Computational Linguistic Models and Language Technologies for Croatian

Bojana Dalbelo Bašić*, Zdravko Dovedan†, Ida Raffaelli†, Sanja Seljan†, Marko Tadić†

Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
bojana.dalbelo@fer.hr
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{marko.tadic, zdravko.dovedan, ida.raffaelli, sanja.seljan}@ffzg.hr

Abstract. *This paper gives an overview of the scientific program "Computational linguistic models and language technologies for Croatian" that has been launched recently. Its short and long term goals, its composition as well as methodology and expected results are presented.*

Keywords. computational linguistics, natural language processing, Croatian language, corpus linguistics, POS/MSD tagging, lemmatization, parsing, WordNet, machine translation, machine aided translation, computer assisted language learning, automatic document indexing, document classification, document summarization.

1. Introduction

Nationally funded projects from the domain of computational linguistics and corpus linguistics (CL), natural language processing (NLP) and language technologies (LT) have already existed for quite a while in Croatia. It would be sufficient to mention only the most prominent ones.

CL: the corpus linguistics in the Institute of Linguistics of the Faculty of Philosophy (today Faculty of Humanities and Social Sciences) already has a long tradition starting with the first Croatian computer corpus in 1967 [1], followed by the first one-million Croatian corpus and Croatian Frequency Dictionary [2] and by Croatian National Corpus [3]. During '90 an intensive research was going on within the project "Computational Processing of Croatian" [4] which was followed by a project "Development of Croatian Language Resources" [5] at the turn of the century at the same Institute and at the Department of Linguistics of the same faculty.

NLP: During '90 at the Department of Information Sciences of the same faculty existed projects "Models of Knowledge and Natural Language Processing" [6], "Computer Understanding of the Croatian" [7], "Tagging and Recognition of Croatian Words" [8] where different NLP methods for processing Croatian were developed. Croatian language resources were digitized within the project "Croatian Dictionary Heritage and Dictionary Knowledge Representation" [9]. At the Faculty of Electrical Engineering and Computing within the i-project (i.e. project that support the usage of IT in research) "Text Mining System" [10] also a series of NLP methods and research procedures were developed.

LT: At the turn of the century a project "Language Technologies for Croatian" was funded by the Ministry of the Science and Technology of the Republic of Croatia and it resulted with a portal for Croatian LT [11]. In 2003 the first book [12], which presented the situation with Croatian LT, defined the problems and proposed the possible solutions was published. Also in mid '90 the first project "Croatian Standard Speech in Speech Technology" [13] dealing with Croatian speech processing was going on and its cooperation in an international project MBROLA [14] resulted in the Croatian diphone database. This first project dealing with speech processing was followed by another project "Digital Speech Processing in Contemporary IT" which lasted from 2002 to 2005 [15] at the Faculty of Electrical Engineering and Computing.

In spite of all this projects and activities which resulted with many publications, there has been no coordinated scientific project (or program) which would try to encompass all activities from these fields. In this way valuable human and financial resources were unnecessary spent because in a number of cases different projects were engaged in investigation of the same phenomena without knowing for each other and

usually getting similar results in different, incompatible data formats.

Since development of LT for any language is too expensive and time-consuming process it should be carefully planned and monitored in its realization particularly in the case of lesser spread languages like Croatian is. In order to facilitate such preplanned and structured approach, a scientific program has been launched within the last national call for projects in 2006.

The rest of this paper will describe this program where in the second section its composition and in the third section expected overall results will be presented.

2. Composition of the program

Since in the national call for projects in 2006 interdisciplinary program were explicitly encouraged, we wanted to launch a proper interdisciplinary program consisting of five projects each covering another subfield and trying to tackle one of the burning problems in Croatian LT. The projects (P1-P5) are distributed between three different scientific fields: P1 and P3 are situated in humanities (linguistics), P2 and P4 are from social sciences (information sciences) while P5 is from technical sciences (computing).

The motivation for such an overall "umbrella" program was simple: to organize and coordinate as many projects from these fields. Croatia does not have too many LT researchers and it would be terrible waste not to organize them in a joint and coordinated manner. This certainly does not mean that this program intended or managed to collect all projects dealing with Croatian LT in this project call. There are other projects, but we believe that this program will lead the mainstream of Croatian LT development in next five years.

At this point of development LT for Croatian we did not include dictionary digitalization nor speech processing in this program because they are already included in other programs with which we are already establishing cooperation as well.

The composition of this program will be explained in next five subsections.

2.1. P1: Croatian Language Resources and their Annotation

This project will primarily deal with augmenting and building Croatian computer corpora. Interpersonal verifiability and exact measurability of corpus data enables the linguistics a scientific approach similar to one in natural science. This project tends to achieve such a level of approach with several objectives:

1) expand the existing Croatian National Corpus (HNK) [16] from 101 million up to 200 million tokens. The balance and sampling of texts according to different text types, media, genres and domains will be particularly observed;

2) linguistically annotate HNK with tags on morphological (PoS, MSD, lemmas) [18], partly syntactic (chunks, clause and sentence structures) producing a Croatian Treebank and semantic (word sense tags on the basis of Croatian WordNet);

3) realize basic statistics on occurrence, frequency and distribution of linguistic units on several language levels on the basis of such a large corpus;

4) compile a certain number of smaller Croatian corpora for special purposes as needed;

5) for the development of multi-lingual LT, a series of parallel corpora "Croatian—language X" will be compiled and processed similar to already existing Croatian-English Parallel Corpus [17].

Each new corpus compilation and its study reveals new, often unexpected results and methods. Such insights into systematized language material are often complementary to intuitive and introspective study of language phenomena. A whole range of new theoretical insights, research and methodological approaches that have not been verified in the processing of such a large quantity of language material so far, are expected to be tested and applied within this project.

The purpose of the project is to build a representative corpus of Croatian in accordance with the recent advances in corpus and computational linguistics. This corpus would serve as a systematically compiled source of language material for all kinds of theoretical and practical studies in Croatistics, general and computational linguistics dealing with Croatian as well as its NLP. Such a corpus is at the same time fundamental language resource for the development of LT for Croatian so its close connection with other projects in the program (P2-P5) is obligatory.

2.2. P2: Computational syntax of the Croatian Language

The project "Computational syntax of the Croatian Language" is a continuation of the project "Natural Language Machine understanding"

where some research has been conducted concerning the applicability of Lexical Functional Grammar and Case Grammars for the Croatian language analysis. The research results have been presented in [19] and [20]. Our experience, gained through the realization of the previous project, gives us the needed competence to conclude that cited methods clearly show 'the school examples' of syntax analysis for simple Croatian sentences but are not suitable for descriptions of lexical and syntax structures of Croatian language over the larger corpus which must include complex sentence structures as well.

The research in the project will follow two main directions: constituency oriented and dependency oriented line of research on Croatian computational syntax.

The first one will depend on the basic grounds that have been established for dealing with the problems of syntax analysis of formal and programming languages. Also, conclusion has been reached that application of automata theory and its upgrade might solve the problem of the description of syntactic structures in Croatian language in a relatively simple and quite efficient way, and that the parser can be built with the ability to parse real sentences validated by the corpus.

Since the early works by N. Chomsky and introduction of transformational generative grammar, numerous models of syntax analysis have also been developed (LFG, HPSG, Tree-adjoining Grammars, Phrase Structure Grammars, etc.) but all have mainly been used for non-Slavic languages. Croatian, as a Slavic language, imposes some additional problems to this formalisms such as long-distance dependencies, branch crossing, relatively free order etc.

On the other hand, we have succeeded in realizing several attempts of syntactic analyses for context free languages and in building our own original model of syntactic analysis of artificial (i.e. programming) languages [21, 22, 23]. An effective deterministic automaton has been produced. The basic starting point, a hypothesis, is that, when extended, this automata may be used for solving problems of natural language syntactic analysis, especially of Croatian. Parser will be build with Croatian words presented as objects and at the same time as function states of automata transitions. Transition from the current state to the next state would be conditioned or constrained by the context of the next word's features. The recognition procedure of the input sentence resembles the process of building a puzzle so it will be called the Puzzle Parser. This computer model will be checked on (a least a part of) a large corpus (HNK) in order for the Croatian Treebank to be built.

The fundamental hypothesis for parser realization is well defined lexical structure of Croatian language. Each word will, next to its standard features (PoS and MSD), have additional properties added to it. These additional properties would include sets of possible words preceding and following each word, and additional contextual properties. Parallel to the process of defining the features, i.e. building a lexicon, computer program for word and it's features insertion will be developed based on already existing Croatian Morphological Lexicon [12, 24].

During the first phase of the research, the prototype parser will be tested over the set of approximately 10 000 most frequently used Croatian words. Afterwards, the parser will be used over a much larger lexical set.

The goal of this research is to find formal models for describing characteristics of Croatian words, to define lexical and syntactical analysis of Croatian sentences, to establish and maintain the system for data entries into the lexicon and parser, and present them as one of the models for syntactic analysis of Croatian sentences.

The second line of research, the dependency oriented one, will try to describe syntactic relations in Croatian sentences using dependency formalism. It represents the continuation of work on the Croatian Dependency Treebank [25, 26] that started in 2005 and which tried to apply the approach used in Prague Dependency Treebank (PDT) [27] to Croatian. In this respect very important reference is [28] which elaborated thoroughly all theoretical prerequisites for this approach to Croatian dependency syntax analysis thus giving theoretical basis for Croatian dependency treebank building.

In order to build the Croatian dependency parser a collaboration with a team building the Slovenian Dependency Treebank [29, 30] is also planned. Desired goal, which has not been explicitly expressed in project proposal, would be to build a Croatian treebank which would feature both annotation formalisms as results of two different parsers: a constituency and dependency.

2.3. P3: Lexical Semantics in Building Croatian WordNet

The project “Lexical Semantics in Building Croatian WordNet” is based on two different but

complementary frameworks under which a research activities will be organized:

1) theoretical framework of lexical semantics;

2) practical work concerning the construction and design of a lexical database.

There are various theoretical and methodological studies in Croatian lexicon, but a systematic and detailed semantic research of wide areas of Croatian vocabulary is still missing. There is also no computational lexical database that would serve as the basis for such wide-coverage studies of vocabulary and at the same time provide data for the development of NLP tools. In such a way, the future Croatian WordNet (CroWN) should have an impact not only on NLP for Croatian, but also in lexical semantics theory and methodology.

In the first phase of the project a set of 1300 base concepts will be translated and adjusted. This set will serve as a starting point for the development of Croatian ontologies. In order to retain the compatibility with other WordNet projects, CroWN will be based on the common set of base concepts, whereas the building of ontologies will reflect conceptualization and lexicalization characteristic for the Croatian language. To enable connecting the CroWN to other WordNets developed in EuroWordNet 1 and 2 and BalkaNet, the interlingual index (ILI) will be assigned to members of Croatian synsets as well.

Apart from the necessity to be compatible with other WordNets, CroWN should preserve and maintain language specificity of Croatian lexical system in order to be a computational lexical database which reflects all semantic and morphological particularities of lexical structures in Croatian that will especially become prominent in the construction of synsets. Each lexical entry will be accompanied by a short definition of its meaning and a set of synonyms.

The next step of the project is the construction of Croatian synsets (in the first step synsets of nouns, later verbs, and eventually adjectives and adverbs). When building verb synsets a systematic attention will be paid to verb valencies. Each verb synset will contain its respective valency frame displaying the information about the morphosyntactic and semantic features of its arguments, like in [31, 32, 33]. In order to provide a full morphological data on Croatian lexical units, they will be linked to the Croatian Morphological Lexicon [12, 24].

Designed in such a manner, the CroWN has its application in:

1) semantic and lexicological studies, and – due to its potential multilinguality – in contrastive studies as well as in foreign language learning and learning of Croatian as a foreign language;

2) development of NLP tools for Croatian such as:

a) semantic tagging of corpora and word sense disambiguation;

b) construction of a chunker and a parser, especially using verb valency frames from CroWN;

c) development of a machine translation system;

d) information retrieval, data mining, retrieval of conceptual relations, document classification, document summarization.

Such a project is the first effort in the field of large semantic network (i.e. lexical database) development for Croatian. In this sense, this project is also the first step in the development of a Croatian thesaurus. The digital storage of language data enables its search via Internet and as well as its application in various area of NLP. Its multilingual compatibility in terms of format and language data makes this resource a potential dictionary of Croatian and all the other languages coded according to the already established guidelines for developing WordNets.

2.4. P4: IT in Translation of Croatian and in Language e-Learning

Integrating into European environment through numerous agreements and the EU accession, linguistic tools and resources of Croatian should be developed in order to enable multilingual communication. As the EU upholds the principles of open access and multilingualism, the Croatian language should be treated in the same way as the other EU languages. This means that the tools and resources for the e-language learning (eLL) and machine translation (MT) of Croatian should also be developed.

The project forks in two main directions of research:

1) prerequisites (i.e. resources and tools) for machine assisted translation (MAT) and for commencing the building of the machine translation (MT) system where Croatian is source or target language while English is the other;

2) resources and tools for e-language learning (eLL) of Croatian.

It is envisaged that the research in the two directions will go hand in hand, as the resources

and tools are to a large extent common to both sections.

The research would involve insights from various models that already exist for other languages, such as researches in corpus linguistics using statistical machine translation (SMT); example-based machine translation (EBMT) resulting with translation memories (TMs). In addition to the (morphologically sensitive and insensitive) statistical analyses, which have not been done on Croatian texts yet, methods would be examined for word, phrase and sentence alignment necessary for the building of translation memory systems based on parallel corpora. Lexical and syntactic relations between Croatian and English relevant for the building of machine translation systems (e.g. EC Systran) will be also examined.

Resources and tools for Croatian MAT/MT will be developed in highly specialized areas in which the texts have been prepared for this type of translation using the controlled language (terminological consistency, unambiguous sentences, simple syntactic structures, and so on). Attention will be paid to MT and MAT as an integral part of documentation processing, as in localization of industrial products.

The existing models for e-learning of more widely spread European languages, which make use of online multimedia materials, self-assessment tests, assignments with feedback, etc. should be developed for the Croatian language, based on empirical research in the area of computer linguistics and language didactics. This line of research would investigate simulations, models and studies related to the application of ICT in language learning; analyses of traditional and hybrid models; acquisition of linguistic and technical competences in eLL. The research would enable the construction of a prototype model for language learning in e-environment, as well as analysis of a blended model of language learning by use of ICT in learning and teaching process and in the life long learning as well.

The choice of tools and standards for the development of language resources will be governed by the use in the EU, with all the necessary adaptations related to the Croatian language. The findings derived from the research and from the practical application of the tools will be used in the education of students enrolled in computer science and modern language departments.

The proposed project is in line with the EU demands, with the priorities stated in FP7, with the needs of the EU's Directorate-General for Translation of EC for the development of language technologies and resources, as well as with Croatia's short-term and long-term strategic goals in science and technology development.

2.5. P5: Knowledge Discovery in Textual Data

While the quantity of information grows everyday, human capacity to understand and process it remains unchanged. Knowledge discovery methods in textual data have the goal to relieve people as much as possible of processing of saved and previously processed data and enable them to concentrate on making decisions based on results of knowledge discovery procedures in textual data. The field of information retrieval is a new research field, a cross-over between machine learning, natural language processing, computational linguistics, mathematics and statistics, and may be presented as a branch of knowledge discovery in data. The tasks of knowledge discovery in textual data may be formulated as: information extraction, document summarization, automatic document indexing, automatic classification and clustering of documents.

The project “Knowledge discovery in textual data” aims to answer as to how to approach such tasks and which models to apply for individual tasks noted above, particularly with texts in Croatian language. The main hypothesis of this project is that a new level of intelligent systems for knowledge discovery in textual data may be developed by comprehensive interdisciplinary problem analysis, by including language-specific knowledge (specifically Croatian), and by applying the results of fundamental researches in mathematics, cognitive science and machine learning.

One further hypothesis of this project is that the problem of overabundance of textual data and the problem of identifying potentially useful, but hidden information may be tackled by identifying individual problems, formulating them into above listed tasks and, by the applying interdisciplinary approach (computational linguistics, mathematics, statistics, computer and cognitive sciences), forming models that represent solutions to these problems, finally integrating all solutions into a system, representing a contribution to rising from the level of information management to the level of knowledge management and decision making.

This project, together with other projects within the framework of the program, shall encompass both research levels of knowledge dis-

covery methods in textual data, the language independent level and the language specific level. The focus of research of this project is the language independent level. However, particular attention shall be given in this project to the integration of the language independent level (to be developed in this project) and the language specific level (to be developed within other projects of the Program), so as to tackle the problem of underdeveloped resources and tools for Croatian. Here, the concept of systematic inclusion of language specific knowledge shall be respected when developing intelligent systems for the Croatian language, regardless of whether this knowledge is acquired by learning (machine learning methods) or is based on symbolic-logical systems.

In the domain of language independent techniques, this project shall study the dimensionality reduction problem for term-document matrices representing a collection of documents with an extremely large dimensionality. The research shall be based on the assumption that this matrix has redundancy and that there is another matrix which is close to the initial matrix representing the collection, but which better represents the collection. Also, it shall be assumed that it is possible to work with graph representation of a document collection and that the document clustering problem may be solved by means of spectral graph partition.

One further assumption of this research is that some of the tasks set in this project (automatic document indexing and summarization) may be successfully solved by approaches based on known models in the field of cognitive psychology, by researching cognitive processes when dealing with text mining, document summarization and indexing tasks. Systems developed to date, based on statistical, syntactic and semantic methods, attempt to replicate the results achieved by humans, but their success is limited due to insufficient knowledge of cognitive processes related to text mining activities. The most recent scientific researches, in which research results in the field of cognitive science are combined with text mining point in this direction of research.

The described assumptions shall be verified by forming and experimentally evaluating a knowledge discovery model in textual data.

Project "Knowledge discovery and textual data" has four main task, each having a number of subtask:

a) Text preprocessing techniques in Croatian language for machine learning processes.

b) Dimensionality reduction and document clustering in the vector space model.

c) Automatic indexing and summarization of documents.

d) Intelligent, language specific information retrieval and extraction.

Project's aim is achieving scientific excellence in the design of computational-linguistics, mathematical-statistical, symbolical-logical and cognitive models for automatic document indexing, efficient information retrieval (collection of documents and www pages), as well as automatic document summarization and knowledge extraction from textual data.

Connecting basic researches with practice should lead to the development of knowledge extraction models and managing textual data in digital form, with particular interest to developing such systems for Croatian language.

The purpose of this project within the framework of the entire program is to raise the level of computational linguistic resources and tools for Croatian language. This will be achieved by new solutions proposed in tasks.

Finally, the project's aim is to include, through the models and systems developed in the project, the Croatian language in the community of European languages at the level of developed linguistic technologies and systems for knowledge discovery in textual data.

3. Expected overall results

While the expected results for individual projects have been presented in the previous subsections, the overall results of the whole program will be discussed further.

Since the aims of projects P1 to P4 are to develop necessary LR and LT, it directly conditions the aim of the P5 which is to use these LR, LT and developed methods on a document and document-collection level in order to achieve better results than with previous only mathematical and/or computational methods which were used without LT. P5 is expected to come out not only with new methods in NLP of Croatian but also with applications that could be directly used in different areas of information extraction, retrieval, classification, indexing and summarization of (Croatian) documents. These applications could also have a market value.

The overall goal of the program "Computational Linguistic Models and Language Technologies for Croatian" is to rise the general level of LT for Croatian after 5 years. The expected

level is similar to one defined within the BLARK [34] which would enable not only the direct application of developed language resources or tools in existing research, but also further research and advances to new levels of LT for Croatian.

One of the ideas behind this program is also not to have five isolated projects but to interconnect different institutions from all around Croatia and their researchers to a research community which shares not only ideas and common interests but also researchers when needed for completion of a certain task. Having in mind a number of MAs and PhDs which will be produced within this program, it may happen that we will come out with a critical mass of researchers from the field of Croatian LT which would enable more advanced studies and building of more complex LT systems (e.g. MT systems). This yet remains to be proven.

The facilitation of the development of LT for Croatian will indirectly enable the Croatian language to participate on equal terms in the process of building of information and knowledge society in EU.

4. Acknowledgements

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-0645, 130-1300646-1776, 130-1300646-1002, 130-1300646-0909, 036-1300646-1986 and partially by joint Flemish-Croatian project CADIAL.

5. References

[1] Bujas, Ž. Ivan Gundulić »Osman«, komputorska konkordancija, Sveučilišna naklada Liber, Zagreb; 1975.

[2] Moguš, M.; Bratanić, M.; Tadić M. Hrvatski čestotni rječnik. Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu-Školska knjiga, Zagreb; 1999.

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

[4] Project info: http://zprojekti.mzos.hr/zProjektiOld/arh_det.asp?sort=3&offset=360&ID=1089.

[5] Project info: http://zprojekti.mzos.hr/zProjektiOld/prikaz_det.asp?sort=3&offset=535&ID=0130418.

[6] Project info: http://zprojekti.mzos.hr/zProjektiOld/arh_det.asp?sort=3&offset=375&ID=1105.

[7] Project info: http://zprojekti.mzos.hr/zProjektiOld/arh_det.asp?sort=3&offset=375&ID=1103.

[8] Project info: http://zprojekti.mzos.hr/zProjektiOld/arh_det.asp?sort=3&offset=345&ID=1067.

[9] Project info: http://zprojekti.mzos.hr/zProjektiOld/prikaz_det.asp?sort=3&offset=475&ID=0130464.

[10] Project web page: http://textmining.zemris.fer.hr/.

[11] Project web page: http://jthj.ffzg.hr/.

[12] Tadić, M. Jezične tehnologije i hrvatski jezik. Exlibris, Zagreb; 2003.

[13] Project info: http://www.mzos.hr/svibor/6/03/010/proj_h.htm.

[14] Project web page: http://tcts.fpms.ac.be/synthesis/mbrola.html.

[15] Project info: http://zprojekti.mzos.hr/zProjektiOld/result_det.asp?trazi=govor&gdje=2&Submit=Pretrazi&ID=0036054.

[16] Croatian National Corpus web page: http://hnk.ffzg.hr/.

[17] Tadić, M. Building the Croatian-English Parallel Corpus. In: Proceedings of the LREC2000, ELRA-ELDA, Athens-Paris; 2000. http://hnk.ffzg.hr/txts/mt4LREC2000.pdf.

[18] 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.

[19] Seljan, S. Leksičko-funkcionalna gramatika hrvatskoga jezika: teorijski praktični modeli. Ph.D. dissertation, Faculty of Philosophy, Univ. of Zagreb; 2003.

[20] Vučković, K. Padežne gramatike i razumijevanje hrvatskoga jezika. M.A. thesis, Faculty of Philosophy, Univ. of Zagreb; 2004.

[21] Dovedan, Z. Formalni jezici: sintaksna analiza. Zavod za informacijske studije Odsjeka za informacijske znanosti, Zagreb; 2003.

[22] Dovedan, Z.; Paić, G.; Seljan, S. Konačni automat i izvođenje gramatike linearne zdesna. In Lasić-Lazić, J. (ed.) Informacijske znanosti u procesu promjena. Zavod za

informacijske studije Odsjeka za informacijske znanosti, Zagreb; 2005.
[23] Dovedan, Z.; Stojanov, T.; Vučković, K. Syntax Analysis Directed by Transition and Action Table. In Lasić-Lazić, J. (ed.) Informacijske znanosti u procesu promjena. Zavod za informacijske studije Odsjeka za informacijske znanosti, Zagreb; 2005.
[24] 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.
[25] Tadić, M. Starting with Croatian Dependency Treebank. Suvremena lingvistika 63; 2007 (in press).
[26] Croatian Dependency Treebank web page: http://hobs.ffzg.hr/.
[27] Prague Dependency Treebank web page: http://ufal.mff.cuni.cz/pdt/.
[28] Vuković, P. Prednosti dvorazinske valencijske sintakse u sintaktičkom opisu slavenskih jezika – na primjeru češkoga i hrvatskoga jezika. Ph.D. dissertation, Faculty of Humanities and Social Sciences, Univ. of Zagreb, Zagreb; 2007.
[29] Džeroski, S.; Erjavec, T.; Ledinek, N.; Pajas, P.; Žabokrtsky, Z.; Žele, A. Towards a Slovene Dependency Treebank. In: Proceedings of the LREC2006, ELRA-ELDA, Genoa-Paris; 2006.
[30] Slovenian Dependency Treebank web page: http://nl.ijs.si/sdt/.
[31] Pala, K.; Ševeček, P. Valence českých sloves. In Sborník prací FFBU, Masarykova univerzita Brno, Brno; 1997.
[32] Pala, K. Semantic Annotation of (Czech) Corpus Texts. In Proceedings of the Second Workshop on Text, Speech and Dialogue. Springer Verlag, Berlin; 1999.
[33] Žabokrtský, Z. Valency Lexicon of Czech Verbs. Ph.D. dissertation, UFAL MFF, Charles Univ., Prague; 2005.
[34] D'Halleweyn, E.; Dewallef, E.; Beeken, J. A Platform for Dutch in Human Langue Technologies. In: Proceedings of the LREC2000, ELRA-ELDA, Athens-Paris; 2000.