Journal of Computing and Information Technology - CIT 18, 2010, 4, 295–301
doi:10.2498/cit.1001914

Croatian Language Resources for NooJ

Kristina Vučković, Marko Tadić and Božo Bekavac

Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

This paper presents the Croatian module for NooJ. The module includes the novel “Posljednji Stipančići” by Vjenceslav Novak as a corpus with fully covered dictionary (i.e. zero unknowns). Examples of morphological and syntactic grammars are presented together with few examples of dictionary entries and their inflectional and derivational paradigms.

Keywords: NooJ, Croatian language, morphological grammar, syntactic grammar, corpus, dictionary, NP chunk, inflection, derivation, multi-word units

1. Introduction

The paper presents the Croatian module for NooJ development environment. NooJ is a development environment used to construct large-coverage formalized descriptions of natural languages, and apply them to large corpora, in real time. The descriptions of natural languages are formalized as electronic dictionaries, as grammars represented by organized sets of graphs [4]. According to the author of this tool, the word ‘NooJ’ is not an acronym, as it might be suspected, and has no additional meaning. We chose NooJ for text processing mainly because of its robustness and simplicity, but also for reasons expressed in detail in [11]. One of the important and useful features of NooJ, regarding Croatian, is its simple description of morphological phenomena. Efficient morphological processing is required for morphologically rich languages like Croatian which, for example, for nouns only has 7 cases, 3 genders and 2 numbers. Beside Nouns (N), Croatian has eleven more parts of speech: Adjective (A), Conjunction (C), Interjection (I), Numeral (M), Pronoun (P), Particle (Q), Adverb (R), Adposition (S), Verb (V), Residual (X) and Abbreviation (Y). Descriptions of some paradigms are briefly demonstrated in Sections 2 and 3.

Section 2 of the paper describes selected corpus and lexical data needed for its description.

Section 3 of the paper describes in detail three different types of morphological grammars that are included in the module.

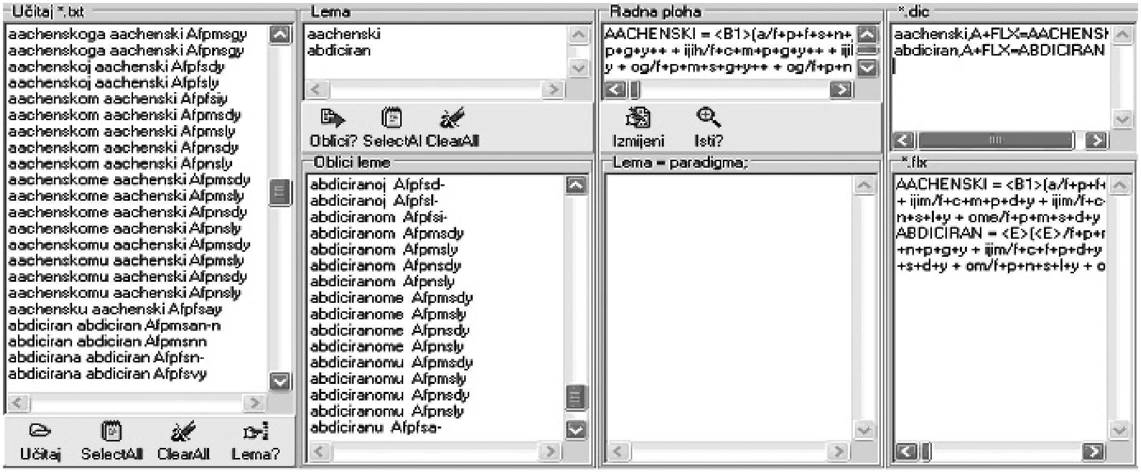
Syntactic grammars are graphs for detecting certain <*NP*> chunks and are described in Section 4 together with graphs for detecting multiword units which are described in Section 5. The plans for our future work are briefly discussed in the final Section of this paper.

2. Corpus and Lexical Data

Croatian module for NooJ uses Vjenceslav Novak’s novel “Posljednji Stipančići” as a working corpus. In order to have zero unknowns in the text, the main dictionary has all the words that are used in the novel.

We wanted to use the existing paradigm descriptions (i.e. inflectional patterns) from [6] and Croatian Morphological Lexicon (CML) [7] so we would not have to start building the computational model of Croatian inflection from the scratch since the data was already there. However, due to the different description principles used by NooJ and CML, the simple, automatic conversion of prototypical examples has produced too many inflectional errors. It was clear that looking for mistakes and fixing them manually, or manual rewriting of existing paradigms using NooJ inflectional formalism would be too time consuming as well as error prone.

This resulted in a new strategy which was to build an external program for conversion [11] of the Croatian Morphological Lexicon i.e. a list



```
aachenskoga aachenski Afpmsgy
aachenskoga aachenski Afpnsgy
aachenskoj aachenski Afpfsdy
aachenskoj aachenski Afpfsly
aachenskom aachenski Afpfsiy
aachenskom aachenski Afpmsdy
aachenskom aachenski Afpmsly
aachenskom aachenski Afpnsdy
aachenskom aachenski Afpnsly
aachenskome aachenski Afpmsdy
aachenskome aachenski Afpmsly
aachenskome aachenski Afpnsdy
aachenskome aachenski Afpnsly
aachenskomu aachenski Afpmsdy
aachenskomu aachenski Afpmsly
aachenskomu aachenski Afpnsdy
aachenskomu aachenski Afpnsly
aachensku aachenski Afpfsay
abdiciran abdiciran Afpmsan-n
abdiciran abdiciran Afpmsnn
abdicirana abdiciran Afpfsn-
abdicirana abdiciran Afpfsvy
```

```
aachenski
abdiciran
```

```
abdiciranoj  Afpfsd-
abdiciranoj  Afpfsl-
abdiciranom  Afpfsi-
abdiciranom  Afpmsdy
abdiciranom  Afpmsly
abdiciranom  Afpnsdy
abdiciranom  Afpnsly
abdiciranome  Afpmsdy
abdiciranome  Afpmsly
abdiciranome  Afpnsdy
abdiciranome  Afpnsly
abdiciranomu  Afpmsdy
abdiciranomu  Afpmsly
abdiciranomu  Afpnsdy
abdiciranomu  Afpnsly
abdiciranu  Afpfsa-
```

```
AACHENSKI = <B1>(a/f+p+f+s+n+
p+g+y++ + ijih/f+c+m+p+g+y++ + ijil
y + og/f+p+m+s+g+y++ + og/f+p+n
```

```
aachenski,A+FLX=AACHENSK
abdiciran,A+FLX=ABDICIRAN
```

```
AACHENSKI = <B1>(a/f+p+f+
+ ijim/f+c+m+p+d+y + ijim/f+c
n+s+l+y + ome/f+p+m+s+d+y
ABDICIRAN = <E>(<E>/f+p+r
+n+p+g+y + ijim/f+c+f+p+d+y
+s+d+y + om/f+p+n+s+l+y + o
```

Figure 1. Application PrepareForNooj used for conversion of data from MULTEXT East to Nooj notation.

of more than 4 million entries of a MULTEXT East conformant inflectional lexicon [3]. The program, symbolically named *PrepareForNooJ* [Figure 1], is divided into several steps. The first one filters out the lemmas and extracts word endings for all its word-forms together with their description (e.g. *Noun* + *common* + *feminine* + *Nominative* + *singular*). The next step filters out matching sets of descriptions and leaves only one occurrence of each. At the same time, dictionary and inflectional grammar files are formed. The last step converts the order and values of features (i.e. tag structure) into those used in NooJ.

The paradigm for nouns gives endings for 7 singular and 7 plural cases (nominative, genitive, dative, accusative, vocative, locative and instrumental). The verb paradigm gives endings in singular and plural and for all three persons for 8 simple verb forms (Long and Short Infinitives, Present, Imperative, Imperfect, Aorist), as well as Passive and Active participles, adding to them also all three genders. Since many paradigms share the same sets of endings, embedded graphs [4] were used in order to reduce the number of repetitions.

The novel "*Posljednji Stipančići*" has 14,944 different tokens. Ten most frequent ones (with the number of their occurrences) are given in Table 1.

The dictionary has 7,594 lemmas with the distribution as shown in Table 2.

All the foreign expressions (mostly Italian and German) used in the text are marked as frozen expressions <FXC> and are not further parsed inside the expression itself.

Token (hr)	**Token (en)**	**# of occurrences**
je	is	2,374
i	and	2,341
se	oneself	1,748
u	in	1,475
da	yes|to give	1,391
na	on	759
a	and|but	656
Što	what	622
Bi	would	558
s	with	521

Table 1. 10 most frequent tokens in the novel "Posljednji Stipančići".

Category	Notation	Distribution
Nouns	<N>	2,764
Verbs	<V>	2,109
Adjectives	<A>	1,945
Numerals	<M>	34
Pronouns	<PRO>	68
Prepositions	<S>	90
Adverbs	<R>	393
Conjunctions	<C>	97
Particles	<Q>	39
Exclamations	<I>	24
Frozen expressions	<FXC>	31

Table 2. Distribution of lemmas in Croatian dictionary.

Since Croatian language is highly inflectional, the number of word forms in the compiled dictionary is much bigger than the number of lemmas that are in the main dictionary. So, out of 407,761 generated word-forms, which is the maximal number of word-forms generated from the main dictionary and paradigm descriptions, there are 38,931 noun forms, 187,593 verb forms, 176,871 adjective forms, 2,981 forms for pronouns, and 382 forms for numerals, 90 forms for prepositions, adverbs, conjunctions, particles, exclamations and frozen expressions. The larger number of forms than it might be expected is due to multiple inflectional forms that some words can have in different cases (e.g. nouns and adjectives) or different tenses (e.g. verbs). For example, noun *crkva* (*church*) can have two plural forms in genitive case: *crkava* or *crkvi*.

Each entry (i.e. lemma) in the main dictionary is accompanied by its PoS category, some main additional features and link to the word which is used as the humanly readable representative for the type of inflection (*FLX*) and derivation (*DRV*) that an entry may have. This is explained in more detail in the following section.

Thus, nouns have additional features for type (common or proper) and gender (masculine, feminine, neutral); verbs may have the feature *+pov* if they are reflexive; adjectives have the features for type (qualificative, possessive) and definiteness, pronouns have the feature for type (personal, demonstrative, indefinite, possessive, interrogative, relative, reflexive), numerals have the feature for type (cardinal, ordinal, roman), prepositions have the feature for the case that is required by that preposition (genitive, dative, accusative, locative, instrumental). Remaining parts of speech have only the PoS category information.

All the descriptions inside the inflectional grammars are given in the form of rules. These rules are invoked by the property "+*FLX*=" that each inflective word has within its lexicon description. So, for example, the dictionary entry for the headword *dječak* (*boy*) is:

dječak,N+c+m+FLX=PROLAZNIK.

Inflectional grammar is looking for the paradigm named *PROLAZNIK* in order to generate all the forms of a headword.

Descriptions of paradigms are written after the following model explained in detail in [4] and [11]:

```
PROLAZNIK = <E>/Nom+s + a/G+s
+ u/D+s + a/Acc+s + <B1>če/Voc+s
+ u/L+s + om/I+s + <B1>ci/Nom+p
+ a/G+p + <B1>cima/D+p + e/Acc+p
+ <B1>ci(<E>/Voc+p + ma/L+p
+ ma/I+p);.
```

This paradigm, written in a textual mode, consists of a number of pairs describing all the possible forms. The first part of this pair describes a change on the word (e.g. <E>/ – no change happened, a/ – suffix 'a' was added, <B1>če/ – the last letter was deleted, then the suffix 'če' was added, etc.) while the second part describes features that the newly made word is given (e.g. /Nom+s – word is added description that it is in Nominative and singular, /D+p – word is in Dative and plural, etc.).

3. Morphological Grammars

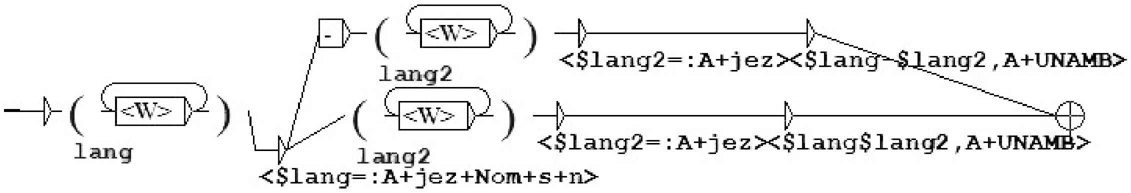
Morphological grammars demonstrate both inflectional paradigms and derivational patterns and can have graphical or textual presentations that are compiled in the finite state transducer. In the previous section, we have shown an example of textual presentation for the paradigm *PROLAZNIK*. In the similar manner, we can use the textual presentation to describe the derivation of a word. For example, the word *tri* (*three*) in the main dictionary has the following information:

tri,M+g1+FLX=TRI+DRV=AM3:
MUZIČKI+DRV=NM3:BRNJICA

The *+DRV* feature first invokes the derivation pattern named *AM3* after which the newly derived word is inflected as the paradigm named *MUZIČKI*. In this example, there are two possible derivations for the word *tri*:

AM3 = <B1>eći/A+po;
NM3 = <B1>ojica/N+f;

After the first derivation, the newly derived word *treći* (*third*) is marked as an adjective



<W>
lang
-
<W>
lang2
<W>
lang2
<$lang=:A+jez+Nom+s+n>
<$lang2=:A+jez><$lang-$lang2,A+UNAMB>
<$lang2=:A+jez><$lang$lang2,A+UNAMB>

Figure 2. Morphological grammar for recognizing new compound words.

while the second derivation produces the word *trojica* (*three people*) that is marked as a noun. Each new word is then inflected using the inflectional pattern names that follow the name of each derivational pattern, *:MUZIČKI* and *:BRNJICA* respectfully.

The following two morphological grammars use graphical presentation. The first one (see Figure 2) describes how from existing words describing a language that are in the dictionary, new compound word forms are built. In our graphical presentations, brackets refer to variables, symbols inside the rectangles are states of final state transducers, dollar sign ($) is used for the variable names ($N) and also variable attributes ($N$Case). These and remaining symbols are explained in more detail in [4].

The grammar in Figure 2 checks if there is a word in the compiled dictionary that is an adjective in Nominative Singular Neutral with semantic attribute +*jez* <$lang=:A+jez+Nom+s +n>. If such a word exists, then another word that is an adjective with semantic attribute +*jez* can come after it. Between these two words an '-' can be found. The newly made compound word takes the same annotation as the second word that it is made of. Thus, if the dictionary contains the following words:

- *engleski (English)*,
- *njemački (German)*,
- *hrvatski (Croatian)*,
- *francuski (French)*;

then the grammar in Figure 2 recognizes the following words as well:

- *englesko-njemački, engleskonjemački*;
- *njemačko-hrvatski, njemačkohrvatski*;
- *hrvatsko-francuski, hrvatskofrancuski*;
- *francusko-engleski, francuskoengleski*;
- *etc.*

and also all other forms of these words, considering different genders, cases and numbers.

This grammar saves time and space since not all language combinations have to be included in the main dictionary this way.

The second grammar (see Figure 3) describes the numerals written in letters. It uses embedded graphs and consists of five subgraphs.

Each word recognized by this grammar is annotated as a cardinal number (<*M*+*g*>). This grammar is adopted from [1] and it recognizes numerals like:

- *dvadesetpet (25)*
- *dvjestodvadesetpet (225)*
- *dvijetisućedvjestodvadesetpet (2,225)*
- *etc.*

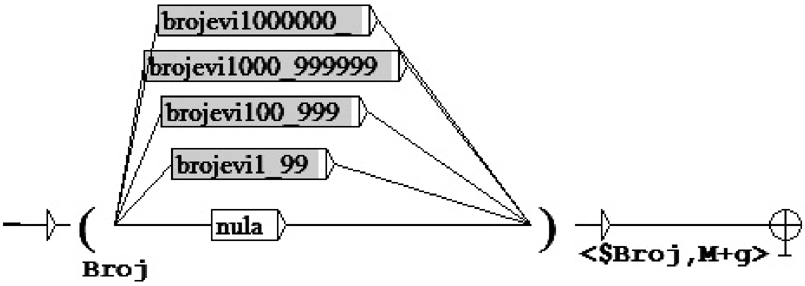


Figure 3. Morphological grammar for recognizing numerals written in letters.

4. Syntactic Grammars

Syntactic grammars work just like the morphological ones with the exception that syntactic ones can recognize multiple word strings or frozen expressions. Thus we might say that the typical usage of such grammars would be to recognize combinations of words such as frozen expressions [2] or named entities [1]. Since Croatian numerals may also be written in the following form:

- *dvjesto dvadeset pet (225)*
- *dvjesto dvadeset i pet (225)*
- *dvije stotine dvadeset pet (225)*
- *dvije stotine dvadeset i pet (225)*
- *etc.*

syntactic grammars are ideal means for representation of their structure as well.

Another example of using syntactic grammar in Croatian text is noun phrase (<*NP*>) chunks detection explained in more detail in [11] and [12].

Each <*NP*> chunk is made of one noun acting as the head of a chunk (see Figure 4) and any number of adjectives, pronouns and numerals preceding the head noun and agreeing with it in case, gender and number (see Figure 5).

The main grammar is shown in Figure 6 and it consists of two subgraphs (*HeadNoun subgraph* and *Add subgraph*).

The recognized <*NP*> chunk inherits all the information about *case, number* and *gender* from its head noun <NP+Case=$N$Case#$N$ ALLF>.

The grammar recognizes the following examples from the novel "*Posljednji Stipančići*":

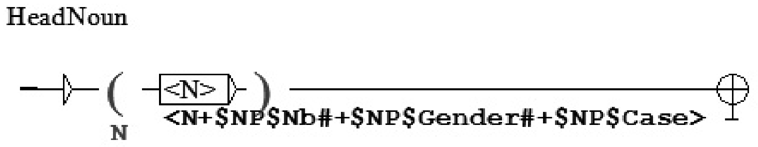


Figure 4. The subgraph for recognizing nouns.

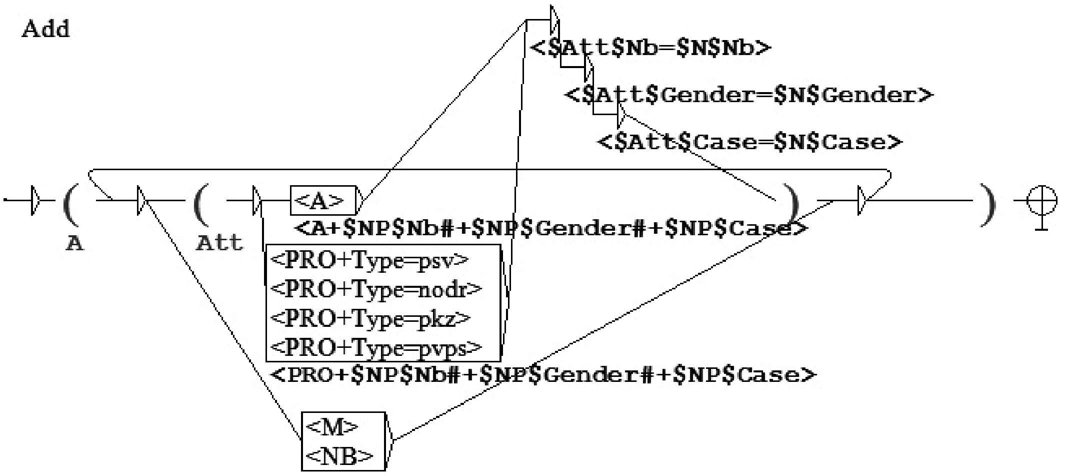


Figure 5. The subgraph for recognizing adjectives, pronouns and numerals preceding the head noun shown in Figure 4.

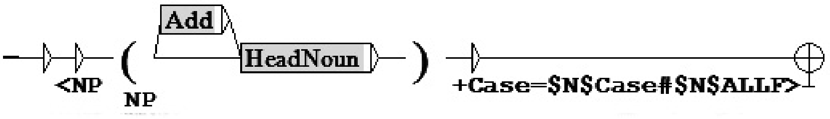


Figure 6. The main grammar for <NP> detection.

- *strmih kamenitih stuba*
- *uzidana nakazna ljudska glava*
- *ozbiljnim tamnosmeđim očima*
- *šesnaestogodišnje djevojče*
- *strogog ženskoga instituta*
- *ocrtanim obrvama*
- *takvim portretom*
- *šest tamnocrvenih mekanih stolica*
- *ogromnim starinskim naslonjačem*
- *Lucijina mati.*

5. Collecting Multi-word Units

NooJ dictionaries consist mainly of simple words (basic word forms or lemmas), but also of multi-word units or compound words, as they are frequently termed. Multi-word units are considered to be expressions whose linguistic behavior is not predictable from the linguistic behavior of their component words. Since some multi-word units can exhibit quite productive inflection and/or syntactic flexibility, particularly in languages that are known for high degree of inflection and relatively free word order, they are sometimes hard to discriminate from generated strings of words that can occur by chance.

Multi-word units, such as *red tape* or *sea cloud* are often language and culture dependent. So if we choose to use literal rewriting from some English or German lexicon of compound words, it would yield poor results in Croatian. On the other hand, if compound words are to be collected from a very large corpus by semiautomatic extraction, we believe that much better results would be achieved. Methodology for semiautomatic extraction of multi-word units from Croatian version of Wikipedia is described in more detail in [2].

Proposed methodology serves as basis for inclusion of multi-word units as lexical entries in NooJ dictionary.

6. Conclusion and Future Directions

We have presented the basic resources for processing Croatian language using NooJ development environment. These resources provide the foundations for all future analyses of Croatian texts inside the NooJ development environment. All additional word descriptions, as well as multiple-word patterns, can easily be added to the system as an upgrade to these resources.

At the moment, we are working on the expansion of simple words dictionaries, but also on different grammars for recognizing derived forms from existing lexical entries as well as on multiword units depending on different contexts. Parsers for simple sentences [10] and detection of clauses [9] are under development as well and so far are giving promising results. None of this work would be possible if the basic resources, as presented here, did not exist.

7. Acknowledgments

This work has been completed within the projects supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants 130-1300646-1776, 130-1300646-1002 and 130-1300646-0645.

References

[1] B. Bekavac, Strojno prepoznavanje naziva u suvremenim hrvatskim tekstovima. PhD. Thesis, Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia, 2005.

[2] B. Bekavac, M. Tadić, A Generic Method for Multi Word Extraction from Wikipedia. In *Proceedings of the 30th International Conference on Information Technology Interfaces ITI 2008*, (2008) Cavtat, Croatia, pp. 663-668.

[3] T. Erjavec, Multext-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, ELRA, Lisbon-Paris*, (2004) pp. 1535–1538.

[4] M. Silberztein, NooJ manual. Available at the web site `http://www.nooj4nlp.net` (200 pages), 2003.

[5] M. Silberztein, NooJ's Dictionaries. In *Proceedings of LTC*, (2005) Poznañ University.

[6] M. Tadić, Računalna obrada morfologije hrvatskoga književnoga jezika. PhD. Thesis, Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, 1994.

[7] M. TADIĆ, The Croatian Lemmatization Server. *Southern Journal of Linguistics*, 29 (2007), pp. 206–217.

[8] S. TEŽAK, S. BABIĆ, *Gramatika hrvatskoga jezika*, Školska knjiga, Zagreb, 1992.

[9] K. VUČKOVIĆ, Ž. AGIĆ, M. TADIĆ, Sentence Classification and Clause Detection for Croatian. In *FASSBL7 Proceedings*, (2010) Dubrovnik – Zagreb, Croatia pp. 131–138.

[10] K. VUČKOVIĆ K, B. BEKAVAC, Z. DOVEDAN, SynCro – Parsing Simple Croatian Sentences. In *Finite State Language Engineering: NooJ 2009 International Conference and Workshop*, (2010) Centre de Publication Universitaire, Touzeur, 207-217.

[11] K. VUČKOVIĆ, M. TADIĆ, Z. DOVEDAN, Rule Based Chunker for Croatian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC2008*, (2008) Marrakesh-Pariz; ELRA, pp. 2544–2549.

[12] K. VUČKOVIĆ, Model parsera za hrvatski jezik. PhD. Thesis, Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia, 2009.

Received: June, 2010
Accepted: November, 2010

Contact addresses:

Kristina Vučković
Department of Information Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3
10000 Zagreb
Croatia
e-mail: kvuckovi@ffzg.hr

Marko Tadić
Department of Linguistics
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3
10000 Zagreb
Croatia
e-mail: marko.tadic@ffzg.hr

Božo Bekavac
Department of Linguistics
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3
10000 Zagreb
Croatia
e-mail: bbekavac@ffzg.hr

KRISTINA VUČKOVIĆ, teaching and research assistant at the University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences. She began to work as a research fellow at the Department of Information Sciences, Faculty of Philosophy in Zagreb in November 2000 on the project "Machine Understanding of Natural Languages". She received her PhD (2009) at the Faculty of Humanities and Social Sciences, University of Zagreb with the dissertation "Model of a Parser for Croatian Language". She has participated in several international and Croatian conferences and has more than thirty published papers in the areas of natural language processing and usage of new technologies in education.

MARKO TADIĆ, linguist, professor at the University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics. He is the head of the Chair of Algebraic and Computational Linguistics at the same department since 2001 and an associated member of the Croatian Academy of Sciences and Arts since 2008. Published more than 60 papers and a book "Jezične tehnologije i hrvatski jezik" (Language Technologies and Croatian Language) and he is also one of the authors of the first Croatian Frequency Dictionary (Hrvatski čestotni rječnik). His interests are in corpus linguistics (Croatian National Corpus, hnk.ffzg.hr), computational linguistics (Croatian Morphological Lexicon, hml.ffzg.hr), language technologies (Computational Linguistic Models and Language Technologies for Croatian, rmjt.ffzg.hr), research infrastructures (NCP for Croatian in project FP7 project CLARIN and ACCURAT and ICT-PSP project LetsMT!).

BOŽO BEKAVAC was born on July 31st 1972, in Split. In April 1997, he obtained B.A. degree in general linguistics and information science from the Faculty of Philosophy, University of Zagreb. He began to work as a research fellow at the Institute of Linguistics, Faculty of Philosophy in Zagreb, in September 1997 on the project "Machine Processing of the Croatian Language". His scientific work focuses on corpora and computational tools for processing of the Croatian language. He received his PhD (2005) at the Faculty of Humanities and Social Sciences, University of Zagreb with the dissertation "Automatic Named Entities Recognition in Croatian Texts". He participated in several international and Croatian conferences and he has around twenty published papers in the fields of named entity recognition and classification (NERC), corpus linguistics, linguistic tools (Intex, NooJ...), computational linguistics, mark-up languages (XML, SGML). He became assistant professor in 2007. He is a member of Croatian Philological Society, Croatian Association for Applied Linguistics, Slovenian Society for Language Technologies and ACL.