Building the Croatian Morphological Lexicon

Marko TADIĆ
Department of Linguistics, Faculty of Philosophy, University of Zagreb
Ivana Lučića 3
Zagreb, Croatia, HR-10000
marko.tadic@ffzg.hr

Sanja FULGOSI
Institute of Linguistics, Faculty of Philosophy, University of Zagreb
Ivana Lučića 3
Zagreb, Croatia, HR-10000
sanja.fulgosi@ffzg.hr

Abstract

The paper presents the work being done so far on the building of the Croatian Morphological Lexicon (CML). It has been collected since 2002 in the Institute of Linguistics, Faculty of Philosophy, University of Zagreb. The CML is planned to have two sub-lexicons: derivative/compositional and inflectional, both produced by a generator. The result of generation is lexicon as two distinct lists of generated combinations of morphemes and complete word-forms both with additional data that can be used in further processing. The inflectional component is presented more in detail in the second part of the paper. At the end, the several possible applications of CML are discussed.

Introduction

Our aim was to make a model of Croatian morphological system in the form of lexicon stored in a database. The lexicon, named Croatian Morphological Lexicon (CML), would include all combinations of morphemes according to morphotactic rules and generated by two morphological generators.

1 Theoretical background

The very idea of generating all possible combinations of morphemes appeared in Halle (1973) for the first time in the framework of generative grammar. Although at that time it seemed that morphology could be represented as a compact subcomponent of the lexicon in the generative grammar, that concept posed several theoretical problems. The most severe one was defining the criteria for filtering-out non-established, yet possible combinations of morphemes. In fact this very approach negated the compactness of morphology since the derivative morphology was separated from inflectional.

Later research in generative and lexical phonology (cf. in Mihaljević 1991:85-86) separated derivative subcomponent from inflectional even more by introducing several layers where different types of rules appeared in different contexts (prefixation, suffixation, composition, inflection...). Not only rules but also the mode of their application resulted in different patterning schemes, particularly on two poles of that continuum of layers (rigidness of paradigms vs. non-rigidness of derivative system, symmetricity of paradigms vs. non-symmetricity of derivative system, cyclic rules in derivation vs. non-cyclic rules in inflection etc.).

Recently, there has been several contributions that were aimed at modeling the morphological systems of other Slavic languages with lexicons like Vetulani (2000) for Polish, Klímova and Kocek (2000) and particularly Osolsobě et al. (2002) for Czech as well as Rojc et al. (2002) for Slovenian.

Our idea of modeling the morphology of Croatian was somewhere along that track but we wanted to keep the model as simple as possible and use the computational data from the research already completed for Croatian.

2 Structure of the CML

Already described concept of dividing a model of the morphological system of Croatian was kept along but in its simplified form i.e. consisting of only two components:

1. **derivation/composition:** modeled as a list of lexical and a list of derivative morphemes

with rules for their combining.[1] A derivational generator could produce all possible combinations of lexical with derivative morphemes regardless of their existence in real texts. Each combination is not independent combination of morphemes (it equals the stem in traditional grammar), but it serves as the input for the second component. The cumulative result of generation of all possible combinations on that level could be called “derivative capacity” of a language (see Tadić 1994:40).

2. **inflection:** modeled as a list of generated stems and a list of inflectional morphemes with rules for their combining. The inflectional generator produces the final word-forms along the paradigms and with regard to inflectional patterns for Croatian as defined and described in detail in Tadić (1994).

Each generated morpheme combination (from both components) should be accompanied by additional data which could give us the possibility to backtrack the generation or use that data for further processing.

Each of the sub-lexicons of CML has different format covering different linguistic units and accompanying data.

2.1 Stem lexicon

For stem lexicon the format is still under consideration but it could look like this:

`pro|da|v|ač|ic-` `da-` $p_1ls_1s_2s_3$ `NAf`

where `prodavačic-` (roughly: *selleress-*) is generated combination of morphemes, `da-` is lexical morpheme or root, $p_1ls_1s_2s_3$ is a derivational pattern describing the morphotactic ordering of morphemes. The p_1, `l`, s_1, s_2 and s_3 could serve in database as pointers to the lists of morphemes. The `NAf` is a tag denoting *nomen agentis* of feminine gender. This system of derivational tags is still under construction and it should be submitted to further discussion and possible cross-Slavic-language examination and, hopefully, some standardization as well. Having in mind the closeness of Slavic languages, it seems that some kind of common recommendation may be achieved at least at the very basic level of unified list of derivational patterns and general semantic categories which are, in some languages more and in some less systematically, represented by certain derivational affixes. Introducing additional semantic information to that kind of lexicon could also be important because overlapping of derivational and semantic system is not always uniform and there is no exception-free 1:1 mapping. This kind of resource could be of a help to lexicographers, lexical semanticians, terminologists etc.

2.2 Inflectional lexicon

Unlike derivation, for inflection there is already established standard for inflectional lexicon and tag system in the form of MulTextEast v2 recommendation (Erjavec 2001) with definition for Croatian dated from 1998.

```
= abeceda Ncfsn
abecede abeceda Ncfsg
abecedi abeceda Ncfsd
abecedu abeceda Ncfsa
abecedo abeceda Ncfsv
abecedi abeceda Ncfsl
abecedom abeceda Ncfsi
abecede abeceda Ncfpn
abeceda abeceda Ncfpg
abecedama abeceda Ncfpd
abecede abeceda Ncfpa
abecede abeceda Ncfpv
abecedama abeceda Ncfpl
abecedama abeceda Ncfpi
= abolicija Ncfsn
abolicije abolicija Ncfsg
aboliciji abolicija Ncfsd
aboliciju abolicija Ncfsa
abolicijo abolicija Ncfsv
aboliciji abolicija Ncfsl
abolicijom abolicija Ncfsi
abolicije abolicija Ncfpn
abolicija abolicija Ncfpg
abolicijama abolicija Ncfpd
abolicije abolicija Ncfpa
abolicije abolicija Ncfpv
abolicijama abolicija Ncfpl
abolicijama abolicija Ncfpi
```

Sample from MTE conformant generated list of word-forms for nouns *abeceda* and *abolicija*.

[1] Composition in this case is treated as special variety of derivation because the combination of two (or in some cases more) lexical morphemes behaves like a single lexical morpheme in further derivative process.

3 Inflectional generation

While the derivational generator remains to be constructed, the inflectional generator for the whole inflectional system was constructed and described in Tadić (1994). It consists of three lists: lemmas (lexicon), endings, and transformations.

3.1 Lexicon

In the lexicon, lemmas (as headwords) are accompanied by minimal amount of data: only stems and inflectional pattern number(s). Additional data, when needed, are gender and information on *singularia/pluralia tantum*.

```
abdikacija abdikacij 343/0/0
abeceda abeced 343/0/0
abolicija abolicij 343/0/0
abrazija abrazij 343/0/0
adaptacija adaptacij 343/0/0
admiral admiral 1/0/0
adoracija adoracij 343/0/0
adresat adresat 1/0/0
adventizam adventizm 16/0/0
advokat advokat 1/0/0
advokatura advokatur 343/0/0
afera afer 343/0/0
afirmacija afirmacij 343/0/0
aforizam aforizm 16/0/0
Afričanin Afričan 3/0/0
```

Sample from the lexicon with lemma, stem and inflectional pattern where `x/y/z` denotes the number of pattern: `x` represents declension, `y` conjugation and `z` comparison. In total there are 613 inflectional patterns where 404 patterns are for nouns, 42 for adjectives, 12 for comparison and 155 for verbs. Up to three different inflectional patterns (or paradigms) of the same type are allowed per verbal lemma because there are verbs which can change its word-forms according to different patterns. Up to four different patterns are allowed for lemmas that are nouns and adjectives because there are instances of parallel word-forms with the same MSD. For adjectives with comparison, `x` and `z` are both obligatory.
In order to keep the model flat and easily accessible for further linguistic refinement, no optimization has been done from the computational point of view.

3.2 List of endings

Endings are represented like ordered n-tuples including morphosyntactic categories with their exact values realizing at certain position in inflectional pattern. For substantives it looks like this:

```
ending = <pat, par, num, cas>
```

where pat is pattern (1-404), par is paradigm (1-4), num is number (1-2) and cas is case (1-7).
For adjectives it is a bit more complicated:

```
ending = <pat, par, gen, num, cas>
```

with pattern (1-42) and addition of gender (1-3).
For verbs the endings are defined like this:

```
ending = <pat, ten, num, per/gen>
```

with pattern (1-155), tense (1-9) which includes participle forms, and person/gender (1-3) with person appearing in finite forms (tenses 2-5) and gender in participles (6-7).
The comparison endings are defined with:

```
ending = <pat, gra>
```

with pattern (1-12) and grade (1-3).

Endings are distributed by types in different tables which are being accessed by usage of coordinates described in n-tuples.

3.3 List of transformations

The transformations are allowed only on stems and are closely connected with inflectional patterns. Precise classification of patterns according to the phonological composition of the stems enables the controlled context of application of quite simple transformational rules and keeps their number as low as 35 for the whole inflectional system. The order of applying transformation rules is also strictly defined.
The lists of transformations are stored in tables, which are isomorphous to the tables with endings and are being accessed the same way. Since the transformations of the stems are pattern and ending dependent, they can be done before final concatenation with endings.
This system of transformations, coupled with precise classification of inflectional patterns,

covered all allomorphy of Croatian inflectional stems including the most complicated cases (e.g. verbs of 1st class).

4 Implementation

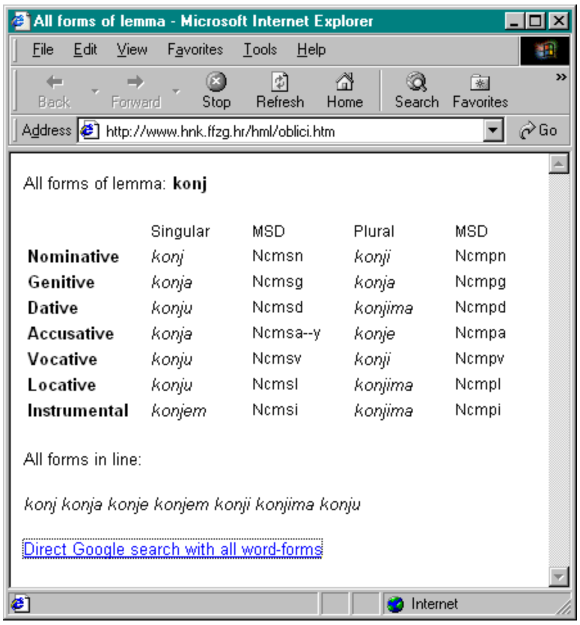
For generation of inflectional lexicon the headword list of Anić (1991) *Croatian dictionary* has been used. Each headword was associated with inflectional pattern thus forming a lexicon with ca 36.000 lemmas (18.019 substantives, 7735 verbs, 5504 adjectives, 64 pronouns, 6517 adverbs, etc.) The periphrastic verbal tenses were not being generated in whole but only their participle parts. Also no reflexive pronouns were included in generation of reflexive verbs so on that level there is no distinction between non-reflexive and reflexive verbs.[2] The main reason for exclusion of periphrastic tenses and reflexive verb word-forms was the in the idea to use inflectional lexicon in initial POS tagging of Croatian texts. Coping with this verbal word-forms would complicate the task at this point.
The generation of word-forms yielded 171380 word-forms for nouns, 232276 word-forms for verbs, 1207786 word-forms for adjectives, and 11706 word-forms for adverbs.

4.1 Usage

The fully generated inflectional lexicon is being submitted to thorough inspection in order to correct possible errors. It will be used in POS/MSD tagging of Croatian National Corpus (HNK).[3] The subset of HNK of 1 million tokens is composed and it will be matched with inflectional lexicon. Since there is no data of that kind for Croatian yet, in this way all possible MSD readings at the unigram level of each token will be made available. The “inflectional weight” and “homographic weight” could be calculated for each token giving additional data for further processing. After (semi-)manual disambiguation and correction, that corpus will be used as a training corpus for a tagger.[4]

4.2 Availability

The inflectional lexicon of CML will be available for querying at the Croatian HLT web-portal[5] giving all word-forms of requested lemma. Beside linguists and other researchers, it will be useful to students of Croatian who want to find or check the proper inflection of words. It will be useful even to a wider public as a tool for web search because the output will allow simple link to all word-forms in web-search query window thus enabling the inflectionally sensitive search of Croatian web-pages.



http://www.hnk.ffzg.hr/hml/oblici.htm

All forms of lemma: **konj**

	Singular	MSD	Plural	MSD
Nominative	*konj*	Ncmsn	*konji*	Ncmpn
Genitive	*konja*	Ncmsg	*konja*	Ncmpg
Dative	*konju*	Ncmsd	*konjima*	Ncmpd
Accusative	*konja*	Ncmsa--y	*konje*	Ncmpa
Vocative	*konju*	Ncmsv	*konji*	Ncmpv
Locative	*konju*	Ncmsl	*konjima*	Ncmpl
Instrumental	*konjem*	Ncmsi	*konjima*	Ncmpi

All forms in line:

konj konja konje konjem konji konjima konju

Direct Google search with all word-forms

Sample window with results of querying CML

Conclusion

The paper presented the work done so far on building the Croatian Morphological Lexicon. While derivative sublexicon was covered only

[2] The idea to exclude the periphrastic part of the verbal paradigm from inflection completely and to treate it on different level i.e. syntactically may not sound so utterly heretic at this point. Since Slavic languages are more-or-less free order languages, the position of auxiliary is sometimes far away from the participle and this can only lead to inflectional analysis which has to take in account the sentence structure. This idea has been tacitly adopted in producing the most of inflectional generators.

[3] See in Tadić (2002:445).

[4] See the results for Slovene in Erjavec et al. (2000). The adaptation of tagset in several iterations or levels of complexity will probably be necessary.

[5] At http://www.hnk.ffzg.hr/jthj.

with theoretical discussion, which needs further cross-Slavic languages refinement and possible standardization, the inflectional sublexicon was presented in detail. This sublexicon is MULTEXT-East conformant and uses MSD tags defined in the scope of that project. Several possible uses of inflectional sublexicon were also suggested, among which the most widely useful can be the possibility to use all word-forms of a lemma in the web-search engine such as Google, AltaVista etc.

Acknowledgements

The work on CML presented here has been done in the frame of the i-project 2002-100 granted by the Ministry of Science and Technology of the Republic of Croatia.

References

Anić, V. (1991) *Rječnik hrvatskoga jezika*, Novi liber, Zagreb, 887 p.

Erjavec, T., Džeroski, S., Zavrel, J. 2000. *Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets*. In LREC2000 Proceedings, M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhaouer, ed., ELRA, Paris, pp. 1099-1104.

Erjavec, T. (2001) *MULTEXT-East Resources, Concede Edition*. (http://nl.ijs.si/MTE/V2 and http://nl.ijs.si/ME/ V2/msd/html/).

Halle M. (1973) *Prolegomena to a Theory of Word Formation.* Linguistic Inquiry, 4/1, pp. 3-16.

Klímová, J. and Kocek, J. (2000) *Derivation in Czech National Corpus*. In LREC2000 Proceedings, M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhaouer, ed., ELRA, Paris, pp. 1463-1466.

Mihaljević, M. (1991) *Generativna i leksička fonologija.* Školska knjiga, Zagreb, 135 p.

Osolsobě, K., Pala, K., Sedláček, R. and Veber, M. (2002) *A Procedure for Word Derivational Processes Concerning Lexicon Extension in Highly Inflected Languages*. In LREC2002 Proceedings, M. González Rodrígez and C. P. Suarez Araujo, ed., ELRA, Paris, pp. 998-1003.

Rojc, M., Kačić, Z. and Verdonik, D. (2002) *Design and Implementation of the Slovenian Phonetic and Morphology Lexicons for the Use in Spoken Language Applications*. In LREC2002 Proceedings, M. González Rodrígez and C. P. Suarez Araujo, ed., ELRA, Paris, pp. 1296-1300.

Tadić, M. (1994) *Računalna obradba morfologije hrvatskoga književnoga jezika*. Ph.D. thesis, University of Zagreb. (http://www.hnk.ffzg.hr/txts/mt_dr.pdf).

Tadić, M. (2002) *Building the Croatian National Corpus*. In LREC2002 Proceedings, M. González Rodrígez and C. P. Suarez Araujo, ed., ELRA, Paris, pp. 441-446.

Vetulani, Z. (2000) *Electronic Language Resources for POLISH: POLEX, CEGLEX and GRAMLEX.* In LREC2000 Proceedings, M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhaouer, ed., ELRA, Paris, pp. 367-374.