Using crowdsourcing in building a morphosyntactically annotated and lemmatized silver standard corpus of Croatian

Filip Klubička, Nikola Ljubešić

Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3, HR-1000 Zagreb
{fklubick,nljubesi}@ffzg.hr

Abstract

This paper describes the creation of a morphosyntactically tagged and lemmatized silver standard corpus by using crowdsourcing. A data set containing 50.322 tokens compiled from the Croatian web corpus hrWaC was annotated using TreeTagger and HunPos taggers trained on the SETimes.HR corpus. Tokens that the tools annotated differently were passed on to the crowd. The crowd looked through contested nouns, verbs and adjectives, while experts checked and corrected those that the crowd decided were incorrect, along with the remaining parts of speech the two taggers did not agree on. The evaluation of the crowdsourcing yielded a single worker's accuracy to be ∼90%, and that of the majority answer of three workers to be ∼97%. While intrinsic evaluation of the resource by calculating accuracy of morphosyntactic tags showed an improvement of 8%, extrinsic evaluation of the corrected corpus on the task of morphosyntactic tagging produced an accuracy increase of little over 1%. The results point to the conclusion that the use of crowdsourcing in creating and improving language resources is indeed useful, but in the case of using the improved resource for enhancing morphosyntactic tagging, given the amount of already available gold corpus data, accuracy should be improved by developing a lexicon.

Uporaba mnoienja pri izdelavi oblikoskladenjsko oznaenega in lematiziranega korpusa hrvaine kot srebrnega standarda

V prispevku opišemo postopek izdelave oblikoskladenjsko označenega in lematiziranega korpusa hrvaščine z uporabo množičenja. Podatkovna množica, ki vsebuje 50.322 pojavnic, je bila vzorčena iz hrvaškega korpusa spletnih besedil hrWaC in oznaena z označevalnikoma TreeTagger in HunPos, ki sta se naučila modela jezika iz korpusa SETimes.HR. Pojavnice, ki sta jih programa označila različno, so bile z uporabo platforme za množičenje ffzgMnoštvo posredovane množici anotatorjev, ki so izmed obeh izbrali pravilno oznako. Množica je pregledala sporne samostalnike, glagole in pridevnike, medtem ko so eksperti pregledali in popravili tiste oznake, za katere se je množica odločila, da so napačne pri obeh označevalnikih, kot tudi preostale besedne vrste. Evalvacija množičenja je pokazala, da je natančnost posameznega anotatorja v povprečju ∼90%, večinska odločitev treh anotatorjev pa ∼97%. Medtem ko je intrinzina evalvacija vira z izraunom natančnosti oblikoskladenjskih oznak pokazala izboljšanje za 8%, je ekstrinzična evalvacija popravljenega korpusa pri nalogi oblikoskladenjskega označevanja povečala natančnost označevanja za malo več kot 1%. Rezultati kažejo, da je uporaba množičenja za izdelavo in izboljšanje jezikovnih virov koristna, vendar pa ne za izboljšanje oblikoskladenjskega označevanja, kjer bi bilo, glede na količino že dostopnih korpusnih podatkov kot zlatega standarda, moči bolje usmeriti v izdelavo leksikona.

Key words: crowdsourcing, silver standard, morphosyntactic annotation, lemmatization, Croatian language

1. Introduction

Crowdsourcing is a method that has lately been used more and more as a means for collecting data, as well as other kinds of organized effort in reaching certain goals. Whether it is crowdfunding, crowdvoting, crowdtagging or microworking,[1] the basic idea behind this method is that a vast number of people can contribute to a larger goal by doing little work individually. Due to its wide, interdisciplinary applicability, crowdsourcing is used more and more in the field of computer science for tagging data as a prerequisite for machine learning, a method applied in many fields, of which natural language processing is one.

The basic goal of this paper is to minimize the effort of building a large linguistic resource of acceptable quality, representative of the Croatian web. This silver standard corpus would be both lemmatized and morphosyntactically tagged. Though it need not improve any particular application, as simply creating a fresh linguistic resource is a worthwhile goal, we also evaluate the resource on certain natural language processing tasks.

The motivation behind including crowdsourcing into the procedure of building an annotated corpus is to simplify and speed up the process of checking and correcting the tags. The idea is that the crowd, whose work is time efficient and normally cheap or even, as in our case, free, can confirm which tags are correct. Consequentially, the expert, whose work is time consuming and, by comparison, expensive, needs to invest less time into checking the tags, focusing only on correcting those that are incorrect and problematic.

This approach represents a kind of middle ground between two approaches to tagging a corpus for machine learning - the classic approach, which is usually done so that an expert manually annotates raw data with no help (thus creating a so-called gold standard), and the more automated approach, where specialized tools automatically annotate the data, and the expert then later corrects the most probable mistakes (thus creating a silver standard). On the

[1] https://sites.google.com/site/crowdsourcewiki/

one hand, the problem is that manual annotation is time-consuming, tiresome and exhausting, but yields high accuracy rates, whereas on the other hand, tools for automatic tagging, though time-efficient, are imperfect and not precise enough. By putting crowdsourcing into the mix, the latter approach is enhanced, speeding up the expert's job.

The paper is structured as follows: after an overview of related work, follows a section that describes the workflow, covering sample selection and data preparation and our crowdsourcing tool. Section 4 describes the crowdsourcing and expert checks. In Section 5 we give intrinsic and extrinsic evaluation of the produced resource while we end with a conclusion in Section 6.

2. Overview of related work

As far as English is concerned, the problem of (statistical) morphosyntactic annotation is considered solved, as a very high per-token accuracy rate of 97.5% has been achieved (Søgaard, 2011), and though this is a recent development, it is not dramatically higher than the results reached by research in the decade preceding it. However, this is not the case for languages like Croatian, which are morphologically richer and have a looser sentence structure. The problem is actively being worked on and quite some progress has been made by following the statistical modeling paradigm: while earlier work (Agić et al., 2008b) achieved a 86.05% accuracy rate at the morphosyntactic level, but was not made available, the most recent work on the problem reaches 87.72% (Agić et al., 2013), resulting in the SETimes.HR corpus, an annotated corpus of Croatian language, which is publicly available[2], as are the models and test sets used in the paper.[3]

Alongside that, in another paper (Agić et al., 2010) the problem of MSD (morphosyntactic description) tagging is approached a bit differently, by using tagger voting, where the results of about a dozen automatic annotation tools are used as votes for the most likely morphosyntactic description, so that the answer given by the most taggers is considered correct. There is also the work of Peradin and Šnajder that approaches the problem from a different angle, by building rule-based grammars, which achieve an accuracy of 86.36%, but the systems are still in the prototype phase and not available as a ready-to-use tool (Peradin and Šnajder, 2012). Yet a third angle from which to approach the issue of lemmatization and MSD tagging is the approach of using a morphosyntactic lexicon during the annotation process(Agić et al., 2008a), but the result of this research is not publicly available.

Turning to languages related to Croatian, a few papers (Gesmundo and Samardžić, 2012b; Gesmundo and Samardžić, 2012a) dealt with lemmatization and tagging using a statistical approach. The models have been trained on the Serbian Multext East 1984 corpus and achieve an accuracy of 86.65% at the MSD level, but they are limited to the domain they were built on. Work has also been done in the past decade that provides an overview of a rule-based approach to the problem by utilizing NooJ and other similar tools.

However, when it specifically comes to using crowdsourcing for creating language resources and tools, and gathering linguistic data, aside from using it to clean up SloWNet[4] (Fišer and Tavčar, 2013), the Slovenian version of WordNet, such work has not been done on Croatian or other related languages. It is not very widespread when it comes to English either, especially if narrowed down to morphosyntactic tagging and lemmatization. There is a paper (Callison-Burch and Dredze, 2010) that provides an overview of the possibilities that Amazon's Mechanical Turk[5] offers in the field of (computational) linguistics, while there is also an effort to approach the crowdsourcing aspect of gathering linguistic data from a completely different angle – namely turning it into a game. Thus, Phrase Detectives[6] (Chamberlain et al., 2008) was designed, the first game created for collaborative tagging of language data on the internet.

3. Description of the workflow

3.1. Research outline

The basic corpus of text used in this research is a randomly selected sample of 5000 sentences from hrWaC 2.0, the second version of the Croatian Web corpus built from the .hr top-level domain, the construction of which is described in (Ljubešić and Klubička, 2014), and that encompasses some 1.9 billion tokens. Given that the idea is to build a high-quality linguistic resource for standard Croatian, but considering that sentences from the whole of the Croatian web vary in their quality and not all are standard Croatian sentences, the first step was to filter the sample. This task was delegated to the crowd and also served as a pilot testing of the crowdsourcing process, a detailed description of which is contained in section 4.

After the crowd completed the pilot task, the chosen standard sentences were annotated using two tools for automatic lemmatization and morphosyntactic tagging – TreeTagger[7] (Schmid, 1994; Schmid, 1995) and HunPos[8] (Halácsy et al., 2007). The tools were trained on the SETimes.HR corpus (Agić et al., 2013), a gold-annotated corpus of texts[9] collected from the Southeast European Times website[10] and the revised MULTEXT-East v4 morphosyntactic specifications that are used on the SETimes.HR corpus [11] were also used here as the annotation standard. The assumption was that those tokens that the tools annotated identically were tagged correctly, while those that the two taggers disagreed on were problematic. Out of 50322, there were 9965 such problematic tokens and they were passed on to the crowd for tagging in three phases – first the nouns, then the adjectives and finally the verbs. Based on the number of answers and the accuracy of the annotators, 3350 were declared correct.

[2] https://github.com/ffnlp/sethr

[3] http://nlp.ffzg.hr/resources/corpora/setimes-hr/

[4] lojze.lugos.si/darja/slownet.html

[5] https://www.mturk.com/

[6] http://anawiki.essex.ac.uk/phrasedetectives/index.php

[7] http://www.cis.uni-muenchen.de/˜schmid/tools/TreeTagger/

[8] http://code.google.com/p/hunpos/

[9] http://nlp.ffzg.hr/corpora/setimes

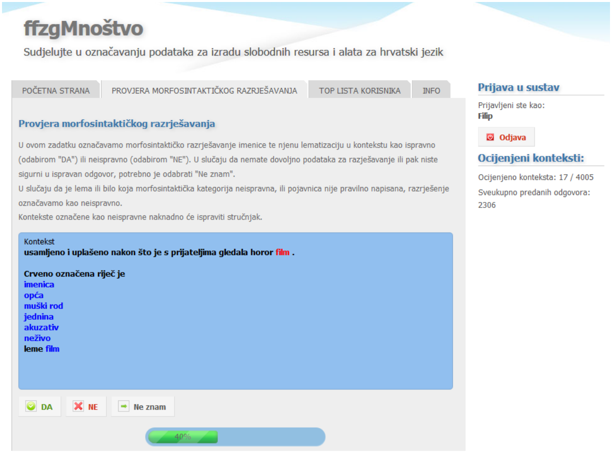
[10] http://www.setimes.com/

[11] http://nlp.ffzg.hr/data/tagging/msd-hr.html

Afterwards, two experts looked over the lemmas and tags that the crowd had tagged as incorrect, as well as those of word classes the crowd did not annotate (such as adverbs, pronouns, prepositions, etc.) and corrected all those that needed correcting. This same approach was used in the creation of the Slovene jos1M corpus (Erjavec and Krek, 2008), only sans crodsourcing. Our corpus was also checked for non-existing tags and for non-agreement between adjectives and nouns, as well as between prepositions and following adjectives or nouns. With that, the silver standard corpus was completed and was then evaluated.

3.2. The ffzgMnoštvo crowdsourcing platform

The ffzgMnoštvo[12] crowdsourcing platform was used as a tagging tool to be used by the crowd. It is actually an adapted version of sloWCrowd (Fišer et al., 2014), adapted by Nikola Ljubešić for the purpose of this research.



ffzgMnoštvo

Sudjelujte u označavanju podataka za izradu slobodnih resursa i alata za hrvatski jezik

Provjera morfosintaktičkog razrješavanja

U ovom zadatku označavamo morfosintaktičko razrješavanje imenice te njenu lematizaciju u kontekstu kao ispravno (odabirom "DA") ili neispravno (odabirom "NE"). U slučaju da nemate dovoljno podataka za razješavanje ili pak niste sigurni u ispravan odgovor, potrebno je odabrati "Ne znam".

U slučaju da je lema ili bilo koja morfosintaktička kategorija neispravna, ili pojavnica nije pravilno napisana, razrješenje označavamo kao neispravno.

Kontekste označene kao neispravne naknadno će ispraviti stručnjak.

Kontekst
usamljeno i uplašeno nakon što je s prijateljima gledala horor film .

Crveno označena riječ je
imenica
opća
muški rod
jednina
akuzativ
neživo
leme film

Prijava u sustav

Prijavljeni ste kao:
Filip

Ocijenjeni konteksti:

Ocijenjeno konteksta: 17 / 4005

Sveukupno predanih odgovora: 2306

Figure 1: ffzgMnoštvo user interface

After registering to the system, users can begin solving the task. They are offered a context that they need to judge, and the possible answers they can choose from are “Yes”, “No” and “Don’t know”. Having registered to the system, users can also see how many answers they’ve given, as well as how many tasks are left in the database to be solved. To make things more interesting, a few gamification elements have been implemented into the platform, such as a progress bar and a hall of fame, which ranks users based on how much they contributed to the project. This is a way to add a healthy dose of competition between the annotators, which further motivates them to participate and solve tasks more regularly, while at the same time making the project more attractive (Chamberlain et al., 2008; Von Ahn, 2006).

4. Crowdsourcing linguistic data

Crowdsourcing via ffzgMnoštvo was done in four phases: 1. checking sentence standard and checking MSDs and lemmas of 2. nouns, 3. adjectives and 4. verbs.

4.1. Annotators

The annotators were exclusively students of the Faculty of Humanities and Social Sciences in Zagreb, attendees of the graduate course Selected Chapters from NLP at the Department of Information and Communication Sciences. The number of annotators varied from phase to phase, depending on how regularly they attended class. A quick demographics overview shows that most of them studied Informatics, Research Track, as a single major at the Department, and three of them had completed their undergraduate studies outside the Faculty. Those who studied a double major, along with Informatics, Research Track, also studied a philological program, be it Linguistics, English, Croatian, etc. Thus the annotators could initially be divided into two groups – those with a formal linguistic education and those without one. But it should be noted that the program at the Department significantly deals with language technologies, so even those who were only a single major actually had some of the required background knowledge. Possible discrepancies were made up for in the course itself, so all the annotators were adequately prepared for solving the task.

This kind of annotator demographic might be considered atypical of crowdsourcing, which usually includes several hundred, if not thousands of annotators, often with much more diverse backgrounds and expertise. However, this research was not intended to be large-scale crowdsourcing, but rather an experiment to measure how well a student group intrinsically motivated to take part in such a project, which is a feasible working force for the future, can solve such problems in a crowdsourcing environment.

4.2. Pilot phase – checking sentence standard

The data preparation phase contained a crowdsourcing pilot test, which, aside from filtering the initial sample, was done to try out the platform and familiarize the users with the system, concept and work principle, as well as to gain insight into how the crowd works and how to best utilize it in the following main phases.

Thirteen annotators were given 5000 sentences to annotate. The question they were asked was “Is the proposed sentence a standard Croatian sentence?” Each user annotated roughly 1000 questions, making up a total of 13167 answers.

One hundred sentences in the database were annotated ahead of time, representing a small gold standard set. The users sometimes got to annotate these golden sentences as well, not knowing the answer was predetermined. Their answers to the golden sentences served to calculate their accuracy and so provide feedback on their reliability – if they answered the gold standard sentences correctly, it can be concluded that they understand the task and that their answers to unannotated sentences are reliable. Calculating their accuracy was simply a matter of dividing the number of correctly answered gold standard sentences with the total number of gold standard sentences the users answered.

It is also important to see how many times a sentence has been tagged, as the goal is to gather as much data as possible by tagging as many sentences as many times possible. In an ideal scenario, each sentence would be tagged by at least two annotators.

The distribution shown in Table 1 shows that sentences were annotated quite a different number of times, i.e. that the variance of the distribution is quite high. This was due

[12] http://faust.ffzg.hr/ffzgmnostvo/

# of answers	# of sentences
0	285
1	941
2	1368
3	1246
4	698
5	462

Table 1: The distribution of the number of sentences by the obtained number of answers

to the algorithms of the ffzgMnotvo platform, and as it is obviously not the most economic way of collecting user responses, an additional intervention was made to the system by defining the number of maximum number of answers per task.

After the crowd tags the data, a final decision has to be made for each sentence on whether or not the crowd deems it standard or non-standard. The decision did not only take into account the distribution of answers, but also the user accuracy rates calculated for this task. So for each sentence that was answered more than 2 times, and there were 3774 such sentences, the accuracy values of the annotators that gave an answer were summed up in favor of that answer. For example, if two users, with accuracy rates of 0.72 and 0.65 said "Yes" to a sentence, and two, with accuracy rates of 0.88 and 0.86 said "No", then the final call for that sentence is "No" (not standard) because 0.88+0.86 $>$ 0.71+0.65. Thus, the crowd decided that 2831 sentences from the initial 5000 sentence set were standard, and these sentences made up the 50322 token corpus. The rest of the sentences were either judged as non-standard (1866 of them), tagged as "Don't know" (only 18), and due to the imperfections of the system, 285 sentences were not tagged even once.

We made an inquiry in the inter-annotator agreement between the users by calculating the Cohen's kappa (Berry and Mielke, 1988), which does not only take into account the observed agreement (Pr(a)), but also accounts for chance agreement (Pr(e)), as seen in equation 1.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \qquad (1)$$

The overall mean of IAA at the sentence standard task is 0.5169, while the mean of observed IAA is 0.7614, showing that there is quite some disagreement between annotators on the issue of sentence standard.

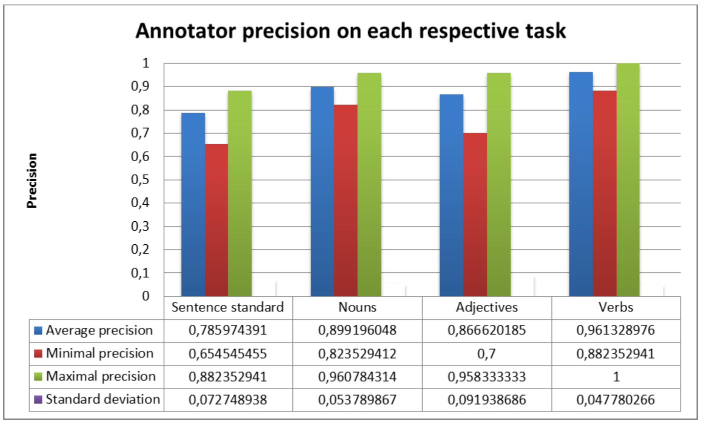
In the end, the sentence standard pilot testing confirmed that the system is applicable to the task, but it also showed that the crowd does not really agree on what a standard sentence is. It seems that whether a sentence is standard or not is open to interpretation and could create an issue when it comes to the quality of the final result. This could be prevented by giving very detailed instructions on what constitutes a standard sentence and by making the question completely unambiguous. Either that, or simply use crowdsourcing for gathering data on issues narrow enough to leave little to no room for interpretation.

4.3. Crowdsourcing MSD and lemmatization

The procedure as described in the former section, but adjusted in accordance with the insight gained from the pilot study, was repeated three more times on three new data sets: 14 annotators tagged 4896 nouns, 8 annotators tagged 2152 adjectives and 6 annotators tagged 478 verbs. The task was presented so that a context was given wherein the token of interest was marked in red, and the annotators were asked to judge the provided morphosyntactic description and the lemma of that token as correct, incorrect or unknown, as depicted earlier in Figure 1.

Of course, the question might arise of why the crowd did not do the whole annotation in the first place, but instead only judged the tags as correct or incorrect. We felt that the task for the crowd cannot be too complex, otherwise the feedback would be too slow and, probably, of low quality. Accordingly, we anticipated that there would not be much gain from delegating the difficult task of MSD tagging to the crowd workers, who are not experts, but rather decided to streamline to process. Furthermore, given the limits of the platform, coupled with the many grammatical categories in play, it would be near-impossible to implement and to properly adjust the interface.

A condensed annotator accuracy analysis shows that a single annotator's accuracy, on average, was about 90%, while that of the crowd collecting three answers was about 97%.



	Sentence standard	Nouns	Adjectives	Verbs
🟦 Average precision	0,785974391	0,899196048	0,866620185	0,961328976
🟥 Minimal precision	0,654545455	0,823529412	0,7	0,882352941
🟩 Maximal precision	0,882352941	0,960784314	0,958333333	1
🟪 Standard deviation	0,072748938	0,053789867	0,091938686	0,047780266

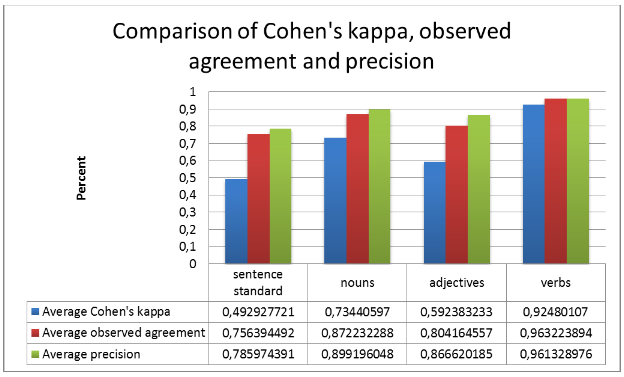
Figure 2: Annotator accuracy on each respective task

A calculation of the average IAA on the morphosyntactic disambiguation and lemmatization task (sentence standard was ignored due to the difference in the nature of the tasks) shows that the average kappa was about 75.05%, while the average observed agreement was 87.99%, as seen in Figure 3.

Annotators agreed much better when it came to MSDs and lemmas, because the task was a lot more unambiguous – if only one of the grammatical categories, or the lemma, was incorrect, the whole thing was to be declared incorrect.

4.4. Expert annotation

After the crowdsourcing was completed, the annotated corpus was split between two experts, whose task was to check and correct the tags that the crowd declared incorrect (2783 nouns, 1416 adjectives and 399 verbs), as well as



	sentence standard	nouns	adjectives	verbs
🟦 Average Cohen's kappa	0,492927721	0,73440597	0,592383233	0,92480107
🟥 Average observed agreement	0,756394492	0,872232288	0,804164557	0,963223894
🟩 Average precision	0,785974391	0,899196048	0,866620185	0,961328976

Figure 3: Accuracy and IAA comparison

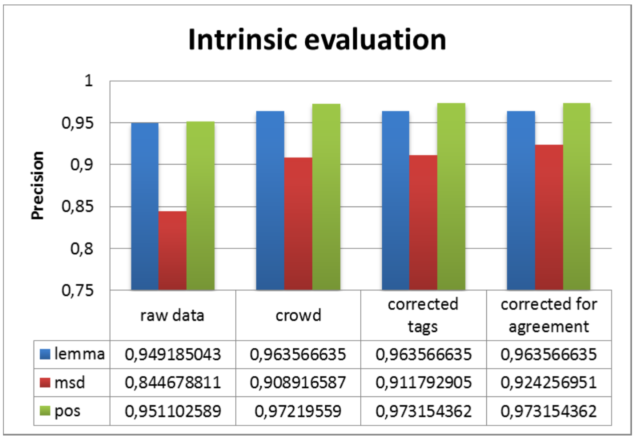
the 2017 disputed tags of word classes the crowd did not annotate (pronouns, adverbs, prepositions, etc.) – making a total of 6615 tags to be checked and, if needed, corrected by the experts. These figures show a 42.12% reduction in the expert's workload thanks to the preceeding crowdsourcing step.

Following the experts' corrections of the tags, two additional steps were taken to further improve the tags – first, the corpus was checked for nonexisting tags (referred to as *corrected tags* in Figure 4), thus ruling out any typos or mistakes the experts might have made, as well as discrepancies and inconsistencies between the annotation standard in the SETimes corpus and the current data set. Second, the corpus was checked for non-agreement – noun phrases that had adjectives that did not agree in gender, number or case with the adjectives or nouns that follow them, as well as prepositions followed by adjectives or nouns that did not share their case (referred to as *corrected for agreement* in Figure 4).

5. Final resource evaluation

The result of the procedure is a lemmatized and morphosyntactically annotated corpus with a total of 50,322 tokens, which has been published and made publicly available on GitHub, along with the accompanying test sets described below.[13] To determine the quality of the final data, the corpus was evaluated on two levels – intrinsic and extrinsic. Intrinsic criteria are those connected to the goal of the system, whereas the extrinsic ones are connected to the system's function (Mollá and Hutchinson, 2003). So by doing an intrinsic evaluation of the corpus, the analysis would look at its accuracy in relation to itself – a sample from the corpus would be manually annotated, representing a gold standard, and it would be compared to that same segment taken from each phase of the corpus construction. Meanwhile, extrinsic evaluation analyzes the corpus' efficiency in a broader context of application, seeing how well it performs in use on some kind of NLP task. Such extrinsic evaluation can be done in at least two ways; either to use the corpus on its own as a resource for building a statistical tagging model, or to merge it with an already existing corpus and analyze its impact in a broader context, as an extension of already existing data for statistical modeling.

For the intrinsic evaluation, 50 sentences were chosen from the corpus of raw data. These sentences were annotated as a gold standard and were compared to the same sentences from every phase of the whole corpus annotation procedure. Three subtasks were taken into account – lemmatization, MSD tagging (providing the full grammatic description) and part-of-speech tagging (providing only the word class), and accuracy served as the evaluation metric. The evaluation showed that the accuracy of the corpus rose by 7.96% at the morphosyntactic level, 1.44% on the level of lemma and 2.2% on the part of speech (POS) level. A more detailed overview by each of the development phases can be seen in Figure 4 showing that the crowdsourcing and expert checking procedure produced most (∼80%) of the overall gain in accuracy.



	raw data	crowd	corrected tags	corrected for agreement
🟦 lemma	0,949185043	0,963566635	0,963566635	0,963566635
🟥 msd	0,844678811	0,908916587	0,911792905	0,924256951
🟩 pos	0,951102589	0,97219559	0,973154362	0,973154362

Figure 4: Results of intrinsic evaluation

The extrinsic evaluation was done in two rounds – first, statistical models were built using the HunPos(Halácsy et al., 2007) tagger, and this was done using data from the corpus in each of its development phases. The models were tried out on a separate test set of 6,429 tokens, or rather 300 sentences, of which 100 were taken from SETimes.HR, 100 from the Croatian Wikipedia and 100 from hrWaC.[14] Again, we used the same three tasks as during the intrinsic evaluation with accuracy as our evaluation metric. When comparing the models' initial and final accuracy on the test corpus, it rose by 0.4% at the part of speech level and 1.09% at the morphosyntactic level, but fell by 0.08% at the level of lemma.

	Raw data	Crowd	Final corrections
lemma	0.925	0.925	0.924
MSD	0.801	0.813	0.812
POS	0.952	0.957	0.952

Table 2: Results of standalone extrinsic evaluation

Second, the raw and final versions of the corpus were merged with an already existing annotated corpus, the SE-

[13] https://github.com/ffnlp/sethr/

[14] The first two data sets were built for (Agić et al., 2013), while web.hr.test was built for the purpose of this paper

Times.HR+[15] corpus and new models were trained on these data sets. Its accuracy on the aforementioned test set at the part of speech level rose by 0.1%, and by 0.17% at the morphosyntactic level, but fell by 0.73% on the level of lemma.

	SETimes+	Raw data and SETimes+	Final corrections and SETimes+
lemma	0.96	0.952	0.952
MSD	0.865	0.853	0.867
POS	0.97	0.969	0.971

Table 3: Results of extrinsic evaluation of expanded corpus

These results show that significant improvement can be achieved by using crowdsourcing for cleaning automatically annotated corpora, but that for the task at hand, given the amount of available gold standard data, minor or no improvement can be achieved.

6. Conclusion

The aim of this research has been fulfilled, as using crowdsourcing has shown itself to be a viable method for creating a silver standard dataset. This claim is backedup by the results of the evaluation performed on the resource. The intrinsic evaluation has shown a great rise in the accuracy of the data, so the positive effect of the crowdsourcing procedure is twofold – along with resulting with a high quality dataset, it also takes some of the weight off of the work the expert taggers do, making the procedure more economical.

The extrinsic evaluation is consistent in different environments – when merged with already existing corpora, the accuracy slightly grows at the MSD and POS levels, while it slightly falls at the level of lemma. These results suggest that the models have reached a plateau and that accuracy will not rise further if the quantity of training data is increased. At such a high accuracy level, it seems that there are so little inaccurate descriptions remaining that they become exceptions which statistical modeling alone cannot handle. Thus, the next logical step is to create a morphological lexicon and pair it with the annotation process, which would improve accuracy significantly.

Concerning adjusting the problem presentation on the crowdsourcing platform to the worker's perspective future work might deal with enhancing and speeding up the process by modifying the order of tasks – an error analysis can be done by looking at the differences between the tags in the initial and final stage of the corpus and then classifying the errors (whether the difference is only in the lemma/gender/case/a certain combination of categories). The tagging could thus be framed as solving groups of similar problems. Such an approach would take less cognitive effort from the annotators and would thus speed up the crowdsourcing.

[15] The SETimes.HR+ corpus is actually the SETimes.HR corpus (Agić et al., 2013) expanded with newspaper articles from various domains, amounting to a total of 135k tokens.

7. References

Ž. Agić, M. Tadić, and Z. Dovedan. 2008a. Combining part-of-speech tagger and inflectional lexicon for Croatian. *Proceedings of IS-LTC*.

Ž. Agić, M. Tadić, and Z. Dovedan. 2008b. Improving part-of-speech tagging accuracy for Croatian by morphological analysis. *Informatica*, 39(32):445–451.

Ž. Agić, M. Tadić, and Z. Dovedan. 2010. Tagger voting improves morphosyntactic tagging accuracy on Croatian texts. In *Information Technology Interfaces (ITI), 2010 32nd International Conference on*, pages 61–66. IEEE.

Ž. Agić, N. Ljubešić, and D. Merkler. 2013. Lemmatization and morphosyntactic tagging of Croatian and Serbian.

C. Callison-Burch and M. Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, page 112. Association for Computational Linguistics.

J. Chamberlain, M. Poesio, and U. Kruschwitz. 2008. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the i nternational conference on semantic systems (I-Semantics 08), Graz.*

T. Erjavec and S. Krek. 2008. The JOS morphosyntactically tagged corpus of Slovene. In *Proceedings of LREC*. Language Resources and Evaluation Conference.

D. Fišer and A. Tavčar. 2013. Več glav več ve: uporaba množičenja za čiščenje sloWNeta.

D. Fišer, A. Tavčar, and T. Erjavec. 2014. sloWCrowd: A crowdsourcing tool for lexicographic tasks. In *Proceedings of LREC*. Language Resources and Evaluation Conference.

A. Gesmundo and T. Samardžić. 2012a. Lemmatising Serbian as category tagging with bidirectional sequence classification. In *Proceedings of LREC*. Language Resources and Evaluation Conference.

A. Gesmundo and T. Samardžić. 2012b. Lemmatisation as a tagging task. In *Proceedings of ACL*. Association for Computational Linguistics.

P. Halácsy, A. Kornai, and C. Oravecz. 2007. HunPos: An open source trigram tagger. In *Proceedings of ACL*, pages 209–212. Association for Computational Linguistics.

N. Ljubešić and F. Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of EACL 2014*. Association for Computational Linguistics.

D. Mollá and B. Hutchinson. 2003. Intrinsic versus extrinsic evaluations of parsing systems. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: Are Evaluation Methods, Metrics and Resources Reusable?*, pages 43–50. Association for Computational Linguistics.

H. Peradin and J. Šnajder. 2012. Towards a constraint grammar based morphological tagger for Croatian. *Text, Speech and Dialogue*, 14:174–182.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

H. Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Association for Computational Linguistics.

A. Søgaard. 2011. Semisupervised condensed nearest neighbor for part-of-speech tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 48–52. Association for Computational Linguistics.

L. Von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.