From a multilingual parallel corpus to the micro-typology of the PERFECT in Baltic and Slavic

All over the world people use around 6500 languages. Obviously, languages vary, but this variation is systematic and largely predictable. Languages from different language families and from geographically unrelated areas share many phonological, structural and semantic properties. Uncovering universal properties and predictable patterns of variation among human languages and understanding how they follow from the organization of our linguistic knowledge in the mind is at the center of linguistic research. In contrast to language universals and parameters of phonological or morphosyntactic variation, little attention has been paid to semantic universals. Recently semantics has gradually come into focus in studies of cross-linguistic variation. Scholars suggest that potential semantic universals may be found at the level of function morphemes (small elements which function as syntactic glue in sentences). The list of such morphemes in each language is limited in contrast to open-ended list of content items. Moreover, the World Atlas of Language Structures shows that such elements tend to exist in many languages of the world independently of their genetic relatedness and geographical distance. For example, numerous languages have function morphemes expressing past test, negation, number, questions. Therefore, it is reasonable to assume that we should look for semantic universals by comparing the meaning of function words across different languages, hoping that it will bring us closer to understanding the mysterious nature of the cognitive architecture of our human language.

Even though the goal of linguistics is to uncover language universals and tendencies, most theoretical research is based on random data and constructed examples. A different perspective is taken by computational linguists, who create language corpora with millions of words and constructions. Such approaches to language are changing the field of linguistics. However, even though quantitative corpus methodology has proven fruitful in language typology, it is still difficult to bridge computational and theoretical linguists. Recently, progress has been made by scholars who apply quantitative corpus methods in the field of semantic micro-typology by exploiting the possibilities of parallel corpora. The goal of this project is to join this line of research and apply quantitative corpus methods in the field of semantic micro-typology of Slavic and Baltic languages by focusing on one of the most cross-linguistically challenging tense-aspect categories: the PERFECT. According to Dahl and Vlelupillai (2013) the 'PERFECT is used to express events that took place before the temporal reference point but which have an effect on or are in some way still relevant at that point'. However, the semantics of the PERFECT is subject to considerable cross-linguistic variation. The goal of this project is to use an innovative methodology dubbed Translation Mining in collaboration with our partners from Utrecht University to systematize this variation in Baltic and Slavic languages by comparing the original version of selected novels and their translations to Russian, Ukrainian, Belarussian, Polish, Czech, Slovak, Serbian, Slovene, Macedonian, Bulgarian, Lithuanian and Latvian. Based on the inductive quantitative research, we will develop the descriptive statistics of tense and aspect use in the investigated languages and we will visualize it using the computational technique called Multidimensional Scaling, which generates temporal maps of hundreds of datapoints connected with the underlying contexts of use of the investigated constructions in individual languages in a multilingual dataset. This method of visualizing big data (not easily grasped by the human brain) considerably facilitates testing hypotheses from semantic and typological literature. This method has been successfully applied in research on the competition between PERFECT and PAST tenses in English, German, Dutch, French, Spanish, Italian and Greek. We think that the Balto-Slavic language group constituting the second largest language family in Europe deserves its place in this dynamically developing research trend. Moreover, Baltic languages have preserved numerous properties of their ancestor Proto-Indo-European and they constitute a window to the genesis of the Indo-European language family.

Our research will be synchronic in nature but we will attempt to gain insights concerning the observed patterns of variation by connecting them with relevant facts about language change in tense-aspect grammars of Slavic and Baltic languages. Additionally, even though theories based on inductive corpus research are insightful, they need to be further verified using experimental methods and controlled data. To adequately interpret the data and to draw reliable conclusions, we will plan interviews with native speaker consultants and crowdsourcing-based acceptability rating and scenario-based elicitation experiments.