CROATIAN LEMMATIZATION SERVER

Marko Tadić

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

ABSTRACT

The need for lemmatization in inflectionally rich languages is indisputable: it is applicable for the whole range of procedures — from text-search, up to parsing. From two predominant approaches to lemmatization: 1) algorithmic (generally rule-based and realized with FSA) and 2) relational (generally data-driven and realized with databases), this paper opted for the latter. The reason is that formal-grammar approaches to Croatian morphology are rare and limited just to a part of morphological system. The other reason is that the generator for Croatian has already been developed (Tadić 1994) as well as Croatian Morphological Lexicon (CML) (Tadić & Fulgosi 2003). The idea was to offer an on-line lemmatization, POS/MSD service using CML v 4.5 as the back-end. The Croatian Lemmatization Server (CLS) is available at http://hml.hnk.ffzg.hr and it offers lemmatization and POS/MSD tagging at unigram level for now. For each token in submitted text, the server delivers all possible lemmas of which this token may be a word-form. For homographic tokens, each lemma is accompanied with all possible POS/MSD tags which are compliant to MulTextEast v3 specifications for Croatian . The CLS can also be used for generation: when lemma is inputted and marked, all its possible word-forms are being retrieved and delivered.

1. Introduction

The need for lemmatization in the majority languages, particularly inflectionally rich ones, is indisputable. It is applicable in the whole range of procedures — from simple text-search, up to complex parsing. The lemmatization enables the researchers and/or users of HLT resources, tools or products the precise and yet extensive access to data at the word-level in a way that appears natural to humans. The traditional educational system has taught us from the beginning of our school days that a lemma is the most important word-form (WF) which is a representative of the whole set of possible WFs (a paradigm) of a lexeme. So every speaker of an inflectionally rich language, intuitively think of words (or lexemes) in their lemma (or basic, or normal) form. Lexeme, treated as a listeme (i.e. item on the list of lexical entries in a lexicon of a language), in lexicons is always presented as lemma or headword.

How is the strength of lemma powerful in human mental lexicon can be well described by the following scenario. When a speaker of inflectionally rich language sits in front of an English-based web-search engine, s/he inputs the lexeme as lemma (e.g. noun in nominative singular). The frequency data, which we know only for Croatian (partially from Moguš et al. 1999), and we have no reason to believe that it would be seriously different for other Slavic languages excluding Bulgarian, have shown that genitive and accusative WFs of a noun are more frequent than nominative. The immediate result is that our lemma-based query yields unsatisfactory document retrieval results: we are missing all the documents where lexeme didn't appear in lemma WF (i.e. nominative singular), but in other more frequent WFs only. Lemmatized query could help us overcome this limitation since it is the natural way users think about lexemes and what they convey: their meaning. The similar problem of a lemma strength appears in other scenarios wherever lemma/WF relation can be established.

The lemmatization in general can be seen as the relation between this selected, representative WF and the rest of possible WFs of a lexeme. The result of lemmatization procedure in computational linguistics and/or NLP could be defined as establishing the explicit relation between lemma and other WFs of lexeme in a corpus, text, structure etc. There are two main predominant approaches to lemmatization: 1) mostly algorithmic (generally rule-based and most efficiently realized with FSA and their derivatives) and 2) mostly relational (generally data-driven and realized usually with databases).

The first approach has it's most successful representative in Two-level morphology (Koskenniemi, 1983), a formalism for describing morphology of natural languages that has been applied to a wide range of typologically different languages (including ones with non-concatenative morphology). Its advantages are very efficient processors (transducers) of moderate formal complexity, fast processing, possibilities of parallel processing etc. The systems build featuring this approach still need the input list of lexical morphemes (or stems) with possible continuation classes (i.e. endings) and rules of their combining. This part of the whole system is also relational and at that level resembles to the second approach.

The second approach was more bottom-up oriented i.e. models of morphology, predominantly inflection only, were build by using morphological (inflectional) lexicons. Instead of building a complex rule-based system for analysis and synthesis of

different WF of a lexeme, the simple list of all possible WFs of tens of thousands of lexemes (or lemmas) can be generated (manually or automatically). In that case the morphological lexicon is the resource/tool which is used for modeling of the morphology of a certain language.

In this paper, where we describe the web-based application of a model of Croatian inflectional system, we opted for the second approach. The second section of the paper describes previous attempts to model Croatian inflectional system, the third section describes the Croatian Morphological Lexicon (CML), while its usage in a web-based lemmatization service is described in the fourth section. The final conclusions also suggest some possible directions for further improvement.

2. Previous attempts to computationally model Croatian inflectional system

Relational databases which represent the model of Croatian inflection such as described in (Kržak & Boras 1985) and (Boras 1990) could have been also used as possible solution. While this relational database model is efficient for generating the WF's beginnings and endings, it does not always respect the proper linguistic morphological boundary. In fact, it does not have to do that completely as long as it provides the correct final results of all WFs of a lemma. But the main weakness of this systems is that they were not publicly available.

Having in mind the problems mentioned, an algorithmic approach to Croatian inflection such as described in (Lopina, 1992) and (Lopina, 1993) could have been chosen as natural solution. It is highly efficient, covers even unknown words since it deals with endings and all possible morphonological alternations using rules. But it was developed only for a smaller part of Croatian inflectional and derivational morphology and certainly could not cover the whole inflectional system.

The next reason for taking the relational based approach in our work was that the comprehensive formal grammatical descriptions of Croatian morphology did not exist at all. There was only one monograph (Babić, 1991) which theoretically covered just a part of verbal inflection. The development of an algorithmically based system begs for such an overall grammatical description which we didn't have. The subsequent reason was that the generator for Croatian WFs has already been developed (Tadić 1994) and with such a system a wide-coverage lexicon of WFs could be generated automatically.

The first version of CML was described in (Tadić 2002) and (Tadić & Fulgosi 2003) where also the prototype version of Croatian Lemmatization Server was described. Here we will present the present of CML briefly.

3. Croatian Morphological Lexicon (CML)

The CML is built upon Croatian inflectional generator *GenOblik* developed and presented in (Tadić 1994) and partially (Tadić 2002). It is a classification based generator which models the whole Croatian inflection with 614 different inflectional patterns i.e. types of inflectional behavior of different POS (nouns are covered by 404, verbs with 155, adjectives with 43 and comparison with 12 patterns). Such a system is able to generate all WFs for single lemma or it can be run on a whole list of lemmas providing that each lemma is accompanied by the number of its characteristic inflectional pattern.

```
abeceda 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
```

Figure 1: Sample from CML v 4.6 of lemma *abeceda* with all its WFs and MSDs

Currently CML (v 4.6) encompasses 45,000+ general language lemmas, 15,000+ male and female personal names and 57,000+ surnames registered in the Republic of Croatia. From this list of lemmas, generator produced 3,800,000+ WFs (Figure 1) with appropriate morphosyntactic description tags (MSDs) in accordance with MULText East recommendations for Croatian in (Erjavec et al. 2003) and (Erjavec 2004). The CML is accessible via CLS which is presented in the next section.

4. Lemmatization as a web-based service

Having CML accessible as on-line lexicon for almost 3 years, gave us a thought to offer an on-line lemmatization service for Croatian texts using CML v 4.6 as the back-end. The prototype version of CLS was developed in order to facilitate the processing of Croatian inflectional morphology for different users in both directions: recognition and generation of WFs together with lemmatization, POS and MSD tagging. The CML as the list of lemmas accompanied by their WFs and respective MSDs is stored in a database which is publicly accessible via CLS at address http://hml.ffzg.hr for guests (userid: *proba*, password: *proba*) or registered users (academic or commercial). The CLS offers the web-service for: 1) generation of all WFs of a single or multiple lemmas, and 2) recognition and lemmatization of single or multiple WF.

4.1. Generation

After typing desired lemma(s) in a HTML-form or uploading a list of lemmas in verticalized UTF-8 encoded text-file, the system will return all possible WFs of lemmas accompanied by full MSD information on the web-page (Figure 2).

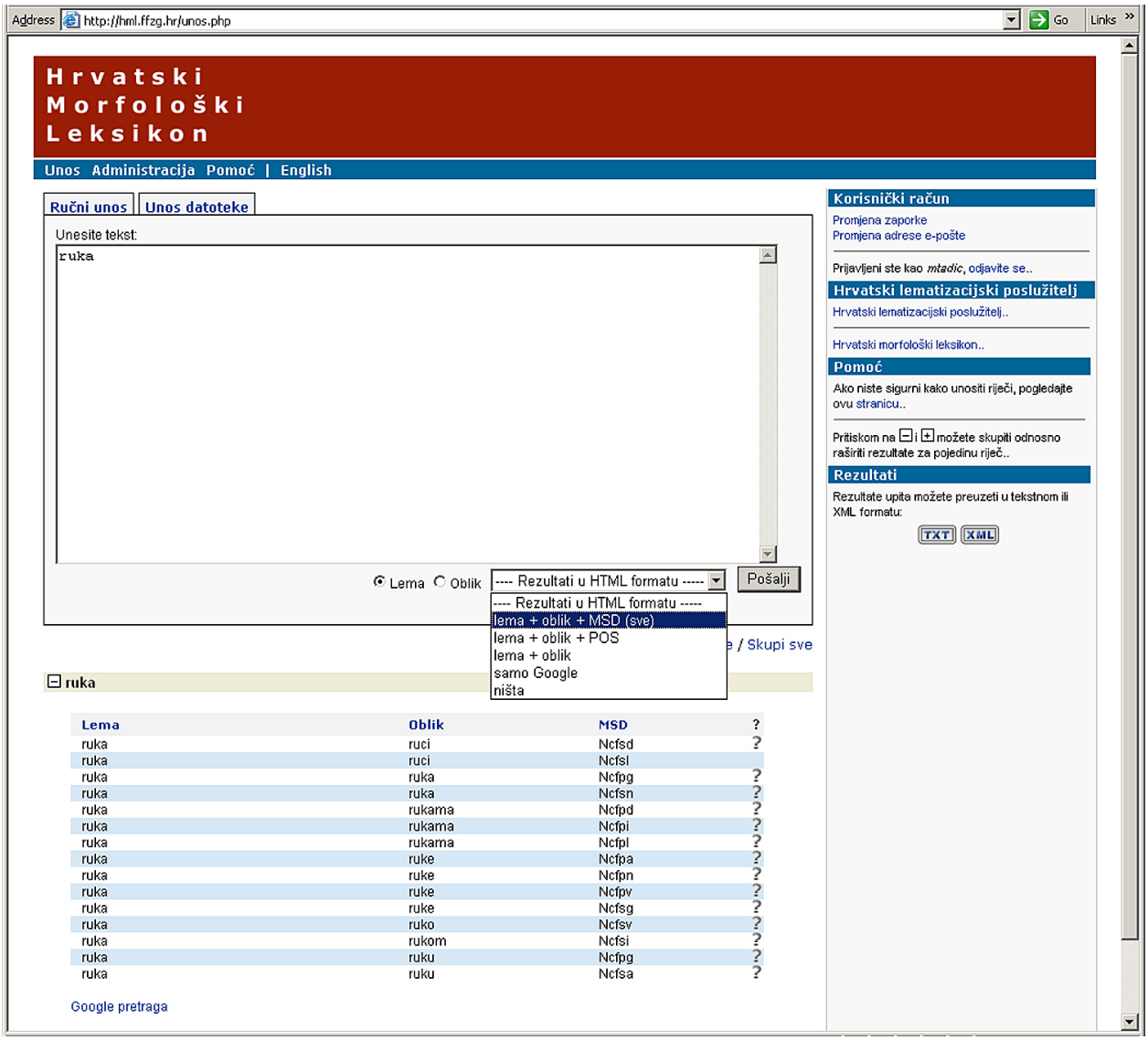

Figure 2: Sample of a single lemma (*ruka*) submitted and the result of generation with all WFs and respective MSDs

4.2. Recognition

After typing desired WF(s) in a HTML-form or uploading a list of WFs in verticalized UTF-8 encoded text-file, the system will return lemma and POS/MSD for them. Each WF gets recognized in all its possible inflectional interpretations at unigram level i.e. without taking into account the preceding or the following words in the submitted text. Effectively, this means that CLS can be used for POS and/or MSD tagging, as well as lemmatization of Croatian texts but without morphosyntactic disambiguation. This represents the first level of processing of Croatian inflection and it can be a useful starting point for all other types of processing at higher linguistic levels (syntactic, semantic, information extraction, etc.) (Figure 3).

When the results are downloaded, they are available in a CSV file in a ZIP file created at the server side for faster download.

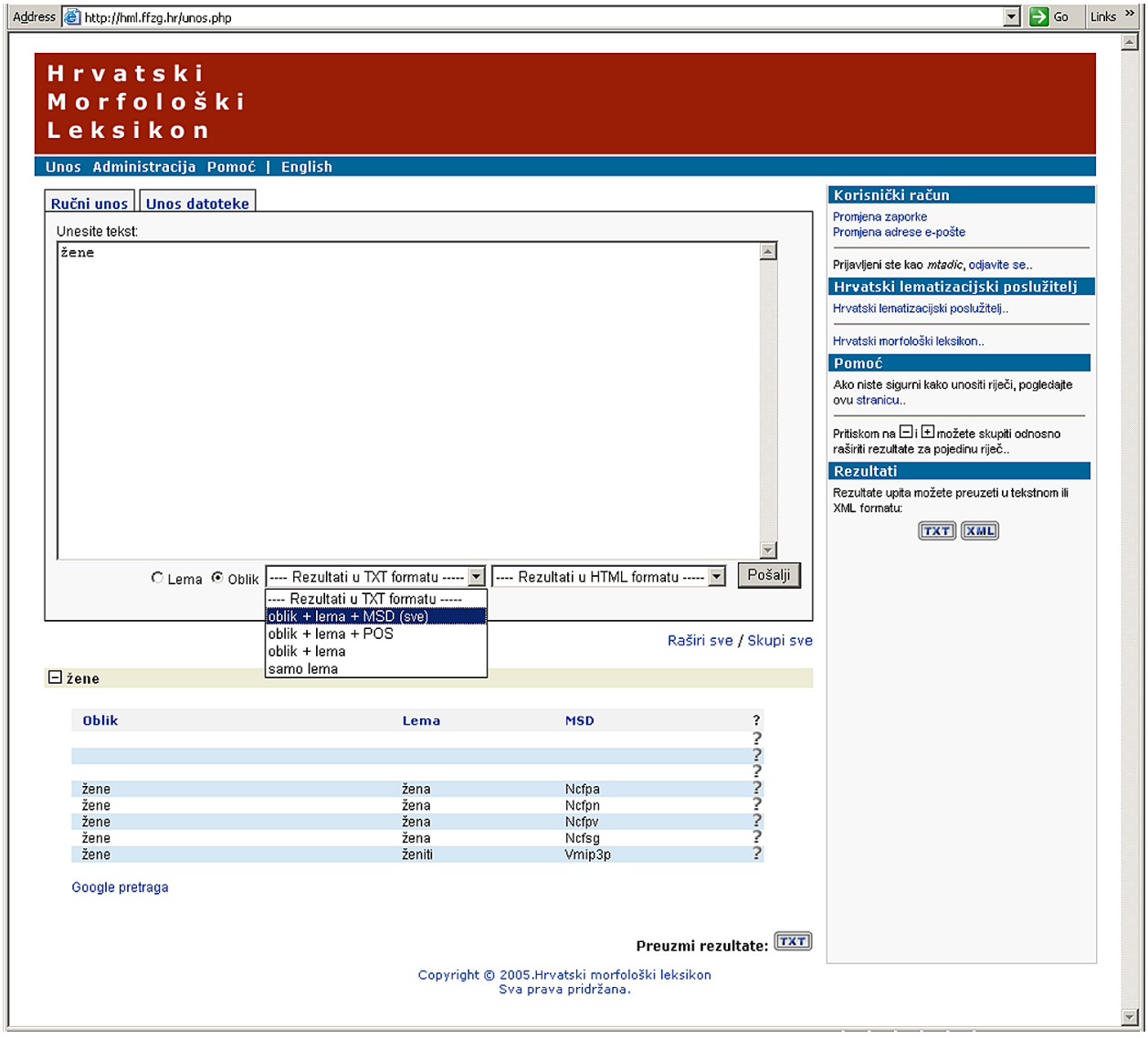

Figure 3: Sample of a single WF (*žene*) submitted with all possible interpretations included

4.3. Features of CLS

The current version of CLS features several output formats.

The results of processing that are displayed at the web-page can be in the format: 1) lemma + WF + full MSD; 2) lemma + WF + only POS; 3) lemma + WF; 4) only Google search; 5) no web-page output.

The results of processing of each submitted WF in a text file can be in the format: 1) WF + lemma + full MSD; 2) WF + lemma + only POS; 3) WF + lemma; 4) only lemma replacing the original WF.

One of the features of the CLS web-service is an automatically generated link to a page with result of the Google query with all and exactly all WFs of the desired lemma (Figure 2: *Google pretraga* link at the lower left corner of the window and Figure 4: results). For each lemma, a list of unique WFs is generated and transformed to an already performed Google query. In fact it looks like a simple series of different WFs connected with Boolean OR. The same task could certainly be done manually, but it would be too time consuming and human-error sensitive. This service is also available to registered users as an URL call to a server-side php script. In this way it can be pipe-lined with other processes. Using CLS as web search engine module is thoroughly described in (Tadić & Bekavac, 2006).

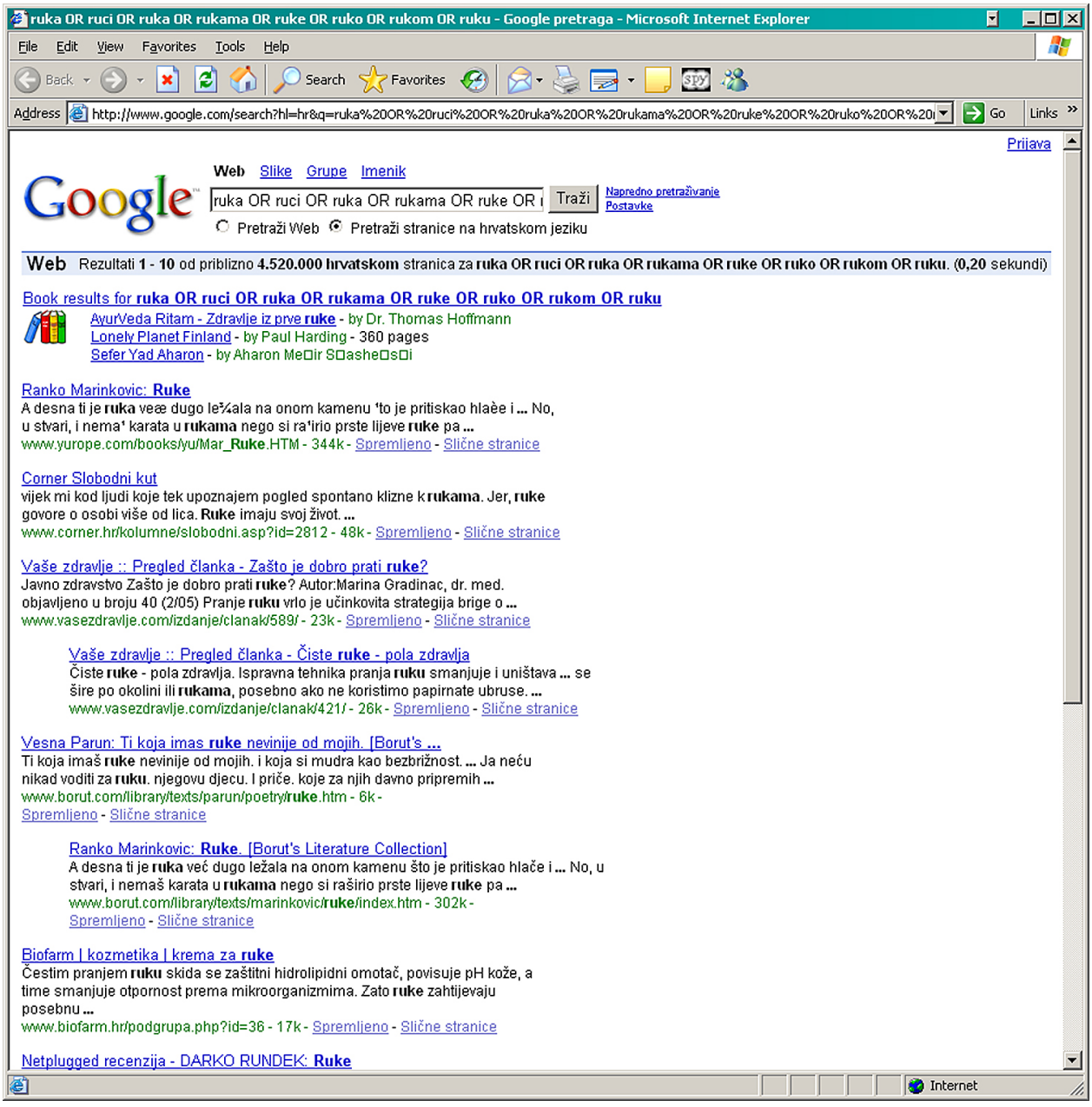

Figure 4: Sample of a generated Google query featuring all WFs of a lemma

The coverage of the CML v 4.6 was tested on a 46 million tokens large Croatian newspaper corpus and it yielded the result

of 96,4% coverage. Besides, all traffic on CLS is being logged so that we have the full statistics of access. Also words unknown to CML are tracked automatically and stored in a separate log file for further processing and updating the CML which is being done on the regular (bi-)monthly basis.

The CLS is used by academic users for lemmatization of Croatian texts during the process of corpus building (Croatian National Corpus in particular), text-mining research, language learning (Croatian for foreigners particularly). There are also several commercial users who are using CLS as lemmatized search engine for their portals, as a tool for downsizing the indexes in huge textual databases etc.

5. Conclusion and future perspectives

We have shown the web-based service for processing Croatian inflection, namely lemmatization, POS/MSD tagging at unigram level and without further morphosyntactic disambiguation. While it has been developed primarily having in mind the usage by computational linguists and computer scientists for research purposes, it turned out to be commercially attractive service also.

There are several directions for possible enhancements of CLS. The first step should be at the lemmatization and POS/MSD tagging task: the system should be able to deliver not just all possible lemma/POS/MSD interpretations (with high homography) but it should also offer certain level of disambiguation i.e. to function as an on line tagger for Croatian with the precision above 90% for texts of a general type. The preliminary tests with some statistical taggers (TnT) have shown promising results as it has been presented in (Agić & Tadić, 2006).

Another possible direction of development could be dealing with multiple word queries. This would require not only defining the Boolean relations between words in a query, but also possible overlapping of their grammatical features (i.e. values in MSD tags). This would allow the system to generate queries with all WFs not of individual words but of syntagms (Figure 5) obeying rules of congruency (e. g. NPs of Adj + N having the same values for morphosyntactic categories like gender, number and case).

```
najplavija zastava
najplavija zastavo
najplavije zastave
najplavijih zastava
najplavijim zastavama
najplavijima zastavama
najplavijoj zastavi
najplavijom zastavom
najplaviju zastavu
plava zastava
plava zastavo
plave zastave
plavih zastava
plavija zastava
plavija zastavo
plavije zastave
plavijih zastava
plavijim zastavama
plavijima zastavama
plavijoj zastavi
plavijom zastavom
plaviju zastavu
plavim zastavama
plavima zastavama
plavoj zastavi
plavom zastavom
plavu zastavu
```

Figure 5: Suggestied form for results of possible two-lemma query with respective WFs

Regarding the updating of CML via CLS, an algorithm for finding lemmas for Croatian unknown words could be developed similar to one applied to Slovenian (Erjavec & Džeroski, 2004).

References

Agić, Ž. & M. Tadić. 2006. Evaluating Morphosyntactic Tagging of Croatian Texts. In *Proceedings of the LREC2006*, Genoa-Paris: ELRA-ELDA. (URL: http://hnk.ffzg.hr/txts/zamt4LREC2006.pdf)

Babić, Z. 1991. *Generativni opis konjugacijskih oblika*, Zagreb: Znanstvena biblioteka HFD-a.

Boras, D. 1990. Rječnička baza kao osnova za izradu automatskog detektora pogrešaka teksta na hrvatskom jeziku pisanog pomoću kompjutora. In S. Tkalac; M. Tuđman, ed., *Informacijske znanosti i znanje*, Zagreb: Zavod za informacijske studije.

Erjavec, T.; C. Krstev; V. Petkevič; K. Simov; M. Tadić; D. Vitas. 2003. The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. In *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*, Budapest: ACL. (URL: http://hnk.ffzg.hr/txts/mte-mpsl03_(erjavec_et_al).pdf)

Erjavec, T. & S. Džeroski. 2004. Machine Learning of Morphosyntactic Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence 18*(1), 17-40. (URL: http://nl.ijs.si/et/Bib/AAI-18-Erjavec-Dzeroski.pdf)

Erjavec, T. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the LREC2004*, Lisbon-Paris: ELRA-ELDA. (URL: http://nl.ijs.si/Bib/LREC04/lrec04-mte109.pdf)

Koskenniemi, K. 1983. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Helsinki: Publications of the Department of General Linguistics, University of Helsinki.

Kržak, M. & D. Boras. 1985. Rječnička baza hrvatskog književnog jezika. *Informatologia Yugoslavica, 17* (3-4), 223-242.

Lopina, V. 1992. Dvorazinski opis morfonoloških smjena u pisanome hrvatskom jeziku. *Suvremena lingvistika 34*, 185-194.

Lopina, V. 1993. Dvorazinski model morfološkoga opisa. In S. Tkalac; M. Tuđman, ed. *Obrada jezika i prikaz znanja*. Zagreb: Zavod za informacijske studije.

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

Tadić, M. 1994. *Računalna obrada morfologije hrvatskoga književnoga jezika*, PhD dissertation, Zagreb: University of Zagreb. (URL: http://hnk.ffzg.hr/txts/mt dr le.pdf)

Tadić, M. 2002. Building the Croatian National Corpus. In *Proceedings of the LREC2002*, Las Palmas-Paris: ELRA-ELDA. (URL: http://hnk.ffzg.hr/txts/mt4LREC2002.pdf)

Tadić, M. & S. Fulgosi. 2003. Building the Croatian Morphological Lexicon. In *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*, Budapest: ACL. (URL: http://hnk.ffzg.hr/txts/mtsf4EACL2003.pdf)

Tadić, M. & B. Bekavac. 2006. Inflectionally Sensitive Web Search in Croatian using Croatian Lemmatization Server. In V. Lužar-Stiffler; V. Hljuz Dobrić, ed. *Proceedings of ITI2006 Conference*, Zagreb: SRCE.