Evaluating Morphosyntactic Tagging of Croatian Texts

Željko Agić*, Marko Tadić**

* Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture
University of Split
Ruđera Boškovića bb, HR-21000 Split, Croatia
zeljko.agic@gmail.com

** Department of Linguistics, Faculty of Philosophy
University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
marko.tadic@ffzg.hr

Abstract

This paper describes results of the first successful effort in applying a stochastic strategy – or, namely, a second order Markov model paradigm implemented by the TnT trigram tagger – to morphosyntactic tagging of Croatian texts. Beside the tagger, for purposes of both training and testing, we had at our disposal only a 100 Kw *Croatia Weekly* newspaper subcorpus, manually tagged using approximately 1000 different MULTEXT-East v3 morphosyntactic tags. The test basically consisted of randomly assigning a variable-size portion of the corpus for the tagger's training procedure and also another fixed-size portion, sized at 10% of the corpus, for the tagging procedure itself; this method allowed us not only to provide preliminary results regarding tagger accuracy on Croatian texts, but also to inspect the behavior of the stochastic tagging paradigm in general. The results were then taken from the test case providing 90% of the corpus for training purposes and varied from around 86% in the worst case scenario up to a peak of around 95% correctly assigned full MSD tags. Results on PoS only expectedly reached the human error level, with TnT correctly tagging above 98% of test sets on average. Most MSD errors occurred on types with the highest number of candidate tags per word form – nouns, pronouns and adjectives – while errors on PoS, although following the same pattern, were almost insignificant. Detailed insight on tagging, F-measure for all PoS categories is provided in the course of the paper along with other facts of interest.

1. Introduction

The primary purpose of our experiment, defined before the actual investigation even started, was to inspect whether statistical methods of MSD/PoS tagging would really be appropriate to tag a highly inflectional language such as Croatian including the desired accuracy of above 90% correctly tagged tokens on average. History of PoS/MSD tagging of Croatian is rather short one. There has been only one rule-based prototype tagger developed and tested in (Žubrinić, 1995) master thesis, displaying very good properties on known words. To the best of our knowledge this paper presents the first analysis of stochastic tagging strategies on Croatian. Therefore, the experiment has to be regarded as preliminary — any of the provided conclusions should be treated only as input for other, thoroughly implemented investigations which may give the final decision regarding MSD tagging of large-scale Croatian corpora.

In plain words, we did not allow this experiment to state that statistical approach should by all means be applied to tagging of Croatian corpora; first and the most important, we wanted to point out to what – on basis of results presented further in the paper – seemed to us the right way for future research and possible system implementation. Taking into consideration the current state of language technologies development for Croatian, even a small research project like the one presented is also an important one.

In the following sections 2 and 3 the paper describes the corpus used for training and testing and implementation of the TnT tagger on the corpus. The sections 4 and 5 present the evaluation methods and results discussion respectively. The paper is concluded with conclusion and future work.

2. Language resources

In order to MSD-tag a text written in Croatian using a Markov model tagger such as TnT, one has to have pre-tagged corpus at his/her disposal. For the purposes of this experiment, we had at hands only a single 100 Kw newspaper corpus; in this section, we provide a short description of its basic characteristics, PoS distribution and some lexical properties.

2.1. The corpus

The *Croatia Weekly* 100 Kw 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 (MTE v3) morphosyntactic specifications on the top of XCES corpus encoding standard (Ide et al, 2000):

```
...
<w lemma="ipak" ana="Rn">ipak</w>
<w lemma="početi" ana="Vmps-sfa">počela</w>
<w lemma="Hrvatska" ana="Npfsd">Hrvatskoj</w>
...
```

Figure 1: Excerpt from the XML encoding of CW100 corpus

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 MTE v3 MSD tagset by matching the CW100 corpus and the Croatian Morphological Lexicon (Tadić & Fulgosi, 2003) at unigram via the Croatian Lemmatization Server (Tadić, 2003) at (http://hml.ffzg.hr).

Croatian language in general implements 12 out of 14 different PoS categories defined in the MTE v3 specification: Adjective (A), Conjunction (C), Interjection (I), Numeral (M), Noun (N), Pronoun (P), Particle (Q), Adverb (R), Adposition (S), Verb (V), Residual (X) and Abbreviation (Y). However, 11 of them actually do appear in the CW100 corpus (Residual missing out), the fact once again suggesting that it's a rather small resource to operate with, both in quantity and quality, especially when compared to resources available for some other highly inflectional Eastern European languages.

General corpus details are presented in the second column of Table 1; overall token count includes word forms and punctuation and it can easily be seen that the newspaper texts contained approximately 25 tokens per sentence (22 of them being word forms), each of them tagged using 896 different morphosyntactic descriptions.

	Total on corpus	Sentences only
Tokens	118529	112902
Word form	103161	98567
Other	15368	14335
Sentences	4626	4162
Different MSD	896	892

Table 1: Corpus details

However, the first conclusion is somewhat incorrect; the data given in the second column presents all tokens in the corpus, including the ones in header and byline sections of newspaper articles and being that these writings are often not entirely grammatical and therefore should not be treated as sentences by definition, we made another simple calculation by excluding all header and byline data; the results are given in the third column of the table and so the corpus has approximately 27 tokens per one grammatical sentence within the articles. Excluding 5627 header and byline tokens would have led us to a loss of 4 different MSD tags and thus we decided to include all corpus data in the tagging process, mainly because of this loss of knowledge and also the nature of tagger operation, which is described further in the text. Moreover, this enabled us to simulate the "real situation" of using this tagging procedure for a large scale newspaper corpus.

PoS	% corpus	Diff. MSD
Noun	30,45%	119
Verb	14,53%	62
Adjective	12,06%	284
Adposition	9,55%	9
Conjunction	6,98%	3
Pronoun	6,16%	312
Adverb	3,88%	17
Numeral	1,84%	48
Abbreviation	1,11%	21
Particle	0,46%	4
Interjection	0,01%	2
Other	12,97%	15

Table 2: PoS distribution on the corpus

Distribution of PoSs in the corpus is in fact exactly what we expected before the experiment actually begun. Common newspaper texts are mainly written in plain Croatian and for news-reporting purposes, most sentences comply to relatively simple subject-verb-object model and therefore nouns and verbs dominate the distribution.

Basic tagging expectations can now be easily inferred directly from data in Table 2; categories Adjective (A) and Pronoun (P) are tagged using a large number of MSD tags and the corpus does not contain them in numbers great enough to contribute statistically to the training procedure. Therefore, we expected the highest error rates to appear upon tagging pronouns and adjectives, followed by the most common ones – nouns and verbs. Actual details on MSD-tagging these parts of speech are discussed in detail throughout the results section.

2.2. The lexicon

In this paper, the term *lexicon* does not represent the before-mentioned Croatian Morphological Lexicon, but a resource built by the tagger's training procedure; we present its details before discussing the TnT tagger itself mainly because they provide additional insight on the nature of the CW100 corpus.

The first column of Table 3 provides us with the number of types that were found on the corpus, while the second one states how many of them had more than one morphosyntactic description. Results state that the CW100 emergent lexicon contains only 13,95% ambiguous entries. Ambiguous token count also suffered an exponential quantitative decrease when presented as a function of MSD per token count; most ambiguous entries have only 2 or 3 MSD tag candidates.

Lexicon entries	Ambiguous
25310	3505

Table 3: CW100 corpus lexical ambiguity

In addition, only 2,01% of the emergent lexicon was ambiguous on PoS; on basis of given information, we reasonably expected good tagging results. However, the nature of available resources and testing methods enclosing the tagger also applied important constraints on interpreting these results; these issues are all thoroughly investigated in sections regarding testing methodology and conclusions.

3. The TnT tagger

The TnT ("trigrams and tags") program, first described in (Brants, 2000), is an excellent implementation of a (hidden) Markov model paradigm of PoS/MSD tagging. TnT is optimized for maximum speed and straightforward usage on virtually any language and tagset, given that both resources are written in a standard file format that the application can understand.

A Markov model approach to part-of-speech tagging basically consists of extracting knowledge from large samples of written language and then representing it as a statistical model consisting of transition and emission probability matrices. A Markov model tagger always tags input text acting as a hidden Markov model (HMM) and using a Viterbi algorithm, while it is trainable on both pre-tagged and untagged corpora, using a simple counting algorithm (VMM method) or a rather complex forwardbackward procedure, also known as the Baum-Welch algorithm (HMM method), to infer its own representation of MSD tags, respectively.

The TnT tagger package consists of four different modules, each of them implementing an important part of the tagging procedure:

- `tnt-para`
  The training module expects pre-tagged corpora as input, and therefore it trains using a VMM method of merely counting occurrences and calculating emission and transition probabilities of a HMM.
- `tnt`
  This module does the actual tagging of input text. It implements unigram, bigram and default trigram tagging method using a Viterbi trellis algorithm. It is also enhanced by linear interpolation as a smoothing method, respective weights are determined by using deleted interpolation and unknown words treated by a suffix trie and successive abstraction.
- `tnt-diff`
  Implements basic methods for result retrieval; the program requires pre-tagged comparison file and a file tagged by the tagging module; it outputs overall tagging accuracy and separate results on known and unknown tokens, along with respective percentages.
- `tnt-wc`
  The simplest module provides users with basic word counts (overall tokens, different tokens and optionally different tags) for a given input file.

During the experiment, additional programs also had to be developed for transforming the CW100 corpus data to a TnT-friendly input file format, creating the desired testing environment and retrieving specific information regarding tagger accuracy that TnT modules could not provide.

The main reasons we have chosen TnT for the experiment were its language independence (i.e. adaptability to virtually any language; the fact being even more important given the current state of language technologies for Croatian language) and overall speed, thus positioning it as currently predominant compared to many other state-of-the-art tagger programs which use different statistical approaches to PoS/MSD tagging.

4. Evaluation methods

The entire testing procedure was subdivided into two major test cases, each of them containing common tagging tests. This segmentation was made in order to present results in a transparent and comparison-friendly manner. The purposes of two major tests are: (Test-1) providing overall tagger accuracy given a tagset and (Test-2) inspecting accuracy over different PoSs given the same tagset. Having the tagset as the only variable in these test definitions, we established two test cases on that basis – one using full MSD and the other using PoS – and applied (Test-1) and (Test-2) on each of them. (Test-1) used a simple mechanism built in the TnT's `tnt-diff` comparison module and incorporated into our testing algorithm:

- 9 training set sizes were defined, each differing in size of the CW100 corpus fragment; the first one had randomly assigned 10% of all corpus sentences, the second one 20% and up to 90% for the ninth training set size.
- For each of the training set sizes, 3 actual training sets were chosen at random from the corpus.
- Every training set was afterwards assigned 4 different test sets and matching compare sets forming pairs, one of these pairs representing the worst case scenario for a given training set (worst case being the one in which sentences are assigned to the test set from what was left of the corpus after assigning sentences to that training set).
- Training procedure was run using all defined training sizes and sets; for each training set size, arithmetic mean of all 12 tests was taken as overall accuracy and also accuracy on known and unknown entries; this test therefore outputs these three numbers, but also a very important worst case scenario characteristic.

(Test-2) was meant to provide us with separate tagging accuracies on different PoS; in order to provide precise descriptions, we had to introduce precision, recall and their harmonic mean, or F-measure, like it was applied, for instance, in (Van Rooy & Schäfer, 2003).

- Given two sets, *B* representing a set of words actually belonging to PoS *X* and *A* representing the tagger's assignment of words into *X*, its precision is defined by:

$$P = \frac{|A \cap B|}{|A|} = \frac{\textit{words correctly assigned to X}}{\textit{words assigned to X}}$$

- Given the same two sets and a PoS, recall is defined as:

$$R = \frac{|A \cap B|}{|B|} = \frac{\textit{words correctly assigned to X}}{\textit{words belonging to X}}$$

Basically, precision by definition measures tagger's ability to correctly assign tags on a chosen set and recall defines how good it is in actually choosing that very set. Precision therefore can suffer over-specialization (the tagger tags the chosen set 100% correct, but it chooses too few candidates for *X*; its precision is perfect, but recall very low) and recall with exactly the opposite problem of over-generalization (the tagger chooses all candidates, but also others not really belonging to *X*; it then shows perfect recall, but poor precision). It is common knowledge that precision and recall cannot be chosen as separate measures of tagger accuracy, but they could form an expression capable of doing that:

$$F(X) = \frac{2P(X)R(X)}{P(X) + R(X)}$$

So the F-measure *F(X)* on PoS *X* is defined as harmonic mean of two natural constraints – recall *R(X)* and precision *P(X)* on that category – and also constrained again to a [0, 1] interval by its well chosen fraction form. This specific formula was derived from the general definition of harmonic mean *H* of *n* real numbers x_i:

$$\frac{1}{H}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i} \qquad n\in N, x_i\in R, \forall i\in\{1,...,n\}$$

The choice of harmonic mean of all other means that could also be used as measures is nothing but a historical consequence, as it has become a *de facto* standard since first introduced in (Van Rijsbergen, 1979).

To summarize before actually presenting the results: (Test-2), the process of identifying F-measure on all PoSs, was run following the same pattern as the one introduced in (Test-1) since both tests used variable training set sizes, their output also contained some interesting figures about the tagger itself.

5. Results

Here we present results of (Test-1) and (Test-2) in detail and for both test cases defined by given tagsets. We also provide additional characteristics on each of the tests in these test cases, such as various diagrams envisioning tagger accuracy as a function of given training sets.

5.1. Results on full MSD tagset

Applying TnT tagger according to rules set in (Test-1) and using a full set of morphosyntactic descriptions for Croatian, as defined in the MULTEXT-East v3 standard, provided us with various data summarized in Table 4 and Figure 2.

Table 4 presents average overall tagger accuracy as a function of training set size and number of unknown tokens that tagger encountered while applying the tagging procedure per training sets. All data is averaged out of 12 runs, including worst case scenario as given previously in (Test-1) definition.

The first two columns of the table describe an important characteristic of the training sets – a number of unknown tokens the tagger encountered on test sets when using provided data to form a given training set; about 30% unknown tokens on average were encountered while tagging with TnT trained on 10% of the corpus, reaching very low 4,51% unknown when training with 90% of the corpus, i.e. the training set size we chose as a keystone of our experiment, being that it provides us with overall tagging accuracy on Croatian texts.

Training set size (% corpus)	Unknown tokens	Overall correct
10%	30,83%	78,48%
20%	22,46%	83,30%
30%	17,41%	86,22%
40%	14,22%	88,16%
50%	11,35%	89,85%
60%	9,09%	91,20%
70%	7,43%	92,53%
80%	5,67%	93,71%
90%	4,51%	94,83%
(+ worst case)	(13,61%)	(86,24%)

Table 4: Overall accuracy on MSD

Therefore, we can come out with a conclusion and a constraint:

- Overall tagging accuracy using TnT with full MTE v3 MSD on Croatian texts peaks very high, at 94,83%.
- General corpus statistics and unknown token count of only 4,51% on that peak value suggest we should still be cautious with our judgement.

Given this important constraint, we consider the worst case scenario on training set size involving 90% of the corpus; results are displayed inside parenthesis, in the last row of Table 4, showing overall worst case accuracy of 86,24% with 13,61% unknown tokens. When compared to rows 5 and 6 of the same table (both having similar unknown token characteristic), all these fact actually lead us to a conclusion that accuracy of a Markov model tagger, applied to a language of such inflectional complexity as Croatian, is above all a function of unknown token count, and only then a function of a tagset and other variables. Of course, larger training corpora provide the tagging procedure with more knowledge, thus consistently reducing unknown token count so it can be stated that accuracy is also primarily a function of training set size, with language specifics being on the second place.

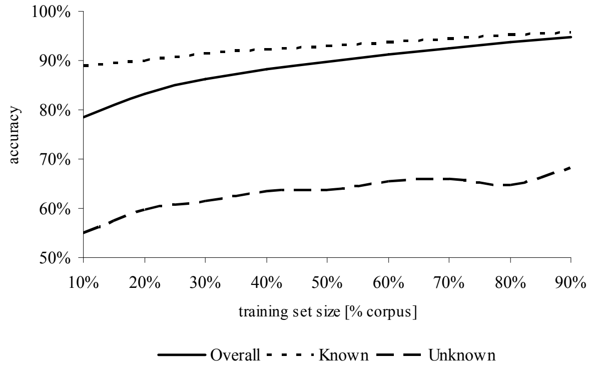

Figure 2: Elements of tagger accuracy

Figure 2 supports previous statements and also gives insight to the nature of function growth; overall accuracy and known token accuracy rise in a clearly logarithmic way, while growth on unknown tokens exhibits same behavior, but with some minor irregularities due to the relative size of the training sets.

Results of (Test-2) are presented in Table 5 and Figure 3 and give detailed insight of tagger accuracy on parts of speech previously marked as being difficult – pronouns, adjectives and nouns – and then we compare them to results achieved on other categories.

% corpus	Pronouns	Adjectives	Other
10%	65,95%	55,77%	86,79%
20%	75,36%	61,13%	90,06%
30%	79,28%	68,07%	92,61%
40%	81,84%	72,77%	94,64%
50%	86,61%	79,64%	94,56%
60%	89,96%	84,47%	96,45%
70%	89,90%	87,75%	96,63%
80%	94,65%	90,11%	97,38%
90%	95,89%	95,54%	97,93%

Table 5: Insight on F-measure

The first couple of rows in Table 5 provide noticeable evidence to support our claims: given an average training set sized at 10% or 20% of the corpus, tagging accuracy on adjectives and pronouns is up to 30% lower then on all other PoS combined and averaged. Moreover, the rise on these two PoSs is much faster than on others, displaying obviously high importance of the quantity of the raw language data quantity for successful tagging of difficult categories.

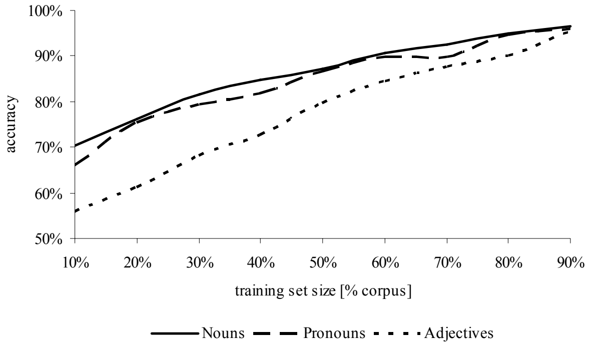

Figure 3: F-measure on nouns, pronouns and adjectives

Along with before-mentioned adjectives and pronouns, Figure 3 also provides MSD-tagging results for nouns; the higher starting accuracy is a direct consequence of their distribution; approximately 30% of the entire CW100 corpus is consisted of nouns and about 10% more than pronouns and adjectives make together.

Given these results on full MSD, it is plain to see that improving accuracy on difficult PoSs would definitely lead to substantial overall improvement and also that additional effort on putting together larger language resource for training is the direction for future work.

5.2. Results on PoS

Results of (Test-1) and (Test-2) are expectedly far better on PoS, being that fewer tags provide the trigram tagger better transition and emission matrices, given the same training set.

% corpus	Overall	Known	Unknown
10%	93,82%	98,46%	82,79%
20%	95,45%	98,26%	84,96%
30%	96,09%	98,55%	83,93%
40%	97,04%	98,52%	86,21%
50%	97,72%	98,76%	86,48%
60%	98,02%	98,71%	89,13%
70%	98,27%	98,64%	90,52%
80%	98,37%	98,78%	84,66%
90%	98,63%	98,77%	90,40%

Table 6: Tagger accuracy on PoS

The main observation is, of course, overall accuracy on PoS peaking at 98,63% and actually reaching human level of error. However, the constraint put upon presenting the corresponding MSD result also applies on PoS.

Additional data in Table 6 and Table 7 also confirms the importance of training procedure, i.e. the quality of HMM matrices, for overall tagging accuracy; the third column of Table 7 shows that PoS accuracy on unknown words is at least 20% higher than the corresponding accuracy on MSD and, given that no lexical entries influence correctness on unknown words, we argue this is a direct consequence of tagset differences, i.e. that 26 PoS tags on approximately 100000 tokens (90% of the corpus) provide TnT with a much more tuned-up transition matrix than 896 MSD tags could ever do.

% corpus	Difference on known	Difference on unknown
10%	9,51%	27,88%
20%	8,16%	25,26%
30%	7,15%	22,54%
40%	6,36%	22,71%
50%	5,69%	22,80%
60%	5,02%	23,73%
70%	4,12%	24,52%
80%	3,54%	20,02%
90%	2,90%	22,11%

Table 7: Absolute PoS-MSD accuracy difference

On the other hand, known token characteristics show that TnT tagger operates amazingly well upon tagging tokens it stored in lexicon during the training procedure; it could therefore be argued that knowing an entry (i.e. having large enough training corpora) is the key point to success, both for PoS and MSD tagging. Since the absolute difference given in Table 7 reaches the lowest value of around 3% in favor of PoS, this result is practically insignificant having in mind the difference between corresponding tagsets.

5.3. Comparison

Three similar up-to-date PoS/MSD tagging research papers were used in order to provide comparison and to place our research results relative to the others; these are:

- Evaluation of various taggers and tagsets of Slovene, as described in (Džeroski et al., 2000); researchers had at hands a Slovene translation of Orwell's famous novel *1984.* for purposes of training and testing the TnT tagger, along with a rule-based tagger, maximum entropy tagger and also a memory-based solution.
- In (Van Halteren et al., 2000), the TnT tagger was tested using three different training corpora, two of them being English – LOB and WSJ corpus – and one Dutch – the Eindhoven corpus.
- A detailed investigation is described in (Hajič, 2000) concerning tagging of Czech, Estonian, Hungarian and Romanian (and also Slovene and English which we ignore as they are already provided in other two papers), using two non-HMM stochastic strategies.

All results, including the ones provided by TnT and the CW100 corpus of Croatian, are presented in Table 8. However, before even considering them, some other facts regarding those other experiments should be pointed out.

First of all, results provided by (Hajič, 2000) are given in form of word-only error rates, meaning that all tokens not representing actual word forms were excluded upon presenting the results. Of course, error rates were easily translated into accuracies, but not the tokens vs. actual words constraint. Secondly, (Hajič, 2000) uses stochastic paradigms other than HMMs – a maximum entropy tagger and an exponential (log-linear) tagger. Research described in (Džeroski et al., 2000) for Slovene is most similar to the one we provide for Croatian, thus Table 9 containing comparison of test details is provided.

	Training set	Diff. MSD on corpus	Result
Croatian	~106000	896	94,83% TnT
Czech	87071	970	82,23%MET
Dutch	~750k	341	92,06% TnT
English	~1M	170	97,55% TnT
Estonian	81383	476	86,05% EXP
Hungarian	102992	401	91,84% EXP
Romanian	104583	486	92,34% MET
Slovene	81805	1004	89,22% TnT

Table 8: Stochastic strategies on various languages

We could argue that overall results of tagging Croatian with TnT are better than ones on Slovene, but two last rows of Table 9 should by all means be taken into consideration before providing actual conclusions. First of these two rows presents the number of unknown tokens TnT encountered when tagging the test case upon which it achieved results presented in Table 8 for Croatian and Slovene; it is plain to see that results for Croatian are around 5% better than the ones for Slovene, but also that the trigram model for Croatian had 7% more “language knowledge”.

	Croatian	Slovene
Total sentences	4626	5855
For training	4163 (90%)	5204 (89%)
For testing	463 (10%)	651 (11%)
Different words	25310	16017
Unknown on test	4,51%	11,75%
Balanced acc.	89,85% / 11,35%	89,22%

Table 9: Croatian vs. Slovene corpus stats

Therefore, in the last row of Table 9, we have chosen tagger accuracy achieved when trigrams for Croatian had approximately the same amount of language information as for Slovene – around 11% unknown tokens encountered – and the two results formed almost a perfect match, thus proving that trigram taggers’ accuracy is highly dependent on the amount of data provided.

We should mention also one more reason for making this balance. While in the Croatian CW100 the test corpus was composed of a number of different articles from newspaper (with relative common vocabulary), in the Slovenian case the test corpus was a single text (a novel) expectedly showing higher percentage of unknown words. This had to be normalized with aforementioned procedure.

With this problem explained and solved, we could finally state that experimental tagging of Croatian texts using the TnT trigram tagger was successful when compared to recent results achieved on other similar languages.

6. Conclusions and future work

There are many directions for further enhancement of the results presented in this paper. Additional annotated corpora could be used in order to test the tagger in an environment differing in distributional properties from the training set, the tagger itself could be altered in order to match specific properties of Croatian language, etc. Given this facts, we consider this paper a baseline for more thoroughly implemented additional research.

The paper presented a full course of experiment of applying a hidden Markov model based TnT tagger on tagging texts of Croatian language using full MTE v3 morphosyntactic descriptions. Obtained results peaked at around 95% correctly assigned tags with around 5% unknown tokens encountered, showing a trigram tagging paradigm would be a reasonable choice for further tagging of larger Croatian corpora in the future.

7. References

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

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. In *Computational Linguistics* 1995, 21(4):543-565.

Džeroski, S., Erjavec, T. & Zavrel, J. (2000). Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets. In *Proceedings of the Second International Conference on Language Resources and Evaluation*. ELRA, Paris-Athens 2000, pp. 1099-1104.

Hajič, J. (2000). Morphological Tagging: Data vs. Dictionaries. In *Proceedings of the ANLP-NAACL*. Seattle, Washington, USA, 2000, pp. 94-101.

Ide, N., Bonhomme, P., Romary, L., (2000). An XML-based Encoding Standard for Linguistic Corpora. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pp. 825-830. (see also at http://www.xces.org).

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

Tadić, M. (2003). *Jezične tehnologije i hrvatski jezik*. Exlibris, Zagreb 2003.

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

Van Halteren, H., Zavrel, J. & Daelemans, W. (2000). Improving accuracy in wordclass tagging through combination of machine learning systems. In *Proceedings of the ANLP-NAACL*. Seattle, Washington, USA, Morgan Kaufman Publishers 2000.

Van Rijsbergen, C. (1979). Information retrieval. Butterworths, London 1979.

Van Rooy, B. & Schäfer, L. (2003). An evaluation of three PoS taggers for the tagging of the Tswana Learner English Corpus. In *Proceedings of the Corpus Linguistics 2003 conference*. Lancaster, UK, 2003.

Žubrinić, T. (1995). Mogućnosti strojnoga označavanja i lematiziranja korpusa tekstova hrvatskoga jezika. Master thesis, Faculty of Philosophy, University of Zagreb 1995.