# Building Named Entity Recognition Models for Croatian and Slovene

**Nikola Ljubešić, Marija Stupar, Tereza Jurić**

Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3
10000 Zagreb
Croatia
{nljubesi,mstupar,tjuric2}@ffzg.hr

**Abstract**

The paper presents efforts in developing freely available models for named entity recognition and classification for Croatian and Slovene. Our experiments focus on the most informative set of linguistic features taking into account the availability of language tools for the languages in question. Beside the classic linguistic features, distributional similarity features calculated from large unannotated monolingual corpora are exploited as well. Using distributional information improves the results for 7-8 points in F1 while adding morphological information improves the results for additional 3-4 points in both languages. The best performing models, along with test sets for comparison with future and existing systems and a HunPos part-of-speech model for Croatian are available for download for academic usage.

**Izgradnja modelov za prepoznavanje imenskih entitet za hrvaščino in slovenščino**

Prispevek predstavi razvoj prostodostopnih modelov za prepoznavanje in klasifikacijo imenskih enot za hrvaški in slovenski jezik. Poskusi se osredotočajo na najbolj informativne jezikoslovne lastnosti, pri čemer upoštevajo dostopnost jezikovnih orodij za jezika. Poleg standardnih jezikoslovnih lastnosti so upoštevane tudi distribucijske lastnosti, ki so bile izračunane iz velikih neoznačenih enojezičnih korpusov. Uporaba distribucijskih lastnosti poboljša rezultate za 7-8 točk v meri F1, uporaba oblikoslovnih informacij pa dodatno za 3-4 točke, in to pri obeh jezikih. Najboljši naučeni model, skupaj s testno množico za primerjavo z obstoječimi in bodočimi sistemi, ter model za oblikoslovno označevanje hrvaščine s programom HunPos so dostopni za prenos za uporabo v znanstvene namene.

## 1. Introduction

Named entity recognition and classification (NERC), nowadays often called just named entity recognition (NER) is a subtask of the information extraction task. It aims to locate and classify text elements into predefined categories, and is regularly applied in many fields, using statistical or rule-based models. State-of-the-art systems tend to be open domain and language independent.

In this paper we present the process of creating NER models for Croatian and Slovene that we publish for free academic use.

The tool we use to build the models is the Stanford Named Entity Recognizer, nowadays a frequently used tool for NER. It is an implementation of Conditional Random Fields sequence models and is available under GNU GPL licence and free for academic use. (Finkel et al., 2005)

Beside many feature extractors that come with this tool, it is designed to work with the clustering method proposed by (Clark, 2003) which combines standard distributional similarity with morphological similarity to cover infrequent words for which distributional information alone is unreliable.

This paper is structured as follows: in Section 2 we give an overview of related work, in Section 3 we present the datasets used in our research, in Section 4 we give an overview of our experimental setup and in Section 5 we present the results of the experiments.

## 2. Related work

To our knowledge, there has been some effort in developing NER systems for south Slavic languages mainly in the direction of building rule-based systems.

A rule-based system for Croatian described in (Bekavac, 2005) uses regular grammars for recognition and classification of names over annotated texts. The system contains the module for sentence segmentation, lexicon of common words, specialized lists of names and transducers for automatic recognition of certain word forms.

A statistical approach described in the diploma thesis (Bošnjak, 2007) uses a semi-supervised method based on lists of names and entity extraction system.

For Serbian a rule-based system (Vitas and Pavlović-Lažetić, 2008) shows that there is a great difference between English and Serbian language, as well as all the other Slavic languages which require a more thorough preparation of the system because of the rich inflectional system.

None of the presented systems are available for academic usage which hinders researchers in looking into higher tasks that require NER as a preprocessing step. One of the main intentions of this paper is to improve this situation.

In the process of building a good NER system, features are considered as important as the selection of algorithm for machine learning. The aim is to find an optimal set of features that will ensure the highest system accuracy with minimum complexity in classifier building. Several NER approaches use a very large number of features (Mayfield et al., 2003), but the inclusion of additional features after a

certain point can yield worse results.

In the students' research paper that precedes this research (Filipić et al., 2012) we have identified properties for the Stanford NER system defining feature extractors that seem to work best for Croatian language. In this paper we use these property files and only vary in training and test sets and the usage of distributional information [1].

The only work we are aware of that examines the usage of distributional features in Stanford NER is (Faruqui and Padó, 2010). The paper describes the process of building and optimizing NER models for German and by using distributional features F1 is improved for 6% in-domain and 9% out-of-domain. Our research is considerably inspired by this paper.

## 3. Corpora

We have built and annotated two corpora, one Croatian and one Slovene. Both copora are built from data taken from specific Internet domains from the Croatian and Slovene web corpora hrWaC and slWaC (Ljubešić and Erjavec, 2011).

The Croatian corpus contains 59,212 tokens taken from four different Internet domains covering two general newspaper portals, nacional.hr and jutarnji.hr, one ICT portal bug.hr and the business news portal poslovni.hr. These data were annotated during a students' project where diversity of data was one of the main points.

The Slovene corpus is almost two thirds the size of the Croatian one containing 37,032 tokens and data from just one general news portal rtvslo.si. While selecting these data the main goal was to build a usable training set with limited annotation capacities.

Beside admitting that these corpora were built opportunistically regarding temporary goals, we want to emphasize that having two corpora of different diversity and size gives us an interesting starting point for our experiments.

The amount of data in both copora is given in Table 1.

| corpus | document # | token # |
|---|---|---|
| hr | 105 | 59,212 |
| bug.hr | 19 | 9,609 |
| jutarnji.hr | 16 | 9,760 |
| nacional.hr | 24 | 20,583 |
| poslovni.hr | 46 | 19,260 |
| sl | 69 | 37,032 |
| rtvslo.si | 69 | 37,032 |

Table 1: Size of the corpora used

The corpora were tagged by the IOB2 standard following the CoNLL-2003 annotation guidelines [2] where each row represents a token in the text with its linguistic annotation and designated predefined named entity category. IOB2 labels show whether a word is at the beginning (B), inside (I) or outside (O) of a named entity. Four categories were annotated - location (LOC), organization (ORG), person (PERS) and miscellaneous (MISC).

| corpus | NE | LOC | MISC | ORG | PERS |
|---|---|---|---|---|---|
| hr | 2,647 | 591 | 632 | 818 | 606 |
| sl | 2,491 | 716 | 378 | 311 | 1,086 |

Table 2: Number of annotated named entities in both corpora

Since for Slovene there are freely available taggers and this is not the case for Croatian, we manually annotated basic part-of-speech (first letter of the Multext-East MSD) on Croatian data as well since related work shows that these features are useful for the task. Slovene data was MSD tagged and lemmatized with the freely available ToTaLe tagger (Erjavec et al., 2005) trained on JOS corpus data (Erjavec et al., 2010).

To be able to use POS information on unseen Croatian data, we trained a simple model for the HunPos tagger (Halácsy et al., 2007) from the Croatian dataset. We performed a simple test of the resulting model by dividing the Croatian dataset into a training and a test set by the ratio of 9:1. Accuracy obtained on the test set was 95.1%. We publish the tagger trained on all available data along with the NER models and the benchmark datasets. To our knowledge, this is the first freely available part-of-speech tagger for Croatian. We are fully aware of more informative MSD taggers being developed years ago and hope that our approach of publishing most of the available results will speed up the process of other researchers releasing existing data and tools under a permissive license.

The amount of annotated named entities in the Croatian and Slovene corpus is given in Table 2. The expected difference in diversity of the data can be clearly observed from these numbers. First of all, although the Slovene corpus has 37% less textual material, it has just 6% less named entities showing a higher density of named entities one would expect from a straightforward newspaper dataset. Furthermore, when we look at the type of named entities, we can observe that the Slovene dataset contains much more person names and slightly more locations while the Croatian dataset contains more organization names and named entities labeled with the miscellaneous category. These data confirm our assumption that the Croatian dataset is much more diverse and will thereby present a harder task for supervised classification.

A final insight in the features and thereby specificities of the two datasets is given by calculating vocabulary transfer between identical portions of development and test sets. The numbers are given in Table 3. The vocabulary transfer is calculated as the token and type percentage of named entities in the test set being already present in the development set.

Two interesting properties can be observed here. First of all, the Slovene vocabulary transfer is higher than the Croatian one pointing at the expected lower content diversity of Slovene data. Secondly, there is almost no difference

[1] The example property file used in this paper can be retrieved from `http://www.nljubesic.net/upload/ner/ner.prop`

[2] See `http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt`

| corpus | token transfer | type transfer |
|---|---|---|
| hr | 10.7% | 10.6% |
| sl | 17.3% | 12.4% |

Table 3: Vocabulary transfer for both corpora on identical portions of development and test set

between token and type transfer on Croatian data showing that the diversity of named entities is really high since almost none of the named entities from the development set present in the test set appears more than once in the Croatian test set which is not the case on Slovene data.

We divided both corpora into development and test sets by shuffling documents and producing test sets of similar size for both languages. The decision to build test sets of similar size was guided by the idea of publishing those test sets as benchmark datasets for both languages. For that reason the Croatian development set contains 53,142 tokens while the Slovene one contain 29,686 tokens, i.e. 56% of the amount of Croatian data.

For calculating distributional similarity of tokens from large monolingual corpora portions of hrWaC and slWaC web corpora were used. For Croatian we built a 100Mw corpus and for Slovene a 50Mw corpus, both containing data from large news portals.

## 4. Experimental setup

Since different annotations on Croatian and Slovene data were available, we evaluated different settings for each language. Beside part-of-speech information for both languages, on Slovene data MSD and lemma information was present as well.

On Croatian data we experimented with POS information ("POS"), distributional information ("DISTSIM") calculated from 10Mw, 50Mw and 100Mw corpora while on Slovene data we experimented with POS, MSD ("MSD") and lemma ("LEMMA") information and distributional information obtained from 10Mw and 50Mw copora. Thereby we performed 8 experiments on Croatian data and 11 experiments on Slovene data (we eliminated the experiments varying with availability of lemma information once it proved to be non-informative).

All the experiments were performed on development sets of both datasets via 5-fold-cross-validation that takes into account document borders. By respecting document borders we were trying to keep the vocabulary transfer as low as possible and thereby obtain the most realistic results, i.e. differences between different experimental settings.

Distributional similarity was calculated by using Clark's `cluster_neyessen` tool (Clark, 2003) with default settings (numberStates=5, frequencyCutoff=5, iterations=10). The number of resulting clusters was set on best-performing values in (Faruqui and Padó, 2010), i.e. for 10Mw corpora 100 clusters and for 50Mw and 100Mw corpora 400 clusters were built. First twenty elements of example clusters calculated from the Croatian 100Mw and Slovene 50Mw corpora are given in Table 4. The Croatian cluster contains exclusively country and city names in the

| |
|---|
| njemačkoj rijeci londonu sarajevu osijeku italiji zadru francuskoj haagu austriji parizu dubrovniku vukovaru španjolskoj milanu bruxellesu rimu beču moskvi berlinu |
| tomaž simon goran martina dejan jan nina tom saša mojca vesna jurij eva nataša maria jernej daniel richard thomas damjan žiga |

Table 4: First 20 elements of sample clusters obtained with Clark's tool on the 100Mw Croatian and 50Mw Slovene corpora

locative (or dative) case. The Slovene cluster contains person first names in the nominative case of both Slovene and English origin. These examples show very clearly how the cluster ID can be used as a very informative feature in the supervised training procedure.

After identifying best performing settings on development sets we calculate our final results by training a system on the whole development set and testing it on the left-out test set.

Finally we calculate learning curves for the best performing settings to identify the gain we can expect from annotating more data.

## 5. Results

The results obtained by 5-fold cross-validation on both development sets are presented for Croatian in Figure 1 and for Slovene in Figure 2. The results of each cross-validation are averaged by calculating the harmonic mean. Regarding the statistical significance of the results, we perform a one-tailed paired t-test over pairs of results we find interesting.

On Croatian results we can observe already in the second experiment that basic morphological information in this simple setting improves F1 for 4.5% ($p = 0,002$). Our third experiment shows that using distributional information obtained from a 10 million token corpus improves the result as much as the part-of-speech information with similar significance ($p = 0.005$). By combining both of these two features we improve our results for 8.5%, highly significantly in comparison to using only one feature ($p < 0.001$). By calculating distributional information on five and ten times more data we get improvements of 2% and 3% when not using part-of-speech information and 1% and 2% improvements when using part-of-speech information. The differences between neighbouring corpus sizes (10 and 50; 50 and 100) are not statistically significant, but the differences between using 10Mw and 100Mw corpora are ($p = 0.007$). We see a steady rise in performance as the unlabeled monolingual corpus size increases motivating us to perform similar calculations on much larger datasets in the future.

The results on Slovene data in the categories present in Croatian data are rather similar backing them up. There are two types of information on Slovene data we did not have for Croatian - MSD and lemma. By using MSD and not only POS information the results do improve for additional 1%, but statistically insignificant ($p = 0.21$). On the contrary, by adding lemma information to the MSD decreases

the result significantly for 5.5% ($p = 0.007$). One could expect such an outcome since lemmatization performs worst on named entities. By adding more distributional information by moving from a 10Mw to a 50Mw corpus we get an improvement even steeper than on Croatian data by getting a 5% improvement, now highly significant ($p < 0.001$). This could be explained by the higher simplicity of this dataset and yield a conclusion that for data from narrower domains additional data sources such as this one give more improvement. We can observe on both datasets that, when using distributional similarity from larger corpora, including additional features like POS or MSD makes the increase in the results lower.

When comparing results on Croatian and Slovene datasets one observes right away that the results on Slovene data are much better although the size of the dataset is under half the size. This can be traced back to the fact that the Slovene dataset has a narrower domain, a higher vocabulary transfer and a higher amount of named entities like person and location which are considered easier to recognize and classify. On the other hand the resulting Croatian module is expected to be more robust and should perform better on different domains.

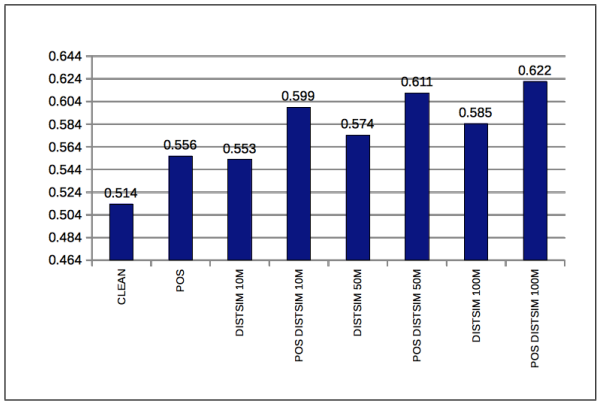


Figure 1: F1 results obtained via 5-fold cross-validation on Croatian development set

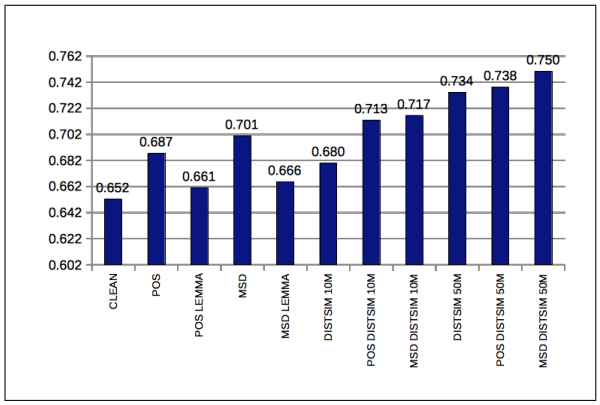


Figure 2: F1 results obtained via 5-fold cross-validation on Slovene development set

We chose two settings per dataset for final testing on the left-out test set. The first one uses distributional information, but leaves out the need for morphological annotation of the data while the second one uses both distributional and morphological information. We present the results of precision, recall, F1, true positives and false positives and negatives by category in Table 5. We consider such an exhaustive data presentation informative since this is the best approximation of the capability of the models we publish alongside this paper.

The number of false negatives shows to be on both datasets and settings higher than the number of false positives with higher percentage than recall as a direct consequence. On Slovene data the best performing categories in reverse order are PERS, LOC, ORG and MISC. On Croatian data LOC tends to perform best, ORG and PERS being a tie and MISC being traditionally the worst category. The somewhat unexpected order of category performance can probably be followed to the wider domain of the Croatian dataset.

| hr DISTSIM 100Mw | | | | | | |
|---|---|---|---|---|---|---|
| Entity | P | R | F1 | TP | FP | FN |
| LOC | 0.8049 | 0.7021 | 0.7500 | 33 | 8 | 14 |
| MISC | 0.7436 | 0.3867 | 0.5088 | 29 | 10 | 46 |
| ORG | 0.6742 | 0.6250 | 0.6486 | 60 | 29 | 36 |
| PERS | 0.9032 | 0.5185 | 0.6588 | 28 | 3 | 26 |
| Totals | 0.7500 | 0.5515 | 0.6356 | 150 | 50 | 122 |
| hr POS DISTSIM 100Mw | | | | | | |
| Entity | P | R | F1 | TP | FP | FN |
| LOC | 0,8293 | 0,7234 | 0,7727 | 34 | 7 | 13 |
| MISC | 0,7778 | 0,4667 | 0,5833 | 35 | 10 | 40 |
| ORG | 0,6989 | 0,6771 | 0,6878 | 65 | 28 | 31 |
| PERS | 0,8500 | 0,6296 | 0,7234 | 34 | 6 | 20 |
| Totals | 0,7671 | 0,6176 | 0,6843 | 168 | 51 | 104 |
| sl DISTSIM 50Mw | | | | | | |
| Entity | P | R | F1 | TP | FP | FN |
| LOC | 0,7423 | 0,7273 | 0,7347 | 72 | 25 | 27 |
| MISC | 0,5000 | 0,2143 | 0,3000 | 15 | 15 | 55 |
| ORG | 0,8947 | 0,3617 | 0,5152 | 17 | 2 | 30 |
| PERS | 0,8966 | 0,8509 | 0,8731 | 234 | 27 | 41 |
| Totals | 0,8305 | 0,6884 | 0,7528 | 338 | 69 | 153 |
| sl MSD DISTSIM 50Mw | | | | | | |
| Entity | P | R | F1 | TP | FP | FN |
| LOC | 0,7957 | 0,7475 | 0,7708 | 74 | 19 | 25 |
| MISC | 0,4688 | 0,2419 | 0,3191 | 15 | 17 | 47 |
| ORG | 0,8947 | 0,3617 | 0,5152 | 17 | 2 | 30 |
| PERS | 0,8619 | 0,8400 | 0,8508 | 231 | 37 | 44 |
| Totals | 0,8180 | 0,6977 | 0,7531 | 337 | 75 | 146 |

Table 5: Test results on the four best performing models (P - precision, R - recall, F1 - F1 measure, TP - true positives, FP - false positives, FN - false negatives)

With the final set of experiments we wanted to examine the learning curves of the best performing approaches to see how much we could benefit in the future by just annotating more data.

The four learning curves were calculated using distributional information and varying the usage of available

morphological information, for Slovene the MSD, and for Croatian part-of-speech information. The curves are calculated by enlarging the training data in ten steps by shuffling the development set data and testing on the test set. The experiment for each training set size was repeated four times to obtain a better estimate of the curve shape. The learning curves are depicted in Figure 3.

The Slovene curve rises much steeper than the Croatian one which is in accordance to all other information pointing to the fact that the Slovene dataset is much easier than the Croatian one. Both learning curves have finished the steepest phase, but are still climbing which shows that the process could further benefit from larger amounts of labeled data. While building the Croatian dataset we actually calculated learning curves during the annotation process to assess if annotating larger amounts of data would prove to be very beneficial. For Slovene data we did an educated guess based on our insights on Croatian data and the fact that this dataset covers a narrower domain.

From these curves no conclusions about the informativeness of morphological information should be drawn like in case of Slovene data where the results not using morphological information seem better than those that use that information. The learning curves are produced by testing the built models on just one dataset while previous results given in Figures 1 and 2 are obtained via cross-validation by evaluating five models built on different data on five different evaluation sets.

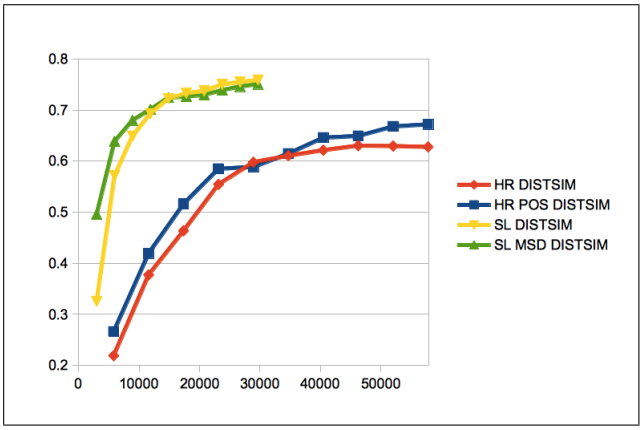


Figure 3: Learning curves calculated on portions of development sets and test sets by using distributional similarity and varying in usage of available morphological information (x axis represents token count, y axis the F1 measure)

## 6. Conclusion

In this paper we have presented the process of building freely available models for named entity recognition and classification for Croatian and Slovene. We have built two datasets, one for Croatian which is larger and covers a broader domain and one for Slovene which is smaller but covers just the general news domain.

We were searching for the optimal set of features on the development set via five-fold cross-validation. Lemmata have shown to be of no use for a morphologically complex language such as Slovene since lemmatization tends to work worst on word classes such as named entities. On the other hand morphological information such as POS tags or full MSD tags proved to be valuable with the latter being more informative. That type of information improved the F1 measure in a 3-5% window. Clustering tokens from a large monolingual corpus by features such as contextual and morphological properties has proven to be beneficial improving the results by using 10Mw corpora for 3-4%. With clustering results from larger corpora the results continue to improve steadily. Combining both morphological and clustering information proved to be the winning combination with an overall improvement of 10% on datasets of both languages. By omitting morphological information for which some preprocessing is required we still get an improvement of 8%.

We are releasing four best performing models free for academic purpose, two for each language - one that uses morphological annotation, and one that does not require such information. Additionally, we release the two test sets as potential benchmarks for future work on named entity recognition and classification for these two languages. The models and datasets can be found on

- `http://www.nljubesic.net/resources/data/croatian-ner/` for Croatian and
- `http://www.nljubesic.net/resources/data/slovene-ner/` for Slovene.

The HunPos part-of-speech model for Croatian can be obtained from

- `http://www.nljubesic.net/resources/data/croatian-pos-tagger/`.

For the future our plan is to increase the amount of annotated data for training by exploiting semi-supervised approaches. Additionally we plan to calculate distributional similarity on larger corpora and take under consideration variations of the method used in this paper.

## Acknowledgement

Research reported in this paper has been supported by the CESAR project within the EU 7th Framework Programme and the ICT Policy Support Programme, grant agreement no. 271022.

## 7. References

Bekavac, Boško, 2005. *Strojno prepoznavanje naziva u suvremenim hrvatskim tekstovima*. Ph.D. thesis, University of Zagreb.

Bošnjak, Matko, 2007. *Strojno prepoznavanje naziva tehnikama strojnog učenja*. Master's thesis, University of Zagreb.

Clark, Alexander, 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary.

Erjavec, Tomaž, Darja Fišer, Simon Krek, and Nina Ledinek, 2010. The JOS Linguistically Tagged Corpus of Slovene. In *International Conference on Language Resources and Evaluation*.

Erjavec, Tomaž, Camelia Ignat, Bruno Poliquen, and Ralf Steinberger, 2005. Massive multilingual corpus compilation: Acquis communautaire and totale. In *The 2nd Language & Technology Conference - Human Language Technologies as a Challenge for Computer Science and Linguistics*. Association for Computing Machinery (ACM) and UAM Fundacja.

Faruqui, Manaal and Sebastian Padó, 2010. Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of the 10th Conference on Natural language processing (KONVENS) 2010*. Saarbrücken, Germany.

Filipić, Lobel, Tereza Jurić, and Marija Stupar, 2012. Strojno prepoznavanje naziva u tekstovima pisanima hrvatskim jezikom. Students' paper awarded with the Rector's award.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning, 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*.

Halácsy, Péter, András Kornai, and Csaba Oravecz, 2007. HunPos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07. Stroudsburg, PA, USA: Association for Computational Linguistics.

Ljubešić, Nikola and Tomaž Erjavec, 2011. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In Ivan Habernal and Václav Matousek (eds.), *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*. Springer.

Mayfield, James, Paul McNamee, and Christine Piatko, 2003. Named entity recognition using hundreds of thousands of features. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03. Stroudsburg, PA, USA: Association for Computational Linguistics.

Vitas, Duško and Gordana Pavlović-Lažetić, 2008. Resources and methods for named entity recognition in Serbian. *INFOTHECA : Journal of Informatics and Librarianship*, 9(1-2).