Original scientific paper

Tagset Reductions in Morphosyntactic Tagging of Croatian Texts

Željko Agić [*], Marko Tadić [**], Zdravko Dovedan [*]
[*] Department of Information Sciences, [**] Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{zeljko.agic, marko.tadic, zdravko.dovedan}@ffzg.hr

Summary

Morphosyntactic tagging of Croatian texts is performed with stochastic taggers by using a language model built on a manually annotated corpus implementing the Multext East version 3 specifications for Croatian. Tagging accuracy in this framework is basically predefined, i.e. proportionally dependent of two things: the size of the training corpus and the number of different morphosyntactic tags encompassed by that corpus. Being that the 100 kw Croatia Weekly newspaper corpus by definition makes a rather small language model in terms of stochastic tagging of free domain texts, the paper presents an approach dealing with tagset reductions. Several meaningful subsets of the Croatian Multext-East version 3 morphosyntactic tagset specifications are created and applied on Croatian texts with the CroTag stochastic tagger, measuring overall tagging accuracy and F1-measures. Obtained results are discussed in terms of applying different reductions in different natural language processing systems and specific tasks defined by specific user requirements.

Keywords: morphosyntactic tagging, part-of-speech tagging, stochastic tagger, Multext East tagset, tagset reductions, Croatian language

Introduction

A typical usage cycle for a majority of stochastic morphosyntactic taggers found today consists of sequentially applying two procedures: the training procedure and tagging procedure. The training procedure takes a previously annotated training corpus of a certain language as input, which it derives into an output language model readable by the tagging procedure. The tagging procedure is fed afterwards with unseen sentences of that language and it uses the language model in order to assign the most probable tags to word forms in the input sentences. Types of these language models and assignment algorithms vary in state-of-the art solutions: from hidden Markov models (Brants 2000; Halácsy et al. 2007) and support vector machines (Giménez and Márquez 2004) to cyclic

dependency networks (Toutanova et al. 2003) and bidirectional perceptron learning (Shen et al. 2007). The tagging accuracy of these methods peeks between 96 and 98 percent on the task of tagging English. Due to such high scores on English, the morphosyntactic tagging task is often considered as a closed or resolved issue in the computational linguistics and natural language processing communities. However, when using these procedures in tagging languages other than English, namely highly inflectional languages such as Czech, Croatian, Slovene and other Slavic languages, the tagging accuracy decreases (cf. Agić et al. 2008a and 2008b) to a point from which the given task does not seem as resolved as it did from the viewpoint of English language.

There are basically two issues that emerge when focusing on Slavic languages rather than English: the size of available corpora and the size of the tagset. On one side, rich morphology demands a more complex tagset in order to describe all the morphosyntactic phenomena. For example, the Penn Treebank is tagged using only 36 morphosyntactic tags (or part-of-speech tags, as it is perhaps better suited in this case), while the experiment with tagging Croatian texts using the TnT tagger (Agić and Tadić, 2006) utilized around 900 different morphosyntactic tags out of the overall 1475 tags that occur in the Croatian Morphological Lexicon (Tadić and Fulgosi 2003, Tadić 2005). And on the other side, lesser spread languages such as Croatian usually do not have at their disposal the person-months required to develop large manually annotated corpora such as, e.g., the Prague Dependency Treebank (Böhmová et al. 2003) for Czech. Even though the 100 Mw Croatian National Corpus does exist (Tadić 2002; Tadić 2006), only its minor part, the Croatia Weekly 100 kw subcorpus was manually annotated with morphosyntactic tags in order to train and experiment with stochastic taggers.

There are basically two separate approaches to improving morphosyntactic tagging accuracy that can be found in the field today:

1. Combining various taggers with each other or with other available language resources and language processing tools. For example, (Rupnik et al. 2008) combines a hidden Markov model tagger with a support vector machine tagger in the task of tagging Slovene, while (Sjöbergh 2003) utilized seven different taggers that implemented six different stochastic tagging paradigms in order to raise overall tagging accuracy for Swedish. For Croatian, an approach with combining the existing hidden Markov model tagger CroTag and the Croatian Morphological Lexicon was undertaken (Agić et al. 2008b), based on the experience of the HunPos tagger of Hungarian texts (Halácsy et al. 2006 and 2007). These approaches are said to either create hybrid taggers – such is the case with CroTag and HunPos when coupled with inflectional lexica – or voting taggers, using additional stochastic for deciding on the best of outputs provided by different taggers, hoping for a divergence of those towards the actual solution. Voting taggers are considered to have an advantage over hybrid taggers when adaptability to various languages is re-

quired, while hybrid taggers are usually more finely tuned for tagging a single specific language.

2. Manipulating the language model. These approaches mainly focus on reducing the tagset to a size desirably comparable to that of the e.g. Penn Treebank in order to downgrade the tagging problem for a given rich morphology language to that of tagging English. Reducing the tagset targets the language model directly, as stochastic taggers are based on counting occurrences of tags in the training corpus: the lower the overall tag count, the finer grained their distributions in the resulting language model. Notable approaches include the so-called tiered tagging approach (Tufiş 1999, Tufiş and Dragomirescu 2004), which compresses or maps the actual tagset into a hidden layer of tags with which the tagging is performed. The real tags are afterwards restored from the hidden layer using a lexicon and a set of handwritten rules. The approach has been shown to work well with different tagging paradigms (cf. Ceauşu 2006). The idea of tiered tagging can be traced back to (Brants 1995), a similar approach that did not yield significant improvements over the baseline tagging accuracy, unlike the tiered tagging approach.

In hindsight, all of these approaches are strictly scientific and task-oriented, as they aspire towards the ideal solution of approaching 100% tagging accuracy for a given language (or any language) while using the full morphosyntactic tagset for that language. However, keeping in mind that morphosyntactic taggers are generally not utilized as standalone applications, but rather as one of many modules in assembling larger natural language processing systems such as named entity recognizers or document classifiers, it should be considered – and this is of special importance for processing languages with sparse language resources and tools, such as Croatian – when and how to reduce the complexity of the tagging task in terms of user- or system-specific requirements. This paper investigates the specific user-oriented approach in which the full morphosyntactic tagset used for tagging Croatian corpora is mapped or split into several meaningful subsets from which the prospective user can choose a language model that is best suited for a specific natural language processing task.

Further sections of the paper describe this generally set research plan in more detail, including short descriptions of the corpus and tagger used in the experiment, along with the setup of the experiment itself. Results are afterwards discussed along with future work plans in the ending section.

Experiment

In the task of reducing a full morphosyntactic tagset into subsets for tagging Croatian texts, three modules must be observed in more detail: the tagger, the corpus from which the language model of the tagger is constructed and finally the tagset itself. The first two modules – the tagger and the corpus – are thoroughly described in previous publications (Agić and Tadić 2006) and (Agić et

al. 2008b) and therefore we present them here in a short overview, focusing afterwards on the morphosyntactic tagset.

The Croatia Weekly 100 Kw manually tagged newspaper corpus (the CW100 corpus further in the text) consists of articles extracted from seven issues of the Croatia Weekly newspaper, which has been published from 1998 to 2000 by the Croatian Institute for Information and Culture (HIKZ). This 100 Kw corpus is a part of Croatian side of the Croatian-English Parallel Corpus (CW corpus) described in detail in (Tadić 2000). The CW100 corpus was pre-tagged using the Multext-East version 3 morphosyntactic specifications (Erjavec 2004) on top of the XCES corpus encoding standard. The whole CW corpus was in fact built in two separate processing stages, as described in (Tadić 2000): firstly, the raw text data was automatically converted into XML format and afterwards tokenized in order to be semi-automatically tagged using full Multext-East version 3 tagset by matching the CW100 corpus and the Croatian Morphological Lexicon (Tadić and Fulgosi 2003, Tadić 2005) at unigram level via the Croatian Lemmatization Server (http://hml.ffzg.hr). The corpus consists of exactly 118529 word forms in 4626 different sentences, tagged by 896 different morphosyntactic tags. Nouns make for a majority of corpus word forms (approximately 30%), followed by verbs (~15%) and adjectives (~12%) which is in fact a predictable distribution for a newspaper corpus.

CroTag is a hybrid tagger consisting of two modules: the second order hidden Markov model training and tagging module (often called the trigram tagger, even though hidden Markov model tagging and trigram tagging are not necessarily the same procedures) and the inflectional lexicon module for boosting the tagger accuracy on unknown word forms. Its description is given in (Agić et al. 2008b) and error analysis provided in (Agić et al. 2009). The tagger uses the second order Viterbi algorithm with beam search to do the actual tagging, while language model sparseness is handled by linear interpolation smoothing at model building time and suffix tries with successive abstraction at runtime, i.e. upon encountering unknown and unhandled word forms. Its accuracy is obviously input dependent as it is a stochastic tagger: it yields an overall accuracy score of approximately 85 percent on a test corpus containing approximately 15 percent unknown word forms. Accuracy rises when decreasing the number of unknown word forms to ~95% correctly assigned tags with ~5% unknown word forms. With such figures, CroTag can be considered a state-of-the-art morphosyntactic tagger.

As mentioned before, Croatian texts are tagged using morphosyntactic tags from the Multext-East version 3 tagset specification for Croatian. As described in detail in (Agić et al. 2009), the tagset is positional, with each of the positions inside tags representing a single morphosyntactic category using different alphabetical characters for denoting different category values. For example a tag Ncmsn would denote a {Noun, common, masculine, singular, nominative} token. Position zero always represents part of speech information (PoS), while

other tag positions represent morphosyntactic categories and their values belonging to this part of speech (MSD). Querying the database backend of the Croatian Lemmatization Server (Tadić 2005) revealed a total of 1475 different Multext-East v3 morphosyntactic tags that are currently instantiated from this tagset in the Croatian Morphological Lexicon, i.e. on approximately 110.000 different lemmas and more than 4 million corresponding word forms.

Table 1. Properties of reduced tagsets on the CW100 corpus

Reduction	Type	Number of tags
$subset_0$	Full Multext-East v3 tagset	896
$subset_1$	Removes all MSD information for all non-inflective parts of speech and numerals	800
$subset_2$	Removes all MSD information for all non-inflective parts ($subset_1$) of speech, numerals and verbs	739
$subset_3$	Uses $subset_2$ and removes all other MSD information except gender, number and case on nouns, pronouns and adjectives and type on nouns	243
$subset_4$	Uses $subset_3$ and removes information on case from all remaining MSD information	48
$subset_5$	Uses $subset_4$ and removes information on gender and number from remaining MSD information	15
$subset_6$	Part of speech information only	13

Now that the modules are presented, tagset reductions must be introduced. Each of the reductions made for this experiment introduces another tagset, i.e. a specific subset of the full Multext-East v3 for Croatian. Obvious enough, the subsets will always impose fewer tags on the corpus than the original tagset. They will be named as subseti, the subscript i indicating depth of the reduction: the higher the index, the stricter the reduction and fewer the number of tags in the subset. Overview of the reductions is given in table 1 and a more elaborate description follows the table.

The first reduction in the table is not a reduction at all: subset0 represents the full tagset and is provided as a reference point or baseline figure. Similar to that, subset6 is a trivial reduction in which all information except the one about the part of speech is discarded. The reductions that can be found in between these upper and lower bounds are designed considering two viewpoints: the error analysis for CroTag in (Agić et al. 2009) and some basic intuition on system- and user-requirements. Namely, the above-mentioned experiment found that approximately 85 percent of all tagging errors occur on nouns, adjectives, pronouns and verbs and that approximately 50 percent of these are, in fact, incorrect assignments of case values. Therefore, the subsets are constructed by first dropping all the information on morphosyntactic categories of non-inflective parts of speech and verbs, eliminating the noise and focusing the analysis on the most difficult categories of the most difficultly tagged inflective parts of speech:

adjectives, nouns and pronouns. In addition, type and degree are stripped from adjectives and type and person from pronouns. Furthermore, case is stripped from these three parts of speech in subset4 and gender and number in subset5, leaving only morphosyntactic category of type for nouns (reminder: a noun can be common or proper and type denotes this). A common guideline for these reductions, besides the error analysis, was – as mentioned before – intuition on user and system requirements. This basically means that amount of information carried by a morphosyntactic category was considered from an average user and system viewpoint. From this perspective, it could be argued that, for example, information on noun type (common or proper) encodes more information – and in addition, information that is more valuable to the natural language processing system or its user – than information on noun case (nominative, genitive, etc.). As an illustration of this argument, consider a named entity detection and classification (NERC) system such as (Bekavac and Tadić 2007). In order to implement a normalization feature that would normalize various types of named entities occurring in the text to their normal (singular, nominative) form, one would require a morphosyntactic tagger able to correctly discriminate between common and proper nouns and male and female gender than e.g. between cases of adjectives and pronouns. Otherwise, the user might end up with a system that would normalize the entity Ive Sanadera as Iva Sanader (female) rather than the obvious choice Ivo Sanader (male) for example. Avoiding or encountering such an error in this framework depends exclusively on morphosyntactic tagging module and hence the intuition that led to these specific tagset reductions.

The data in table 1 is self-explanatory. However, it is rather interesting to note that maintaining gender, number and case for adjectives, nouns and pronouns and type for nouns and removing all other information from the tags induces a serious drop in the number of tags from subset2 to subset3. Removing case information expectedly reflects in overall tag numbers roughly as division of subset3 cardinality by seven as there are seven distinct cases in the Croatian language. The gaps in tag-space between subset2 and subset3 and also subset3 and subset4 should by all means be noted as they indicate there are many other options than only these presented in this paper. All of them should be considered for detailed sub-tagset design on basis of specific user or system requirements.

The experiment setup was also taken from the (Agić et al. 2009) experiment with CroTag error analysis. More specifically, the CW100 corpus is split into ten different parts, equal in number of sentences contained. Nine parts are used for creating the language model for the tagger and the tenth is always used for validating that model. The training sets had ca 106.676 tokens on average (average 23.426 types), while the testing sets had average 11.852 tokens (average 4.638 types). All counts and results are tenfold cross-validated. This procedure is repeated for each of the reduced tagsets subseti. Overall tagging accuracy is provided for the subsets and separate F1-measures are given on adjectives,

nouns and pronouns, i.e. the most difficult parts of speech for tagging Croatian texts. The following section provides experiment results and discussion.

Discussions of results

The results of the experiment are presented in condensed form by tables 2 and 3. Table 2 provides information on overall tagging accuracy achieved by the CroTag tagger on all the tagset reductions. For each of these subsets, the tagger was first trained on 90 percent of the CW100 corpus – the full tagset of the corpus reduced beforehand, corresponding to the subset in question – and then tested on the remaining 10 percent. The procedure was repeated ten times for each of the subsets, i.e. it was tenfold cross-validated. In the table, overall accuracy is given as a function of the number of different morphosyntactic tags found in each subset (see table 1). The tagging accuracy itself is presented by stating the average accuracies for each of the reduced tagsets, followed by their 95 percent confidence intervals. The table is accompanied by a simple histogram in figure 1 in order to indicate the functional dependency between the number of tags and overall tagging accuracy.

Table 2. Overall tagging accuracy with reduced tagsets

Reduction	Number of tags	Accuracy
$subset_0$	896	84.80±1.62
$subset_1$	800	85.35±1.86
$subset_2$	739	85.77±1.76
$subset_3$	243	86.18±1.94
$subset_4$	48	90.35±1.69
$subset_5$	15	96.02±1.00
$subset_6$	13	96.23±0.97

Both table and figure indicate an expected behaviour of the stochastic tagger: accuracy steadily rises with the decrease of the tagset size. More precisely, this dependency is expected due to the sparseness issue in the contextual probability matrices of second order hidden Markov model taggers (cf. Agić et al. 2008a). However, with respect to goals of this experiment, it should be noted that the decrease in tagset size gained when moving from subset2 to subset3 – amounting to a difference of 496 morphosyntactic tags – is shown here to provide only a slight gain of 0.41 percent in tagging accuracy while dropping 195 tags when moving from subset3 to subset4 caused the tagger to be a substantial 4.17 percent more accurate. Moreover, moving from subset4 to subset5, thereby dropping 33 tags also resulted in a substantial accuracy increase of 5.67 percent (i.e. accurately tagging 1 or 2 more word forms in a sentence with 25 word forms!), indicating that the stochastic tagger gains more accuracy when decreasing in the region of smaller tagsets. Therefore, tagset design should be approached with

caution between these margins when keeping in mind overall goals of specific natural language processing system design.

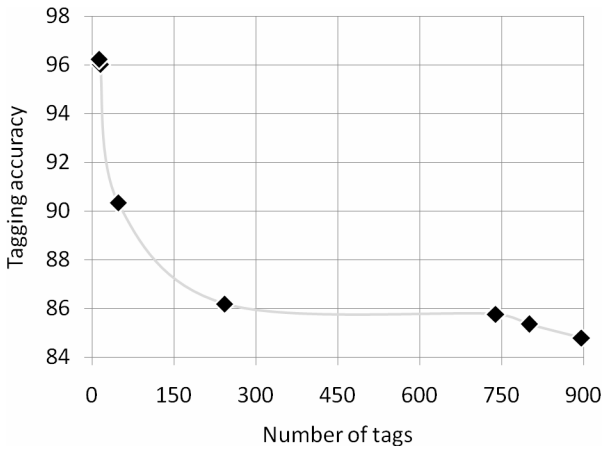

Figure 1. Tagging accuracy as a function of tagset size

Table 3 provides F1-measures on the most difficultly tagged parts of speech in Croatian: adjectives, nouns and pronouns. Recall and precision are left out of the table for conciseness and also because they were so narrowly tied with each other, thus rendering them uninteresting.

Table 3. F1-measures on adjectives, nouns and pronouns

	$subset_0$	$subset_1$	$subset_2$	$subset_3$	$subset_4$	$subset_5$	$subset_6$
Adj	0.64±0.04	0.63±0.04	0.63±0.04	0.65±0.05	0.74±0.05	0.92±0.02	0.91±0.03
Noun	0.79±0.03	0.78±0.03	0.78±0.04	0.78±0.04	0.86±0.03	0.95±0.01	0.97±0.01
Pro	0.76±0.03	0.75±0.04	0.75±0.05	0.76±0.05	0.87±0.04	0.99±0.01	0.99±0.01

As in previous experiments with tagging Croatian texts, adjectives are shown to be the most difficult of Croatian parts of speech, followed by pronouns and nouns. As with the previous table, notable accuracy increases can be seen between subset3 and subset4 and also subset4 and subset5 on all three parts of speech. Consulting the descriptions of reductions in table 1, it is clear that the first increase occurs when these parts of speech are stripped of the category of case, shown in (Agić et al. 2009) to be the most difficultly tagged category in Croatian. The other increase occurs when subset5 virtually becomes a part-of-speech-only tagset, removing information on gender and number and keeping only the type of nouns.

Conclusions and future work

Using the CroTag stochastic morphosyntactic tagger and the Croatia Weekly 100 kw manually tagged corpus of Croatian, this experiment has shown how tagset design or, more specifically, tagset reductions influence the accuracy of morphosyntactic tagging of Croatian texts. Its results may be used in other, more elaborate sub-tagset designs based on the Multext-East version 3 tagset specifications, with respect to overall goals of the resulting system and the requirements of the end user.

Acknowledgements

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-1776 and 130-1300646-0645.

References

Agić, Željko; Tadić, Marko. Evaluating Morphosyntactic Tagging of Croatian Texts // *Proceedings of the 5th International Conference on Language Resources and Evaluation* / ELRA, Genoa-Paris, 2006.

Agić, Željko; Tadić, Marko; Dovedan, Zdravko. Investigating Language Independence in HMM PoS/MSD-Tagging // *Proceedings of the 30th International Conference on Information Technology Interfaces* / Lužar-Stiffler, Vesna; Hljuz Dobrić, Vesna; Bekić, Zoran (ed.). Zagreb, SRCE University Computer Centre, University of Zagreb, 2008. pp 657-662.

Agić, Željko; Tadić, Marko; Dovedan, Zdravko. Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. // *Informatica*. 32 (2008), 4; pp. 445-451.

Agić, Željko; Tadić, Marko; Dovedan, Zdravko. Error Analysis in Croatian Morphosyntactic Tagging // *Proceedings of the 31st International Conference on Information Technology Interfaces* / Lužar-Stiffler, Vesna; Jarec, Iva; Bekić, Zoran (ed.). Zagreb, SRCE University Computer Centre, University of Zagreb, 2009. pp. 521-526.

Bekavac, Božo; Tadić, Marko. Implementation of Croatian NERC system // *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007, Special Theme: Information Extraction and Enabling Technologies* / Piskorski, Jakub; Tanev, Hristo; Pouliquen, Bruno; Steinberger, Ralf (ed.). Prague, ACL, 2007. pp. 11-18.

Böhmová, A.; Hajič, J.; Hajičova, E.; Hladká, B. The Prague Dependency Treebank: A Three-Level Annotation Scenario. // *Treebanks: Building and Using Parsed Corpora* / A. Abeillé (ed.), Kluwer, 2003, pp. 103-127.

Brants, Thorsten. Tagset Reduction Without Information Loss. // *Proceedings of ACL-95 student session* / Association for Computational Linguistics, 1995. pp. 287-289.

Brants, Thorsten. TnT - a statistical part-of-speech tagger // *Proceedings of ANLP 2000*.

Ceaușu, Alexandru. Maximum Entropy Tiered Tagging // *Proceedings of the 11th ESSLLI Student Session* / Janneke Huitink and Sophia Katrenko (ed.), June 20, 2006, Malaga, Spain, pp. 173-179.

Erjavec, Tomaž. Multext-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora // *Proceedings of the Fourth International Conference on Language Resources and Evaluation* / ELRA, Lisbon-Paris 2004, pp. 1535-1538.

Giménez, J.; Márquez, L. SVMTool: A general POS tagger generator based on Support Vector Machines // *Proceedings of the 4th International Conference on Language Resources and Evaluation* / Lisbon, Portugal, 2004.

Halácsy, P., Kornai, A., Oravecz, C., Trón, V., Varga, D. Using a morphological analyzer in high precision POS tagging of Hungarian. // *Proceedings of 5th Conference on Language Resources and Evaluation* / ELRA, 2006, pp. 2245-2248.

Halácsy, P., Kornai, A., Oravecz, C. HunPos - an open source trigram tagger // *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* / Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 209-212.

Rupnik, Jan; Grčar, Miha; Erjavec, Tomaž. Improving morphosyntactic tagging of Slovene by tagger combination. // *Proceedings of the Slovenian KDD conference – SiKDD 2008.* / Ljubljana, Slovenia, 2008.

Shen, L.; Satta, G.; Joshi, A. Guided learning for bidirectional sequence classification. // *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* / Prague, Czech Republic, 2007. pp. 760-767.

Sjöbergh, J. Combining POS-taggers for improved accuracy on Swedish text // *NoDaLiDa 2003, 14th Nordic Conference on Computational Linguistics*. / Reykjavik, 2003.

Spoustová, D., Hajič, J., Votrubec, J., Krbec, P., Květoň, P. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech // *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*. / Prague, Czech Republic. Association for Computational Linguistics, 2007.

Toutanova, K.; Klein, D.; Manning, C.D.; Yoram Singer, Y. Feature-rich part-of-speech tagging with a cyclic dependency network // *Proceedings of HLT-NAACL 2003* / pp. 252-259.

Tadić, Marko. Building the Croatian-English Parallel Corpus // *Proceedings of the Second International Conference on Language Resources and Evaluation* / ELRA, Paris-Athens 2000, pp. 523-530.

Tadić, Marko. Building the Croatian National Corpus. // *Proceedings of LREC 2002* / ELRA, Pariz-Las Palmas 2002, Vol. II, str. 441-446.

Tadić, Marko; Fulgosi, Sanja. Building the Croatian Morphological Lexicon // *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages* / Budapest, ACL, 2003. pp. 41-46.

Tadić, Marko. The Croatian Lemmatization Server // *Southern Journal of Linguistics*. 29 (2005), 1/2; pp. 206-217.

Tadić, Marko. Developing the Croatian National Corpus and Beyond // *Contributions to the Science of Text and Language. Word Length Studies and Related Issues* / Grzybek Peter (ur.), Kluwer, Dordrecht 2006, str. 295-300.

Tufiş, Dan. Tiered Tagging and Combined Classifiers // *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692* / F. Jelinek, E. Nöth (eds.), Springer, 1999, pp. 28-33.

Tufiş, Dan, Dragomirescu, Liviu. Tiered Tagging Revisited. // *Proceedings of the 4th LREC Conference* / Lisbon, Portugal, 2004, pp. 39-42.