

Building statistical models to compare rival word-formation strategies: the case of Russian compound agent nouns

Chiara Naccarato^{1,2}

¹ University of Pavia ,

² University of Bergamo

Abstract. The aim of the talk is to discuss the validity of different statistical methods that can be employed to compare rival word-formation strategies. The object of the research is constituted by rival constructions that form compound agent nouns in Russian and, specifically, the focus is restricted to verb-based suffixed compounds such as *basn-o-pisec* ‘fable writer’. The rival constructions under investigation are formed with the following agentive suffixes: *-ec*, *-lec*, *-tel’*, *-nik*, *-čik/ščik*, *-l’ščik*, *-ka*, *-lka*, *-ø*. To understand in what ways these constructions differ one from the other, the compounds are analyzed according to a number of parameters, i.e. the part of speech and the semantic role of the first member of the compound, the transitivity and the formal aspect of the verbal element of the compound, the animacy of the referent denoted by the compound, and the semantics of the compound. Different statistical methods are employed to determine what parameters contribute most to distinguishing between the rival constructions under examination. First, a binomial logistic regression analysis is performed to compare the constructions formed with an expressed suffix against the zero-suffix construction. Then, to compare each construction against one another, I resort to two different statistical methods: a) multinomial logistic regression, and b) conditional inference trees and random forests. The analysis shows that the latter two methods appear to give us a better understanding of the behavioral profile of the rival constructions examined compared to binomial logistic regression, which suggests that the main differences among the rival constructions are not to be found in the opposition “expressed suffix vs. zero suffix”. In fact, it is the comparison among all the rival word-formation strategies that seems to produce more significant results. A number of issues related to the validity of the results obtained also need to be taken into consideration: What happens when we have too many points in our dataset? Does the big size of the dataset negatively affect the reliability of the results obtained? These (still open) questions will also be addressed during the talk.