerimisoskust.

Viited:

- Muischnek, Kadri, and Kaili Müürisep. "Named Entity layer in Estonian UD treebanks." In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 179-184. 2023.
- Sirts, Kairit. "Estonian Named Entity Recognition: New Datasets and Models." In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 752-761. 2023.
- Orasmaa, Siim, Kadri Muischnek, Kristjan Poska, and Anna Edela. "Named entity recognition in Estonian 19th century parish court records." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 5304-5313. 2022.