# Building a morphosyntactic lexicon for Serbian using Wiktionary

Aleksandra Miletic

# Building a morphosyntactic lexicon for Serbian using Wiktionary

**Aleksandra Miletic**
UMR 5263 CLLE, CNRS & University of Toulouse
Maison de la Recherche, 5, allée Antonio Machado
31 000 Toulouse, France
aleksandra.miletic@univ-tlse2.fr

## Abstract

Creation of MS lexica often relies on exploiting existing traditional lexical resources or extracting lexical information from raw or annotated corpora. However, this approach is problematic in the case of under-resourced languages like Serbian, for which the starting points for these methods are by definition scarce. An alternative method consists in using new collaborative resources made possible by the evolution of the Internet. The lexicon described in this paper was derived from one such resource: Wiktionary for serbo-croatian. We show that, although the resulting lexicon does not have perfect coverage, this approach allowed us to have a relatively rapid building process resulting in a resource under a non-restrictive license. Some enrichment techniques have also been used in an effort to extend the lexicon coverage.

**Keywords:** morphosyntactic lexicon, under-resourced languages, Serbian, Wiktionary

## 1 Introduction

This paper presents the creation of a morphosyntactic lexicon for Serbian derived from several resources: the Wiktionary edition for Serbo-Croatian, a manually POS-tagged corpus, and specialized preposition lists. This work is part of a larger effort to transform ParCoLab (Stosic, 2015), a parallel corpus of Serbian, English, and French, into a syntactic treebank.

English and French already boast a variety of different NLP resources: lexical and morphological resources (e.g. Clément et al., 2004, Romary et al., 2004, Sajous et al., 2013 for French; Fellbaum, 1998, Brierly et al., 2008 for English), POS-tagging methods (e.g. Shen et al., 2007 for English, Denis & Sagot, 2009 for French) and parsers (McDonald et al., 2006, Nivre et al., 2006 for English and French; Urieli 2013 for French). This makes parsing English and French subcorpora of ParCoLab much more straightforward than the annotation of their Serbian counterpart. Serbian is a Slavic language with rich inflectional morphology and flexible word order. This type of languages typically represents a challenge for NLP, and Serbian is not an exception. Despite the recent developments in POS-tagging and lemmatization (Gesmundo & Samardzic, 2012) and first experiments in parsing (Jakovljevic et al., 2014), it can still be considered as an under-resourced language: no training corpora for parsing are available and, to the best of our knowledge, the only freely available morphosyntactically tagged training corpus is still the *cesAna* corpus from the MULTEXT-East project (Krstev et al., 2004a). Although morphological and lexical resources for Serbian are referenced in previous works (Krstev et al., 2004a, 2004b), up to now we have been unable to gain access to them[1]. Consequently, the project of transforming ParCoLab into a syntactically annotated corpus demands an intensive resource-building campaign, in which the creation of a lexicon containing the information relevant to morphosyntactic and syntactic analysis is one of key parts.

The lexicon presented here contains 1 226 638 million wordforms for 117 445 lemmas, corresponding to a total of 3 066 214 unique triples <*wordform*, *lemma*, *morphosyntactic description*>. It is thought for NLP applications such as POS-tagging and parsing. It is downloadable under the Creative Commons BY-SA 3.0 license at the following address: http://redac.univ-tlse2.fr/.

## 2 Related works

Creation of morphosyntactic lexica often relies on exploiting existing lexical resources or extracting lexical information from raw or annotated corpora (cf. Clément et al., 2004, Sagot, 2005). Although previous works reference a morphosyntactic lexicon (Krstev et al., 2004a) and a morphological dictionary in Intex format (Krstev et al., 2004b) for Serbian, we have been unable to gain access to them up to now. Furthermore, they are subject to a no redistribution license, and our goal is to create a freely accessible set of NLP tools for Serbian. As for corpora-based methods, ParCoLab was not large enough for this method to be efficient (1,6M tokens in the Serbian subcorpus) at the beginnig of this work. We therefore looked for alternative methods, which led us to an existing lexical resource for Serbian: Wiktionary.
Wiktionary is a collaborative dictionary launched in 2002. Today it exists in 158 languages, and entries can

[1] It should be noted that a new lexicon for Serbian was published (Ljubesic et al., 2016) after the completion of the work presented here. It will be taken into account in our future work.

contain definitions, as well as information on pronunciation, inflection, semantically related words, translations into other languages, etc. This makes Wiktionary a valuable resource for NLP, but the fact that it is created through crowd-sourcing can put in question the quality of its content and the quality of the resources derived from it. However, several works have shown that resources based on crowd-sourcing can yield results that are competitive or even better than those obtained from resources built by experts (Strube & Ponzetto, 2006, Gabrilovich & Markovitch, 2007, Zesch et al., 2007, Zesch & Gurevych, 2010). Since the first works on Wiktionary in 2008, its suitability for NLP research seems to have become an accepted fact: it has been used to measure semantic relatedness between words (Zesch et al., 2008), create synonymy networks (Navarro et al., 2009), build or enrich ontologies (Meyer & Gurevych, 2012, Pérez et al., 2011), and derive morphosyntactic lexicons (Sajous et al., 2013, Sagot, 2014, Senrich & Kunz, 2014).

As this approach is low-cost compared to a manual creation process, it is especially useful where time and human resources are scarce. It is even more so for low-density languages, for which other starting points for resource derivation can be difficult to find. Both conditions apply to our case. Using Wiktionary also allowed us to have a relatively rapid building process and a resulting resource under a non-restrictive license.

## 3 Building process

The base for our lexicon was derived from the Wiktionary edition for Serbo-Croatian. Two Wiktionary editions treating Serbian content exist: the Serbo-Croatian one (sh.wiktionary.org) and the Serbian one (sr.wiktionary.org). This seems to be due to extra-linguistic rather than linguistic factors. We chose the Serbo-Croatian edition for its size: 850 000 entries vs. 45 000 in sr.wiktionary.org. Since the lexicon will be used to parse ParCoLab, we focused on getting morphosyntactic information, especially the case, number and gender, as they are essential for syntactic analysis of Serbian.

Wiktionary is made publicly available through periodic XML data dumps. We used the sh.wiktionary.-com dump from October 02 2015. It should be noted that only the macrostructure of the pages is encoded in XML, whereas the page content is rendered in wikicode, a very flexible, under-specified text-based format. Since no systematic description of the wikicode syntax is available, building a parser for wikicode needs to be done through meticulous observation of the pages. As noted in (Navarro et al., 2009, Sajous et al., 2013), the page structure in different language editions varies substantially and a wikicode parser developed for one language cannot be simply transported to a different wiktionary edition. This is why it was necessary to develop a new parser for the Serbo-Croatian Wiktionary.

Another difficulty in parsing Wiktionary stems from the fact that different encoding conventions can coexist within one dump. For example, there are two main page types in the Serbo-Croatian Wiktionary: lemma-based, which gives the complete inflectional paradigm of a lemma in a table (cf. Figure 1), and wordform-based, in which the entry is an inflected wordform, for which all the possible morphosyntactic interpretations are given (cf. Figure 2).

In the first format, the morphosyntactic properties of each form are either given through codes or need to be deduced from the position of the form in the table (typically the case of nouns, cf. Figure 1). This is possible because the tables follow the generally accepted case ordering for Serbian. However, some articles were found where the instrumental and the locative forms switched places. In order to ensure the correct case information, our parser performs a rule-based check to verify that the supposed case corresponds to the wordform ending.

```
==== Deklinacija ====
{{sh-imenica-deklinacija2
|jezik|jezici
|jezika|jezika
|jeziku|jezicima
|jezik|jezike
|jeziče|jezici
|jeziku|jezicima
|jezikom|jezicima
}}
```

Figure 1: Lemma-based article

In the wordform-based format, the information is given in the form of textual descriptions (cf. Figure 2). The order of the elements is not fixed, and some pieces of information can be missing. The parser needed to be flexible enough to manage this diversity in order to extract as much data as possible.

```
=== Flektirani oblici ===
'''guvernerskim'''

# instrumental množine ženskog roda pozitiva
određenog vida pridjeva
[[guvernerski#Srpskohrvatski|guvernerski]]
# lokativ množine ženskog roda pozitiva
određenog vida pridjeva
[[guvernerski#Srpskohrvatski|guvernerski]]
# dativ množine muškog roda pozitiva
određenog vida pridjeva
[[guvernerski#Srpskohrvatski|guvernerski]]
# instrumental množine muškog roda pozitiva
određenog vida pridjeva
```

Figure 2: Form-based articles

Other resources at our disposition were used to improve the coverage. 107 prepositions were imported from manually created lists resulting from previous theoretic work on spatial relations in Serbian (Stosic, 2001). 76 additional prepositions, 43 conjunctions, 33 interjections and 868 adverbs were extracted from the manually POS-tagged part of the ParCoLab corpus (Miletic, 2013) and integrated in the lexicon.

## 4 Quality of the lexicon

The lexicon presented here contains 1 226 638 million wordforms for 117 445 lemmas, corresponding to a total of 3 066 214 unique triples <*wordform, lemma, morphosyntactic description*>. For comparison, GLAFF, a lexicon for French derived from the wiktionary, contains 1 425 848 wordforms corresponding to 186 082 lemmas. Our lexicon is in plain text format as illustrated in Figure 3. The first column contains the inflected wordform followed by one or more morphosyntactic descriptions (MSDs). The structure of the MSDs is POS-specific, but in each case the first slot indicates the POS, and the last one the lemma, while the intervening slots encode values of different morphosyntactic properties.

```
trag    N_m_nom_sg_trag N_m_acc_sg_trag
traga    V_Present_3_sg_0_tragati
N_m_gen_sg_trag
tragah  V_Imparfait_1_sg_0_tragati
tragahu V_Imparfait_3_pl_0_tragati
tragaj  V_Imperatif_2_sg_0_tragati
```

Figure 3: Lexicon format

The values of the MS properties are given in a relatively explicit format in order to facilitate the manual verification. Another version of the lexicon with MSDs in the more standard MULTEXT-East format (Krstev et al., 2004a) will also be made available.
The MS properties for inflected categories are given in Table 1.

| POS | MS properties encoded in lexicon |
|---|---|
| Verb | verb form, person, number, gender |
| Noun | gender, case, number |
| Pronoun | case, number, gender |
| Adjective | case, number, gender, degree of comparison |
| Adverb | degree of comparison (if applicable) |

Table 1: Morphosyntactic properties for inflected classes

In order to evaluate the lexicon, we calculated its coverage over a portion of ParCoLab. The texts used in the test come from 3 contemporary novels, containing 150 000 tokens equivalent to 28 980 unique wordforms. The coverage was calculated for all wordforms, and then for those appearing at least 2, 5 and 10 times in the subcorpus (cf. Table 2). Eliminating wordforms that occur only once improves the coverage for 4.7%, but these wordforms constitute more than 50% of the identified unique wordforms (cf. number of unique wordforms for thresholds 1 and 2). This is probably due to the relatively small size of the subcorpus used for coverage calculation. In order to have more reliable results, the test will be repeated with a larger portion of ParCoLab.

| frequency treshold | # of unique wordforms | Found in lexicon | Coverage |
|---|---|---|---|
| 1 | 28 980 | 20 808 | 71.80% |
| 2 | 10 630 | 8 136 | 76.53% |
| 5 | 2 946 | 2 328 | 79.02% |
| 10 | 1 241 | 990 | 79.77% |

Table 2: Lexicon coverage

These results also show that although the lexicon is a solid starting point, the resource needs to be developed further. One of the possibilities is to develop a parser for sr.wiktionary.com, which could contain valuable additions to the exissting ressource. We are also considering the possibility of ranking the wordforms found in ParCoLab but not in the lexicon by frequency and adding the most frequent ones manually.

We also performed a quantitative analysis of the lexicon, which gave us insight into ambiguity of Serbian. For 1.2 million wordforms in the lexicon, there are more than 2.5 million MSDs (2.1 MSD per wordform). 727 000 wordforms (60%) are ambiguous. Furthermore, the number of MSDs per wordform can be very high: more than 37 000 wordforms have 10 or more associated MSDs, with 5 wordforms reaching a maximum of 43 MSDs. Although wordform ambiguity in Serbian is intuitively high, the existence of wordforms with 15 or more MSDs seems noteworthy. A manual evaluation of these highly ambiguous wordforms will be performed to exclude errors due to the extraction method or to the quality of the source articles.

These results incited us to try to identify different types of ambiguity in the lexicon. We distinguished 4 categories: i) wordforms corresponding to different lemmas belonging to different POS categories (cf. *krilo*, which can be a nominative/accusative singular of the noun *krilo* ‘lap’, or neuter singular of the past participle of the verb *kriti* ‘to hide’), ii) those corresponding to lemmas having the same form but belonging to different POS categories (cf. *blizu*, which can be a preposition ‘near’ or an adverb ‘nearby’), iii) those corresponding to different lemmas, but having the same POS category (cf. *vrata*, genitive singular of the noun *vrat* ‘neck’, or nominative/accusative plural of the noun *vrata* ‘door’), and iv) ambiguous forms belonging to the paradigm of the same lemma (cf. *jastucima*, which can be dative, instrumental or locative plural of the noun *jastuk* ‘pillow’). The results of this

analysis show that 95% of the ambiguous wordforms belong to the last category (cf. Table 2). This indicates that a large part of ambiguity in Serbian is due to the syncretism in inflectional paradigms.

| | # of word-forms | % of all ambiguous wordforms |
|---|---|---|
| Ambiguous POS and lemma | 15 496 | 2.13% |
| Ambiguous POS, unambiguous lemma | 303 | 0.04% |
| Unambiguous POS, ambiguous lemma | 19 822 | 2.72% |
| Unambiguous POS and lemma, ambiguous MS properties | 691 814 | 95.10% |

Table 3: Ambiguity anlaysis

## 5 Conclusions and future work

This work presents a new lexicon for Serbian containing 1.2 million wordforms corresponding to over 126 000 lemmas. The ressource was created through the use of complementary resources: the greater part of the content was derived from Wiktionary for Serbo-Croatian and subsequently completed with closed-class words from manually created lists and a manually POS-tagged subcorpus. This approach allowed us to reach a solid coverage on a contemporary literary corpus, but the results also showed that there is still room for improvement. For this, we are considering two main approaches: we will explore the possibility to exploit the Serbian edition of Wiktionary (sr.wiktionary.com), which could contain valuable additions to the existing resource. We will also test a semi-manual enrichment process based on frequency lists of wordforms found in ParCoLab, but missing from the lexicon.

## References


Clément, L., Lang, B., and Sagot, B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC2004), p. 1841–1844, Lisbon, Portugal.

Brierley, C. and E. Atwell. (2008). ProPOSEL: a Prosody and POS English Lexicon for Language Engineering. In *Proceedings of LREC'08 Language Resources and Evaluation Conference*, Marrakech, Morocco. May 2008.

Denis, P., & Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation.* Hong Kong.

Fellbaum, C., editor. (1998). WordNet: An Electronic Lexical Database. MIT Press.

Gabrilovich, E., and Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1606–11, Hyderabad, India.

Gesmundo, A., & Samardžić, T. (2012). Lemmatising Serbian as a category tagging task with bidirectional sequence classification. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation.* Istanbul.

Jakovljević, B., Kovačević, A., Sečujski, M., & Marković, M. (2014). A Dependency Treebank for Serbian: Initial Experiments. *Speech and Computer Lecture Notes in Computer Science, 8773*, pp. 42-49.

Krstev, C., Vitas, D., & Erjavec, T. (2004a). MULTEXT-East resources for Serbian. In *Proceedings of 7th Informational Society - Language Technologies Conference*, pp. 108-114. Ljubljana.

Krstev, C., Vitas, D., Stanković, R., Obradović, I., & Pavlović-Lažetić, G. (2004b). Combining heterogeneous lexical resources. In *4th International Conference on Language Resources and Evaluation (LREC'04)*, pp. 1103-1106.

Ljubešić, N., Klubička, F., Agić, Ž., Jazbec, I.-P. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, p. 4264-4270.

McDonald, R., Lemran, K., & Pereira, F. (2006). Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X).*

Meyer, C. M. and Gurevych, I. (2012). OntoWiktionary – Constructing an Ontology from the Collaborative Online Dictionary Wiktionary. In Pazienza, M. T. and Stellato, A., editors, *Semi-Automatic Ontology development: Processes and Resources*, chapter 6, pages 131–161. IGI Global, Hershey, PA, USA.

Miletic, A. (2013). Annotation semi-automatique en parties du discours d'un corpus littéraire

serbe. *Mémoire de Master*, Université Charles de Gaulle Lille 3.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P., and Huang, C.-R. (2009). Wiktionary and NLP: Improving synonymy networks. In *Proceedings of the 2009 ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pp. 19–27, Singapore.

Nivre, J., Hall, J., & Nilsson, J. (2006). MaltParser A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*.

Anton Pérez, L., Gonçalo Oliveira, H., and Gomes, P. (2011). Extracting Lexical-Semantic Knowledge from the Portuguese Wiktionary. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, EPIA 2011, pp. 703–717, Lisbon, Portugal.

Romary, L., Salmon-Alt, S., and Francopoulo, G. (2004). Standards going concrete: from LMF to Morphalou. Zock, M. and Saint-Dizier, P., editors, *COLING 2004 Enhancing and using electronic dictionaries*, pp. 22–28, Geneva, Switzerland.

Sagot, B. (2005). Automatic acquisition of a Slovak lexicon from a raw corpus. In *Text, Speech and Dialogue*, pp. 156-163, Springer, Berlin Heidelberg.

Sagot, B. (2014). DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland.

Sajous, F., Hathout, N., and Calderone, B. (2013a). GLÀFF, un Gros Lexique À tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pp. 285–298, Les Sables d'Olonne, France.

Sennrich, R., & Kunz, B. (2014, May). Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Shen, L., Satta, G., & al. (2007). Guided learning for bidirectional sequence classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 760-767. Prague.

Stosic, D. (2001). Le rôle des préfixes dans l'expression des relations spatiales. Eléments d'analyse à partir des données du serbo-croate et du français. *Cahiers de Grammaire 26*, p. 207-228.

Stosic, D. (2015). ParCoLab (beta), A Parallel Corpus of French, Serbian and English. Toulouse, France: CLLE-ERSS, CNRS & Université de Toulouse 2. (http://parcolab.univ-tlse2.fr)

Strube, M., and Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on ArtificialIntelligence (AAAI-06)*, pp. 1419–24, Boston, MA.

Urieli, A. (2013). *Analyse syntaxique robuste du français : concilier méthodes statistiques et connaissances linguistiques dans l'outil Talismane.* PhD thesis. Université Toulouse II le Mirail.

Zesch, T., and Gurevych, I. 2007. Analysis of the Wikipedia category graph for NLP applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*, pp. 1–8, Rochester, NY. Association for Computational Linguistics.

Zesch, T. et Gurevych, I. (2010). Wisdom of Crowds versus Wisdom of Linguists - Measuring the Semantic Relatedness of Words. *Journal of Natural Language Engineering*, 16(01):25–59.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC2008)*, Marrakech, Morocco