Directions of Dependency Structures in the Czech National Corpus SYN2020: Application to Genre Classification

Xiyning Chen[1], Miroslav Kubát[2], Ján Mačutek[3]

[1]University of Ostrava – cici13306@gmail.com

[2]University of Ostrava – miroslav.kubat@gmail.com

[3]Mathematical Institute, Slovak Academy of Sciences / Constantine the Philosopher University in Nitra – jmacutek@yahoo.com

Abstract

This study looks into the features of syntactic structures within the Czech National Corpus SYN2020, a balanced corpus with 100 million words of contemporary written Czech from 2015 to 2019. SYN2020 is segmented into three major text-type groups: fiction, non-fiction, and journalistic texts. Each group further divides into several subcategories. The focus of this research is on examining four pivotal dependency structures, namely, subject, object, attribute, and adverbial, across these diverse text-types. Our objective is to analyze the dependency directions (head-initial and head-final percentages) of these structures and investigate whether their usage differs among the various text-types and subcategories. Such differences could pave the way for utilizing dependency directions as a syntactic index, offering a novel methodology for distinguishing between text-types. This approach has the potential to enhance our understanding of written Czech's syntactic details and may serve as a stylometric tool for classification.

Keywords: stylometry, genre, dependency syntax, Czech.