Error Analysis in Croatian Morphosyntactic Tagging

Željko Agić[*], Marko Tadić[**], Zdravko Dovedan[*]

[*]*Department of Information Sciences*
[**]*Department of Linguistics*
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
{zeljko.agic, marko.tadic, zdravko.dovedan}@ffzg.hr

Abstract. *In this paper, we provide detailed insight on properties of errors generated by a stochastic morphosyntactic tagger assigning Multext-East morphosyntactic descriptions to Croatian texts. Tagging the Croatia Weekly newspaper corpus by the CroTag tagger in stochastic mode revealed that approximately 85 percent of all tagging errors occur on nouns, adjectives, pronouns and verbs. Moreover, approximately 50 percent of these are shown to be incorrect assignments of case values. We provide various other distributional properties of errors in assigning morphosyntactic descriptions for these and other parts of speech. On the basis of these properties, we propose rule-based and stochastic strategies which could be integrated in the tagging module, creating a hybrid procedure in order to raise overall tagging accuracy for Croatian.*

Keywords. Morphosyntactic tagging, part-of-speech tagging, error analysis, error distribution, Croatian language, hybrid tagging

1. Introduction

By definition, morphosyntactic taggers based on stochastic models, such as trigram taggers implementing second order hidden Markov model algorithms (cf. [4]), induce tagging errors. Assigning an incorrect morphosyntactic tag to a wordform given as input occurs for two main reasons (cf. [2]): sparseness of n-gram data in the contextual probability matrix and lack of lexical coverage in the lexical probability matrix. Both of these factors are highly dependent on training corpus size, which could be compensated only by a small margin by using smoothing and unknown word handling methods.

Given certain language and morphosyntactic tagset by which its corpus was annotated, it could be argued, and perhaps more formally investigated, at which point increasing the training corpus (which is a slow and demanding process, requiring expert knowledge) ceases to be economical in terms of increasing overall tagging accuracy. However, learning from our own experience with implementing and utilizing the CroTag trigram tagger [3] and developing natural language processing systems in general (and also not having the luxury to use human resources for further manual morphosyntactic annotation of Croatian corpora), we decided to undergo an experiment which would provide us with a proof for the planned course of action: integrating the core HMM-based tagging module and rule-based (or perhaps even other stochastic) error-correcting procedures into a modular hybrid tagger. One of such courses of action is described in [3].

This experiment approaches the problem from another perspective, which we consider to be somewhat more systematic. We have chosen not to develop generic accuracy-boosting modules which are proven to be valuable to languages similar to Croatian (such is the case with morphological analysis of unknown tokens at trigram tagger runtime in [3]) by default, but instead, we wanted to thoroughly investigate all various properties of errors induced by CroTag tagger when running as HMM. Only then we will choose an appropriate strategy or a set of strategies for handling and correcting those specific errors. These strategies would then become procedures developed specifically for tagging Croatian texts, thus having the advantage of being more finely-tuned than the generic ones.

However, this approach does imply, and moreover rely on, a certain expectation, stating that errors generated by a stochastic tagger indicate certain erroneous patterns, systematic manifestations of flaws contained in the language model created from the training corpus. The remainder of the paper seeks to provide evidence of this statement and consequentially to justify

possible future implementations of modules for handling such manifestations.

Similar research plans might not be especially meaningful for languages with relatively poor morphology and small morphosyntactic tagsets, such as English. However, it was thoroughly conducted for languages similar to Croatian with the same goal of reaching higher overall tagging accuracy, e.g. in [6] and [7] for morphosyntactic tagging of Czech language.

The next section provides short descriptions of language resources and standards used in the experiment along with basic layout of the test cases. Section 3 discusses results we obtained by the experimental framework and section 4 indicates future work directions in terms of strategies for improving overall tagging accuracy in our annotation framework on the basis of presented results.

2. Experiment setup

As in previous experiments with stochastic tagging and improving tagging accuracy for Croatian, the CW100 newspaper corpus was also used in this experiment. Detailed description of the corpus can be found in [1], while table 1 provides only a short overview.

Table 1: Overview of corpus subsets

Set	Tokens	Unique	Tags
Training	106676.10	23426.40	879.60
Testing	11852.90	4638.60	473.20

The corpus is split into ten different parts, equal in number of sentences contained. Nine parts are used for creating the language model for the tagger and the tenth is always used for validating that model. All counts and results are tenfold cross-validated. Table 1 thus states that test sets had 4638.60 unique tokens on average, annotated by 473.20 different morphosyntactic descriptors.

CW100 is annotated using the Multext-East version 3 morphosyntactic tagset specification [5] for Croatian. The tagset is positional, with each of the positions inside tags representing a single morphosyntactic category using different alphabetical characters for denoting different category values. For example a tag Ncmsn would denote a {noun, common, masculine, singular, nominative} token. Position zero always represents part of speech information (PoS), while other tag positions represent morpho-syntactic categories or subpart of part of speech information (sub-PoS). Further in the text, especially in tables, position zero or PoS information is also represented as MSD_0, while other positions or sub-PoS information are referred to as MSD_{1n}.

Table 2 provides a distribution of parts of speech for the cross-validated test sets, indicating their usual distribution in Croatian newspaper texts. Note that other parts of speech in the table also include punctuation, which accounts for their substantial overall count.

Table 2: Distribution of parts of speech on test sets

Type	Count	Percentage
Noun	3547.10	29.93%
Verb	1734.60	14.64%
Adj	1421.20	11.99%
Adp	1135.40	9.58%
Other	4014.60	33.86%

Error analysis is conducted by inspecting differences in morphosyntactic tags that were manually assigned to test sets and the entire CW100 corpus by human annotators and those automatically assigned by CroTag. Investigation encompasses differences in PoS and sub-PoS in general and differences in specific sub-PoS values for the most frequent and the most frequently mistaken parts of speech.

The tagger is trained as a second order HMM, i.e. with the first Markov assumption extended to two discrete time units and with output symbol emissions depending only on current state of the model. We chose this default setting in order to eliminate accuracy bias induced by unknown wordform handlers as described in [3].

3. Results

Experiment results are here presented and discussed in ascending order with regards to their specificity, from the most general to the most specific ones. Table 3 therefore provides overall error count for the test sets.

Table 3: Error count overview

MSD_0 errors	MSD_{1n} errors	Overall
366.50	1510.50	1877.00
3.09%	12.74%	15.83%

Overall tagging accuracy of 84.17 percent for a trigram tagger is as expected. The remaining difference presents overall error count of 15.83

percent, 3.09 being errors on part of speech, i.e. incorrectly assigned PoS value. Once again as expected, a majority of overall errors, more than 80 percent, falls under sub-PoS errors, i.e. errors involving incorrect assignment of values of morphosyntactic categories.

Table 4: Error counts for known and unknown tokens

Type	MSD_0 errors	MSD_{1n} errors
Known	8.84%	50.94%
Unknown	10.69%	29.53%

Table 4 is also used to set the stage for more thorough analysis, as it indicates whether errors occur more often on tokens that were included in the language model of the tagger at training or on those that were not encountered.

It can be clearly seen, somewhat surprisingly, that a majority of sub-PoS errors occur on tokens seen by the tagger during training. This suggests that additional fine-tuning modules for raising tagger accuracy should now emphasize refining and complementing the language model and not dealing with unknown words anymore, as they are already to some extent appropriately handled by module described in [3].

Table 5: Error distribution for parts of speech in Croatian

PoS	MSD_0	MSD_{1n}	All errors
Noun	4.62%	36.04%	40.66%
Adj	4.65%	22.94%	27.59%
Pro	0.40%	8.88%	9.28%
Verb	3.40%	4.77%	8.17%
Adv	3.27%	0.79%	4.06%
Adp	0.49%	3.13%	3.62%
Other	2.68%	3.93%	6.60%
Total	19.53%	80.47%	100.00%

Table 5 presents the distribution of errors in tags in the Croatian language with respect to the parts of speech. Consistent with results in [1], tagger yields a majority of incorrect MSD tag assignments for nouns, adjectives, pronouns and verbs in that descending order, more than 85 percent when combined.

However, perspective gained in [1] is here broadened by counts, indicating that contribution of nouns and adjectives to overall error rate is much more significant than the one of pronouns and verbs due to overall occurrence counts of these parts of speech in the corpus, as already given in table 2. It is also interesting to note, as a side-effect, that most errors in nouns and adjectives and especially pronouns are almost exclusively sub-PoS errors and for verbs the contribution of PoS errors is also substantial with respect to their overall count.

Table 5 sets another course for future handler module implementation, as it is clear from the data how nouns and adjectives should be paid special attention with almost 70 percent of all tagging errors occurring when tagging these parts of speech.

Table 6: Occurrences of sub-PoS errors by position in tag

Position	Count	Percentage
1	115.70	5.99%
2	178.50	9.24%
3	520.00	26.92%
4	653.80	33.85%
5	375.20	19.42%
Other	88.40	4.58%

On the basis of previous sets of conclusions, stating that sub-PoS errors on known wordforms, especially nouns and adjectives, should be given an emphasis in implementing error-correction procedures, table 6 provides an additional perspective on the nature of errors occurring on specific sub-PoS values. In this table, counts and corresponding percentages representing fractions of overall sub-PoS error count are given as a function of position of erroneous value inside MSD tags.

It should be noted once again that position zero represents part of speech value and, as such, it is not explained here but rather separately, in table 8. Distribution in table 6 clearly indicates that a majority of sub-PoS errors, almost 90 percent, occurs on tagset positions 2 to 5.

This table sets another milestone for future work plans concerning tagger improvement, as these tagset positions – position 2 to 4 for nouns and 3 to 5 for adjectives and pronouns – denote morphosyntactic categories of gender, number and case in the Multext-East tagset for Croatian, respectively.

Table 7 contains a short digression from the path set by the previous table, as it presents the distribution of error counts inside single MSD tags. Counts and percentages are given here dependent on number of different errors that occur inside tags. However, this information is also important with regards to experiment goals,

as it states how many of the incorrectly assigned morphosyntactic tags are likely to have a single error inside them and how many could contain multiple errors.

Table 7: Number of errors occurring on single MSD tag

Errors in tag	Count	Percentage
1	1018.30	67.41%
2	323.60	21.42%
3	77.80	5.15%
4	7.00	0.46%
5	0.90	0.06%
Other	82.90	5.49%

It can be seen that the functional dependency is exponentially decreasing, with tags containing only one or two errors making up for almost 90 percent of errors. We could also theoretically combine results given in tables 6 and 7 to state that, even if multiple errors do occur on a single morphosyntactic tag, they are most likely to be distributed on positions 2 to 5, making it easier to handle them. Errors falling further away from the fifth MSD tag position could also be considered less important from a perspective of developing natural language processing systems, as they encode more specific and generally less required morphosyntactic categories.

Table 8 deals with incorrect PoS assignments, i.e. occurrences of incorrect values at position zero in assigned tags. More specifically, as table 5 has shown that a large majority of PoS errors is shared between adjectives, adverbs, nouns and verbs, incorrect assignment map is given only for these parts of speech here.

Table 8: Mapping of incorrectly assigned parts of speech

Error	Adj	Adv	Noun	Verb
Adj	/	44.49%	34.43%	58.14%
Adv	32.58%	/	7.97%	4.46%
Noun	31.53%	10.18%	/	33.91%
Verb	32.77%	3.86%	28.61%	/
Other	3.12%	41.47%	28.99%	3.49%

Some conclusions indicated by this table are rather straightforward. Adverbs are most often mistaken for adjectives (44.49%), nouns for adjectives (34.43%) and verbs for adjectives (58.14%) and nouns (33.91%). PoS errors on adjectives are almost evenly spread between adverbs, nouns and verbs. Incorrect assignments of nouns for other PoS are usually residuals such as foreign names in a large majority of occurrences, while adverbs in this category most often fall under conjunctions. It was also noted that errors from this table are sometimes caused by incorrect tags appearing in the language model, which is in turn caused by errors in manual annotation of the training corpus, making it easy to either link troublesome wordforms to corresponding parts of speech at tagger runtime or maybe semi-automatically correct the training corpus before utilizing it in the training procedure. Even though PoS errors make up for only 3 percent of overall errors, they are the most significant in terms of transferring incorrect information to the user or another system, as it is intuitively clear that saying an adjective is a noun introduces more noise than saying that a specific noun is in nominative case when it is actually in accusative case. However, this is in fact possible to define more precisely only with regards to specific user or system requirements and all errors should receive equal treatment in this experiment in order to enable specific treatments for specific users or systems afterwards.

Table 9: Error distribution for several morphosyntactic categories

Category	Adj	Noun	Pro
Type	1.49%	6.83%	2.17%
Gender	32.30%	15.90%	24.75%
Number	18.40%	22.71%	14.92%
Case	37.07%	54.56%	43.83%
Other	10.74%	0.00%	14.33%

Table 9 provides for sub-PoS what table 8 provided for PoS: a distribution of errors in morphosyntactic categories for most error-prone parts of speech. With regards to table 5, these are adjectives, nouns and pronouns. Errors are here distributed over several specific morphosyntactic categories, which fortunately have identical meanings for the given parts of speech, making it easier to present and discuss them together.

As expected on basis of table 6, most of the category value errors occur on gender, number and case. For adjectives, gender and case equally dominate the distribution, while values presented as other most often indicated errors in category called animateness, since even human annotators often dropped it from annotation. Majority of errors in nouns occurs in case category, similar to pronouns. Case is followed by gender and only then by number in descending order for

adjectives and pronouns, while for nouns number preceded gender. This data also implies certain strategies with regards to specific requirements, as focus could be given to case over gender by default and otherwise if needed in a specific application.

Table 9 is complemented by distributions of specific errors for each of the categories from this table. More precisely, another important deliverable of this experiment is a set of tables indicating incorrect mappings of one category value to another for each of the morphosyntactic categories. For example, these mappings contain information on how often nominative is mistaken for accusative in noun case category and how often if a masculine adjective said to be feminine and neuter. However, given the large size of these mappings and tight space constraints for this paper, we choose not to provide the entire distribution here. Instead we discuss observations we consider to be the most important given our specific future intentions and provide a sample distribution for morphosyntactic category of case for nouns in table 10.

Table 10: Sample error distribution of case category pairs for nouns

Correct value	Incorrect assignment	for Noun
Accusative	Genitive	9.09%
	Nominative	16.15%
Genitive	Accusative	6.04%
	Nominative	7.36%
Nominative	Accusative	20.77%
	Genitive	9.75%
Dative	Locative	9.75%
Instrumental	Locative	3.86%
Locative	Dative	2.84%
	Instrumental	2.59%
Other		11.78%

In gender category in adjectives, errors are most often encountered on the masculine-feminine pair of values in both directions, followed by masculine-neuter pair with incorrect assignments of masculine to feminine and neuter to masculine adjectives being the most frequent ones, but only by a small margin. On the other hand, these figures are somewhat different for nouns, where the masculine-feminine pair is more accentuated in the distribution, always making up for more than 50 percent of all gender errors. Gender distribution for pronouns is the least useful as it is flat, with practically identical counts for all value pairs. Regarding the number category on all three parts of speech, incorrect assignments of plural to singular occur more often than in the other direction for that pair, especially for nouns. Case is the most indicative in terms of invalid assignment pairs and it follows the same pattern for all three parts of speech. On average, more than 70 percent of such error pairs are distributed within a 3-tuple containing nominative, genitive and accusative case, with incorrect mappings of what should be nominative case into accusative and genitive case governing the distribution for nouns and adjectives. For pronouns, all these distributions, including the one for case value, are generally more sparse and inconclusive, most probably due to overall frequency of pronouns in the corpus and test sets.

Experiment [6] and especially [7] conducted for morphosyntactic tagging of Czech language, using various tagsets and taggers differing from the pair utilized in our experiment with Croatian texts, provided highly correlated distributions of errors for adjectives, nouns and pronouns, with a high majority of errors occurring precisely on values denoting their case, gender and number in that particular order. This fact in turn implies another hypothesis requiring verification, stating that similar distributions of error occurrences in morphosyntactic tagging do propagate through similar languages, regardless of tagsets and morphosyntactic taggers used in processes of their annotation. Also, from another perspective, high correlation of these results for Croatian and Czech language indicates the applicability of method and software developed for purposes of this experiment in conducting morphosyntactic tagging error analyses for other languages.

4. Conclusions and future work

Stochastic morphosyntactic tagging, namely trigram tagging or second order hidden Markov model tagging as implemented by the CroTag tagger, is governed and limited by probability matrices, smoothing procedures and unknown wordform handlers. Generic approaches to improving its efficiency in terms of achieving higher overall accuracy figures, generally include (a) tagger output combination and tagger module integration, creating either stochastic cascades or hybrid combinations of stochastic and rule-based procedures and (b) additionally complementing or improving the language model by more fine-tuned smoothing or unknown wordform handling

procedures, possibly implemented for specific languages or sets of language, as is the case for morphological analysis module described in [3].

Results of this experiment might suggest other improvement options available for tagging the Croatian language exclusively, but probably also extendable to other languages implementing similar tagsets. Both stochastic and rule-based approaches could be implemented for handling various observed regularities in error-yielding behavior of our trigram tagger, always on basis of specific requirements. Additional stochastic modules might include training a second order HMM module on sequences of morphosyntactic category values and using it for calculating probabilities of, for example, gender or case sequences in a sentence or text and replacing subsequences of low probability with the ones that are more likely to occur. Rule-based handlers might be implemented to deal with certain patterns of specific wordforms or wordform n-tuples causing specific n-tuples of errors on morphosyntactic categories to appear. For example, specifically for Croatian and on basis of figures provided in this experiment, occurrences of adjective and noun sequences could be forced to agree in gender, number and case by an external procedure if incorrectly assigned by the tagger. Also, as mentioned in the previous section, the software developed for purposes of this experiment could be further improved, documented and made available to other researchers having the same objectives as presented here for Croatian language tagging. These directions are all ready for future research.

5. Acknowledgements

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-1776 and 130-1300646-0645.

6. References

[1] Agić Ž, Tadić M. (2006). Evaluating Morphosyntactic Tagging of Croatian Texts. Proceedings of the Fifth LREC. ELRA, Genoa-Paris 2006.

[2] Agić Ž, Tadić M, Dovedan Z. (2008). Investigating Language Independence in HMM PoS/MSD-Tagging. Proceedings of the 30th ITI. Cavtat, Croatia, pp. 657-662.

[3] Agić Ž, Tadić, M, Dovedan Z. (2008). Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. Informatica, 32:4, pp. 445-451.

[4] Brants T. (2000) TnT – A Stochastic Part-of-Speech Tagger. Proceedings of ANLP.

[5] Erjavec T. (2004). Multext-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the Fourth LREC. ELRA, Lisbon-Paris 2004.

[6] Hajič J, Vidova-Hladka B. (1998). Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. Proceedings of COLING-ACL Conference, pp. 483-490.

[7] Vidova-Hladka B. (2000). Czech Language Tagging. Doctoral thesis, Charles University, Prague, 2000.