One way to provide deeper insight into data and to build more powerful, robust models is bridging between linguistic knowledge and statistical learning. I'll present model-based approaches that incorporate linguistic knowledge in novel ways. First, I'll show how linguistic knowledge comes to the rescue in processing languages which lack large data resources. I'll describe a new approach to cross-lingual knowledge transfer that models the historical process of lexical borrowing between languages, and I will show how its predictions can be used to improve statistical machine translation systems. In the second part of my talk, I'll argue that insight into linguistic coherence, prototypicality, simplicity, and diversity of data helps improve learning also in resource-rich conditions. I'll present a novel method that optimizes—using Bayesian optimization—linguistic content and structure of training data to find a better curriculum for learning distributed representations of words. With improved word representations, we improve part-of-speech tagging, parsing, named entity recognition, and sentiment analysis. I'll conclude with future research questions.

Bio:
Yulia Tsvetkov is a PhD candidate in the Language Technologies Institute at Carnegie Mellon University. She works with professor Chris Dyer on using machine learning to tackle cross-lingual and cross-domain problems in natural language processing, focusing on computational phonology and morphology, distributional and lexical semantics, and statistical machine translation of both text and speech. Yulia completed her undergraduate degree in Technion, the Israeli Institute of Technology, and her master's degree in the University of Haifa, where she received several excellence awards.