The Automatic Morphological Analysis of the Croatian Language- The Verbal, Nominal and Adjectival Inflections within the Morphological Parser HUMOR

Abstract

The present paper discusses the linguistic problems concerning the automatic morphological analysis of the Croatian language, namely the problems that concern the verbal, nominal and the adjectival inflections that arise when trying to develop a new application of the existing morphological parser HUMOR to the Croatian language. HUMOR, standing for High- Speed Unification Morphology, developed by MorphoLogic, has already successfully been implemented among others for a highly agglutinative language- the Hungarian, and an inflectional language- the Polish. Since the creators of MorphoLogic have argued that the analyzer had been suitable to all kinds of languages and all kinds of operating systems due to its unique system of operation, a new application of HUMOR has undergone further development, namely its implementation to another highly inflectional language- the Croatian. HUMOR itself, as the authors have already pointed out, has several applications. The main goal is not the development of industrial spelling checkers, hyphenators and thesauri, since these modules have been on the market for several years, but the linguistic parsing of lemmas for searching purposes, as well as the shallow or full parsing in translational supporting systems. The Croatian version of HUMOR, however, will undergo another application, namely the categorization of verbal, nominal and adjectival inflections, which, when summarized, will provide an additional help for the learning and teaching of Croatian as a second or foreign language.

When trying to develop a Croatian version of HUMOR, it has been essential to use the existing engine**,** namely the programme itself, but to produce a new database, which has encountered numerous problems concerning the discrepancy between the Croatian language policy and the status of the Croatian standard language. It has been essential to make up a Croatian lexicon, consisting of a minimum number of entries, but this course encountered numerous problems. Croatian language, belonging to the Slavic language group, has most of the time in history been paired with the Serbian and was therefore, until 1991, categorized only as a Serbo-Croatian language. Until that time, there had not been a contemporary Croatian dictionary or a contemporary codification. Due to the fact that the 19th ct and earlier codifications of the Croatian language have nowadays been considered archaic, it has been essential to decide upon the fact whether all the lexical entries in this project should belong to the contemporary standard language, or whether the programme itself should be able to analyze lexical entries of the 19th century Croatian language as well. The basis of the lexical part of HUMOR is the newest Anić's *Rječnik hrvatskoga jezika* (The Dictionary of the Croatian Language), which encounters 60,000 lexical entries. The categorization of the grammatical and inflectional rules has been made upon the Barić's *Hrvatska gramatika* (The Croatian Language Grammar) and the spelling rules upon the Babić- Finka- Moguš's *Hrvatski Pravopis*. Nevertheless, the linguistic dilemma concerning the authenticity of the further language corpus still remains.

After creating the lexical database, there had to be some actions done in order to obtain the linguistic categories, which will later be used by the parser itself. The lexical basis, therefore, as common to HUMOR, consists of a range of specially handled and categorized roots - *stems* and affixes- *terms*. The traditional expressions *root* and *affixes* are deliberately not used, since our definitions do not comply with the traditional categories. In our case the

stem is not the linguistically considered root of a certain word, to which e.g. suffixes can be added, but the part of the word which remains unchanged during the inflections.

The main linguistic problems and the dilemmas concerning the adaptation of HUMOR to the Croatian language, lie mostly in the insufficient information provided by the grammar textbooks and the definitions of certain grammatical and inflectional rules which have proven to be ambiguous, and therefore making the learning of the Croatian language as well as the clarification of certain conjugational and inflectional rules very difficult. According to the Croatian grammar, there are 6 main classes of conjugational verb groups in the Croatian language with more than 100 verb forms. The conjugational forms of verbs also include some archaic forms that are nowadays not being used, which makes the morphological clarifications more difficult due to the lack of the information needed. Whereas the grammar of e.g. the Hungarian language describes only three tenses- the present, past and the future, the Croatian language operates (although in written form only) with six tenses, three of which describing the Past Tense only- *perfekt, aorist* and *imperfekt* and one the Past Perfect- *prošlo svršeno vrijeme.* The problem mentioned and considering only the past tenses lies in the fact that *aorist* and *imperfekt* have nowadays been considered archaic, and therefore out-of-use, but have still been implemented into the grammatical descriptions of the Croatian language. Also, unlike the agglutinative languages with the defined position of affixes, the inflected verbs in the Croatian language make up different cases with the help of prepositions as well. The analysis of a syntactical concordance of e.g. a PREP + VERB, however, belongs to further applications of HUMOR.

The nominal and adjectival inflections have also had to undergo certain adaptations when implementing them to HUMOR. The main problems here lie in the fact that there are certain discrepancies between the nominal and adjectival categories, i.e. the discrepancies between the "textbook" and the actual, inflectional noun or adjectival genders. Unlike the traditional categories, the adjectival inflectional forms can be reduced to merely 16 different items, which then make up different adjectival inflectional groups.

When trying to implement HUMOR to the morphological parsing of the Croatian language, several problems have occurred, all of which considering the linguistics. The issues described in this paper have been discussed until the present stage of the whole project. Nevertheless, although encountering several linguistic dilemmas when implementing HUMOR to the Croatian language, one should bare in mind the benefits of such a morphological analyzer and the linguistic uses not only for parsing, but for the development of translational systems for Croatian and other minor languages as well.