Improving Chunking Accuracy on Croatian Texts by Morphosyntactic Tagging

Kristina Vučković*, Željko Agić*, Marko Tadić**

*Department of Information Sciences
**Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{kvuckovi, zeljko.agic, marko.tadic}@ffzg.hr

Abstract

In this paper, we present the results of an experiment with utilizing a stochastic morphosyntactic tagger as a pre-processing module of a rule-based chunker and partial parser for Croatian in order to raise its overall chunking and partial parsing accuracy on Croatian texts. In order to conduct the experiment, we have manually chunked and partially parsed 459 sentences from the Croatia Weekly 100 kw newspaper sub-corpus taken from the Croatian National Corpus, that were previously also morphosyntactically disambiguated and lemmatized. Due to the lack of resources of this type, these sentences were designated as a temporary chunking and partial parsing gold standard for Croatian. We have then evaluated the chunker and partial parser in three different scenarios: (1) chunking previously morphosyntactically untagged text, (2) chunking text that was tagged using the stochastic morphosyntactic tagger for Croatian and (3) chunking manually tagged text. The obtained F1-scores for the three scenarios were, respectively, 0.875 (P: 0.826, R: 0.930), 0.900 (P: 0.866, R: 0.937) and 0.930 (P: 0.912, R: 0.949). The paper provides the description of language resources and tools used in the experiment, its setup and discussion of results and perspectives for future work.

1. Introduction

Implementing procedures for automatic processing of morphology and syntax are seen today as one of the most important milestones in enabling advanced language technologies for any language (cf. Krauwer, 2003). This is mainly due to the fact that:

- language processing modules such as current state-of-the-art morphosyntactic taggers or partial and deep parsers (cf. ACLWiki, 2010) make possible the fast creation of large quantities of annotated language resources without imposing high demands on manual annotators and
- these modules are very efficient pre-processing tools in large-scale natural language processing systems as their demands on processing time and space are small when compared to more complex systems.

These facts are especially true for languages with less complex morphology and syntax, such as English, where tasks like morphosyntactic tagging and parsing are today considered as more or less resolved issues. However, morphologically and syntactically complex languages – such as Slavic languages or, more specifically, Croatian language – still pose a challenge, in terms of both accuracy and overall system efficiency, even in these basic tasks (cf. Agić et al., 2009; Buchholz and Marsi, 2006; Nivre et al., 2007). In addition, pipelining natural language processing tools has shown to be a well-investigated and straight-forward way to increase the performance in many basic tasks such as morphosyntactic tagging and parsing.

With an overall goal of enabling advanced language technologies for Croatian language (cf. Dalbelo Bašić et al., 2007) we have developed, among other modules, a stochastic morphosyntactic tagger (cf. Agić et al., 2008) and a rule-based chunker and partial parser (Vučković et al., 2008; Vučković, 2009). In this paper, we present results of an experiment in the attempt to improve the performance of the chunker by using a morphosyntactic tagger as a pre-processing module for the chunker.

Development and improvement of parsers brought a belief that the role of morphosyntactic taggers as pre-processing tools is weakening (cf. Charniak et al., 1996; Charniak, 1997) and that modern parsers do not really benefit from pre-tagging the input, or at least not substantially. However, in our opinion, this claim may be shown to be valid mainly for languages with less complex morphology and syntax and not for morphologically complex languages. From this specific perspective, our experiment was also targeted to show whether such a claim is applicable to Croatian and whether pre-tagging of the input text provide better chunking results when compared to chunking previously untagged text. Other related work might include approaches such as (Pla et al., 2000), (Nasr and Volanschi, 2006) and (Domínguez and Infante-Lopez, 2008).

The following chapter of the paper present**s** the setup of the experiment or, namely, the language resources and tools we used in conducting it. Further, we discuss the obtained results and conclude by stating an outline for our future work plans.

2. The experiment

In this chapter, we provide a brief description of language resources and tools used in the experiment, along with the experiment framework.

2.1 Morphosyntactic tagger

The CroTag morphosyntactic tagger (cf. Agić et al., 2008) is a second order hidden Markov model tagger that implements a linear interpolated trigram contextual model, unigram word-tag probabilities, suffix tries and successive abstraction in combination with some simple regular expression matching for handling unknown words. It is a state-of-the-art stochastic tagger in terms of overall accuracy and F1-measures on difficultly tagged parts of speech, its results virtually identical to those of TnT (Brants, 2000) and HunPos (Halacsy et al., 2007). Accuracy of CroTag on Croatian texts is raised on difficultly tagged parts of speech – namely adjectives, nouns and pronouns – by combining it with the Croatian Morphological Lexicon, an inflectional lexicon of Croatian, serving as an underlying resource for the Croatian Lemmatization Server (Tadić, 2005). A detailed investigation of the CroTag tagger accuracy and error footprints is given in (Agić et al., 2009a). The accuracy of the tagger on the experimental data used in this experiment amounts to ca 86% of accurately assigned morphosyntactic tags (cf. Agić et al., 2008).

2.2 Chunker and partial parser

For developing the chunker, we used NooJ (Silberztein, 2004; Silberztein, 2005; Silberztein, 2006) as a tool for natural language processing that uses formalized descriptions of inflectional and derivational morphology, lexicon, regular grammars and CF grammars. NooJ utilizes electronic lexicons and grammars represented by organized sets of graphs. It integrates morphology and syntax thus enabling morphological operations inside the syntactic grammars. Using morphological and syntactic formal descriptions in NooJ, it is possible to insert or delete additional annotations on different linguistic levels. NooJ uses FSTs, RTNs, ERTNs, CFGs and regular expressions as underlying technologies. In NooJ, grammars could be defined in several ways: writing regular expressions or using graphical interface for drawing the grammar graphs. System then interprets the graphical representation and converts it into an automaton. Cascading grammars and invoking grammars from within each other are completely supported thus leading to a powerful and yet user-friendly development environment. The way in which NooJ's enhanced grammar uses internal variables for storing parts of the recognized sequences in order to use them for constraining the output, greatly increases the functionality of this tool. NooJ not only lets you use derivational and inflectional morphology engine for processing variables' content but also retrieves and extracts values of a variable's property associated with its content. For these reasons, we have chosen NooJ as our development platform for building local grammars that function as a chunker for Croatian (Vučković et al., 2008).

Two separate models for chunk detection were built. The one for detection of pre-tagged text will further in the text be referred as the pre-tagged model and the one for untagged text as the untagged model. Syntactic grammars that both models use are finite state transducers applied to the text in a cascaded manner and are based on the modularity of local grammars.

The pre-tagged model has only two syntactic grammars, first of which recognizes NP and VP chunks. The second grammar recognizes PP, AP and AT chunks. The untagged model, on the other hand, has 13 syntactic grammars or, more precisely, implements 11 additional syntactic grammars prior to the last two, that are identical to those two of the pre-tagged model. These eleven grammars are used as a tool for of morphosyntactic disambiguation.

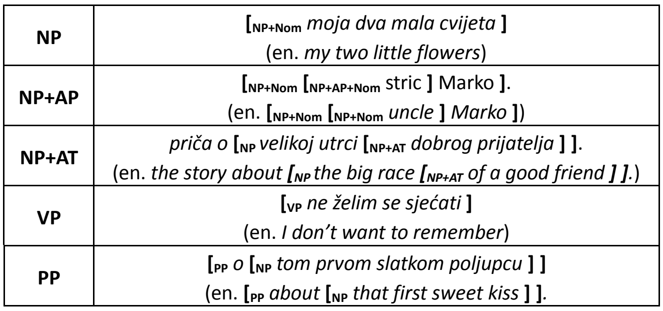
Since the number and type of chunks depends mainly on the language being processed (cf. Abney, 1996), we have defined five types of chunks for Croatian.

Noun phrase (NP) can be simple NP or complex NP (coordination). The simple NP consists of one main noun and any number of pronouns, adjectives and numerals preceding it if they agree in number, gender and case. In the case where the personal name consisting of first and last name (names) is present, both (all) nouns that the name consists of represent the head of a noun phrase. In the cases where there is no main noun, the head of the noun phrase is the last adjective in the chunk or personal, demonstrative, interrogative or indefinite pronoun if standing alone. The complex NP consists of any number of simple NPs if they all agree in case and are separated by a comma except the last two that are connected with any word from the set {*i*, *ili*, *ni*, *niti*, *te*}.

As special types of noun phrases we use apposition and adverb phrases to describe the following occurrences in the language. Apposition phrase (NP+AP) includes two noun phrases first of which is an apposition to the second one but only if they agree in number and case. Attribute phrase (NP+AT) includes at least two NPs where the following NP is an attribute to the NP that it immediately follows and if the second NP is in genitive case.

Verb phrase (VP) has one main verb as a head of the phrase and any of the following additions in any order: one or two auxiliary verbs (depending on the tense), reflexive pronoun *se* if the main verb is reflexive, negation *ne* and one infinitive form of the verb. If there is a one-word adverb inside the VP chunk it is recognized as a part of that VP.

Prepositional phrase (PP) consists of one preposition as a head of a chunk and an NP, NP+AP or NP+AT chunk that follows it and with which it agrees in case. It is important to note that prepositions in Croatian language do not have cases per se. However, there is a strict rule of what preposition can precede a noun phrase concerning the case of the noun phrase. Thus, all the prepositions are additionally marked with that case in the main dictionary.



NP	[$_{\text{NP+Nom}}$ *moja dva mala cvijeta*] (en. *my two little flowers*)
NP+AP	[$_{\text{NP+Nom}}$ [$_{\text{NP+AP+Nom}}$ stric] Marko]. (en. [$_{\text{NP+Nom}}$ [$_{\text{NP+Nom}}$ *uncle*] *Marko*])
NP+AT	*priča o* [$_{\text{NP}}$ *velikoj utrci* [$_{\text{NP+AT}}$ *dobrog prijatelja*]]. (en. *the story about [$_{\text{NP}}$ the big race [$_{\text{NP+AT}}$ of a good friend]].*)
VP	[$_{\text{VP}}$ *ne želim se sjećati*] (en. *I don't want to remember*)
PP	[$_{\text{PP}}$ *o* [$_{\text{NP}}$ *tom prvom slatkom poljupcu*]] (en. [$_{\text{PP}}$ *about* [$_{\text{NP}}$ *that first sweet kiss*]].

Figure 1 Examples of phrases

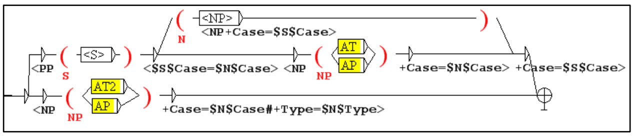


Figure 2 Local grammar sample from the chunker

Figure 1 provides samples for the five types of chunks we implemented in the chunker, while figure 2 gives an illustration of one of the local grammars from the chunker cascade. This specific local grammar is used to recognize PP chunks, attributes and appositions in both the pre-tagged and the untagged model. The yellow colored nodes are also local grammars, illustrating the cascaded design paradigm of the system. The red brackets mark variables, the content of which is checked for case agreement between the main noun of an NP and a preposition inside the PP chunk.

Previously conducted manual evaluation (Vučković et al., 2008) of the chunker provided F1-scores of 0.92, 0.83 and 0.97 on NP, PP and VP chunks, respectively. It was conducted on a corpus of 137 sentences, containing 1150 different NP chunks, 348 PP chunks and 447 VP chunks. Other than sentence segmentation and tokenization, no pre-processing was done with the input sentences. For a detailed description of this experiment, see (Vučković et al., 2008).

2.3 Corpus

The Croatia Weekly 100 kw newspaper corpus (the CW100 corpus henceforth) 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). The CW100 corpus is a part of Croatian side of the Croatian-English Parallel Corpus described in detail in (Tadić, 2000). The CW100 corpus was pre-tagged using the Multext-East v3 morphosyntactic specifications (Erjavec, 2004) on the top of XCES corpus encoding standard. The whole CW100 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 the full MTE v3 tagset by matching the CW100 corpus and the Croatian Morphological Lexicon at unigram level via the Croatian Lemmatization Server (Tadić, 2005; http://hml.ffzg.hr). After that all possible MSD interpretations were manually corrected and only the appropriate one was left in the corpus. Some corpus stats are provided in table 1.

Sentences	4.626
Tokens	118.529
Word forms	103.161
Other tokens	15.368
Different MSD tags	896

Table 1 CW100 corpus stats

Sentences	459
Tokens	10.131
Chunks	5.513
NP chunks	3.332
VP chunks	1.182
PP chunks	999

Table 2 Chunking Gold standard stats

Distribution of parts of speech in the corpus is as expected on a newspaper corpus. Common newspaper texts are written in plain Croatian and for news-reporting purposes so most sentences comply with the relatively simple subject-verb-object model and therefore nouns, verbs and adjectives dominate the distribution. More details on the CW100 corpus can be found in e.g. (Agić and Tadić, 2006).

The gold standard used for this experiment was taken from the CW100 corpus. Actually, the entire CW100 corpus, being previously annotated on various levels and thus very suitable for evaluation of tools for processing Croatian, was designated to become a gold standard for parsing as well. However, at the time of conducting this experiment, 459 of the 4.626 sentences (roughly 10%) of the CW100 were manually chunked so we had no choice but to use these sentences in the evaluation. Some stats regarding the gold standard are given in table 2 and table 3. An interesting side-note regarding the data not displayed in table 2 is that, out of the 3.332 noun phrase chunks, 67 were found to be in the role of noun apposition and 504 were noun attributes. Table 3 shows that a majority of NP chunks were in the nominative case (~32%), followed by genitive (~27%), accusative (~18%) and locative (~13%). On the other hand, preposition phrases followed an expectedly different distribution, dominated by locative (~42%), accusative (~24%) and genitive case (~23%). It should be noted that the high number of occurrences of NP chunks in locative case is due to NPs wrapped within preposition phrases.

Table 4 provides additional data for noun phrase chunks. Namely, it focuses on the spread across different cases for noun phrases that were annotated as attributes or appositions.

Case	NP chunks	PP chunks
Nominative	1.075	11
Genitive	908	229
Dative	112	5
Accusative	586	238
Vocative	2	0
Locative	436	421
Instrumental	189	89
Not assigned	24	6

Table 3 Distribution of cases on NP and PP chunks

Case	NP attribute	NP apposition
Nominative	0	38
Genitive	496	12
Dative	4	7
Accusative	2	3
Vocative	0	0
Locative	0	1
Instrumental	0	6
Not assigned	2	0

Table 4 Distribution of attributes and appositions

The table expectedly indicates that a large majority of noun attributes (98.80%) was in genitive case, while appositions were mostly found in nominative (~55%) and genitive (~20%), followed by dative and instrumental.

2.4 Experiment setup

The experiment was conducted as follows. The 459 sentences of the gold standard were stripped of the XML annotation and written to a file. Three copies of this file were created. The first one was forwarded to the chunker without preprocessing. The second one was paired with the morphosyntactic annotation and lemmatization that was done earlier for the entire CW100 corpus in the manner already described above. Finally, the third one was first forwarded to the CroTag tagger and afterwards to the chunker, thus carrying morphosyntactic annotations that were assigned by the tagger. Precision, recall and F1-scores were then calculated overall and separately for NP, PP and VP chunks. For NP and PP chunks, the scores were also calculated for each case.

The following section provides a discussion of the results. However, before proceeding, another important remark should be noted. Namely, being that the chunker module also operates as a partial parser – assigning dependent NPs within PPs or other NPs – a notion of partial parsing evaluation was also included in the experiment. Although the emphasis of the evaluation and the entire experiment is placed on chunking Croatian texts, the partial parsing procedure it is also within the focus of our discussion. For a brief discussion regarding more advanced steps in parsing Croatian texts, see future work.

3. Results and discussion

The presentation of the results is distributed across the following three tables in a manner ranging from general to specific observations regarding the performance of our chunker and partial parser when (and when not) combined with the CroTag morphosyntactic tagger.

Table 5 presents overall scores (recall, precision and F1-measure) of the chunker on noun phrases (NP), preposition phrases (PP), verb phrases (VP) and overall, with the chunker running in three different modes: with untagged text, CroTag-tagged text and manually tagged text provided as input. With the exception of preposition phrases (see table 6 and comment), the table shows a very obvious and consistent increase in F1-scores when moving from untagged to statistically and manually tagged input text. Overall scores indicate that this increase in overall F1-score is achieved by consistently raising both precision and recall.

		Untagged	CroTag	Manual
NP	P	0.789	0.816	**0.877**
	R	0.954	**0.965**	0.953
	F_1	0.864	0.884	**0.913**
PP	P	0.952	0.955	**0.958**
	R	0.898	0.827	**0.912**
	F_1	0.924	0.886	**0.934**
VP	P	0.824	0.929	**0.970**
	R	0.892	0.952	**0.970**
	F_1	0.857	0.941	**0.970**
All	P	0.826	0.866	**0.912**
	R	0.930	0.937	**0.949**
	F_1	0.875	0.900	**0.930**

Table 5 Chunk assignment accuracy

Regarding the overall scores, both precision and recall are thus the lowest on untagged input text, with both of them steadily increasing proportionally to the quality of tagging, benefiting from resolved ambiguity of the input. An overall F1-score for chunking of 0.930 is achieved when using manually annotated input text, while a score of 0.900 is obtained by using a morphosyntactic tagger. From a system design point of view, the second score (chunking with stochastic tagging) is somewhat more interesting as it is directly applicable to real-world tasks and systems.

Table 6 increases the difficulty of the chunking tasks by demanding that the module also assigns a correct case to the recognized phrases. The first set of numbers to observe is the one indicating a steep decrease in noun phrase chunking accuracy for the chunking module that operates on untagged input. This is due to the fact that the module assigns all possible interpretations (i.e. all seven Croatian cases) to the ambiguous phrases. In this phase of the evaluation, we chose to consider the ambiguous output of this module as incorrect, with regards to overall applicability of the combined module. We were guided by the general argument for robust disambiguation by the parsing modules given in (Nivre, 2006). Aside from this decrease in NP-chunking quality for untagged text, other figures are consistent with the ones in table 5. It should be noted that the accuracy on PPs does not follow the pattern of the one for NPs, as cases of PPs are more easily hard-coded to rules (see table 3). Furthermore, the module does not even benefit from stochastic tagging for PPs, as the CroTag tagger is shown to erroneously assign these specific cases – especially locative and instrumental – somewhat more frequently (cf. Agić et al., 2009a).

		Untagged	CroTag	Manual
NP	P	0.271	0.613	**0.780**
	R	0.327	0.724	**0.848**
	F_1	0.297	0.664	**0.813**
PP	P	0.902	0.933	**0.938**
	R	0.851	0.807	**0.893**
	F_1	0.876	0.866	**0.915**

Table 6 Chunking accuracy with case assignment

Level	Untagged	CroTag	Manual
1	0.797	0.833	**0.874**
2	**0.754**	0.671	0.722
3	0.605	0.612	**0.636**

Table 7 F1-scores for partial parsing of NPs with case

Table 7 deals with the task of partial parsing. Namely, as previously stated, besides detecting disjoint surface phrases (NP, PP, VP), the chunker also assigns certain dependent noun phrases to the higher level NPs and PPs. This feature is implemented to the depth of five layers, i.e. the layer of the chunk and four cascades below. In table 7, F1-scores are given for NPs in the first three annotation layers following the chunk layer, i.e. layer zero. The last two layers are excluded from the table because of the decreasing distribution of NP counts per layer, i.e. because the data was too sparse. Data in the table reveals the accuracy of noun phrase partial parsing to be reversely proportional to the depth of the layer.

4. Conclusions and future work

In this paper, we have presented the results of an experiment with pipelining a stochastic morphosyntactic tagger CroTag (cf. Agić et al., 2008) with a chunker and shallow parser for Croatian (cf. Vučković et al., 2008). We have shown that the chunker benefits from using the disambiguation provided by the tagger and that this benefit – quantified in terms of precision, recall and F1-score – increases with the increase of morphosyntactic tagging accuracy. The results and conclusions provided in this paper might prove beneficial to other languages sharing properties, namely rich morphology and relatively free word order, with Croatian.

Future research of this matter will probably be spread along the following general guidelines.

Currently, the partial parser does not recognize appositions inside the NP chunk where the main NP chunk is not a proper name. Also, if an attribute is in a case other than genitive, which is a rare but possible phenomenon in Croatian, it is not recognized by the existing rules. Other patterns not recognized by the rules include embedded attributes and appositions, i.e. presence of both apposition and attribute inside an NP in the following manner: (1) one NP is an apposition to the second NP, both of which are an attribute to the third NP and (2) one NP is an attribute to the second NP, both of which are an apposition to the third NP. These problems will hopefully be resolved in future versions of the module.

Experiments with reducing the full Multext-East tagset for purposes of raising the accuracy of the CroTag tagger with regards to the requirements of the chunker could be conducted. Results of an experiment dealing with tagset reductions for CroTag are already available in (Agić et al., 2009b). Furthermore, an influence of the tagset and training set size of the tagger to the chunking and shallow parsing accuracy on Croatian texts should be investigated in more detail, in order to detect the points of cost and benefit for possible applications of this pipeline. Our gold standard corpus for chunker evaluation should also be expanded and possibly linked with the construction of the Croatian Dependency Treebank (cf. Tadić, 2007). Experiments similar to the one presented here should be conducted for state-of-the-art dependency parsers (cf. Buchholz and Marsi, 2006; Nivre et al., 2007) using the Croatian Dependency Treebank as a training corpus in the task of dependency parsing of Croatian.

5. Acknowledgements

The research within the project Accurat leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347. This work was also supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-1776 and 130-1300646-0645.

6. References

Abney, S. (1996). Chunk Stylebook, Retrieved February 20, 2006 from: http://www.vinartus.net/spa/96i.pdf, 1996.

ACL Wiki: State of the art. (2010). Available at URL http://aclweb.org/aclwiki/index.php?title=State_of_the_art (last accessed 2010-03-18).

Agić, Ž., Tadić, M. (2006). Evaluating Morphosyntactic Tagging of Croatian Texts. *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, ELRA, 2006.

Agić, Ž., Tadić, M., Dovedan, Z. (2008). Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. *Informatica* 32:4, 2008, pp. 445-451.

Agić, Ž., Tadić, M., Dovedan, Z. (2009). Error Analysis in Croatian Morphosyntactic Tagging. *Proceedings of the 31st International Conference on Information Technology Interfaces*. Zagreb, SRCE University Computer Centre, University of Zagreb, 2009. pp. 521-526.

Agić, Ž., Tadić, M., Dovedan, Z. (2009). Tagset Reductions in Morphosyntactic Tagging of Croatian Texts. *The Future of Information Sciences: Digital Resources and Knowledge Sharing*. University of Zagreb, pp. 289-298.

Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. Proceedings of the Sixth Conference on Applied Natural Language Processing. Seattle, Washington 2000.

Buchholz, S., Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. *Proceedings of the 10th Conference on Computational Natural Language Learning*. New York, NY, pp. 149-164.

Buitelaar, P., Declerck, T., Sacaleanu, B., Vintar, Š., Raileanu, D., Crispi, C. (2003). A Multi-Layered, XML-Based Approach to the Integration of Linguistic and Semantic Annotations. *Proceedings of the EACL2003 Conference, Workshop on NLP and XML Language Technology and the Semantic Web*. EACL, Budapest, 2003, pp. 9-16.

Charniak, E., Carroll , G., Adcock , J., Cassandra , A., Gotoh , Y., Katz , J., Littman , M., McCann, J. (1996). Taggers for Parsers. *Artificial Intelligence* 85:1-2, *Special Volume on Empirical Methods*. Elsevier Science Publishers Ltd, Essex, UK, pp. 45-57.

Charniak, E. (1997). Statistical Parsing with a Context-free Grammar and Word Statistics. *Proceedings of the 14th National Conference on Artificial Intelligence*. AAAI Press/MIT Press, 1997, pp. 598-603.

Dalbelo Bašić, B., Dovedan, Z., Raffaelli, I., Seljan, S., Tadić, M. (2007). Computational Linguistic Models and Language Technologies for Croatian. *Proceedings of the 29th International Conference on Information Technology Interfaces*. Zagreb, SRCE, 2007. pp. 521-528.

Domínguez, M. A., Infante-Lopez, G. (2008). Searching for Part of Speech Tags That Improve Parsing Models.

Lecture Notes In Artificial Intelligence 5221. *Proceedings of the 6th International Conference on Advances in Natural Language Processing (ANLP)*. Gothenburg, Sweden, pp. 126 - 137.

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

Halácsy, P., Kornai, A., Oravecz, C. (2007). 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.

Krauwer, S. (2003), The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. *Proceedings of SPECOM 2003*. Moscow, 2003.

Nasr, A., Volanschi, A. (2006). Integrating a POS Tagger and a Chunker Implemented as Weighted Finite State Machines. *Lecture Notes in Computer Science 4002 – Finite-State Methods in Natural Language Processing*. Springer, pp. 167-178.

Nivre, J. (2006). Inductive Dependency Parsing. *Text, Speech and Language Technology*, Vol. 34. Springer, Dordrecht, The Netherlands.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. *Proceedings of the CoNLL Shared Task of EMNLP-CoNLL 2007*. Prague, Czech Republic, pp. 915-932.

Pla, F., Molina, A., Prieto, N. (2000). Improving Chunking by Means of Lexical-Contextual Information in Statistical Language Models. *Annual Meeting of the ACL*. *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*, pp. 148-150.

Silberztein, M. (2004). NooJ: an Object-Oriented Approach. *INTEX pour la Linguistique et le Traitement Automatique des Langues*. Muller, C., Royauté, J., Silberztein, M. (eds), Cahiers de la MSH Ledoux. Presses Universitaires de Franche-Comté, pp. 359-369.

Silberztein, M. (2005). NooJ's Dictionaries. *Proceedings of the 2nd Language and Technology Conference*. Poznan University, 2005.

Silberztein, M. (2006). NooJ Manual. Available at URL http://www.nooj4nlp.net/NooJ%20Manual.pdf.

Tadić, M. (2000). 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, M. (2002). Building the Croatian National Corpus. *Proceedings of the Third International Conference on Language Resources and Evaluation*. ELRA, Paris-Las Palmas 2002, Vol. II, pp. 441-446.

Tadić, M., Fulgosi, S. (2003). Building the Croatian Morphological Lexicon. *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*. Budapest, ACL, 2003. pp. 41-46.

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

Tadić, M. (2006). Developing the Croatian National Corpus and Beyond. *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Kluwer, Dordrecht 2006, pp. 295-300.

Tadić M. (2007). Building the Croatian Dependency Treebank: the initial stages. *Suvremena lingvistika* 63, pp. 85-92.

Vučković, K., Tadić, M., Dovedan, Z. (2008). Rule Based Chunker for Croatian. *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech-Paris, ELRA, 2008.

Vučković, K. (2009). Model parsera za hrvatski jezik. PhD Thesis, Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, 2009.