# Morphosyntactic Tagging of Slovene Legal Language

Tomaž Erjavec
Dept. of Knowledge Technologies, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
E-mail: tomaz.erjavec@ijs.si, http://nl.ijs.si/et/

Bence Sárossy
Pázmány Péter Catholic University
2087, Piliscsaba, Egyetem út 1., Hungary
E-mail: steksz@freemail.hu

**Keywords:** human language technologies, part-of-speech tagging, Slovene language, legal language

**Received:** November 7, 2006

*Part-of-speech tagging or, more accurately, morphosyntactic tagging, is a procedure that assigns to each word token appearing in a text its morphosyntactic description, e.g. "masculine singular common noun in the genitive case". Morphosyntactic tagging is an important component of many language technology applications, such as machine translation, speech synthesis, or information extraction. In the paper we report on an experiment on morphosyntactic tagging of Slovene, on a sample of Slovene legal language. We evaluate the accuracy of the TnT tagger, which had been trained on the MULTEXT-East language resources for Slovene. The test data come from the freely available parallel English-Slovene corpus SVEZ-IJS, which contains the Slovene translation European Union legal acts. Presented are the details of the manually corrected test corpus and an analysis of the tagging errors. The paper also discusses a simple transformation-based program that fixes some of the more common errors, and concludes with some directions for future work.*

*Povzetek: V prispevku je opisan poskus oblikoslovnega označevanja na vzorcu slovenskih pravnih besedil.*

## 1 Introduction

Morphosyntactic tagging, also known as part-of-speech tagging or word-class syntactic tagging (van Halten, 1999) is a process in which each word appearing in a text is assigned an unambiguous morphosyntactic tag. This process is, in general, composed of two parts: the program first assigns, on the basis of a morphological lexicon, all the possible tags that a word form can be associated with, and then chooses the most likely tag on the basis of the context in which the word form appears in the text. For instance, the Slovene word form hotel has three possible tags: two (nominative and accusative singular) of the noun lemma hotel, and one verbal (masculine past participle), of the lemma *hoteti* (to want). Yet in the sentence *Šel je v hotel* (*He went to a hotel*), the token hotel should be tagged as a noun in accusative case.

Morphosyntactic tagging was first developed for the English language, where the set of morphosyntactic tags is relatively small (~50, depending on the specific tagset used). English is an inflectionally poor language, so problems arise mainly in connection with ambiguities at the word class (part-of-speech) level, e.g. in determining whether "left" should be tagged as an adjective (my left hand), a noun (on your left), or a verb (he left early). Taggers and (manually) tagged corpora were later developed also for morphologically richer languages, such as Czech (Hajič and Hladka, 1998) and Slovene (Erjavec et al., 2000). Such languages typically distinguish more than a thousand morphosyntactic tags, and the largest problem, at least at first sight, is caused by having to disambiguate between the large number of syncretic inflectional forms within word classes. For example, nouns can be four ways ambiguous regarding their inflectional properties: the word form *človeka* (from the lemma *človek / man*) can function either as singular genitive or accusative, or the dual nominative or accusative.

Most contemporary taggers learn the model of a given language from a manually tagged corpus, possibly supported by a morphosyntactic lexicon. Such programs are robust, but they do make mistakes. The accuracy of tagging depends on the properties of the language, the tagset used, size of learning corpus, the similarity of the training corpus with the text to be tagged, and of course the particular tagger.

Our attempts regarding automated tagging of Slovene were connected to the morphosyntactic resources developed in the MULTEXT-East project (Erjavec, 2004), http://nl.ijs.si/ME/, which contain a morphosyntactic specification (defining the tagset), a morphological lexicon and a small (100,000 words) manually tagged corpus, which contains the novel „1984" by G. Orwell. The first experiments (Erjavec et al., 2000) showed that from four publicly accessible

taggers the best results were achieved by TnT (Brants, 2000). TnT is a Hidden Markov Model tri-gram tagger, which also implements an unknown-word guessing module. It is fast in training and tagging, and is able to accommodate the large tagset used by Slovene. In subsequent work (Erjavec and Džeroski, 2004) we also tackled lemmatisation (so, *hotela* → *hotel* or *hoteti*, depending on the tag), concentrating esp. on unknown words. For this, we used the program CLOG, based on Inductive Logic Programming, which had been trained on the MULTEXT-East lexicon. The program learns rules (decision lists) for each morphosyntactic tag separately, and is thus dependent on prior morphosyntactic tagging.

In the present paper we describe an evaluation of tagging (and lemmatisation) completed on another dataset, namely on a sample from the SVEZ-IJS corpus of legal language (Erjavec, 2006). We were interested in the accuracy of tagging on a corpus that is very different from the training corpus, as this shows how best to improve the tagging accuracy in the future. We wanted to know what kind of errors are the most frequent ones, and whether it is possible - and if yes, to what extent - to correct them in a simple way.

In the remainder of this paper we first introduce the experiment set-up, i.e. the tagger and the dataset. This is followed by the analysis of errors and the description of a transformation-based program that corrects some most frequent errors, a comparison of accuracy levels reached in different experiments and finally, some conclusions.

# 2 Test data

The experiments use the "totale" program (which invokes the TnT tagger) and a sample of the Slovene part of the SVEZ-IJS corpus. The sample was first tagged automatically, and the results manually corrected.

## 2.1 Tagging with totale

For linguistic tagging we use the program "totale" program (tokenisation, tagging, and lemmatisation) (Erjavec et al., 2005), which:

1. tokenizes the text, that is, it splits it into words, punctuation marks and sentences
   (with the m1Token module, a part of totale)
2. assigns morphosyntactic tags to words
   (with the TnT tagger, (Brants, 2000))
3. lemmatizes the text
   (with CLOG (Erjavec and Džeroski, 2004))

Both TnT and CLOG are programs that learn language models from previously prepared data, namely from a manually tagged corpus and a morphosyntactic lexicon. Our morphosyntactic tagging model was learned on the MULTEXT-East corpus, the „1984" (100,000 tokens), and a small sample of IJS-ELAN corpus (5,000 tokens). The lemmatiser was trained on the MULTEXT-East morphosyntactic lexicon (the complete inflectional paradigms of 15,000 lemmas).

## 2.2 The SVEZ-IJS corpus

The SVEZ-IJS parallel English-Slovene corpus, http://nl.ijs.si/svez/ (Erjavec, 2006) contains EU legal texts, the so called Acquis Communautaire. Version 1.0 of this corpus contains 2×5 million words and was made in 2004 on the basis of the translation memory produced by the Translation Department at SVEZ (The Office of the Government for European Affairs) (Erbič et al., 2005). The corpus was compiled from the parallel English- Slovene translation units, where each such unit typically contains one sentence or a part of a sentence, e.g. an item in a list.

## 2.3 The sample

For the evaluation of tagging we made a sample of the totale automatically tagged corpus, in which we included 3 consecutive Slovene segments out of every 1000 segments; this gave us 3‰ of the Slovene part of the corpus. The sample was then converted into an Excel table and was manually corrected, while preserving the automatically assigned tags and lemmas. This file serves as the dataset from which the numbers given in the present paper were extracted.

| Unit | n | Ratio | |
|---|---|---|---|
| Characters | 513.650 | | A |
| Segments | 821 | 625 A/B | B |
| All tokens | 15.765 | 19 C/B | C |
| Punctuation (tokens) | 2.346 | 15 % C | D |
| Words (tokens) | 13.419 | 85 % C | E |
| Words (types) | 5.189 | 2.59 E/F | F |
| Lemmas (types) | 3.062 | 4.38 F/G | G |
| Morph. tags (types) | 452 | 29.69 E/H | H |

Table 1. Test data, basic statistics

An analysis of the sample size is given in Table 1, which shows, e.g., that the sample contains around half a million characters and 15,000 tokens, of which 13,000 are words. These consist of around 5,200 word forms or 3,000 lemmas. All the lemmas are written in lower-case, therefore e.g. the word forms *koren* and *Koren* have the same lemma, although the second can be a proper name. The last line in the table shows the test set contains a surprisingly small number of tags, less than 500.

The column Ratio shows the proportions between various measures and contains e.g. the average segment length in characters (821) and tokens (19), and the average number of different word forms per lemma (4.4).

Table 2 shows the distribution of words in more detail. We can see that around 15% of all the tokens and more than 18% of the words are unknown to the tagger, which highlights the difference between MULTEXT-East and SVEZ-IJS corpora, but is also the result of the small size of the MULTEXT-East corpus used for training. The table also shows the statistics over the word classes: most frequent words in the text are nouns, adjectives and prepositions, together covering around half of all the tokens. This means that the overall tagging

accuracy depends largely on the ability of the tagger to correctly interpret these three word classes, esp. the nouns.

The last two word classes given in the table are important for two reasons. First, abbreviations (Y) and residuals (X) and are not parts-of-speech; from a (morpho)syntactic point of view Y typically covers nouns (e.g. *Dr.*), although it can describe whole phrases (e.g. *etc.*). X (residual) is used to tag foreign words, and often appears successively, e.g. *carte de séjour de résident privilégié de Monaco*, so, from the morphosyntactic point of view, a series of X tags functions as a noun phrase. The second characteristic of these two categories is the relatively large number of tokens they cover (4.5%) in the SVEZ-IJS sample. As discussed later, these two categories are responsible for a significant part of errors in the automated tagging.

| | n | tokens | words |
|---|---|---|---|
| Words | 13.419 | 85.1 % | 100 % |
| Known | 10.996 | 69.7 % | 81.9 % |
| Unknown | 2.423 | 15.4 % | 18.1 % |
| Noun (N) | 4.928 | 31.3 % | 36.7 % |
| Verb (V) | 1.287 | 8.2 % | 9.6 % |
| Adjective (A) | 1.694 | 10.7 % | 12.6 % |
| Adverb (R) | 373 | 2.4 % | 2.8 % |
| Numeral (M) | 795 | 5.0 % | 5.9 % |
| Pronoun (P) | 743 | 4.7 % | 5.5 % |
| Conjunction (C) | 1.102 | 7.0 % | 8.2 % |
| Preposition (S) | 1.787 | 11.3 % | 13.3 % |
| Particle (Q) | 107 | 0.7 % | 0.8 % |
| Abbrev. (Y) | 474 | 3.0 % | 3.5 % |
| Residual (X) | 128 | 0.8 % | 1.0 % |

Table 2: Test data, (un)known words and distribution per word class.

# 3 Analysis of automated tagging

On the basis of the manually tagged sample we evaluated the accuracy of automated tagging with the MULTEXT-East trained totale. Table 3 shows the absolute number of various types of errors, as well as giving them as a percentage of tokens or words respectively. We further split each error type according to the overall error, as well as the error for known and unknown words separately.

Table 3 gives the precision for three types of linguistic annotation performed by totale. The first is the error rate of the morphosyntactic tagging itself, where, according to the strictest metric, the system achieves an 89.6% overall accuracy. The second type is the accuracy of tagging for the word class only. This means that the tagger might have assigned the wrong tag but did at least manage to correctly identify the word class, i.e. the first letter of the tag. It is useful to distinguish these two types of errors, as many applications or users require only the word category, and do not make use of, say, inflectional features.

| | n | Token acc. | Word acc. |
|---|---|---|---|
| **Wrong m.s. tag** | 1,799 | 88.6 % | 86.6 % |
| For known words | 950 | 92.9 % | 91.4 % |
| For unknown words | 849 | 65.0 % | 65.0 % |
| **Wrong word class** | 748 | 95.3 % | 94.4 % |
| For known words | 155 | 98.8 % | 98.6 % |
| For unknown words | 593 | 75.5 % | 75.5 % |
| **Wrong lemma** | 220 | 98.6 % | 98.4 % |
| For known words | 88 | 99.3 % | 99.2 % |
| For unknown words | 132 | 94.6 % | 94.6 % |
| For wrong tag | 217 | 87.9 % | 87.9 % |
| For correct tag | 3 | 99.8 % | 99.8 % |

Table 3: Accuracy of automated tagging.

The third type of annotation we analyse is the lemmatisation. It is interesting to note that the accuracy of lemmatization is higher than for morphosyntactic tagging, which means that tagging errors do not necessary influence the lemmatization. Nevertheless, as the last two rows show, the errors of lemmatization are almost exclusively due to erroneous morphosyntactic tags: there are only three instances, where the morphosyntactic tag is correct, but the lemma is wrong.

## 3.1 Errors in word class tagging

Because of the importance of word class tagging, we will discuss this topic separately from errors of morphosyntactic description. In Table 4 we give a matrix showing errors according to actual word class (horizontally) and according to the word-class assigned by the tagger (vertically). The diagonal thus gives the numbers for errors which happen internally to a word class and do not affect the word class accuracy, while the other cells give the confusions between different parts-of-speech; they show, e.g. that nouns were interpreted as verbs in 95 cases.

The table shows that the tagging of open word classes (written in bold letters) is significantly less successful than tagging of function words, which is understandable as the most of the latter group is known to the tagger. To a certain extent pronouns are an exception, but only regarding error rate within word class. The reason for small absolute accuracy of tagging of pronouns is their especially rich inflectional structure: pronouns cover around half (more than thousand) of all the morphosyntactic tags.

In most of the cases erroneous interpretation is assigned to nouns, numerals, residuals and abbreviations. In case of nouns the relative number of errors is small, however due to their large number, the effect on the overall accuracy is significant. Misinterpretation of the nouns as verbs is possibly due to the nature of learning data base.

| | ***N*** | ***V*** | ***A*** | ***R*** | ***M*** | *P* | *C* | *S* | *Q* | ***I*** | ***X*** | ***Y*** | * |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***N*** | **609** | 6 | 9 | 4 | 47 | 0 | 1 | 1 | 0 | 0 | 69 | 241 | 987 |
| ***V*** | 95 | **18** | 2 | 1 | 28 | 2 | 2 | 0 | 0 | 0 | 35 | 17 | 200 |
| ***A*** | 28 | 1 | **275** | 12 | 8 | 3 | 0 | 0 | 0 | 0 | 14 | 9 | 350 |
| ***R*** | 14 | 1 | 4 | **15** | 0 | 1 | 1 | 1 | 0 | 0 | 6 | 11 | 54 |
| ***M*** | 0 | 0 | 1 | 0 | **11** | 6 | 0 | 0 | 0 | 0 | 0 | 18 | 36 |
| *P* | 1 | 0 | 1 | 0 | 2 | **105** | 0 | 0 | 0 | 0 | 0 | 1 | 110 |
| *C* | 0 | 1 | 0 | 3 | 0 | 0 | **0** | 0 | 10 | 0 | 0 | 11 | 25 |
| *S* | 1 | 0 | 0 | 0 | 1 | 0 | 0 | **18** | 0 | 0 | 1 | 6 | 27 |
| *Q* | 0 | 0 | 0 | 3 | 0 | 0 | 2 | 0 | **0** | 0 | 0 | 0 | 5 |
| ***I*** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** | 1 | 0 | 2 |
| ***X*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** | 0 | 0 |
| ***Y*** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | **0** | 3 |
| * | 749 | 27 | 292 | 38 | 98 | 117 | 6 | 20 | 10 | 0 | 128 | 314 | 1799 |

Table 4: Confusion table of word class errors. N = nouns, V = verb, A = adjective, R = adverb, P = pronoun, S = preposition, C = conjunction, Q = particle, I = interjection, M = numeral, Y = abbreviation, X = residual.

The other three problematic word classes (numerals, residuals and abbreviations), have similar reasons for causing such a large number of errors. On the one hand, words of these classes are almost always unknown, as they are present neither in the training corpus nor in the lexicon, except for a limited number of numerals, on the other hand they do not have a consistent morphosyntactic interpretation, especially true for residuals and abbreviations, which makes them harder to predict. We come back to this problem in sections 4 and 5.

### 3.2 Errors within word classes

In this section we will take a closer look at errors within word classes. As seen in the diagonal of Table 4, most of these errors appear with nouns, which include, according to the MULTEXT-East specification, five attributes for Slovene: type, gender, number, case and animacy. Around 85% of the errors are connected to case combined with number, and similar behaviour can be observed with adjectives and pronouns. A more detailed analysis of these errors showed that in most of the cases it is impossible to assign correct tags on the basis of the local morphosyntactic context, as used by TnT. The tagging of other word classes is less problematic. In case of verbs, mostly gender and number are erroneously tagged.

## 4 Rule-based transformation tagging

The main question is, of course, how to improve tagging accuracy. As an attempt in this direction we implemented a program, which corrects some errors made by the TnT tagger. In this section we describe this program and the improvement on accuracy when using it.

The program is written in Perl and takes automatically tagged text as input data. The program has access to data about the form of the word, its tag assigned by TnT, and whether the word is known to TnT. For each word, the program runs a cascade of hand-written rules, where rules have the following format: „if *condition* then *assign a morphosyntactic tag*, else *next rule*.” In the conditions we use a function called feature, which takes a feature for its first argument and a token as second, and returns the value of the feature for the token. We give the first two rules as an example:

```
...
① elsif ($freq == 0  and
        feature("idwrd",$sent[$focus])=~/^[IVX]+$/)
    {$outmsd="Mc---r"}
② elsif ($freq == 0 and
        feature("case",$sent[$focus]) eq 'uc' and
        not (feature("case",$sent[$focus-1])  eq  'uc' or
        feature("case",$sent[$focus+1]) eq  'uc'))
    {$outmsd="Y"}
...
```

The first rule deals with Roman numerals, as their mis-recognition was one the largest problem of tagging numerals. The condition says that the word is required to be unknown ($freq == 0), and the form (feature idwrd) of the focus token ($sent[$focus]), has to be composed only of characters *I, V* and *X* (regular expression /^[IVX]+$/). The rule thus fires for tokens such as *MCMLXX*, and will change their tag (whatever it was) into Mc---r, which stands for word class=numeral, type=cardinal, form=roman.

The second rule corrects the word tag by changing it for Y, i.e. it tags the word as an abbreviation if the word is unknown, contains only capital letters, and the word immediately to its left or right is not capitalized. The rule will thus apply to cases such as: *Čist dobiček ECB se prenese …*, (*The net profit ECB is transferred…*) but will not incorrectly tag unknown words like *RAZČLENITEV PO ODDELKIH …* (*BREAKDOWN BY DIVISION*).

Currently we have implemented five rules, based on the analysis of some frequent and also easily correctable errors. The first two rules have already been described. The third changes the tag to abbreviation, if the unknown word includes numbers and not more than three letters (e.g. *2002/917/ES*), regardless of context. The fourth changes the tag of all supines to nominal masculine

nominative, and the last changes the tag of *a* (which was always tagged as a conjunction) to abbreviation, if it is followed by a punctuation mark, e.g. *Annex IV a. OJ No L 71*.

Table 5 gives the results for the dataset first tagged by TnT and then corrected by the program implementing the above five rules. The first column gives the numbers of tokens that had their word classes changed, and the second of tokens with changed morphosyntactic tag. The first line shows the number of tags that were wrong, but the program changed to the correct ones, the second gives the numbers of those tokens which TnT tagged correctly, but the Perl program subsequently corrupted. The third line shows the number of tokens that had an incorrect tag assigned by TnT, were changed by the Perl program, yet the changed tag was also wrong. The last line shows those instances where the TnT tag was wrong and was subsequently "changed" to the same tag, i.e. a rule fired, but to no effect. The values shown in Confused and Identical rows do not influence tagging accuracy, although it is preferable to have a small number of confusions, as the new errors are likely to be more complex than original ones. The absolute number of corrected errors by the Perl tagger comes from subtracting the second line from the first; the overall improvement is given in the last line.

| | Word class | Morpho-syntactic tag |
|---|---|---|
| Corrected | 291 | 289 |
| Corrupted | 4 | 4 |
| Confused | 14 | 16 |
| Identical | 2 | 2 |
| Improvement | 287 | 285 |

Table 5: Result of automated error correction.

The numbers show that, for the case of full morphosyntactic tags, the relative error decreases by 16% and the tagging accuracy grows from 86.6% to 88.9%. This difference is not high, however, it was not our aim to maximize the accuracy of the morphosyntactic tagging; it will have been noted that all the rules strictly correct the word class tags; and the improvement of accuracy for word class tagging is much more significant: using only five transformation rules the accuracy grows by 38.4% relative, from 94.4% to 96.6% absolute accuracy.

# 5 Comparison of tagging accuracy

In Table 6 we give a short summary and comparison of morphosyntactic and word class tagging accuracies for the various experimental settings and compare the results from this paper to previous research on Slovene. The first line gives the results reported in Erjavec et al. (2000), in which the MULTEXT-East corpus, i.e. "1984" was used (with ten-fold cross validation) both for training and testing. The second line shows the evaluation of tagging as presented in this paper, therefore with on a corpus significantly different from the training one. The Tnt+Trans gives the results obtained after the application of the transformation program described in the previous section.

| | Morpho-syntactic tag | Word class |
|---|---|---|
| 1984: TnT | 89.2 % | 96.6 % |
| SVEZ-IJS: TnT | 86.6 % | 94.4 % |
| SVEZ-IJS: TnT+Trans | 88.9 % | 96.6 % |
| SVEZ-IJS - X,Y: TnT | 89.4 % | 97.6 % |
| ZRC SAZU: TreeTagger | 83.6 % | ? |

Table 6: Overview of tagging accuracies for Slovene

It should be noted that the very common errors of abbreviations and residuals (foreign words) are caused not so much by the tagger, but rather in the tokenization. A robust solution to the problem of tagging X and Y would thus be rather in adding to the tokenisation a special module which would identify abbreviations and foreign words and add them to the lexicon used by the tagger. From this perspective it is interesting to take a look at the accuracy rates obtained by omitting X and Y tokens from the evaluation. Line four (SVEZ-IJS - X,Y) of the table shows that under these conditions accuracy of TnT tagging would reach 89.4% for morphosyntactic tagging and 97.6% for word class tagging, i.e. would be greater than on "1984" itself.

Finally, the last row of the table shows the results of tagging Slovene as presented in Lönneker (2005), which is, to our knowledge, the only other research aiming at the evaluation of automated morphosyntactic tagging for Slovene. Lönneker describes the usage of TreeTagger (Schmid, 1994) on the ZRC SAZU manually tagged corpus (Jakopin and Bizjak, 1997) of one million words. This experiment differs from ours in a number of parameters: the tagger used, the tagset, the size of learning corpus and the structure of test corpus. It is therefore difficult to make a direct comparison, nevertheless, the difference in the results is surprising, especially with regard to the fact that the ZRC SAZU training corpus contains more than million words of mixed genre texts, while "1984" has only 100,000 and contains one novel only. Lönneker (2005) makes some hypotheses as to why the accuracy is lower in her tests then in the ones reported in Erjavec et al. (2000): one reason could be the more detailed ZRC SAZU tagset, which e.g. distinguishes different types of names (personal, country, mythological), the other less consistency in the manual tagging of the ZRC SAZU corpus, which was performed by different people over a long period of time, and without detailed guidelines or a firmly fixed tagset. A further reason could be that the TnT tagger is better than TreeTagger, esp. at tagging unknown words.

# 6 Conclusions

In the paper we analyzed the accuracy of automated morphosyntactic tagging with TnT trained on the

MULTEXT-East morphosyntactic resources for Slovene. The evaluation took a manually corrected sample from the Slovene part of SVEZ-IJS corpus of legal EU texts, which comprised around 15.000 tokens, and includes around 15% words not included in the training set. The evaluation showed that the absolute accuracy regarding word tokens in the sample is 86.6%, for the whole tagging and 94.4% for word class tagging. If we improve tagging with a transformational program, which corrects some frequent but simple errors, the accuracy increases to 88.9% for the morphosyntactic tags and 96.6% for word class tagging.

We have mentioned one way to improve accuracy, which includes pre-processing to identify abbreviations and foreign words. Higher accuracy would also certainly be obtained if we were to use a larger training corpus, consisting of a variety of text types. The main problem to this kind of solution is the lack of available manually tagged corpora for Slovene.

There are several other options on how to improve tagging accuracy. An interesting approach and a publicly accessible program is described by Brill (1992), which was also the inspiration for our transformational program. A significant difference is that we wrote the rules manually, while the Brill tagger learns rules from a training corpus. A different approach used for languages with a rich tagsets and small training corpora is described in Tufiş (2006), which proposes a method for tagset reduction, so improving the data density for the tagger, yet in such a way that the original tags can be reconstructed via the lexicon.

## Acknowledgements

The work of the second author on this paper was supported by grant CMEPIUS RS.

# References

Brill, E. (1992). A Simple Rule-Based Part of Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing, ACL*. Trento, Italy.

Brants, T. (2000). TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference, ANLP-2000*. Seattle, WA. 224--231.

Erjavec, T., Džeroski, S., Zavrel, J. (2000). Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000)*. ELRA, Paris.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004). ELRA*, Paris.

Erjavec, T., Džeroski, S. (2004). Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence*, 18, 17--41.

Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R. (2005). Massive multilingual corpus compilation: ACQUIS Communautaire and totale. In *Proceedings of the Second Language Technology Conference.* April 2004, Poznan.

Erjavec, T. (2006). The English-Slovene ACQUIS corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'2006)*. ELRA, Paris.

Erbič, D., Krstič Sedej, A., Belc J., Zaviršek-Žorž, N., Gajšek, N., Željko, M. (2005). Slovenščina na spletu v dokumentih slovenske različice pravnega reda Evropske unije, terminološki zbirki in korpusu. In Zbornik Simpozija *Obdobja 24: Razvoj slovenskega strokovnega jezika*. Ljubljana.

Hajič, J., Hladka, B. (1998). Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. *COLING-ACL'98.* ACL.

van Halteren, H. (ed) (1999). *Syntactic Wordclass Tagging.* Kluwer.

Jakopin, P., Bizjak, A. (1997). O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična Revija. 45/3-4*. 513-532.

Lönneker, B. (2005). Strojno oblikoslovno označevanje slovenskih besedil: Kako daleč smo. *Slavistična revija 53/2*. 193--210.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing.* Manchester. 44--49.

Tufiş, D. (2006). Tagset Design for High Accuracy POS Tagging and Automatically Building Mapping between Arbitrary Tagsets. *Workshop on Annotation Science: State of the Art in Enhancing Automatic Linguistic Annotation (LREC'06)*. ELRA, Paris.