DISAMBIGUATION OF HOMOGRAPHIC ADJECTIVE AND ADVERB FORMS IN CROATIAN TEXTS

DANIJELA MERKLER
DAŠA BEROVIĆ
ŽELJKO AGIĆ

Abstract

We propose a NooJ morphosyntactic tagging postprocessing module for resolving problems which arise in disambiguation of adjectives and adverbs that, due to morphology of the Croatian language, appear in the same form. Specifically, forms of descriptive adjectives in the nominative singular case in the neuter gender are the same as the forms of the adverbs that are made from those adjectives by suffixation. We have implemented a set of local grammars for correcting such occurrences and applied it on two corpora of Croatian texts and manually evaluated the module. We suggest the inclusion of such a model of disambiguation to existing pipelines of processing Croatian text and developing Croatian language resources in order to improve the efficiency and accuracy of those tools and procedures.

Introduction

Parallel and comparable language resources for Croatian are still not sufficiently developed. Since Croatian language is included in the ACCURAT project (URL http://www.accurat-project.eu) – whose main goal is to overcome problems of machine translation and to account for the lack of linguistic resources for under-resourced areas of machine translation – the lack of resources became even more evident when pairing Croatian with other languages included in the project. The problem is especially apparent in cases when the other language is under-resourced as well. This was the reason why we needed to put special emphasis on high annotation quality for existing language resources for Croatian.

The motivation for the work presented in this paper were manually detected morphosyntactic tagging and lemmatization errors in the existing recourses for Croatian, such as the Croatian National Corpus v2.5 (cf. Tadić 2002; 2009), which was automatically lemmatized and MSD-tagged using the CroTag stochastic tagger and lemmatizer (Agić et al. 2008; 2009), its manually annotated subcorpora, the Croatian Dependency Treebank (Tadić 2007) and other underlying resources. Manual analysis showed regular patterns in these annotation errors. A substantial part of these invalid annotations were errors in lemmatization and MSD-tagging of homographic forms of descriptive adjectives in the nominative singular case in the neuter gender and the forms of adverbs that are made from those adjectives by suffixation. In these cases, adverbs were lemmatized and MSD-tagged as adjectives. Since the realization of adverbs is co-text dependent, we noticed several types of patterns that were typical for the appearance of such adverbs. Therefore we used NooJ (Silberztein 2003, 2004, 2005) grammars as a solution for disambiguation of homographic adverbs and adjectives.

In this paper, we present design and implementation of four NooJ local syntactic grammars from four patterns that we noticed in manual corpus analysis and their application and evaluation and manual evaluation on two selected corpora of Croatian: the Croatia Weekly 100 kw newspaper corpus and Orwell's *1984* corpus (cf. Agić et al. 2011), both manually lemmatized and MSD-tagged using the Multext East morphosyntactic specifications (cf. Erjavec 2010). In designing and implementing the NooJ local grammars, we have consulted Croatian grammars and research related to the topic of adverbs and adjectives in Croatian (Pranjković 1992, Barić et al. 2005, Silić and Pranjković 2005).

Experiment setup

As shortly mentioned in the introduction, for purposes of this experiment, we made available two manually annotated corpora of Croatian texts – the Croatia Weekly 100 kw newspaper corpus (cw100 further in the text) and the Croatian translation of George Orwell's novel *1984* as a part of the Multext East v4 (MTE v4 further in the text) specification and resources. In this section, we present the corpora in more detail, along with a short insight on the experimental setup for testing the local grammars we developed in NooJ for purposes of automatical correction of annotation errors.

The cw100 newspaper corpus consists of articles extracted from seven issues of the Croatia Weekly newspaper, which has been published from 1998 to 2000 by the Croatian Institute for Information and Culture (HIKZ). This corpus is a part of Croatian side of the Croatian-English Parallel Corpus, described in detail in (Tadić 2000). The cw100 corpus was pre-tagged using the MTE v3 morphosyntactic specifications on top of XCES corpus encoding standard.

	cw100	**orwell**
sentences	4,626	6,625
tokens	118,529	106,632
tags	896	802

Table 1. Basic corpora stats

The Croatian translation of the *1984* corpus was just recently made available within the Multext East v4 project (cf. Agić et al. 2011). Similar to the cw100 corpus, it was semi-automatically lemmatized and MSD-annotated according to the MTE v4 specification and is XML-encoded according to the TEI recommendations (URL http://www.tei-c.org/).

Basic corpora stats and distributions of parts of speech for both corpora are given in tables 1 and 2, respectively.

part of speech	**cw100**	**orwell**
abbreviation	1,312	56
adjective	14,300	9,244
adpostition	11,314	7,760
adverb	4,603	7,881
conjunction	8,276	8,553
interjunction	7	45
noun	36,093	20,524
numeral	2,178	1,177
other	15,368	14,426
particle	547	1,680
pronoun	7,307	11,627
verb	17,224	23,659
total	**118,529**	**106,632**

Table 2. Distribution for parts of speech on the available corpora

In order to automatically validate the manual annotation with MSD-tags in these two corpora within NooJ, we had to import both the text layer and the annotation to NooJ. Firstly we converted the corpora to plaintext

three-column format and then created a simple script to wrap it in NooJ XML (cf. Silberztein 2003) and convert the MTE v3/v4 morphosyntactic tags to NooJ-style annotation (cf. Vučković et al. 2010). The resulting corpus encoding is illustrated in Figure 1.

The local grammars developed for automatic correction of annotation on the level of morphosyntactic tags, as described in the following chapter, were evaluated within NooJ on the two corpora. Each of the grammars was run on each of the corpora and their precision was calculated by definition, i.e. as a fraction of valid and all applied corrections.

```
<Z>
  <LU LEMMA="samo" CAT="R" Type="na">Samo</LU>
  <LU LEMMA="X" CAT="X">ne</LU>
  <LU LEMMA="soba" CAT="N" Type="c" Gender="f" Nb="p" Case="G">soba</LU>
  <LU LEMMA="X" CAT="X">sto</LU>
  <LU LEMMA="X" CAT="X">jedan</LU>
  <LU LEMMA="X" CAT="X">!</LU>
</Z>
<Z>
  <LU LEMMA="soba" CAT="N" Type="c" Gender="f" Nb="s" Case="N">Soba</LU>
  <LU LEMMA="X" CAT="X">sto</LU>
  <LU LEMMA="X" CAT="X">jedan</LU>
  <LU LEMMA="reći" CAT="V" Type="main" Tense="PDR" Nb="s">rekao</LU>
  <LU LEMMA="biti" CAT="V" Type="pg" Tense="PR" Person="3" Nb="s">je</LU>
  <LU LEMMA="časnik" CAT="N" Type="c" Gender="m" Nb="s" Case="Nom">časnik</LU>
  <LU LEMMA="X" CAT="X">.</LU>
</Z>
```

Figure 1. Sample from *1984* in NooJ XML format

Results and discussion

In the course of manual analysis of the presented corpora, we have detected four distinct patterns in which adverbs that are homographic with adjectives occur and designed local grammars in NooJ to detect these patterns. In these grammars, we searched for adjectives in the nominative singular case in the neuter gender which should be marked as adverbs. They are determined by their cotextual environment.

The grammars and results of their evaluation are presented further in this section.

Vpg + A* + V

This grammar searches for complex tenses, i.e. tenses that consist of an auxiliary verb and a main verb, but in this case an adverb also occurs between them. We set a constraint that the main verb should not be an

infinitive and that it should not be in any form of the lemma *biti* (en. *to be*) because it would then be a part of a nominal predicate.

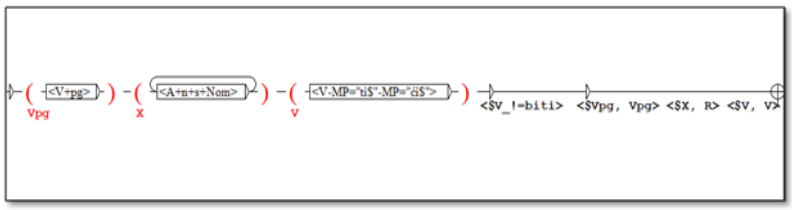

Figure 2. Pattern Vpg + A* + V in NooJ editor

Vpg + A + A*

In this grammar, an adverb occurs between components of a nominal predicate, in this case between an auxiliary verb and an adjective. However, in order for the word form previously annotated as an adjective to have the actual function of an adverb, it should not agree in gender, number and case with the adjective which is a part of the nominal predicate.

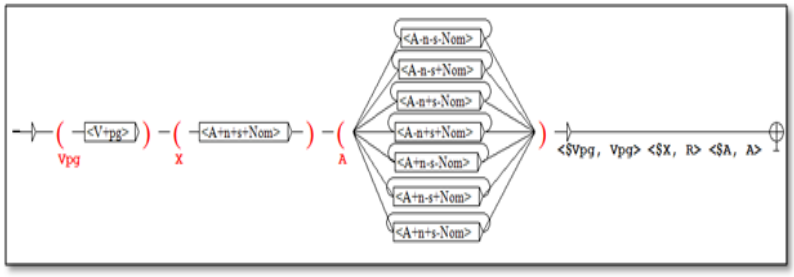

Figure 3. Pattern Vpg + A + A* in NooJ editor

A* + V

With this grammar, we tried to detect what is considered to be the most frequent or common occurrence of an adverb, namely, its occurrence with a verb. We had to set a constraint that a word form must occur before the adverb, i.e. a lexical unit which is not an auxiliary verb must be detected, in order to avoid the results with complex tenses that we obtained with the first local grammar. Besides, we also set a constraint that the verb should not be in infinitive so it would not be part of a nominal predicate.

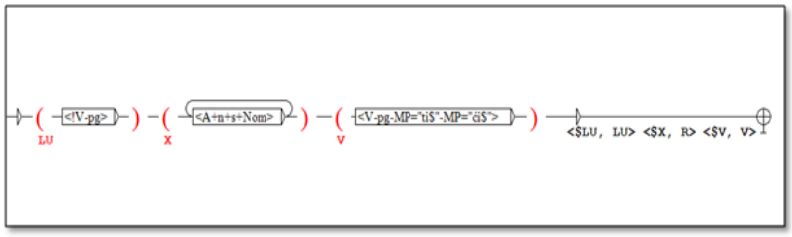

Figure 4. Pattern A* + V in NooJ editor

A + A* + N

This grammar mainly detects the adverbs that occur as intensifiers of an attribute in noun phrases. That is why we set the constraint that the noun and the adjective which constitute the noun phrase should agree in gender, number and case, but at the same time should not agree in these morphosyntactic categories with the previous word form falsely annotated as adjective, i.e. the adverb.

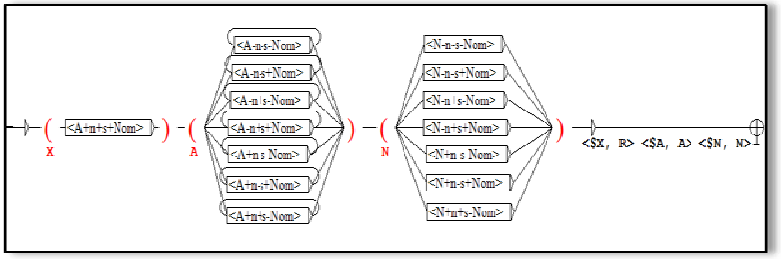

Figure 5. Pattern A + A* + N in NooJ editor

Concordandes obtained by applying these four correction rules were manually evaluated according to the evaluation plan and the obtained results are presented in Table 3.

The table shows that the second rule – the one that detects falsely annotated adverbs appearing as parts of nominal predicates – is 100% correct. The first rule performed poorly due to the fact that the separation of word forms annotated as verbs may appear simply as a consequence of relatively free word order in Croatian and is thus not an exact indicator of false annotation. The third rule performed substantially better on newspaper texts than on fiction being that its generic design was much better suited for non-literary texts. The fourth rule performed consistently on both of the available corpora.

rule	cw100	orwell	cw100 + orwell
Vpg + A* + V	0.64	0.62	0.63
Vpg + A + A*	1.00	1.00	1.00
A* + V	0.82	0.54	0.67
A + A* + N	0.69	0.75	0.70
total	**0.77**	**0.61**	**0.70**

Table 3. Accuracy for first version of grammars

We noticed that many of the incorrect results included the word *sve* – which can be a pronoun, an adjective or a particle in Croatian – so we upgraded all grammars in order not to recognize word *sve*. Beside the target word having to be an adjective in the nominative singular case in the neuter gender, we added the constraint that it must not be the word *sve*. The upgraded grammars are illustrated in Figure 6.

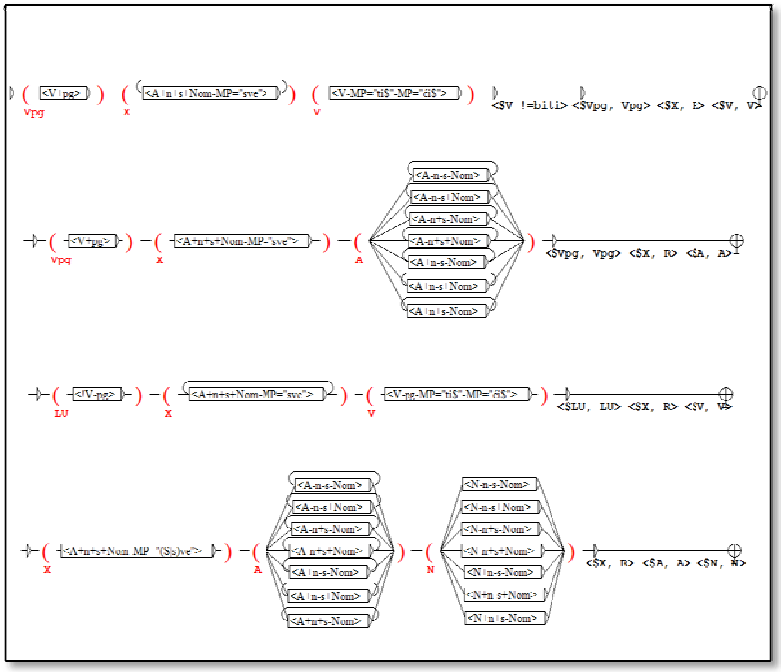

Figure 6. Upgraded grammars

The results after applying new, upgraded grammars are presented in Table 4. They show certain improvements, especially for the first rule, where the obtained results in the newspaper corpus are 100% correct.

Besides, 100% correct result is obtained in the fourth pattern as well, but only for the literature corpus. In total, there is a significant difference between newspaper and literature corpus in favour of newspaper corpus. As previously stated, this is most likely due to the elaborately generic design of the rules, targeting the most common usages of language constructions.

rule	cw100	orwell	cw100 + orwell
Vpg + A* + V	1.00	0.83	0.92
Vpg + A + A*	1.00	1.00	1.00
A* + V	0.87	0.63	0.74
A + A* + N	0.78	1.00	0.82
total	**0.89**	**0.73**	**0.83**

Table 4. Accuracy for upgraded grammars

Conclusions and future work

We have designed, implemented and validated a NooJ module and framework for automatic correction of errors in morphosyntactic annotation of Croatian texts concerning adjectives falsely annotated as adverbs by human annotators and the CroTag tagger and lemmatizer. Overall module performance on both newspaper texts and fictional texts was shown to be of adequate quality for usage in pipelines for automatic processing of Croatian, as well as for purposes of ongoing semi-automatical development of new language resources for Croatian. Future work directions include

(1) wrapping the module with the CroTag tagger/lemmatizer as a post-processing procedure and performing a detailed error analysis of their combination and
(2) refining the prototype rules presented here in order to perform more consistently on different genres and domains.

Acknowledgement

The research within the project ACCURAT leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no. 248347.

References

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

Agić Ž, Tadić M, Dovedan Z. (2009). Evaluating Full Lemmatization of Croatian Texts. *Recent Advances in Intelligent Information Systems*. Warsaw, Academic Publishing House EXIT, 2009, pp. 175-184.

Agić Ž, Merkler D, Berović D, Tadić M. (2011). Development and Applications of the Croatian 1984 Corpus for the MULTEXT-East Resources. *Proceedings of SlaviCorp 2011 – The Second Conference on Slavic Corpora*, 2011, in press.

Barić E and associates. (2005). *Hrvatska gramatika* [Grammar of Croatian]. Zagreb, Školska knjiga.

Erjavec T. (2010). MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Proceedings of the Seventh Conference on International Language Resources and Evaluation – LREC 2010*.

Pranjković I. (1992). Prilozi kao „riječi sviju vrsta" [Adverbs as „words of all types"]. *Suvremena lingvistika*, vol. 18, no. 2, pp. 243-249.

Silberztein M. (2003). *NooJ manual*. Available at the website http://www.nooj4nlp.net (last accessed 2011-06-10).

Silberztein M. (2004). NooJ: an Object-Oriented Approach. INTEX pour la Linguistique et le Traitement Automatique des Langues, C. Muller, J. Royauté M. Silberztein (eds), *Cahiers de la MSH Ledoux*. Presses Universitaires de Franche-Comté, pp. 359-369.

Silberztein M. (2005). NooJ's Dictionaries. *Proceedings of the 2nd Language and Technology Conference*, Poznan University, 2005.

Silić J, Pranjković I. (2005). *Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta* [Grammar of Croatian for High Schools and Universities]. Zagreb, Školska knjiga.

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

Tadić M. (2002). Building the Croatian National Corpus. *Proceedings of LREC 2002*. ELRA, Paris-Las Palmas, vol. II, pp. 441-446.

Tadić M. (2007). Building the Croatian Dependency Treebank: the initial stages. *Suvremena lingvistika*, 33 (2007) , vol. 63, pp. 85-92.

Tadić M. (2009). New version of the Croatian National Corpus. *After Half a Century of Slavonic Natural Language Processing*. Brno, Masaryk University, 2009, pp. 199-205.

Vučković K, Tadić M, Bekavac B. (2010). Croatian Language Resources for NooJ. *CIT – Journal of computing and information technology*, 18 (2010), pp. 295-301.