

Nguy Giang Linh, Zdeněk Žabokrtský

Faculty of Mathematics and Physics, Charles University, Prague

{linh, zabokrtsky}@ufal.mff.cuni.cz

1. Abstract

Our created system consists of handwritten rules developed and tested using the Treebank data, which contain more than 45,000 coreference links in almost 50,000 manually annotated Czech sentences. The F-measure of the system is 74.2%.

2. Layers in PDT 2.0

- **morphological layer (m-layer):** a lemma and a positional morphological tag are added to each token
- **analytical layer (a-layer):** a surface-syntactic dependency tree
- **tectogrammatical layer (t-layer):** a complex deep-syntactic dependency tree, in which only autosemantic words have nodes of their own. It also contains restored 'pro-dropped' subjects.

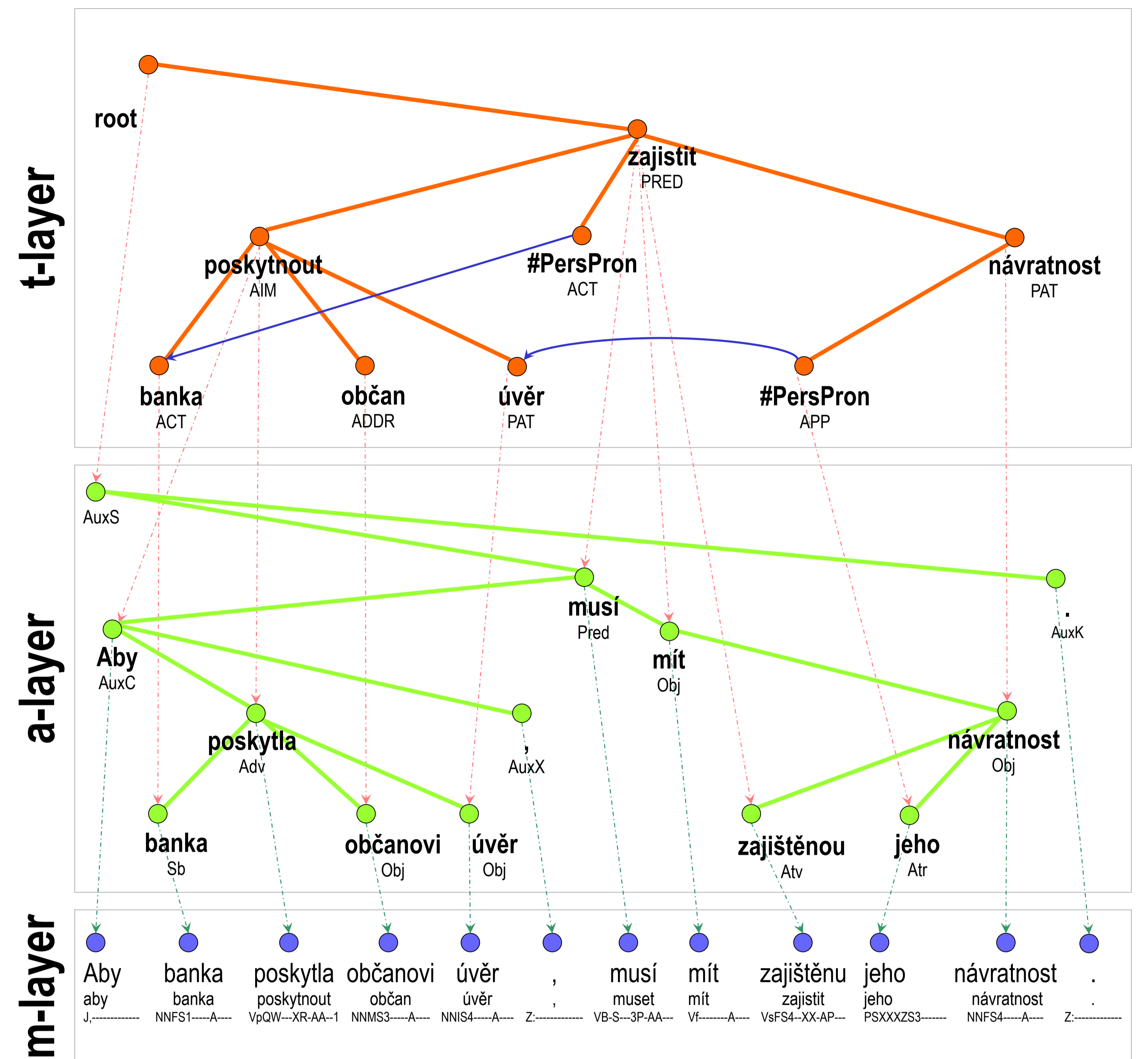


Figure: Linking the layers illustrated on the sentence *Aby banka poskytla občanovi úvěr, musí mít zajištěnou jeho návratnost.* (Lit. In order the bank to offer a citizen a loan, (it) must have collateralized its return.)

3. Coreference in PDT 2.0

PDT 2.0 contains 49,431 tectogramatically annotated sentences of newspaper texts. Coreference has been annotated manually in all this data. There are 46,908 coreference links (counting both textual and grammatical ones).

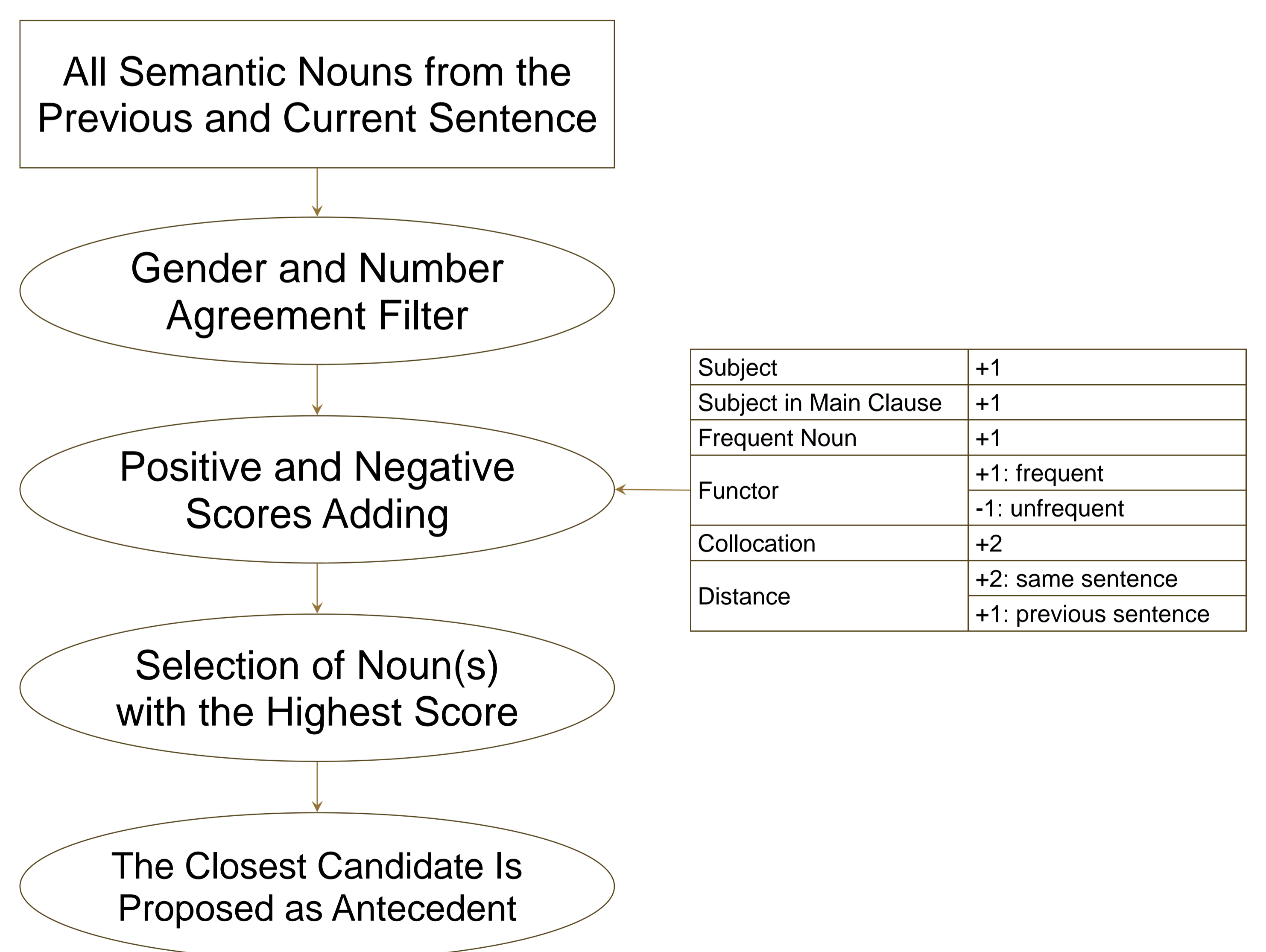
	Referring Expressions	Number
Textual	Personal Pronouns	16,690
	Possesive Pronouns	1,932
	Demonstrative Pronouns	1,494
	Pronouns Referring to Sentence(s)	2,174
Grammatical	Relative Pronouns	8,939
	Reflexive Pronouns	4,404
	Reciprocity Pronouns	1,114
	Verbs of Control	8,379
	Verbal Complements	1,782

5. Evaluation

Our approach was developed and tested on training and development test data. Finally it was tested on evaluation test data for the final scoring and gave the following results: precision 73.9%, recall 74.5%, F-measure 74.2%.

4. Rule-based Approach

In our system, the following procedure is used for each personal (and possessive) pronoun (expressed on the surface or restored during the annotation of the tectogrammatical tree structure – zero pronoun):



6. Final Remarks

In the future we would like to continue on improving Czech anaphora resolution with various statistical methods. We are also going to focus on resolution of bridging anaphora and noun phrase anaphora, the pilot data sets for which have been already annotated.