Recent Advances in Intelligent Information Systems
ISBN 978-83-60434-59-8, pages 175–184

# Evaluating Full Lemmatization of Croatian Texts

Željko Agić[1], Marko Tadić[2], and Zdravko Dovedan[1]

[1] Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
zeljko.agic@ffzg.hr, zdravko.dovedan@ffzg.hr

[2] Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
marko.tadic@ffzg.hr

## Abstract

The paper presents the implementation and evaluation of a module for full lemmatization of Croatian texts. The module implements several lemmatization procedures, all of them based on merging outputs of the previously developed stochastic morphosyntactic tagger CroTag and the inflectional lexicon of Croatian. Evaluation of the lemmatization module on two test cases, simulating realistic and ideal operating conditions, provided full lemmatization accuracy scores of 96.96 and 98.15 percent on a newspaper corpus, respectively. It is also shown that a majority of errors in this framework, 57.14 percent in the realistic testing scenario, occur on word forms with external homography. Moreover, approximately 80 percent of all lemmatization errors occur on nouns, adjectives, verbs and adverbs in that particular order. Language resources, testing environment and procedure descriptions are provided in the paper along with a discussion of obtained results and possible future research directions.

**Keywords:** full lemmatization, morphosyntactic tagging, Croatian language

## 1 Introduction

Previous implementation and evaluation of both inflectional lexicon (Tadić and Fulgosi, 2003; Tadić, 2005) and CroTag (Agić and Tadić, 2006; Agić et al., 2008a,b), a state-of-the-art stochastic morphosyntactic tagger developed for tagging Croatian texts, has enabled development of a full lemmatization module based on simply combining outputs from the two components.

The basic idea supporting this scheme is that valid output of the tagger disambiguates the ambiguous – both in terms of lemmas and morphosyntactic tags – output of the inflectional lexicon in a sentence context. Disambiguation is achieved by selecting the lemma corresponding to the output tag, providing that the same tagset is implemented in both inflectional lexicon and language model of the tagger. An illustration is given in Table 1.

If a tagger would output correct tags only and an inflectional lexicon would exhibit a 100-percent coverage for a given language, the problem of full lemmatization could be easily resolved by the procedure illustrated in Table 1. However, even the most accurate stochastic taggers currently peak at 97-98 percent (cf. Brants, 2000; Gimenéz and Márquez, 2004; Shen et al., 2007; Toutanova et al., 2003) correctly assigned tags while the nature of natural language itself prevents inflectional lexicons from achieving complete language coverage.

TABLE 1: Tagger indicates correct lemma

| | |
|---:|:---|
| Input wordform | da |
| Tagger response | da **Css** |
| Inflectional lexicon | da da2 Qr dati Vmia2s dati Vmia3s dati Vmip3s **da1 Css** |
| Resulting lemma | da1 |

Therefore, additional heuristic procedures should be implemented so that the full lemmatization module in the presented paradigm could achieve satisfying accuracy and robustness on unrestricted texts. This is of special importance for texts written in Croatian, being inflectionally rich and relatively free order language like other Slavic languages.

Related work on morphosyntactic tagging, morphological analysis and lemmatization for other Slavic languages encompasses many research experiments. However, very few of them approach the problem of full lemmatization by sequentially running and then merging outputs of taggers and inflectional lexica. The reason for this in our case is most probably because of specifics of Croatian language resources and natural language processing tools development, having separately implemented first the Croatian inflectional lexicon and afterwards the CroTag tagger. Lemmatization of Croatian texts was also approached from a normalization perspective in (Šnajder et al., 2008), reporting peak lemmatization accuracy of 92.82 percent. Procedures in lemmatizing the Slovene language are also highly relevant for Croatian and several successfull approaches exist, utilizing both rule-based and machine learning techniques and their combinations (cf. Džeroski and Erjavec, 2000; Erjavec and Džeroski, 2004; Juršič et al., 2007). An approach similar to the one taken for Croatian can be found in (Halácsy et al., 2006), resulting in the development of HunPos trigram tagger (Halácsy et al., 2007), which in turn inspired the CroTag tagger.

Lemmatization procedures are presented in section 2, followed by test environment features such as corpus details, test cases, utilized tools which are covered in section 3. Results discussion, conclusions and future improvement plans are situated in section 4.

## 2 Lemmatizer

Using the lemmatization paradigm defined in the previous section, the lemmatizer could be basically regarded as a set of procedures for combining outputs of tagger and lexicon, implementing relatively simple merging rules for solving two basic problems of these modules: errors produced by the tagger and the lack of lexical coverage of the lexicon, i.e. missing lexical entries with regards to unrestricted corpora.

Two elementary courses of action were considered in this implementation, i.e. two sets of procedures for dealing with erroneous tags yielded by the CroTag tagger and insufficiencies of the Croatian inflectional lexicon. These sets and specific procedures they contain are described in the following subsections.

### 2.1 Baselines

In this subsection, baseline approaches to full lemmatization are described. They serve as an illustration of what can be achieved in terms of full lemmatization accuracy on Croatian texts without using full morphosyntactic disambiguation provided by the tagger. Also, as a consequence of cascaded fallbacks given in their descriptions, all three baselines are used as default fallback procedures in the merge procedures that utilize both the inflectional lexicon and the tagger.

`Baseline 1` is the simplest approach to full lemmatization taken in this experiment and arguably the simplest approach to full lemmatization in general. Here, outputs of the tagger and the inflectional lexicon are not even considered, as this naïve approach always assigns the wordform as the lemma. This approach serves only as a reference point for evaluating other lemmatization procedures, indicating what might be the worst possible performance of full lemmatization of Croatian texts.

`Baseline 2` also does not require a morphosyntactic tagger to operate, as it only deals with the unambiguous output of the inflectional lexicon. Here, a lemma is chosen from the output of the inflectional lexicon if and only if it is the only lemma the lexicon provided. Otherwise, the procedure falls back to `Baseline 1`, simply choosing the wordform as the lemma and signalling that in the output. Although this baseline is more refined than `Baseline 1`, namely by using the lexical coverage of the inflectional lexicon, even the basic intuition indicates how it might fail on practically every occurrence of external (or lexical) homographs in the text (this being particularly important for highly inflective languages such as Croatian).

`Baseline 3` is a naïve attempt in addressing the issue of lemmatizing lexical homographs without utilizing the disambiguation module, i.e. the CroTag morphosyntactic tagger. When the procedure encounters a wordform covered by more than one lemma in the inflectional lexicon, a single lemma is randomly chosen from this pool and assigned to the wordform. Otherwise, the procedure falls back to `Baseline 2` and possibly `Baseline 1`.

### 2.2 Merge procedures

The so-called merge procedures are defined by different approaches to combining or merging output of the inflectional lexicon and output of the morphosyntactic tagger for a given wordform into a single 3-tuple (wordform, lemma, morphosyntactic tag). Each of the three following merge procedures use `Baseline 3` as a default fallback option.

Merge 1 reduces Multext-East v3 (Erjavec, 2004) morphosyntactic tags to part of speech information only. The procedure iterates over lemmas provided by the inflectional lexicon, comparing their part-of-speech with the part of speech tag assigned to the wordform by the tagger. If the two parts of speech match, the corresponding lemma is assigned to the wordform. Otherwise, the default fallback procedure is called.

Merge 2 is a straightforward and easily implementable upgrade of Merge 1. Instead of matching part of speech information only, entire morphosyntactic tags are compared here. If the morphosyntactic tag provided by the tagger equals one of the morphosyntactic tags provided by the inflectional lexicon, the corresponding lemma is assigned to the wordform. Otherwise, the procedure falls back to Baseline 3. Although Merge 2 might appear to display an advantage over the previous procedures, both Merge 1 and the baselines, it actually introduces drawbacks because possible tagging errors are not accounted for. The method exclusively trusts statistical tagger over hand-made inflectional lexicon by default and, as such, it is expected to introduce noise. Note that this is significantly more relevant for Merge 2 than for Merge 1 being that errors in morphosyntactic tagging of Croatian texts using the full Multext-East v3 morphosyntactic tagset are much more likely to appear on more specific morphosyntactic features (such as gender, number and case of adjectives, nouns and pronouns) than on part of speech alone, as described in (Agić et al., 2009).

Merge 3 is implemented to account for problems raised by Merge 2. It is a simple tweak compensating for minor errors introduced by the tagger. It relies on a before-mentioned observation, stating that stochastic taggers are more likely to introduce errors in more specific morphosyntactic features rather than in part of speech alone. In positional morphosyntactic tagsets such as Multext-East v3 – used in the Croatian inflectional lexicon, the CroTag tagger language model and in this experiment – it would mean that errors are more likely to occur further away from the first letter of the morphosyntactic tag, which encodes part of speech information, as verified by (Agić et al., 2009). Therefore, Merge 3 removes the strict demand on equality defined in Merge 2 and replaces it with a demand on similarity. In other words, it looks up a list of (lemma, morphosyntactic tag) pairs given by the lexicon and chooses the lemma for which the corresponding morphosyntactic tag is the most similar to the tag provided by the tagger for a given wordform. This method obviously also prefers tagger over lexicon, but still considers and handles the possibility of tagger making an error. Once again, the default fallback for this procedure is Baseline 3.

## 3 Experiment

Beside the before-mentioned trigram tagger CroTag and the inflectional lexicon for Croatian, the Croatia Weekly 100 kw (CW100 further in the text) newspaper corpus was available and used in this experiment. CW100 is XCES-encoded, automatically matched with the Croatian inflectional lexicon at unigram level, afterwards manually disambiguated and made compliant with Multext-East v3 specifications.

It contains 118,529 tokens, with 103,161 of them being actual wordforms in 4,626 sentences and annotated using around 900 out of 1,475 different morphosyntactic tags found in the inflectional lexicon. Other corpus details are given in (Agić and Tadić, 2006). The CW100 corpus is currently the only manually annotated gold standard corpus available for experiments involving morphosyntactic tagging and lemmatization of Croatian texts. Therefore, even though bias might be placed here on basis of the corpus size and domain specificity, it should be duly noted that other resources of similar quality and reliability were unavailable at the time of conducting these experiments.

For experimental purposes, sentences of CW100 corpus were assigned into ten disjunct subsets, roughly equal in wordform counts, by random sentence sampling. The training sets had encompassed 10 percent or approximately 11,853 wordforms on average and were used in tenfold cross-validation of lemmatization procedures. The other 90 percent of sentences was used in training the CroTag tagger.

Two test scenarios were envisioned, relating to tagger accuracy on test sets – the realistic and the idealistic one.

In the realistic scenario, CroTag was trained on nine test sets and used for tagging the one remaining test set, i.e. the one not used by the training procedure. The realistic scenario allowed observations of full lemmatization accuracy when tagger encountered unknown wordforms and subsequently returned wrong tags relatively often.

The idealistic scenario considered a know-it-all tagger, trained on the entire CW100 corpus and utilized in tagging its subsets. This scenario ensured the highest possible tagging accuracy and enabled insight on what was expected to be the highest possible full lemmatization score, i.e. it indicated boundaries of this paradigm of full lemmatization and also properties of errors that could not be corrected by simply combining merging procedures. It subsequently requires additional work in developing other, more refined procedures. Therefore, realistic testing scenario served to indicate whether the `Baseline{1,2,3}` and `Merge{1,2,3}` procedures could be utilized in natural language processing systems for Croatian as they are, while the idealistic testing scenario was used to explore limitations of such combinations and possibilities of creating new ones. It should be noted that both testing frameworks included tenfold cross-validation for purposes of comparison, regardless of the different purposes of these two sets of experiments and the obviously higher importance of the realistic scenario in drawing general conclusions about full lemmatization of Croatian texts.

### 3.1 Realistic scenario

In the first of the test cases (a realistic one), all of the `Baseline` and `Merge` procedures are evaluated on all test sets and averaged in order to detect which one represents the best full lemmatizer for Croatian. Table 2 provides the results.

Results indicate that best choice for full lemmatization of Croatian is procedure `Merge 3`, utilizing occurrences of lexical unambiguity and falling back to morphosyntactic tag similarity stochastics when necessary. However, it should be noted that `Merge 1` and `Merge 3` differ in only 0.65 percent, implying that stochastic procedure implemented in `Merge 1` was able to disambiguate lemmas

Table 2: Lemmatization procedure accuracy

| Procedure | Base1 | Base2 | Base3 | Merge1 | Merge2 | Merge3 |
|---|---|---|---|---|---|---|
| Accuracy overall | 57.53% | 87.92% | 88.44% | 96.31% | 95.51% | 96.96% |
| Errors overall | 42.47% | 12.08% | 11.56% | 3.69% | 4.49% | 3.04% |

solely by means of part of speech equality, compensating for errors introduced by tagging. This is an important note with regards to properties of stochastic taggers, namely the increase of performance that is achieved by reducing the morphosyntactic tagset size. Given these observations, the other results were as expected: `Merge{1,2,3}` procedures outperformed `Baseline{1,2,3}` and procedure `Baseline 3` outperformed both `Baseline 1` and `Baseline 2`. It is also interesting to note how overall error rate reduces significantly from `Baseline 1` to `Baseline{2,3}` simply by choosing a lemma when there is only one lemma to choose anyway or by randomly choosing a lemma if there is more than one of them in the pool. One could argue that such a procedure is quite accurate and robust at the same time since it does not require a morphosyntactic tagger at all, contributing to overall speed and memory requirements of the module.

Table 3: Test environment details

| | |
|---|---|
| Wordforms overall | 11852.90 |
| Known to tagger | 9906.60 (83.58%) |
| Unknown to tagger | 1946.30 (16.42%) |
| Realistic tagger accuracy | 84.75% |
| Accuracy on known words | 88.68% |
| Accuracy on unknown words | 65.79% |
| Idealistic tagger accuracy | 98.76% |

Full lemmatization results are accompanied by Table 3, providing insight on the test environment in which the discussed accuracies were obtained. The tagger encountered 16.42 percent unknown wordforms among 11,853 wordforms per test case on average, resulting in an accuracy loss, i.e. a rather expected error rate of 15.25 percent, given the common properties of the trigram tagging paradigm (Agić et al., 2008c). For example, given an average Croatian example sentence counting 27 wordforms, the tagger would return an incorrect morphosyntactic tag for 4 wordforms and procedure `Merge 3` would still assign a wrong lemma to only one of these wordforms on average. These figures and this example indicate that the full lemmatization system implementing the `Merge 3` procedure could be utilized in larger natural language processing systems for Croatian with an expected very high accuracy, at least on texts from the same domain or newspaper texts in this particular case.

Table 4 provides the results of a more detailed inspection for full lemmatization implemented by procedures Base{1,2,3} and Merge{1,2,3} in terms of their error distributions. When concentrating on the overall error rate only, it can be divided into following components: (a) errors on wordforms known to both inflectional lexicon and tagger lexical database acquired at training, (b) errors on wordforms unseen by the tagger and known to the lexicon, (c) errors on wordforms unknown to the lexicon and yet seen by the tagger at training and (d) wordforms unknown to the lexicon and unseen by the tagger. Attention was also given to errors on homographic wordforms, i.e. the wordforms for which the inflectional lexicon provided more than one candidate lemma.

TABLE 4: Overall error rate by components

| Procedure | Base1 | Base2 | Base3 | Merge1 | Merge2 | Merge3 |
|---|---|---|---|---|---|---|
| Known by both (a) | 78.28% | 80.88% | 82.08% | 52.70% | 56.68% | 46.18% |
| Unknown by tagger (b) | 21.44% | 10.56% | 8.98% | 19.24% | 20.28% | 19.86% |
| Unknown by lexicon (c) | 0.07% | 2.69% | 2.81% | 8.81% | 7.23% | 10.66% |
| Unknown by both (d) | 0.21% | 5.87% | 6.14% | 19.25% | 15.81% | 23.30% |
| Errors on homography | 25.37% | 89.20% | 88.73% | 64.59% | 70.92% | 57.14% |

It is important to note how the contribution of errors of homography increases with increased algorithm complexity and subsequent reduction of errors caused by naïve algorithms Baseline{1,2,3}, accounting for between 57.14 and 70.92 percent of all lemmatization errors in the Merge{1,2,3} procedures. Regarding contributions of (a)-(d), an emphasis should be placed on correcting errors that occur when (a) the wordform is known both by the lexicon and the language model of the tagger as these are frequent with regards to a high reported lexical coverage of the Croatian inflectional lexicon (more than 96 percent) and also with regards to their share in the overall error rate (between 46.18 and 56.68 percent for the merging procedures).

TABLE 5: Error distribution by part of speech for Merge 3

| Part of speech | Percentage of errors |
|---|---|
| Noun | 31.86% |
| Adjective | 19.38% |
| Verb | 16.97% |
| Adverb | 11.48% |
| Residual | 9.44% |
| Other | 10.87% |

Table 5 is an illustration of what seems as a well-known distribution of errors between parts of speech for morphosyntactic tagging of Croatian texts (cf. Agić et al., 2009). Here, that distribution is shown to be valid for full lemmatization of Croatian, as well. A majority of errors occurs when lemmatizing nouns and adjectives, followed closely by verbs and then by adverbs and residuals. It should be noted that morphosyntactic tag for residual was used in the CW100 corpus for annotating e.g. foreign company names, thus resulting in an increased error rate and occurence count in general for this part of speech, caused once again by the corpus domain.

### 3.2 Idealistic scenario

Table 6 presents full lemmatization results of the idealistic scenario, achieved by using tagger previously trained on the entire CW100 corpus in order to reduce tagging error rate. It is given for procedure `Merge 3` alone, now proven to be the best of given choices for Croatian text in realistic test scenario.

It could be noted in the first place that the test environment had served its purpose: there were no wordforms unseen by tagger and tagging accuracy had a peak at 98.76 percent correctly assigned tags as was shown previously, in table 3. It should also be mentioned that previously relevant categories (b) and (d) with counts for wordforms unknown to tagger provided zero values in this table as no wordforms are unseen by tagger at training here.

Lemmatization accuracy had increased by 1.19 percent when compared to realistic test case scenario while tagging accuracy increase of 14.01 percent was much more substantial. At this point, it could be argued that lemmatization accuracy implemented via procedure `Merge 3` has already peaked and could not grow any further (at least not significantly), thus implying that additional (possibly rule-based) error handling modules would be required to lemmatize Croatian texts with higher accuracy. Such a conclusion is also backed up by results displayed in table 2, namely the small difference between accuracy of procedures `Merge 1` and `Merge 3`, stating that even implementing really simple stochastics for tagging error compensation brings the merging paradigm of lemmatization close to its limits. Also, properties of errors as given in tables 4 and 5 do encourage a change of perspective towards implementing specific, narrowly aimed rule-based module for handling specific error occurences.

TABLE 6: Idealistic test scenario results for Merge 3

| | |
|---|---|
| Lemmatization accuracy | 98.15% |
| Error rate | 1.85% |
| Known by both (a) | 44.18% |
| Unknown by tagger (b) | 0.00% |
| Unknown by lexicon (c) | 55.82% |
| Unknown by both (d) | 0.00% |
| Errors on homography | 29.56% |

## 4 Conclusions and future work

This experiment has shown how a large coverage inflectional lexicon can be combined with a stochastic morphosyntactic tagger in the task of lemmatizing an inflectionally rich and relatively free order language, which Croatian certainly is. Lemmatization accuracy on Croatian newspaper texts reached peak values of 96.96 and 98.15 percent on two different testing scenarios: a realistic one and an idealistic one.

The results obtained, namely the error distributions and the lists of errors from lemmatizer output, will be used in implementing simple rule-based correcting modules for the described lemmatizer. It is estimated that corrections of several remaining errors in the manual annotation of the CW100 corpus would push overall lemmatizer accuracy above 99 percent in this testing framework when combined with the rule-based error handlers.

## 5 Acknowledgements

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-1776, 130-1300646-0645, 036-1300646-1986 and partially by Flemish and Croatian governments within the joint CADIAL project (KRO/009/06).

## References

Agić Željko, Tadić Marko (2006). Evaluating Morphosyntactic Tagging of Croatian Texts. In *Proceedings of LREC 2006*.

Agić Željko, Tadić Marko, Dovedan Zdravko (2008a). Combining Part-of-Speech Tagger and Inflectional Lexicon for Croatian. In *Proceedings of the 6th Language Technologies Conference*, Ljubljana, Slovenia, pp. 445–451.

Agić Željko, Tadić Marko, Dovedan Zdravko (2008b). Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. In *Informatica*, Vol. 32, No. 4, pp. 116–121.

Agić Željko, Tadić Marko, Dovedan Zdravko (2008c). Investigating Language Independence in HMM PoS/MSD-Tagging. In *Proceedings of the 30th International Conference on Information Technology Interfaces*, pp. 657–662.

Agić Željko, Tadić Marko, Dovedan Zdravko (2009). Error Analysis in Croatian Morphosyntactic Tagging. In *Proceedings of the 31th International Conference on Information Technology Interfaces*, in press.

Brants Thorsten (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of ANLP 2000.*

Džeroski Sašo, Erjavec Tomaž (2000). Learning to Lemmatise Slovene Words. In *Learning Language in Logic, Lecture notes in Computer Science, Lecture notes in Artificial Intelligence, 1925.*, Springer, pp. 69–88.

Erjavec Tomaž (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of LREC 2004*.

Erjavec Tomaž, Džeroski Sašo (2004). Machine Learning of Morphosyntactic Structure: Lemmatizing Unknown Slovene Words. In *Applied Artificial Intelligence, 18.*, Taylor and Francis, pp. 17–41.

Giménez J and Márquez L. (2004). SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of LREC 2004.*

Halácsy Péter, Kornai András, Oravecz Csaba (2007). Using a morphological analyzer in high precision POS tagging of Hungarian. In *Proceedings of LREC 2006*, pp. 2245–2248.

Halácsy P, Kornai A, Oravecz C, Trón V, Varga D (2006). HunPos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 09–212.

Juršič M, Možetič I, Lavrač N. (2007). Learning Ripple Down Rules for Efficient Lemmatization. In *Proc. 10th Intl. Multiconference Information Society IS 2007*, Vol. A, pp. 206–209.

Shen L, Satta G, Joshi A. (2007). Guided learning for bidirectional sequence classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 760–767.

Šnajder Jan, Dalbelo Bašić Bojana, Tadić Marko. (2008). Automatic acquisition of inflectional lexica for morphological normalisation. In *Information Processing and Management, 44.*, Elsevier, pp. 1720–1731. pp. 760–767.

Tadić Marko (2002). Building the Croatian National Corpus. In *Proceedings of LREC 2002*, pp. 441–446.

Tadić Marko, Fulgosi Sanja (2003). Building the Croatian Morphological Lexicon. In *Proceedings of EACL 2003 Workshop on Morphological Processing of Slavic Languages*, pp. 41–46.

Tadić Marko (2005). The Croatian Lemmatization Server. In *Southern Journal of Linguistics*, Vol. 29, pp. 206–217.

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

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