AUTOMATIC MORPHOLOGICAL ANALYSIS OF THE CROATIAN LANGUAGE
The Verbal, Adjectival and Nominal Inflections within the Morphological Parser HUMOR[1]

Melita Aleksa
Faculty of Philosophy Osijek, Croatia

1. Introduction

HUMOR, which stands for High–Speed Unification Morphology, is a unification–based morphological parser developed by Morphologic which has nowadays been used in many branches, primarily for the morphological parsing of different languages and secondly, as a basis for different computer–aided language tools. The programme itself, as the authors have claimed, is universal, that is, it can be applicable to all languages and modified for different tools, it is very fast in performing the analyses and is partially self–correctable. HUMOR has been used and developed among others for the morphological parsing of an agglutinative language — Hungarian, inflectional languages — English and German, and a highly–inflectional language — Polish. It has been implemented into different translational systems and spelling–checkers, the most important probably being the MobiMouse (a language tool used for the translation of words that appear on the user's screen), MobiDic and MobiCat (two–way computer dictionaries) and MetaMorpho (a translational system used for translating simple sentences from English into Hungarian).
The present paper discusses, however, another implementation of HUMOR to a highly–inflectional language — Croatian, and the application of different solutions used in HUMOR to other language spheres. Unlike the issues which occurred when describing the agglutinative languages, when implementing HUMOR to inflectional languages, different kinds of problems have to be solved and different grammatical and language solutions have to be made. The differences in the approaches to describing the languages and making a morphological parser lie not only in the nature of the languages itself, i.e. in the fact they belong to different language groups, but also in the language history and language policy.
The aim of this paper is, among others, to present these issues and the language solutions which have been made up until the present stage of the whole project, namely the descriptions of verbal, adjectival and nominal inflections in the Croatian language.
The final goal of the whole project is not only the development of the above mentioned language tools for the Croatian language, but also the clarification of certain grammatical and inflectional rules of the language, which will then make way to a more effective teaching and learning of Croatian as a second or foreign language.

2. HUMOR

2.1. The working principles of HUMOR

The first demo version of the unification–based morphological parser HUMOR was developed by Morphologic in 1992 with the purpose of morphological parsing of languages and its use and application in different language tools. The main goal has not been 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

[1] Support of Magyar Ösztöndíj Bizottság

purposes, as well as the shallow or full parsing in translational supporting systems (Proszéky–Kis 1999: 266). The usage of HUMOR was best explained by Prószéky–Kis (1999) who said that "the first point of using the morphological analyzer in the parser is to get as much linguistic information about a single word form as possible. The second point is using the basic principles of the morphological analyzer to implement the parser itself. This means that we either collect or generate phrase patterns on different linguistic levels (noun phrases, prepositional phrases, verbal phrases etc.) and compile a Humor–like lexicon of them. On a specific linguistic level, each atomic element of a pattern actually corresponds to a (more) complex structure on a lower linguistic level" (Prószéky–Kis 1999:267).
The shallow and full parsing of the lemmas is performed internally, whereas on the surface level the analyzed forms of words are presented. In the following example the programme was given the task to analyze the inflected form of the Croatian verb *dati (eng. to give)—damo (1st person Plural)*, which gave the following results:

Analysis of "damo":
(1) dama[Sf]=dam+o[51]
(2) dati[Vs]=da+mo[p1]

As it can be seen, the input has not only been associated with the inflected form of the verb *dati*, but another word has been taken into consideration. The form is namely compatible with the vocative singular (51) case of the feminine noun (Sf) *dama (eng. lady)*, and the 1st person plural (p1) of the same verb. There were approximately 25 000 words/sec analyzed on an average computer which again proves the efficiency of HUMOR.

2.2. The lexical basis of HUMOR

Apart from the engine of the programme itself, the crucial part for the parsing of lemmas in the language is summoned in the lexicon. The Croatian lexical database consists of approximately 60 000 lexical entries, the core of it is the newest edition of Anić's (2000) *Rječnik hrvatskoga jezika (Dictionary of the Croatian Language).* The lexical entries from the dictionary have been included in the lexical database, new lemmas can be added or duplicated if necessary. This is especially important if the language itself allows more than one possibility e.g. regarding the orthography of words. The lexical entry has to be duplicated then and the lexicon naturally expands. "The metadictionary mechanism retains many advantages of the two–level systems. It means in the practice that users can add entries to the running system without re–compiling it" (Prószéky–Kis 1999:266).
The categorization of the grammatical and inflectional entries in the lexicon has been made in the following works: Barić et.al. (1995) *Hrvatska gramatika* (*The Croatian Grammar),* Raguž (1997) *Praktična hrvatska gramatika (The Practical Croatian Grammar),* Silić–Pranjković (2005) *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta (The Croatian Grammar for High–Schools and Faculties),* Težak–Babić (2005) *Gramatika hrvatskoga jezika: priručnik za osnovno jezično obrazovanje (The Croatian Language Elementary Grammar)* as well as Težak's works (1991, 1995, 1999, 2000). The spelling rules were taken from Babić–Finka–Moguš's (1996) *Hrvatski Pravopis (The Croatian Orthography)*. These choices and the problems concerning the selected works will be discussed later in this paper.

When considering the different notions used in HUMOR and the providing of grammatical input and markings into the lexical database, the following facts have to be taken into account. "Humor 99 lexicons contain stem allomorphs (...) instead of single stems. Relations

among allomorphs of the same base form (e.g. *wolf, wolv*) are, however, important for syntax, semantics and the end–user. An online morphological parser needs not be directly concerned with the derivation of allomorphs from their base forms, for example, it does not matter how *happi* is derived from happy before *–ly*. This phenomenon — a consequence of the orthographical system — is handled by the off–line linguistic process of Humor 99, which makes the analysis much faster. This method is close to the lexicon compilation used in finite–state models" (Prószéky–Kis 1999:262). According to these facts, the traditional categories of *roots* and *affixes* have not been used, however, new categories of *stems* and *terms* have been introduced: "Concatenation of stem allomorphs and suffix allomorphs is licensed with the help of the following two factors: continuation classes defined by paradigm descriptions and classes of surface allomorphs. The latter is a cross–classification of the paradigms according to phonological and graphemic properties of the surface forms. Both verbal and nominal stem allomorphs can be characterized by sets of suffix allomorphs that can follow them. When describing the behaviour of stems, all suffix combinations beginning with the same morpheme are considered equivalent because the only relevant pieces of information come from the suffix that immediately follows the stem" (Prószéky–Kis 1999: 262).

The main difference between agglutinative and inflectional languages lies mainly in the description and determining of these items, upon which the whole parsing is conducted, i.e. all the words have to be analyzed according to these beneath–the–surface categories (*stems* and *terms*). It is also important to notice that a word can consist of a Ø–term, i.e., similar to the fact that a certain word could have a Ø–morpheme but then again, it is not possible that a word consists of a Ø–stem. The categories of stems and terms do not comply with the traditional categories of roots and affixes and therefore the traditional notions have deliberately not been used. For example, when describing the Croatian noun *brzina (eng. speed)*, unlike in traditional morphology where the word is divided into the root morpheme *brz–*, followed by the affixes *–in–* and *–a* (Silić–Pranjković 2005: 164), HUMOR treats the word as a compound of a stem *brzin–* and a term *–a*. It is simply due to the fact that the part *brzin–* remains unchanged during the inflections and the segmentation of a word–form is based on surface patterns. This means that typical sequences of separate suffix morphemes are analyzed as a whole" (Prószéky–Kis 1999:261).

All of the stems and terms belonging to the Croatian lexicon have been categorized and labelled accordingly, as shown on the example of the surface output of the noun *stranke (eng. parties)*:

Analysis of "stranke"
(3) stranka[Sf]=stran+ke[21]
(4) stranka[Sf]=stran+ke[12;42;52]

The terms have been divided from the stems by the + sign. The first number in the brackets stands for the numbering of the grammatical case (1 standing for the nominative, 2 the genitive, 3 the dative, 4 the accusative, 5 the vocative, 6 the locative and 7 the instrumental case) and the second number for the grammatical number (1– singular, 2– plural). There can be several marks assigned to one term, as shown in (4). Nevertheless, examples (1) and (2) are "a typical stemming problem where the computer is not entitled to choose between the different possible stems. In these cases, all stems must be returned. Choice is a task of either the end–user or a disambiguator module that is based on the context of the word" (Prószéky–Kis 1999:267).

2.3. Testing the system

In order to verify the accuracy of the described data, the whole system had to be tested on a large Croatian language corpus consisting of approximately 50 million words. As previously mentioned, HUMOR is a partially self–correctable system, meaning that mistakes are easily discovered when being tested on a corpus, whereas the lexical entries have to be manually modified. The works included in the corpus belong to five categories, the first consisting of the most important Croatian literary works written in the Croatian language, i.e. the most representative novels, plays and poems. The choice of genres was influenced by the need to cover, if possible, most of the language spheres when compiling a testing corpus. Since the plays represent the most accurate version of the spoken language, they naturally had to be taken into consideration.

The dilemma regarding the choice and ranking of the most important literary works referred in the first place to the language problem and secondly to the subjectivity of the matter. The latter was solved by choosing a compilation of the most important literary works published by Bulaja (1999, 2000, 2002) *Klasici hrvatske književnosti (The Classics of the Croatian Literature)*, who had already done some research and compiled the most important literary works written in the Croatian language from the 16th to the 20th century. Due to presumably legal rights the most contemporary works are not included in the compilation. But from a linguistic point of view, as they were written in the contemporary standard language, the data was compensated for by using other contemporary language sources. The language problem mentioned concerns the question of the parser itself and the archaisms. Due to the facts which will be discussed further on in this paper, only the 20th century literary works have been taken into consideration, namely the works written in contemporary Croatian language.

The second category in the corpus consists of media resources, articles published in recent times, since they presumably reflect the current language situation. The newspaper articles included originate from different resources, including daily and weekly newspapers, women and men magazines and also teenage magazines. The aim was to cover most of the specific language varieties and language codes used by different speaker groups.

The third category consists of texts compiled from the internet, the most important resources being blogs. When taking text from the internet into account, there are two problems which had to be solved. Firstly, not all of the texts have been written in Croatian language, but many of them contain words of Serbian origin, due to the fact that the contemporary Croatian language resolved from the 1990s. The official standard language until that time was the Serbo–Croatian language, the residues of which are still present in the spoken language. The second problem concerning the internet texts is the use of sometimes superficial or inconsistent characters. Therefore they had to be bowdlerised, the unnecessary elements had to be deleted and after that they were included into the corpus.

The fourth category includes texts from textbooks, as these also contain a specific register, as well as texts from schoolbooks, which belong to the last, fifth category.

2.4. Further applications of HUMOR

The universality of HUMOR can be seen in numerous implementations of the programme to different language tools. As Prószéky–Kis (2002) state, "morphological analysis has three main purposes: (a) linguistic stemming for accurate dictionary lookups, (b) spelling correction and (c) preparation of shallow parsing of the context to identify candidates for multi–word expressions" (Prószéky–Kis 2002:3).

HUMOR is a basis for several programmes: the MobiDic dictionaries, now successfully being used among others for English, German and Hungarian languages (the experimental versions also include Spanish, Polish and Japanese) (Prószéky–Kis 2002:3), a translational system MetaMorpho, currently used for translating simple sentences from English into Hungarian, and a MobiMouse, a translational system used to offer translations of words and expression displayed on a computer screen, successfully using among others the English, German and Hungarian languages. The tools that use HUMOR are described as context–sensitive instant comprehension tools, more than a dictionary lookup engine as they tailor dictionary entries to the context of the translation point. The tool is less than a translation engine, as it performs no syntactic processing of the source text, only a series of dictionary lookups (Prószéky–Kis 2002).
Further applications of HUMOR also include publishing the Croatian morphological database consisting of the morphologically parsed lemmas with the grammatical descriptions, enabling searches in the corpus and clarifying the grammatical rules regarding the inflections, which can then provide a great help for the learning and teaching of Croatian as a second or foreign language. It is important to notice that there are no such works in Croatian that would clarify and precisely determine verbal and nominal inflections in the Croatian language, whereas the Hungarians are able to rely on the Papp's (1969) *A magyar nyelv szóvégmutató szótára* and Elekfi's (1994) *Magyar ragozási szótár*. These linguistic works provide additional help for the learners of Hungarian, by providing an insight into the morphological system of the language itself. The Croatian implementation of HUMOR will hopefully provide an important basis for developing similar works in the Croatian language.

3. The Croatian lexical database

When constructing the HUMOR Croatian lexical database, different kinds of problems had to be solved (in opposite to other languages implemented to this kind of morphological parsing), concerning the language policy, the reference works which were taken into account and the issue of parsing the written or the spoken language as well.
After the process of defining the characters belonging to the Croatian alphabet, it has been essential to make up a lexicon consisting of a minimum number of entries, which encountered several problems, the first one concerning the language policy.

3.1. The language policy

As it is commonly known, the Croatian language, similarly to other languages belonging to the South Slavic language group has gone through a long period of codification. It first began in the 16th century, followed by different attempts to unify the orthography on this area. Until the 19th century there were several initiatives to apply a unified orthographical system to all the versions of the Croatian language, meaning the dialects in which the works were being written. The authors at that time applied different orthographical rules, depending on the areas of their habitats; the Dalmatians using the Italian orthography and the scholars living in the northern part of Croatia applying the Hungarian way of marking the palatals. Later, however, there were some attempts to apply the orthographical rule of one phoneme – one grapheme, though unsuccessfully (Moguš 1995:73). When examining texts until the 19th century, one could find the following graphemes used for marking the contemporary ones:

The contemporary graphemes	Some of the graphemes used until the 19th century
č	*ç, ʒ, cs, ts, cz*
ć	*c', ch, tj*
đ	*gh, dy*
lj	*l̓, ļ, ly, gl*
nj	*ñ, ń,, nj, ny, gn*
š	*ſ, sh,*
ž	*ẓ, sh, x*
-je /- ije	*ě*

The major problem when trying to analyze the earlier texts is not only the orthography, but the inconsistency of its use as well. Different graphological rules had been applied according to different dialects of speech, which means that one phoneme had several graphical representations. Until the 19th century, there were several dialects used for the written language, including the *koine* – the mixture of three Croatian dialects (Moguš 1995: 59). Although there are several dictionaries from that period, some of which incorporate the Croatian dialects at that time (e.g. Sušnik–Jambrešić's (1742) *Lexicon latinum interpretationae illyrica, germanica et hungrica locuples*), the major problem if trying to apply HUMOR to the morphological parsing of the earlier Croatian texts would thus lie in the multiplicity of the versions used, followed by different orthographical codifications of the language.

The 19th century, given the orthographical reforms proposed by Ilirici – reformers working among others on unifying the orthographical system and codifying a Croatian standard language with the prevail of one dialect, provides texts with a somewhat unified orthography, but a declensional system not accepted by all Croatian regions. This period can be characterized as an attempt to establish a standard Croatian language, followed by the unifying idea of a brotherhood and a making of a new language standard – the Croato–Serbian or Serbo–Croatian. The contemporary standard Croatian language began its development and struggle in 1971, was interrupted, but continued in the 1990s, clearly distinguishing the Croatian language from the Serbian language (Moguš 1995).

When summarized, the Croatian language can be roughly divided into three major periods – the period until the 1830s, with the not unified orthographical system and no standard written language, the period until the 1990s, having the attribute of a Serbo–Croatian language and the contemporary Croatian standard language from the 1990s until the present day.

When implementing HUMOR to the morphological parsing of the Croatian language the major dilemma concerned the language policy mentioned. Since the Polish version of HUMOR makes it possible to morphologically parse the 18th century texts as well, the question regarding the Croatian version was also whether this should be enabled for this version of HUMOR too. Given the afore–mentioned reasons, the only solution would lie in the possibility to analyze texts written after 1830s, but again with certain discrepancies. The language until the 1990s, Serbo–Croatian, is on the morphological, syntactical and lexical levels different from contemporary Croatian language. The language policy presently considers the Croatian and Serbian languages as separate languages and not dialects. However, there have been a number of linguists who claim that these two languages are merely dialects of the same language, namely the Serbo–Croatian (Wardhaugh 1991). In his work, Wardhaugh (1991: 29) argues that the main differences between the Croatian and Serbian languages lie mainly in word preferences, and that there are no grammatical or phonetical differences between these two languages. This opinion has been criticized by Croatian linguists, who argue that the differences between Croatian and Serbian lie in every

language sphere and include differences in morphology, syntax and semantics as well as phonetics (Težak 2004). Regarding the contemporary language policy, one can conclude that Serbo–Croatian was developed either by the process of synthesis or the analysis from these two languages. Since 1991, many works have been written in order to codify the contemporary Croatian language and point out the differences between Croatian and Serbian, like Brodnjak's (1991) *Razlikovni rječnik srpskog i hrvatskog jezika (Dictionary of Differences between the Croatian and Serbian languages).* Words that have been included in this dictionary had been categorized as belonging unexceptionally to the Croatian or the Serbian language. The problem lies in the fact that the spoken language, as well as some of the contemporary texts written in Croatian language (especially texts from the internet sources – blogs), still contain words that, according to this dictionary, belong to the Serbian language. For example, according to Brodnjak (1991: 411) the Serbian word *ponekad (eng. sometimes)* has its counterpart in the Croatian language, namely *katkad*. Nevertheless, when analysing Croatian internet pages, the word *ponekad* occurs approximately on 110 000 pages (Google 2005), which proves the fact that the everyday use of the language cannot be considered as a standard Croatian language, imposed by the language policy and the media, but rather the remains of the former Serbo–Croatian.

Given all these reasons, if the data containing the traces of the Serbo–Croatian were left out of the lexical database i.e. if the words belonging unexceptionally to the Serbian language were left out of the Croatian lexical database, the parser would not be able to use the corpus consisting of the material published until the 1990s, including also some of the major literary works. It is also important to mention that several 20th century Croatian authors also used the concoction of these two languages and made up a new sort of a literary koine: the Nobel–prize awarded Ivo Andrić invented his own language, using Serbian lexemes and Croatian syntax, due to which his works cannot be included into the testing corpus.

Therefore, if not the "purist" version of HUMOR was chosen, meaning the approach to examine earlier texts written in Serbo–Croatian language, it would then mean, naturally, expanding the current lexical database and including the lexemes from the Serbian language to the existing morphological parser. In other words, it would mean developing a morphological analyser which would recognize a great number of lexical entries in texts with a questionable Croatian / Serbian origin i.e. implementing the parser for the Serbian language as well. The question is, if this approach was chosen, whether the parser could then be referred to as the Croatian morphological parser.

On the other hand, if it was chosen not to incorporate Serbian lexemes and the differences concerning the Serbian language on the morphological level as well, only one part of the 20th century data would be examined due to the missing lexemes. This way the morphological parser would rely only on the contemporary standard Croatian language only, meaning that some of the most important texts from that period would not be included into the corpus.

3.2. Reference works

The present Croatian lexical database, as mentioned in 2.2, consists of language data from several works. The lexical entries, the lemmas of the database, have been taken from Anić's (2000) *Rječnik hrvatskoga jezika (Dictionary of the Croatian language)*, whereas for the grammatical description and references the following grammar textbooks have been used: Barić et al. (1995) *Hrvatska Gramatika (The Croatian Grammar)*, Raguž (1997) *Praktična hrvatska gramatika (The Practical Croatian Grammar*, Silić–Pranjković(2005) *Gramatika hrvatskoga jezika (The Croatian Language Grammar)* and Težak – Babić (2005) *Gramatika hrvatskoga jezika: priručnik za osnovno jezično obrazovanje* (*The Croatian Language*

Elementary Grammar Handbook). The spelling rules, as already mentioned, have been made upon Babić– Finka–Moguš's (1996) *Hrvatski Pravopis (The Croatian Orthography)*. Apart from the testing corpus, for the verification of the data Moguš–Bratanić–Tadić's (1999) *Hrvatski čestotni rječnik (The Croatian Dictionary of Word Occurrences)* has been used.
The reason for using several grammar reference textbooks lies mainly in the difference of the approaches the authors have used when describing the Croatian grammar. The differences can be seen in the use of the terminology the authors impose on readers, as well as the diversity of the language solutions presented in these grammar books, which will be explained later in the succeeding chapters.
When considering the works mentioned, one should conclude that these references provide lots of useful information regarding the Croatian language, but nevertheless have some drawbacks as well, especially considering the clarification of grammar rules and the approaches that have been used. The core of the lexical database in HUMOR is represented by the lexical entries from Anić's (2000) dictionary. The dictionary is said to contain 60 000 lexical entries belonging to the modern standard language, the number of which after eliminating the unnecessary data has been reduced to 56 000. From the aspect of HUMOR, the unnecessary data included among others duplicated lexical entries, as well as nominalised letters of the Croatian alphabet, e.g. *S, n. (the letter "S", neuter)*. The main problem when considering this dictionary is the occasional discrepancy between the dictionary data and the standard language. For example, the orthographical doublets like in the example of the adjective Hungarian (*mađarski* and *madžarski*) have not been included into the dictionary, as well as some words that occur in the testing corpus. When considering the dictionary lemmas there are questionable data concerning the genders of the nouns, which then had to be verified through the grammar books as is explained in 6.1.
The verification of the described data apart from the corpus obtained for the purposes of HUMOR is also done by using the *Hrvatski čestotni rječnik* compiled upon the Croatian National Corpus. The main drawback of the dictionary form the aspect of morphological parsing is the occasionally missing data concerning the exact number of certain word forms occurring in each of the given subcorpora, as well as the non–listing of all word forms of a given lemma.

3.3. The choice of the language variety

The works listed, as well as the compiled lexical database, are used in describing and creating the morphological parsing of the written standard Croatian language. However, there have been certain dilemmas concerning the parsing of the spoken language as well and covering the different versions of the Croatian language, namely the dialects.
When deciding upon the language parsing, the first choice was put on the language codification and the testing corpus. The corpus used for testing the system was compiled of the data belonging to the written language, with some aspects (drama, blogs) being the nearest to the spoken language codification. Secondly, there is a lack of data concerning the codified spoken Croatian language which could be used in building the testing corpus.
Where the spoken language is concerned, the accentuation of the standard language also has to be considered. As there are lots of pronunciation varieties of words in the Croatian language, codifying all the regional varieties would demand additional research, with the uncertainty of further application and use. Furthermore, the standard variety of the prosodic elements of lemmas has thus been accounted for in Anić's (2000) dictionary. The problem concerning the further codification of word forms lies in the insufficiency of the data needed for performing this action. The reasons for choosing the standard written Croatian language

apart from those mentioned also lie in the complexity of the morphological descriptions. When taking the prosodic elements of language into account, the number of different stems and terms in HUMOR grows.
Considering all these elements, the decision was made to develop a parser primarily for the analysis of the standard written Croatian language. The flexibility of the programme itself makes it possible to extend the parser for the processing of the language including the prosodic elements of speech, which is regarded as one of the further aims.

4. Verbal inflection

When implementing HUMOR to the morphological parsing of the Croatian language, several linguistic dilemmas have occurred. The first one concerns verbal inflections.
There are approximately 11 000 verbs in the Croatian lexical database, which are divided according to the traditional grammar into six conjugational groups, each having a number of conjugational classes. The classification of the conjugational classes is also not unified. Barić (1995) and Raguž (1997) for example, mention only 7 conjugational classes in the first group, while Silić–Pranjković (2005) list 18 different conjugational classes in the same group. Since the verbal inflections in HUMOR depend upon the division of verbs according to their stems and terms, HUMOR uses 116 different conjugational types, upon which different paradigms can be generated or analyzed. Tadić (1994) in his work also stated that keeping the traditional system of word classifications when trying to use computational language descriptions leads to extremely complicated, if not impossible processing (Tadić 1994: 45).
The first linguistic dilemma concerning the verbal inflections in HUMOR regards the tenses. Unlike, for example, agglutinative languages that operate only with three tenses (the past, present and future tense), the Croatian language has got seven different tenses. There are four tenses used to describe the past, namely the *perfekt*, *aorist*, *imperfekt* and *pluskavmperfekt*.
The major dilemma when trying to describe the verbal inflections in the morphological parser lies in the usage of the tenses and in the verb forms. The Croatian grammar textbooks provide insufficient information as to the correct forms used in *imperfekt* and *aorist*, due to the fact that these two tenses are nowadays considered somewhat archaic but can still occur in everyday use. Raguž (1997:185) states that aorist occurs not so often, but more often than the textbooks claim it. Imperfekt, on the other hand, has been extinct from the modern language (Raguž 1997:186). Tadić (1994), when mentioning the imperfekt and aorist forms of the verb *zadronjaše* also states that the forms could be easily identified, however, it is not likely that anyone could identify the verb itself (Tadić 1994: 45).
Nevertheless, since they have been present in the Croatian written language and HUMOR is constructed to parse the written standard Croatian, these tenses have to be included in the morphological descriptions of verbal inflections, although the traditional grammar provides ambiguous information concerning them. Silić–Pranjković (2005) when mentioning the verbal forms only state that some of the forms are built by finite and some by infinite verbs (2005: 58). Barić (1995) defines aorist as the tense which can be built by "finite and rarely by infinite verbs"[2] (1995:238) whereas later in the text imperfekt is defined as a tense that can be built by "only the infinite verbs"[3] (1995:238). There is a similar definition in Raguž's (1997: 181ff) book as well. The major problem is that the grammar textbooks do not specify the "rare" infinite verbs which can have both tenses — aorist and imperfekt. It is interesting to note that to these exceptions belong also the most frequently used verbs in Croatian — the

[2] Translated by Melita Aleksa
[3] Translated by Melita Aleksa

infinite verbs *čitati (eng. to read), ljubiti (eng. to kiss)* and *dati (eng. to give)* (Barić et al 1995: 254ff), which again makes the learning of Croatian more difficult.
When considering the fact that the tenses mentioned are considered rather archaic, Silić–Pranjković (2005) take several items into account and state that aorist and imperfekt have been replaced by the perfect tense of finite verbs, although nowadays they have been being reused in communicating through electronic media as they occupy less free space (2005:66ff). Nevertheless, although some of the forms still occur in everyday use, one can not define for certain all their correct forms. Barić (1995) in his book when describing the verbs of the 1st group, 4th class, mentions several double imperfekt forms, like e.g. the forms of the verb *vući (eng. to pull) — vucijah* and *vučah.* However, it is not unanimous, whether other verbs belonging to the same group and class can have double forms as well. For example, is it likely that the verbs *peći (eng. to bake), sjeći (eng. to cut)* and *strići (eng. to shear)* can have both *pecijah / pečah, sjecijah / sječah* and *strigah / strizijah*?

The discrepancies between Anić's (2000) dictionary and Barić's (1995) grammar textbook affect not only the archaic forms of verbs, but the present tense as well. The inflected forms, e.g. of the verb *gnjiti* in the present tense, according to Barić (1995) are *gnjijem, gnjiješ, gnjije, gnjijemo, gnjijete, gnjiju* only, whereas Anić (2000) also mentions an additional form, *gnjim*, which then leads us to the questionable 3rd p. Pl. form of *gnjiti — gnju* or *gnjiju.*
The answers to these questions can be obtained only through testing the parser on a Croatian corpus.

The second linguistic dilemma regarding the verbal inflections in HUMOR concerns the notions of the *glagolska imenica (gerund*) and the *glagolski pridjev ("verbal adjective").*
The question is whether one should treat gerunds as separate lemmas having nominal characteristics or as parts of verbal paradigms. Raguž (1997) in his grammar textbook does not mention the category of gerunds. On the other hand, Silić–Pranjković (2005) do not define the notion separately, but use it when discussing other issues. Since Anić (2000) treats them as separate entries in the dictionary only when they have other meanings as well, this led to a conclusion to treat them accordingly in HUMOR, specifying their verbal origin. There occurred a similar problem when verbal adjectives came into question. Verbal adjectives (active or passive) are adjectives derived from the base forms of the verbs. Silić–Pranjković (2005) state that verbs can be "deverbalized" i.e. turned into verbal adjectives (2005: 383). This led to a possibility to treat them as parts of verbal paradigms in HUMOR. Furthermore, they are being used when deriving verbal forms in perfekt, pluskvamperfekt, both conditionals and futur II and possess all adjectival forms, apart from the definite ones (Raguž 1997: 197). They can have either a predicative (5) or an attributive (6) function, or be used as a special form of perfekt (7). Apart from these functions, a verbal adjective can have another function as well, namely the optative mood (8) (Raguž 1997:197).

(5) Jabuka je *poklonjena*.
The apple is *given away.*

(6) *Poklonjena* jabuka je crvena.
The *given* apple is red.

(7) *Pao* avion.
Fallen a plane.

(8) Bog te *blagoslovio!*
God you *blessed*!

When constructing his GENOBLIK, Tadić (1994) used the traditional solution, which meant that the derivational processes of making verbal adjectives and adverbs were included into the inflection (Tadić 1994:20). Since, from the aspect of morphological parsing in HUMOR, only the morphological characteristics of words are taken into account, the verbal adjectives have been treated as parts of verbal inflections. The actual drawback of this solution is of a statistical nature, as the number of Croatian adjectives (including the category of verbal adjectives) cannot be exactly determined.

One of the further issues regarding verbal inflections lies in the parser's analysis. The empty character i.e. the space or space-like character in HUMOR is defined as the analysis barrier. Due to this fact the Croatian tenses consisting of two words are not recognized as a whole, but each word is treated as a separate lexeme. This is the case with the tense perfekt, which is being inflected in the following way, regarding the different genders:

Perfekt of the verb *jesti (eng. to eat):*

			m	f	n
(9)	Sg.	1.	jeo sam	jela sam	—
(10)		2	jeo si	jela si	—
(11)		3.	jeo je	jela je	jelo je
(12)	Pl.	1.	jeli smo	jele smo	—
(13)		2	jeli ste	jele ste	—
(14)		3	jeli su	jele su	jela su

5. Adjectival inflection

Before describing adjectival inflections when implementing HUMOR to the Croatian language, there are certain facts that have to be clarified: The Croatian lexical database contains approximately 10 000 adjectives, the exact number of which cannot be determined due to the fact that adjectives derived from verbs are not represented as separate lexical entries in the dictionary, as described in Chapter 4.
As in every inflectional language, there are also in the Croatian language declinable and indeclinable adjectives. The latter category includes adjectives like *super (eng. great)*, *top (eng. top)* and *fit (eng. fit*). Since the present paper describes only the declinable adjectives, the category of indeclinable adjectives will be left out of discussion, since from the aspect of morphological parsing they belong to the invariable category of words.
One of the most important issues when describing the adjectives and the adjectival inflections in HUMOR is determining the category of adjectives themselves. As already stated, keeping all the traditional word categories in a morphological parser would lead to sometimes superficially complicated, if not impossible solutions. Therefore the solution implied here makes the boundaries of the adjective category broader, i.e. all the word classes that have adjective–like inflections are here considered as parts of the adjectives group.

5.1. Adjectival inflections according to the traditional grammar

According to the traditional grammar, the adjectives have been divided into several groups regarding their semantic categories, however not unanimously. Raguž (1997) mentions three

groups of adjectives to which all the adjectives in the Croatian language belong, namely the *opisni pridjevi (descriptive adjectives)*, *odnosni pridjevi (relational adjectives)* and *posvojni pridjevi (possessive adjectives)* (Raguž 1997:88). Apart from these categories, he also mentions the category of "definement", according to which there are *određeni (definite)* and *neodređeni (indefinite)* adjectives, which refer to descriptive adjectives only (Raguž 1997:88). Silić–Pranjković (2005) on the other hand divide adjectives into four categories, namely the *kakvoćni pridjevi (qualitative adjectives)*, *posvojni pridjevi (possessive adjectives)*, *gradivni pridjevi (material adjectives)* and *odnosni pridjevi (relational adjectives)* (Silić–Pranjković 2005:133f). The category of definement is here also present, but has not been associated with any of the adjectival groups mentioned above. Both grammar textbooks, however, include also the *living* and *non–living* category, which actually semantically defines the succeeding noun.

Considering the declension of adjectives, Silić–Pranjković (2005) take only the morphological characteristics into account, while Raguž (1997: 89) distinguishes between two types of declensions according to different categories. He states that the descriptive adjectives can have a nominal or a pronominal declension and provides somewhat ambiguous descriptions, which makes the learning of Croatian more difficult. According to the author the "nominal declension is not entirely nominal; it is a mixture of a nominal and a pronominal declension"[4] (Raguž 1997: 89).

When summarized, according to the traditional grammar, in order to inflect an adjective in three genders, singular and plural, one has to categorize it first semantically, then regarding its definement. Furthermore, one should decide upon the succeeding noun which type of declension s/he wants to use, i.e. the living or the non–living paradigm. This all leads to the fact that e.g. the adjective *crven (red)* in written standard Croatian has 123 inflectional cells (altogether 91 different inflectional cells), as shown in Figure 1. When the prosodic elements of the spoken language are taken into account, the number of different inflectional cells is even larger.

Figure 1. The inflectional forms of the adjective *crven (red)*

Sg.	**Masculine, defined living**	**Masculine, defined non-living**	**Masculine, undefined living**	**Masculine, undefined non-living**	**Feminine, defined**	**Feminine, undefined**	**Neutrum, defined**	**Neutrum, undefined**
N	crveni	crveni	crven	crven	cvena	crvena	crveno	crveno
G	crvenog/ga	crvenog/ga	crvena	crvena	cvene	crvene	crvenog/ga	crvena
D	crvenom/me/mu	crvenom/me/mu	crvenu	crvenu	cvenoj	crvenoj	crvenom/me/mu	crvenu
A	crvenog/ga	crveni	crvena	crven	cvenu	crvenu	crveno	crveno
V	crveni	crveni	-	-	crvena	-	crveno	-
L	crvenom/me/mu	crvenom/me/mu	crvenu	crvenu	cvenoj	crvenoj	crvenom/me/mu	crvenu
I	crvenim	crvenim	crvenim	crvenim	cvenom	crvenom	crvenim	crvenim

Pl.	**Masculine, defined**	**Masculine, undefined**	**Feminine, defined**	**Feminine, undefined**	**Neutrum, defined**	**Neutrum, undefined**
N	crveni	crveni	crvene	crvene	crvena	crvena
G	crvenih	crvenih	crvenih	crvenih	crvenih	crvenih
D	crvenim/ma	crvenim/ma	crvenim	crvenim	crvenim/ma	crvenim/ma
A	crvene	crvene	crvene	crvene	crvena	crvena
V	crveni	-	crvene	-	crvena	-
L	crvenim/ma	crvenim/ma	crvenim/ma	crvenim/ma	crvenim/ma	crvenim/ma
I	crvenim/ma	crvenim/ma	crvenim/ma	crvenim/ma	crvenim/ma	crvenim/ma

[4] Translated by Melita Aleksa

The Slavic languages allow in certain cases alternatives, which is why in the Croatian genitive, dative, accusative, locative and instrumental several declensional forms can be found.

When the above mentioned factors are taken into account, the learning and description of the Croatian adjectival declensions proves to be rather complicated and gives way to many questions. The first problem noticed is the semantic categorization of adjectives themselves. The categories like descriptive, possessive or qualitative adjectives are not included in the dictionaries, which makes the automatic, computer–aided parsing of adjectives according to the traditional grammar impossible. Apart from that, foreign language learners should find the generating of adjectival inflectional forms without the semantic background also dubious.
The second problem concerns the categorization of adjectives according to their definement and the living and non–living category. Whereas in HUMOR the space or a space–like character represents the analysis barrier, the categorizations that depend on the succeeding noun represent a rather complicated, if not an impossible action.
Due to all these facts, another system was needed, namely, the possibility to unify all the inflectional forms and generate, that is, analyze adjectival inflections without the use of the semantic background.

5.2. The solution used in HUMOR

All the declensional forms presented in Figure 1 can be reduced to merely 16 different items, as shown in Figure 2:

Figure 2.The inflectional forms of the adjective *crven (eng. red)* as used in HUMOR

1 crven-Ø	9 crven-ome
2 crven-a	10 crven-omu
3 crven-u	11 crven-e
4 crven-im	12 crven-ih
5 crven-i	13 crven-ima
6 crven-og	14 crven-oj
7 crven-oga	15 crven-o
8 crven-om	16 crven-om

The forms 8 and 16 are in the case of this adjective the same, but can differ when other adjectives are considered.

Furthermore, embedding merely these 16 inflectional forms into the declensional paradigm, there is an inflectional matrix of adjectival declensions generated:

Figure 3.The summarized inflectional paradigm of Figure 1 with the above mentioned 16 inflectional forms

Sg.	**masculine, non-living**	**masculine, living**	**neutrum**	**feminine**
N/#V	*1 / # 5		15	2
G	*2 / #6,7			11
D	*3 / #8,9,10			14
A	*1 / #5	*2 / #6,7	15	3
L	*3 / #8,9,10			14
I	4			16

Pl.	**masculine**	**neutrum**	**feminine**
N/#V	5	2	11
G	12		
D	4,13		
A	11	2	11
L	4,13		
I	4,13		

*** undefined, # defined**

The presented system has proven to be useful not only from the aspect of morphological parsing, but from the aspect of foreign language learning as well. There is no need for memorizing all the 91 different inflectional cells, but merely the 16 different inflectional forms (Figure 2) have to be learned and embedded into the inflectional matrix (Figure 3). The system does not rest on semantics, but on the morphological characteristics of adjectives only, which can be seen in the further categorization of adjectives themselves. According to their stems and terms, they can be divided into 21 inflectional types:

Figure 4. The inflectional types used in HUMOR

Infl. type	**The *term* of the adjective, Sg., nominative**		**Example**	**Number of entries in Anić's dictionary**
	Form 1	**Form 5**		
I	Ø	-i	crven	2185
II	-	-i	mačji	654
III	-	-i	hrvatski	4879
IV	-ar	-ri	dobar	14
V	-ao	-li	zao	34
VI	-o	-li	debeo	41
VII	-tao	-li	odrastao	2
VIII	-an-	-ni	čudan	1894
IX	-žak	-ški	težak	2
X	-alj	-lji	šupalj	1
XI	-ak	-ki	plitak	24

Infl. type	**The *term* of the adjective, Sg., nominative**		**Example**	**Number of entries in Anić's dictionary**
	Form 1	**Form 5**		
XII	-tan	-ni	koristan	47
XIII	-zak	-ski	uzak	9
XIV	-dak	-tki	sladak	7
XV	-o	-jeli	cio	1
XVI	-io	-jela	ishlapio	4
XVII	-av	-vi	ovakav	1
XVIII	Ø	-i	krnj	2
XIX	-bak	-pki	gibak	3
XX	-al	-li	obal	7
XXI	-ben	-bni	dioben	1
indekl.	indeclinable adjectives			38

Compared to the traditional grammar descriptions of adjectives, it can be said that the HUMOR–like forms 1 and 5 (base forms, dictionary entries) have proven to be the most important, due to several factors. If the dictionary entry includes both of them, the adjective is semantically and traditionally categorized as descriptive, which means that the base form is the form 1. On the other hand, if the base adjectival form is the form 5 only, the adjective can be then linked to Raguž's category of possessive adjectives. It can be clearly seen that the adjectives derived from nouns (e.g. *Marija (n.)* → *marijin (adj.)*) exclude the defined declension, whereas certain descriptive adjectives have no undefined inflectional forms (e.g. *muskulaturni (eng. muscle-like)*).

5.3. Comparison of adjectives

The Croatian adjectives make up their comparatives with the *–ši,–i(–ji)* suffixes and superlatives with adding the *naj–* prefix to the comparative form. When irregular adjectives are taken into account, their base form in comparative also changes. Since from the aspect of HUMOR these affixes then generate different stems and different terms, enlarging the number of declensional types and inflectional cells, they have been handled as separate lexical entries and included into the lexicon. Naturally, the markings specifying their origin are included into their morphological descriptions.

5.4. Linguistic dilemmas

When describing adjectival declension for the morphological parsing of the Croatian language, several linguistic dilemmas occurred. The first one concerns the problem of

adjectival declensions in the feminine gender only. According to Anić (2000), there are several adjectives that can be inflected only in the feminine gender, which rules out the possibility of inflecting them in other genders as well, like e.g. *trudna (eng. pregnant).* In this case the semantic and not the morphological properties of words have been taken into account. The problem arises when the linguistic usage proves the existence of other, morphological correct declensional forms. According to the Croatian translation of Jaroslav Hasek's *The Good Soldier Svejk: and His Fortunes in the World War*, the adjective pregnant can be used in the masculine gender as well (Hašek 2004:37). The problem concerns the doubt about whether to rely on textbooks or on the actual linguistic usage and allow the morphologically correct forms to be parsed as well. If the latter should be selected, the entry in the lexicon has to be altered as well, making then the masculine form 1 the base form of the adjective.

The second problem concerns the adverbs. According to the traditional grammar, they have been included into different categories regarding their origin – the nominal and adjectival adverbs, the latter being derived from adjectives (Raguž 1997: 271). The problem mentioned concerns the masculine adjectives with the suffix *–i*, adjectives belonging to the inflectional type II and III in HUMOR (Figure 4), like e.g. *mačji (eng. cat's)* and *hrvatski (eng. Croatian)*. According to Raguž (1997:271) adverbs in this group can only be derived from adjectives with suffixes *–ski* / *–čki*. The question is whether it is possible to derive and compare adverbs from other adjectives in this group as well. The answer as to whether these really exist or not cannot be proven by using the corpus analysis either since the adverbial forms coincide with either the adjectival forms 5 or 15.

The third problem includes the orthography and the existence of different orthographical solutions for the same adjective provided by some authors. For example, the adjective Hungarian can have according to Babić–Finka–Moguš's (1995) *Hrvatski pravopis* two orthographical representations, namely the *mađarski* and *madžarski* (Babić et al.1995: 277). These options, however, have not been included in Anić's dictionary — merely the first example has been taken into account. The only solution when constructing this morphological parser lies in the possibility of including both orthographical versions as separate lemmas in the HUMOR database.

6. Nominal inflections

Apart from verbal and adjectival inflections, there have also been some distinctions in the word class of nouns when considering the Croatian and other agglutinative or inflectional languages. Unlike Germanic languages, as well as Hungarian, where the nominal inflections require different approaches to the automatic parsing, the Croatian nouns, similarly to adjectives, are being inflected according to three genders, singular and plural, in seven cases. Although there is another grammatical number, the dual, it has no effect from the morphological point of view, as it coincides with either of the two numbers. The dual is merely evident when numbers are considered (Raguž 1997: 6).

There are approximately 27 000 nouns in the Croatian lexicon. According to the grammar textbook, there are several nominal categories, namely *vlastite imenice (personal nouns)* and *opće imenice (general nouns)* which can be further divided into *konkretne imenice (concrete nouns)*, and *apstraktne imenice (abstract nouns)* (Raguž 1997:4-5). Furthermore, Raguž (1997) also mentions additional categories regarding the further categorization of the mentioned nominal groups, namely the diminutives, augmentatives, as well as the gradation of nouns according to the "personal connection to something" (Raguž 1997:5). Given all this

information, one can conclude that a more formalistic approach would be needed when describing nominal declensions in the morphological parser HUMOR.
Silić–Pranjković (2005) have chosen a morphological approach when describing nouns and therefore categorized them according to their morphological characteristics, taking only the morphemes into account and differentiating the declensions according to the base form alternations (Silić–Pranjković 2005: 98ff). The only declensional category which they take into account is the category of a living and a non–living noun and the declensions are handled accordingly. The only problem that could occur if this categorization would be applied to HUMOR is the absence of the semantic markings from the lexical database, ones relating to the category of living or non–living beings, due to the fact that Anić's (2000) dictionary also lacks them. Raguž (1997) takes a different approach, naming three types of declensions, the *a–declension, e–declension* and *i–declension,* according to the genitive singular suffix. When describing the nouns that are being inflected according to the declensions named, the author introduces a rather complicated way of distinguishing them, relying on genders of the nouns as well as endings in the nominative case. This approach proves to be helpful when there are no alternations of the base form (Raguž 1997: 13). Nevertheless, since the system of the morphological parsing in HUMOR relies merely on stems and terms, the number of declensional categories naturally expands as the number of different terms also enlarges.

6.1. Linguistic problems considering the nominal declensions

The problems considering nominal declensions are the ambiguity of the grammar rules, the most important of which concerns the suffixes. The Croatian grammar, as could be seen earlier, allows alternations in certain cases. The problem occurs when the alternations cannot be used interchangeably, like is the case with the vocative singular in masculine and neuter gender. This case allows two suffixes to be used, namely *–e* or *–u*, depending on certain rules. The suffixes can be used also interchangeably, though not with all words. For example, the nouns ending in *–ar/-er/-ir* can have either *–e* or *–u*: *gospodar (eng. master) → gospodar**e** / gospodar**u***, with the ending *–u* being used more frequently (Raguž 1997:10). Although Raguž in his book specifies certain rules concerning the base form endings and the vocative case, there are still some ambiguities left, i.e. nouns that have not been included into any of the exceptions mentioned, but still forming an exception, e.g. *žal (eng. rief)→ žal**u*** only**,** or other nouns ending in *–st, –t* etc. Silić–Pranjković (2005) in their explanation do not mention the possibility of using alternative forms in the vocative case, but cite few examples in which either the first or the second ending mentioned can be seen. The grammatical definition as to when to use which ending is missing (Silić–Pranjković 2005:98ff).
The solution to the problem cannot be solved by using the corpus analysis either, as these endings coincide with endings in other cases. The partial solution would lie in excluding several cases where the vocative is being followed by the exclamation mark, but this again cannot guarantee a satisfying result.

The second problem regards the singularia and pluralia tantum i.e. the absence of sufficient information regarding declensions of these nouns. Anić (2000) when describing the word *vrata (eng. door)* or *djeca (eng. children)* specifies only the category of a pluralia tantum, which if not processed manually would lead to the false paradigm, namely the neuter singular declension instead of the neuter plural. The doubtful markings are also noticed when other words belonging to this category are being analysed. According to the dictionary, the words *pleća, leđa* have been marked with the neuter gender, whereas *prsa* with the feminine. The nominal declension of the latter word proves, on the other hand, clearly that these nouns all

have the neuter gender. The collective noun *momčad (eng. team of men)* has on one hand been characterized as having a neuter gender (Anić 2000), while the declension ones again proves that the noun is feminine (Silić-Pranjković 2005: 112). The category of collective nouns to which words *djeca (eng. children), braća (eng. brothers), gospoda (eng. gentlemen)* belong, disable the automatic use of the lexical database, as they carry markings of the neuter gender plural (important from the point of view of syntactical concordance), but have feminine singular declensional paradigm. All these cases then have to be manually processed.

The third issue in question concerns the homographic words. Since the Croatian version of HUMOR, as explained earlier, has been constructed for the parsing of the written standard version of the language, the differences in the accentuation have not been taken into account. If the differences do not consider the homonymy only, but the different declensions of the words in question as well (e.g. *pas (eng. dog)→GSg. psa* and *pâs (eng. waist)→GSg. pâsa*) both words are included into the lexical database and processed accordingly.

7. Conclusion

When implementing HUMOR to the morphological parsing of the Croatian language, different problems were encountered which had to be and will have to be solved. When compared to other inflectional languages, the problems that occurred do not consider the morphology of the language only, but the language policy as well. The issues described in this paper have been accounted for until the present stage of the project, namely the descriptions of verbal, adjectival and nominal inflections in the Croatian language. The solutions that have been proposed in HUMOR have not been useful from the aspect of morphological parsing only, but from the aspect of learning, i.e. teaching Croatian as the second or foreign language. In continuing the project, there will be some future descriptions conducted, including among others the description of adverbs, numerals, as well as other still unprocessed word classes. Apart from the description of the written language, the future prospects also include the possible implementation of the prosodic elements of the language, i.e. the implementation of HUMOR to the spoken language as well. Apart from several linguistic dilemmas when implementing HUMOR to the Croatian language, one should always bear in mind the benefits of such a morphological analyzer and its linguistic uses not only for parsing, but for the development of translational systems for Croatian and other minor languages as well.

Summary

The present paper discusses the linguistic problems concerning the automatic morphological analysis of the Croatian language, namely the problems concerning 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. The paper does not describe the theoretical working basis of the programme itself, but concentrates mainly on the linguistic issues that have arisen when implementing the parser to this language. Apart from the descriptions of linguistic problems, there are also some solutions present that also provide additional help when teaching or learning Croatian as a second or foreign language.

8. References

Anić, V. 2000. *Rječnik hrvatskoga jezika*. Zagreb: Novi Liber

Babić et al.1996. *Hrvatski pravopis.* Zagreb: Školska knjiga

Barić E. et. al. 1995. *Hrvatska Gramatika*. Zagreb: Školska Knjiga

Brodnjak, V. 1991. *Razlikovni rječnik srpskog i hrvatskog jezika*, Zagreb: Školske novine

Elekfi L. 1994. *Magyar ragozási szótár*. Budapest: MTA Nyelvtudományi Intézete

Hašek, J. 2004. *Doživljaji dobrog vojnika Švejka za svjetskog rata*. Zagreb: Biblioteka jutarnjeg lista

Moguš M. 1995.*Povijest hrvatskoga književnoga jezika*. Zagreb: Globus

Moguš M., Bratanić, M., Tadić, M. 1999. *Hrvatski čestotni rječnik.* Zagreb: Školska knjiga – Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu

Papp F. 1969. *A magyar nyelv szóvégmutató szótára.* Budapest: Akadémiai Kiadó

Prószeky, G., Kis, B. 1999. A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.* Maryland, USA: College Park. 261-268. Retrieved August 2, 2006 from http://www.morphologic.hu/h_pgpub.htm

Prószéky, G., Kis, B. 2002. Context-Sensitive Dictionaries. In: *COLING-2002*, Taipei, Taiwan. Retrieved August 2, 2006 from http://www.morphologic.hu/h_pgpub.htm

Raguž, D. 1997. *Praktična hrvatska gramatika.* Zagreb: Medicinska naklada

Silić, J., Pranjković, I. 2005. *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*. Zagreb. Školska knjiga

Tadić. M. 1994. *Računalna obrada morfologije hrvatskoga književnog jezika*. Ph.D. Thesis, Manuscript. Zagreb: Filozofski fakultet Sveučilišta u Zagrebu. Retrieved August 02, 2006 from http://www.hnk.ffzg.hr/mt/

Težak, S. 1991. *Hrvatski naš svagda(š)nji.* Zagreb: Školske novine

Težak, S.1995. *Hrvatski naš osebujni*. Zagreb: Školske novine

Težak, S. 1999. *Hrvatski naš (ne)zaboravljeni*. Zagreb: Tipex

Težak, S. 2004. *Hrvatski naš (ne)podobni*, Zagreb: Školske novine

Težak. S., Babić S., 2005. *Gramatika hrvatskoga jezika: priručnik za osnovno jezično obrazovanje*. Zagreb: Školska knjiga

Wardhaugh, R. 1995: *Szociolingviszika.* Budapest: Osiris- Századvég

www.google.com, April 12, 2005