Inflectionally Sensitive Web Search in Croatian using Croatian Lemmatization Server

Marko Tadić, Božo Bekavac

Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{marko.tadic, bbekavac}@ffzg.hr

Abstract. *Web search engines are becoming everyday commodity but still they are mainly oriented towards English. Other languages, with their different word structures, are not represented well with available web search engines.*

We present one solution for a morphologically sensitive search engine which enables the Croatian-speaking users to use query words more naturally and get the results that cover all and only all word-forms of searched words.

Keywords. Web search engine, computational linguistics, NLP, Croatian, morphological processing, inflection.

1. Introduction — defining the problem

Web search engines have become an everyday commodity. They appear not only at the global web level with services such as Google, Yahoo etc., but also almost every portal, or slightly more complex web site, features the local search engine covering its pages. Besides, the approach to so called "hidden web" i.e. vast quantities of textual data stored in databases, is mainly accessible by usage of queries.

What is common to all those systems, is that they are well suited for English language users. What about data in other languages? They could be accessible via multilingual search engines where user starts the search in one language (e.g. English) and gets results in other. But what about searching in a targeted language when it is not English? Should the search engines consider the specifics of their language structure, particularly inflectionally rich languages? While English has quite a limited number of possible word-forms (WF) for each lexeme, languages with significantly more WF for each lexeme are being more frequently used (German, Dutch, Finnish, Hungarian, all Slavic languages etc.) in the web context. Shouldn't the search engine which pretends at the precise hits feature also this particular property of these languages? Number of possible different WFs for a lexeme and the frequency distribution of different WFs in these languages are highly language dependant: the former depends on morphology (i.e. inflectional system), while the latter depends on the morphosyntactic features of a language (analytic vs. synthetic languages and their particularities in encoding syntactic relations in a sentence predominantly with WFs or word-order).

Speakers of inflectionally rich languages, under the influence of their mother tongue education, where they have been taught for years that the basic WF (or lemma) represents all possible WFs of a lexeme, intuitively fire a query on a web search engine choosing only lemma (i.e. nominative singular for nouns). Although they intuitively expect that the usage of lemma should also yield the results covering all other WFs, they usually get only documents where this lexeme appeared only as lemma, thus missing all documents where it appeared in other WFs only.

The frequency distribution of other WFs could also be of importance for inflectionally sensitive search engines since, at least in the case of Croatian nouns (and the same could be expected for other Slavic languages), the accusative and genitive cases (i.e. WFs) are more frequent than nominative singular (lemma). This means that, in the most common search scenario, when a speaker of Croatian language fires a query using lemma (nominative singular) of a noun, (s)he misses all documents where this lexeme does not appear in nominative singular, including accusative and genitive being the more frequent cases. How could this search problem be tackled in order to: 1) offer more user friendly and user language sensitive (user adaptable) search engines; 2) get better recall without the decrease of precision in document retrieval? We will concentrate on a specific solution for one language, namely

Croatian, but we believe that similar solutions could be applicable for other structurally and typologically close languages.

The organization of the rest of the paper is as follows: in the second section we will show related works i.e. possible solutions to a problem. In the third section the detailed description of our suggested solution will be discussed. In the fourth section the results of testing will be presented while the paper will be concluded with further directions.

2. Possible solutions for an inflectionally sensitive search engine in Croatian

In this section we give an overview of existing systems which could be or are already used for inflectionally sensitive search in Croatian.

The first and the crudest solution could be the usage of joker characters (usually Kleenex star or asterisk "*") in queries providing that the search engine supports this kind of advanced search. Unfortunately, for Croatian this solution is inadequate since the query like *glav**, while user expects that (s)he will get all WFs of a noun *glava*, yields not only all WFs of searched lexeme, but also all derivatives, adjectives, compounds etc. in all their WFs such as *glavni, glavarina, glavnina, glavočika, glavnica* etc.

The second possible solution could be the usage of a stemmer such as one described in [1]. This system provides the normalized form of a lexeme, but it could go beyond the inflectional morphology and enter into derivational morphology covering also derivatives (such as adjectives) or compounds. In this way it will offer more generalization than needed for inflectionally sensitive search since users of Croatian usually consider derivatives and compounds as different lexemes. This kind of generalization of useful for other types of applications such as document indexing, classification, summarization etc.

Relational databases which represent the model of Croatian inflection such as described in [2] and [3] could be also used as possible solution. While this relational database model is efficient for generating the WF's beginnings and endings, it doesn't always respect the proper linguistic morphological boundary. In fact, it doesn't have to do that completely as long as it provides the search engine with all WFs. But the main weakness of this systems is that they are not publicly available.

Having in mind the problems mentioned, an algorithmic approach to Croatian inflection such as described in [4] and [5] could be 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 was developed only for a smaller part of Croatian morphology and not for the whole inflectional system.

There is also a commercial application, namely the search engine Pogodak [6]. It features the inflectionally sensitive search which covers Croatian nouns only. While the majority of queries would be expected to be formulated as nouns, the usage of verbs as well as adjectives should not be neglected (particularly in the cases such as possessive adjectives where the most frequent way of expressing possessiveness is by deriving the article, such as *Ivanov* instead of possessive genitive *Ivana*). The Pogodak search engine is a commercial application and modification of well established Slovenian search engine Najdi [7] for Croatian. While it gets decent results for nouns only, its weakest part is a smaller-scale search engine than other globally positioned web search engines (e.g. Google).

3. Our solution

We would like to suggest a generic approach which could be used not just for a inflectionally sensitive search engine by user's choice, but also allows the processing of Croatian inflection in a more complex way covering both directions: recognition and generation of WFs of lexemes.

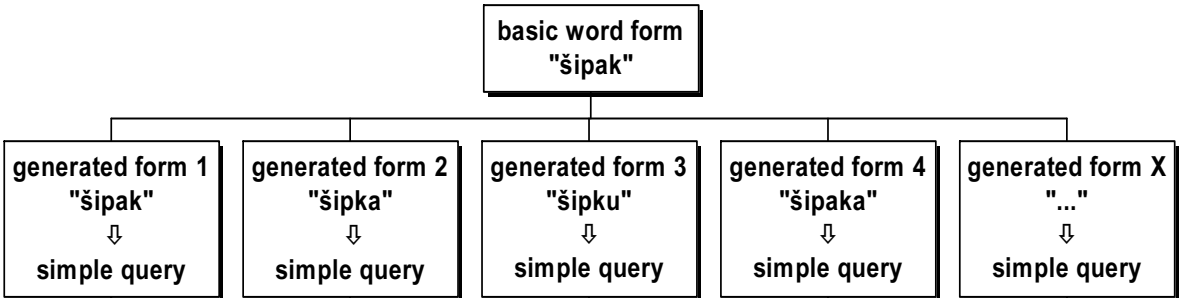

Figure 1. Schematic representation of all WF query generation (adapted from [8])

The starting point for every inflectionally sensitive query should be providing the search engine with a layer of morphological processing. This could be done either as 1) an internal subsystem at the search engine server side, or 2) as a web-based service which is positioned and called in between user query and search engine. Both solutions have advantages and drawbacks.

The first solution could enable industrial-strenght fast-response search engine, but at the same time it has to deal with the problem of words unknown to the system and its updating with new lemmas and their WFs (this is expected to be done by an expert on the regular basis — weekly or monthly — and it can't be done automatically at this point of development).

In the second case the morphological layer is fed with the exact wording of a query, then it generates all WFs of given search-word(s) and returns them to a target search engine thus providing inflectionally sensitive search. This search yields all and only all WFs of a given search-word without any over- and/or undergeneration. An advantage of this second approach is that the centralized web-based service is always up-to-date and there is no need for upgrading as there is with inflectional module as internal system at the server side. This solution could show effects of delay since it introduces the call to a service from different URL, but we believe that in the situation where more and more Internet-based services are introduced, this possible delay could be solved by technical facilities to meet the needs of commercial users. Typical individual user who fires a query with several words is not even aware of different Internet-based service since its call is hidden.

Another possible application of this system, which is in fact being used by one of our industrial users, is the the use of our system for generating lemmatized inverted indexes of the documents in their own textual databases (see in detail in 3.2).

The prerequisite for such a system for Croatian is the Croatian Morphological Lexicon (CML): the v 1.0 was described in detail in [9] and [14] where the first inflectionally sensitive web search for Croatian has been demonstrated. The v 2.0 will be presented here briefly.

3.1. Croatian Morphological Lexicon

The CML is built upon Croatian inflectional generator *GenOblik* developed and presented in [10] and partially [11]. 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 2. Sample from CML v 2.0 of lemma "abeceda" with its all WFs and MSDs

Currently CML (v 2.0) encompasses >45,000 general language lemmas, >15,000 male and female personal names and >57,000 surnames registered in Croatia. From this list of lemmas, generator produced >3,700,000 WFs (Figure 2) with appropriate morphosyntactic description tags (MSDs) in accordance with MULText East recommendations for Croatian [12].

The CML is publicly accessible via Croatian Lemmatization Server which is presented below.

3.2. Croatian Lemmatization Server (CLS)

Croatian Lemmatization Server was developed in order to facilitate the processing of Croatian inflectional morphology in both directions: recognition and generation of WFs together with lemmatization. The CML as the list of lemmas accompanied by their WFs and their respective MSDs is stored in a database which is 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 typed in a HTML-form, but also of list of lemmas from an uploaded verticalized UTF-8 encoded text-file; 2) recognition of single (Figure 3) or multiple WFs typed in a HTML-form or of list of WFs from an uploaded verticalized UTF-8 encoded text-file.

Each WF submitted to the CLS gets recognized in all its possible inflectional interpretations

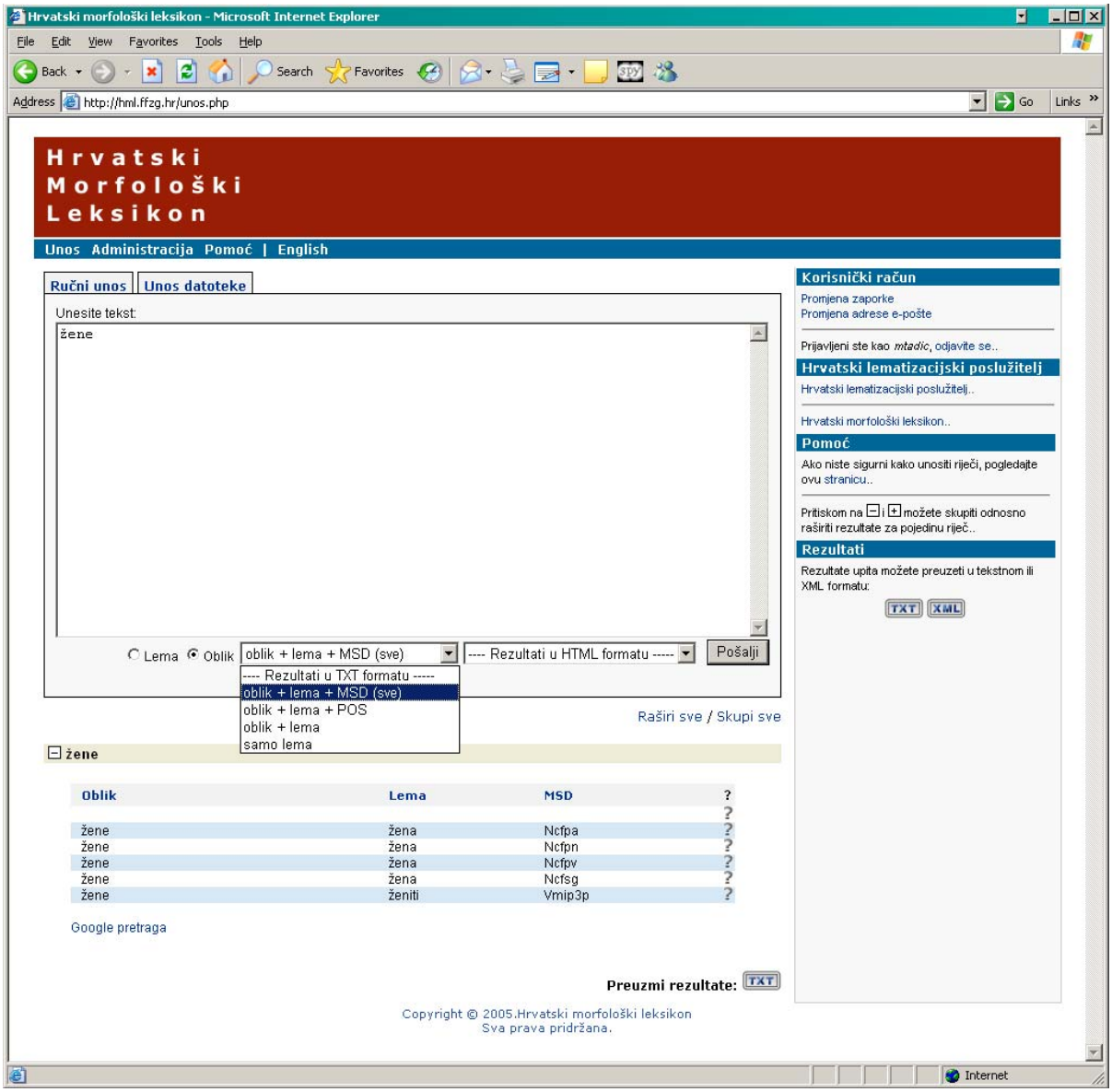

Figure 3. Sample of a single WF submitted with results included

(Figure 3) at unigram level i.e. without taking into account the preceding or the following tokens 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, etc.).

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

The coverage of the CML was tested on a 46 million tokens large Croatian corpus and it yielded the result of 96,4%. All unknown tokens are tracked automatically and stored in a log file for further processing and enhancing of CML.

3.3. Querying Google with Croatian inflection

One of the features of the CLS web-service is an automatically generated link to a result of the Google query with all WFs of the desired lemma (Figure 3: **Google pretraga** at the lower left corner of the window). For each lemma, a list of unique WFs is generated and transformed to a Google generated query. In fact it looks like a simple series of different WFs with Boolean OR between them. The same task could certainly be done manually, but it will 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 pipelined with other processes. This approach does not cover multiple word queries yet, since this will require defining their Boolean relations.

4. Testing the service

In order to get more insight in the coverage of inflectionally sensitive web search with the system described so far, we performed a series of tests using 20 randomly selected nouns from each of four different frequency rankings from the frequency list of Croatian National Corpus v 2.0 [13] which has the size of 101.2 million tokens. We wanted to test whether the inflectionally sensitive query always yields better results and does this difference show a regular pattern (or distance) for nouns with different frequencies of occurrences. Four different rankings were selected with the difference in the order of the magnitude. In the first column a lemma is listed. The second column gives the results of a query of Google search engine with exact WF (i.e. nominative singular only), the third with lemma (i.e. with all WFs generated), and the fourth gives the ratio between retrieved pages from the first and the second query. Using this ratio we wanted to see how much more hits we are gaining with all WFs instead of single exact WF. The Google queries were narrowed to a best-selling Croatian newspaper web-site only in order to keep the numbers without many trailing zeros.

	Google WF	Google lemma	WF lemma
godina	190,000	284,000	1.49
čovjek	16,900	154,000	9.11
kuna	121,000	127,000	1.05
predsjednik	137,000	166,000	1.21
vrijeme	143,000	186,000	1.30
pitanje	98,800	138,000	1.40
vlast	47,200	103,000	2.18
stranka	38,000	89,600	2.36
ministar	84,200	113,000	1.34
mjesto	79,700	154,000	1.93
zemlja	41,800	170,000	4.07
broj	102,000	130,000	1.27
strana	229,000	290,000	1.27
milijun	30,200	129,000	4.27
srijeda	52,300	107,000	2.05
dolar	507	59,600	117.55
zakon	51,800	105,000	2.03
sud	58,100	109,000	1.88
posao	64,800	151,000	2.33
kraj	44,400	150,000	3.38
Sum/Average	**1,630,707**	**2,915,200**	**8.17**

Table 1. Nouns with frequency above 10,000

	Google WF	Google lemma	WF lemma
proizvod	14,300	47,000	3.29
operacija	9,940	26,100	2.63
brod	15,200	34,800	2.29
ocjena	924	39,300	42.53
hotel	10,700	35,100	3.28
gradnja	13,500	37,900	2.81
ljubav	16,400	26,000	1.59
smještaj	12,300	16,200	1.32
pozicija	835	31,100	37.25
prihod	9,750	27,800	2.85
reakcija	938	23,500	25.05
nasilje	9,480	21,900	2.31
ostatak	10,600	24,700	2.33
majka	13,000	26,600	2.05
porez	12,100	25,600	2.12
naziv	9,520	32,900	3.46
prijevoz	9,390	15,000	1.60
povjerenik	813	11,700	14.39
konferencija	15,300	56,700	3.71
ishod	797	11,500	14.43
Sum/Average	**185,787**	**571,400**	**8.56**

Table 2. Nouns with frequency 1000 to 1005

	Google WF	Google lemma	WF lemma
djed	131,000	347,000	2.65
dopust	51,500	238,000	4.62
gostoprimstvo	19,900	36,300	1.82
igračka	64,100	410,000	6.40
krivnja	60,100	272,000	4.53
ljubimac	108,000	1,750,000	16.20
mrlja	58,100	125,000	2.15
naraštaj	27,100	138,000	5.09
neuspjeh	117,000	226,000	1.93
oluja	194,000	357,000	1.84
osovina	41,800	128,000	3.06
paket	476,000	1,240,000	2.61
plinovod	26,100	57,200	2.19
pomoćnik	543,000	670,000	1.23
smrtnost	52,100	83,700	1.61
stol	627,000	1,580,000	2.52
tajnica	574,000	665,000	1.16
tečaj	673,000	1,220,000	1.81
vježba	139,000	651,000	4.68
zbroj	90,200	160,000	1.77
Sum/Average	**4,073,000**	**10,354,200**	**3.49**

Table 3. Nouns with frequency 100

	Google WF	Google lemma	WF lemma
bruka	15	26	1.73
brojač	14	27	1.93
cinik	22	98	4.45
datoteka	40	74	1.85
defenziva	5	53	10.60
dezinsekcija	15	35	2.33
dramatizacija	29	101	3.48
glupan	9	33	3.67
guska	37	150	4.05
instanca	36	240	6.67
jedro	45	236	5.24
kondicija	29	220	7.59
konzola	30	79	2.63
kopriva	12	35	2.92
korigiranje	15	28	1.87
moreplovac	14	34	2.43
navoz	19	122	6.42
odstupnica	5	65	13.00
oganj	25	75	3.00
plodnost	59	115	1.95
Sum/Average	**475**	**1846**	**4.39**

Table 4. Nouns with frequency 10

4.5. Discussion

What can be seen from the four tables is that lemma query on Google search engine always yields more retrieved web pages than the simple WF query. This was certainly expected, but the ratio between those two types of queries is something what could come as a surprise. Theoretically for each noun lemma there are 14 different grammatical WFs while there is maximum 10 different orthographic forms (due to extensive homography in Croatian). The flat expectation would be that lemma query should behave somewhere close to this ratio (1:10) but it shows a significant variation from 1.05 (*kuna*) up to

117.55 (*dolar*). This two extreme cases are in fact present because of the very nature of Croatian inflection i.e. WF *kuna* is lemma (nominative singular feminine), but it is at the same time homographic form of genitive plural feminine which is used as a currency measure when denoting amounts of money. Therefore this is the most expected form. At the other extreme is also a name of the currency: *dolar* is lemma (nominative singular masculine) while the most expected WF is genitive plural masculine (*dolara*) with the same function as *kuna* above. The average overall ratio for all tables is 6.16. There can be also noticed that the average ratio between two types of queries is around 8 for high frequency nouns, while for lower frequency nouns it stays around 4. In general, it can be said that for Croatian nouns inflectionally sensitive Google search should yield around 6 times more web pages than the simple WF search.

5. Conclusion and further directions

We have shown a simple, generic solution for inflectionally sensitive search using Google search engine for web pages in Croatian. This system allows search with all Croatian inflectional POS (nouns, verbs, adjectives, adverbs, pronouns, numbers) and generates Google queries with all and only all of their WFs. It also features the lemmatization, POS/MSD tagging of Croatian texts without disambiguation. It also enables generation of all possible WFs of a given Croatian lemma. The system is available as the web service for guests and registered users (academic and commercial).

This simple inflectionally sensitive search has to be evaluated further in order to see how this improved recall relates to the precision but this is beyond the scope of this paper. Further directions for this type of web service could be connection to other search engines of user's choice.

The second step should be the refinement: at the lemmatization and POS/MSD tagging task, the system should deliver not just all possible lemma/POS/MSD interpretations (with high homography) but it should also offer certain level of disambiguation i.e. function as an on-line tagger for Croatian with the precision above 90% for general texts. The preliminary tests with some statistical taggers (e.g. TnT) show promising results but they is beyond the scope of this paper. Also an algorithm for finding lemmas for Croatian unknown words could be developed similar to one applied to Slovenian in [14].

6. References

[1] Šnajder, J. Rule-based automatic acquisition of large-coverage morphological lexicons for informational retrieval. Tech. Report, MZOŠ 2003-082, ZEMRIS, FER, University of Zagreb; 2005.

[2] Kržak, M.; Boras, D. Rječnička baza hrvatskog književnog jezika. Informatologia Yugoslavica, 17 (3-4); 1985.

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

[4] Lopina, V. Dvorazinski opis morfonoloških smjena u pisanome hrvatskom jeziku. Suvremena lingvistika 34; 1992.

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

[6] Pogodak search engine: http://www.pogodak.hr.

[7] Najdi search engine: http://www.najdi.si.

[8] Tadić, M. Information Retrieval Meets Human Language Technology. Proceedings of the CUC2000, CARNet, 2000. http://www.carnet.hr/CUC/cuc2000/radovi/prezentacije/F/F4/F4_f.pdf

[9] Tadić, M.; Fulgosi, S. Building the Croatian Morphological Lexicon. Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages, ACL, Budapest; 2003. http://hnk.ffzg.hr/txts/mtsf4EACL2003.pdf.

[10] Tadić, M. Računalna obrada morfologije hrvatskoga književnoga jezika. Ph.D. dissertation, Faculty of Philosophy, Univ. of Zagreb; 1994. http://hnk.ffzg.hr/txts/mt-dr-le.pdf.

[11] Tadić, M. Building the Croatian National Corpus. Proceedings of the LREC2002, ELRA-ELDA, Las Palmas-Paris; 2002. http://hnk.ffzg.hr/txts/mt4LREC2002.pdf.

[12] Erjavec, T.; Krstev, C.; Petkevič, V.; Simov, K.; Tadić, M.; Vitas, D. The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages, ACL, Budapest; 2003.

[13] Croatian National Corpus: http://hnk.ffzg.hr.

[14] Erjavec, T.; Džeroski, S. Machine Learning of Morphosyntactic Structure: Lemmatising Unknown Slovene Words. App. AI 18(1), 2004.