

**University of Szeged**  
**Doctoral School In Linguistics**  
**Theoretical Linguistics**

**Pronominal anaphora resolution in Hungarian texts**

**Summary**

**Kovács Viktória**

**Supervisor: Dr. Szécsényi Tibor**

**Szeged**

**2021**

The aim of the dissertation is to examine the results of the currently used supervised machine learning experimental methods for automatic anaphora resolution in Hungarian texts.

I used two corpora for the experiments: the SzegedKoref Corpus, which is the coreference annotated subcorpus of the Szeged Corpus, and for comparison the KorKorpusz. Machine learning experiments were performed using the Weka software, the Mention-pair model, and the Random forest algorithm. In these experiments the classifier makes decisions about pairs of mentions, namely, whether they are anaphorically related to each other or not, so for evaluation I used the MUC task evaluation metrics.

My null hypothesis is that it is possible to automatically resolve pronominal anaphoras in Hungarian texts without semantic information, only based on morphological, syntactical, and other surface structure-based features. My first hypothesis is that models achieve best results if we do not manually reduce the number of positive or negative examples in the training files. My second hypothesis is that selecting the pronoun-antecedent pair with the highest probability value brings greater efficiency. My third hypothesis is that adding the cognitive linguistic-based features to the machine learning experiment improves the success of the model building.

I pointed out that it is important: 1 the type of the text itself, as there are big differences between the machine learning experiments' results of the two corpora, 2 the type of the annotation, as it affects the quantity and quality of positive and negative examples, 3 the type of the pronoun, as pronouns behave differently from each other based on the examined aspects. It has been proved that in case of measuring distance between the two expressions it is important to consider not just the number of clauses but the relationship between the clauses. A further result of my experiments is the finding that the effect of the features I examined may differ when the goal is identifying more antecedents.