**Thesis topics offered by Prof. Goran Glavaš**

**Topic #3: Analysis of Impact of Explicit Syntax in Language Learning and Understanding?**

**Level:** Advanced, appropriate for motivated master or bachelor students, ideally with some experience with NLP and modern machine learning and NLP libraries (e.g., PyTorch, Transformers).

**Short description:** Explicit syntax produced by (constituency or dependency) parsers has long been the backbone of virtually any natural language understanding approach. Recent large-scale pretraining of deep neural language models like BERT [1] and has enabled the so-called pre-training-fine-tuning paradigm in which explicit syntax is no longer needed for successful solving of language understanding tasks [2, 3]. What is more, recent work demonstrates that large pretrained Transformers contain much of the common syntactic knowledge [4, 5], e.g., such as the one encoded by treebanks like Universal Dependencies [6].

While syntax is clearly a type of strong linguistic inductive bias, these recent results would suggest that the knowledge gained from such bias can be compensated from large corpora if such corpora is available for large-scale pretraining. This opens two important research questions to be explored in this thesis: (1) at which pretraining scale (i.e., what corpus size) do language models (reliably) obtain syntactic knowledge of a language? (2) at which pretraining scale does the injection of explicit syntactic knowledge (e.g., in the form of intermediate parsing training on treebanks) becomes irrelevant for the downstream language understanding performance and (3) for which languages in the world do we have large enough corpora so that explicit syntactic knowledge becomes irrelevant?

**References:**


